Academic Education Publishing House

# Intelligent Detection Method for Personal Protective Equipment Based on Improved YOLOv11s

**Yang Liu, Jiangang Zhang**
*Henan University of Technology, Zhengzhou, Henan, China*

**Abstract: In industrial production environments, the standardized wearing of personal protective equipment (PPE) is the key to ensuring safety and reducing the risk of accidents. With the development of video surveillance and intelligent perception technology, automatic PPE recognition based on deep learning has gradually become a research hotspot. However, the common background of industrial scenarios, severe occlusion of dense personnel, large differences in target scale, and unstable lighting conditions still make the existing detection models still face challenges in accuracy and real-time. In order to solve these problems, this paper proposes an improved YOLOv11s detection algorithm that combines SIoU loss function and CBAM attention mechanism by taking PPE intelligent identification and early warning in industrial scenarios as the research object. Based on the public PPE dataset, five algorithms were used to compare: YOLOv11s, YOLOv11s + CBAM, YOLOv11s + SIoU + CBAM, original FasterR-CNN and FasterR-CNN+SIoU. The results show that the introduction of CBAM attention mechanism and SIoU bounding box regression loss can significantly improve the performance of PPE detection in complex scenarios, and the improved YOLOv11s model achieves the best balance between accuracy and real-time.**

**Keywords: Personal Protective Equipment; Object Detection; YOLOv11s; Faster R-CNN; Attention Mechanism**

## 1. Introduction

### 1.1 Background and Significance of the Study

In industrial production, the failure of operators to wear personal protective equipment (PPE) is an important cause of safety accidents. Traditional management relies on manual inspection and post-event supervision, which is inefficient, subjective, and difficult to achieve continuous monitoring. With the development of deep learning and computer vision, automatic PPE identification based on video surveillance provides a new means for industrial safety management. However, the complex background of industrial sites, dense personnel, large changes in target scale and variable lighting make PPE targets have serious occlusion and boundary blurring, which poses a challenge to the accuracy and real-time performance of the detection model. Existing studies mostly focus on a single model or improvement strategy, and lack system comparison under unified datasets and evaluation criteria, and need to weigh between accuracy, speed and computational complexity. In order to solve the above problems, this paper proposes an algorithm based on the intelligent identification and early warning of PPE in industrial scenarios, and compares the original YOLOv11s, YOLOv11s+CBAM, YOLOv11s+SIoU+CBAM, Faster R-CNN and Faster R-CNN+SIoU, and analyzes the performance differences between the first-stage and two-stage detection models under the condition of unified dataset, training strategy and evaluation indicators 。 The purpose of this study is to evaluate the performance of different models in terms of detection accuracy, recall, reasoning speed and model complexity, and to provide a reference for the selection and engineering application of industrial PPE models. The main contributions of this paper include: (1) systematically comparing five typical object detection models under unified experimental conditions, covering one-stage and two-stage frameworks, providing an objective basis for the selection of industrial PPE scenario models; (2) The CBAM attention mechanism and SIoU bounding box regression loss were introduced into YOLOv11s and Faster R-CNN, and the effectiveness of each module in complex industrial environment was verified by ablation experiments. (3) Evaluate the performance from the multiple dimensions of accuracy, recall, inference speed and model complexity, analyze the trade-off between accuracy and real-time, and put forward a reference scheme suitable for actual deployment. Therefore, the study of PPE target detection methods for complex industrial scenarios has certain engineering application value and practical significance for improving the intelligent level of safety supervision on the work site and reducing the risk of safety accidents.

### 1.2 Research Status at Home and Abroad

Aiming at the problem of automatic identification and safety early warning of personal protective equipment (PPE) for workers in industrial scenarios, scholars at home and abroad have proposed multi-level solutions. In China, Wang et al.· [1] improved the YOLO framework to achieve high-precision identification of safety helmets, gloves, and reflective vests, providing automated monitoring means for industrial safety management. Zhou et al.· [2] used a multi-scale feature fusion strategy to improve the recall of helmet detection in complex scenarios. Li et al. [3] analyzed the effects of diverse postures, complex PPE types, and lighting and occlusion on detection performance through the application of deep learning in industrial safety monitoring. Liu et al. [4] optimized feature extraction and multi-level prediction structure to achieve a balance between target positioning accuracy and detection speed. Liu et al.· [5] proposed ConvNeXt to combine convolutional networks with Transformer to enhance the generalization ability of visual feature extraction. Zhao et al.· [6] pointed out that in complex industrial and transportation scenarios, models need to take into account accuracy, speed, and robustness. In terms of international research, Wang et al.[7] proposed YOLOv9, a one-stage object detection model that enhances real-time detection performance through programmable gradient information, achieving a better balance between detection accuracy and inference efficiency. Bochkovskiy et al. [8] optimized the network structure and data enhancement strategy in YOLOv4 to optimize speed and accuracy. Woo et al. [9] proposed a CBAM attention mechanism to improve the perception of key regions through channel and spatial weighting. Zheng et al.· [10] introduced the Distance-IoU loss function to improve the regression accuracy of bounding boxes and alleviate the positioning error of small targets and occluded targets. Zhao et al. [11] proposed RT-DETR, a real-time detection Transformer that integrates end-to-end object detection with efficient attention mechanisms, achieving a better balance between detection accuracy and inference speed. Liu et al. [12] introduced an enhanced feature pyramid network to strengthen multi-scale feature representation, significantly improving small object detection performance in complex scenes. Park et al. [13] proposed ssFPN, which models scale-sequence feature relationships across pyramid levels, enhancing robustness and accuracy for small target detection. Although research has progressed, there are still limitations: the first-stage detection model (such as the YOLO series) has high real-time performance, but it is easy to miss the detection of PPE with small targets or severe occlusion. Two-stage models (such as FasterR-CNN) have high accuracy but slow inference speed, making it difficult to meet the real-time monitoring needs of industrial sites. Therefore, how to ensure the detection accuracy while taking into account the reasoning efficiency, and improve the sensitivity to key PPE features such as hard hats, gloves, and reflective vests is the core challenge of intelligent identification and early warning of industrial PPE. Based on this, this paper proposes a multi-model comparison and improvement method, and systematically evaluates YOLOv11s and FasterR-CNN by introducing CBAM attention mechanism and SIoU bounding box regression optimization, which provides a reference for industrial PPE detection model selection and engineering application.

## 2. Experimental Methods

### 2.1 Object Detection

Target detection methods can be divided into two categories: one-stage and two-stage, focusing on detection speed and accuracy respectively. For industrial PPE scenarios, there are differences in real-time, positioning ability and small target recognition between models at different stages, which need to be comprehensively compared and optimized.

The one-stage object detection method integrates target positioning and category prediction into a single end-to-end network, intensively predicts the whole image, and realizes the simultaneous regression of categories and bounding boxes.

Compared with the two-stage method, the first-stage model has a simpler structure, better inference speed and real-time performance, and the representative models include SSD and YOLO series. YOLO transforms detection into regression problems by dividing the grid to predict target location, confidence, and category probability, significantly reducing computational complexity. Subsequent versions continue to optimize the network structure, feature fusion, anchor frame mechanism and loss function, and the multi-scale detection and lightweight design improve the accuracy and small target recognition ability in complex scenarios. YOLOv11 significantly enhances the recognition effect of dense and small targets while maintaining high-speed detection.

The two-stage object detection method splits the task into two stages: "candidate area generation" and "target classification and bounding box regression", which first locates potential targets and then performs fine identification, with high detection accuracy, and the representative models include R-CNN, Fast R-CNN and Faster R-CNN. Faster R-CNN introduces Regional Suggestion Network (RPN) to realize end-to-end sharing of candidate box generation and feature extraction, improving detection efficiency and reducing computational redundancy. It extracts features through deep convolutional network, generates candidate boxes by RPN, and finally completes classification and regression.

However, the two-stage method has slow inference speed and limited real-time performance, and is difficult to meet the needs of rapid early warning due to the diverse target scale and occlusion in industrial PPE detection. Therefore, under the premise of ensuring accuracy, the structure or loss optimization of Faster R-CNN is carried out, and compared with the single-stage model system.

### 2.2 CBAM

Convolutional Block Attention Module (CBAM) is a lightweight, pluggable feature enhancement module that has been widely used in vision tasks such as object detection. CBAM improves the model's ability to express key target information by adaptively weighting the feature map in the feature extraction stage. Its structure is composed of channel attention module and spatial attention module sequence, which can optimize features from different dimensions. The channel attention module obtains channel-level feature descriptions through global average pooling and global maximum pooling, and learns the importance weight of each channel, thereby strengthening the semantic features related to the target category. The spatial attention module performs feature aggregation on the channel dimension, generates spatial weight distributions, and guides the model to focus on the key location areas of the target in the image. Due to its simple structure and small number of parameters, CBAM has a limited impact on the inference speed of the model, making it suitable for embedding lightweight object detection networks. In industrial PPE testing, CBAM can highlight key areas such as safety helmets and protective clothing, improving the robustness and detection stability of the model in complex environments.

### 2.3 Optimization of SIoU and Bounding Box Regression

The bounding box regression loss function is the key to improve the object detection performance, which mainly measures the geometric difference between the predicted box and the real box. Traditional IoU losses only focus on the overlapping area, which is difficult to reflect the center distance, length-to-width ratio and direction information, resulting in limited positioning accuracy and convergence speed. On this basis, SIoU (Scalable Intersection over Union) loss introduces angular constraints and directional guidance, optimizes the prediction box from multiple dimensions of distance, angle and shape, and quickly approximates the real box along the optimal direction, effectively reducing the invalid regression path. Compared with the traditional IoU loss, SIoU has better training stability and positioning accuracy.

### 2.4 The Overall Process of This Experiment

The overall flow of the intelligent detection method for personal protective equipment (PPE) in industrial scenarios proposed in this paper is shown in Figure 1. The detection method uses the image or video data collected at the industrial site as the input, automatically identifies the wearing status of the operator and his protective equipment through the object detection model, and triggers the early warning mechanism when violations are detected to realize intelligent monitoring of industrial safety.

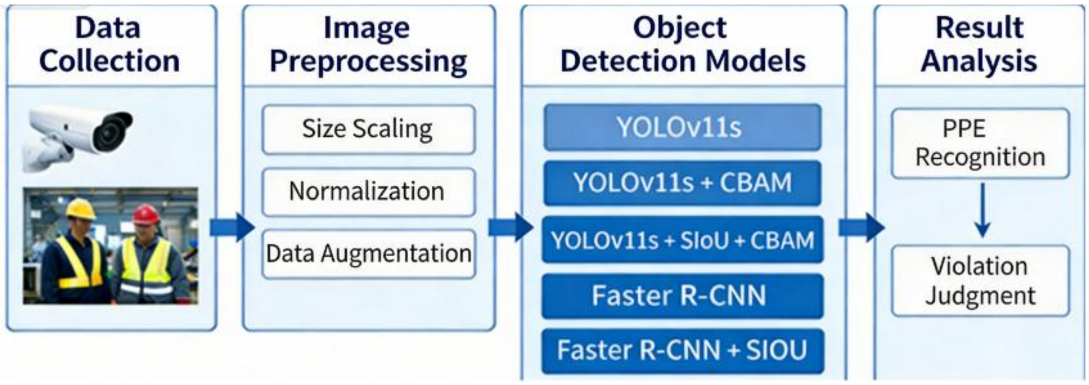


**Figure 1. Overall Flow Chart of PPE Intelligent Detection in Industrial Scenarios**

In the data input stage, the detection method first converts and preprocesses the collected image data in a unified format, including size scaling, normalization and data enhancement, so as to improve the model's adaptability to complex industrial environments. Subsequently, the preprocessed data is fed into the object detection model for inference. This paper focuses on five detection models, namely original YOLOv11s, YOLOv11s + CBAM, YOLOv11s + SIoU + CBAM, FasterR-CNN and FasterR-CNN + SIoU, which are used to locate and classify PPE targets.

In the model output stage, the detection method analyzes the test results to determine whether the personnel are not wearing or incorrectly wearing protective equipment. When the detection results meet the preset violation conditions, the detection method will automatically trigger the alarm module to realize safety warning through interface prompts. This process realizes a complete closed loop from data collection, object detection to security warning.

The overall flow of PPE intelligent identification and early warning algorithm for industrial scenarios proposed in this paper is shown in Figure 2. The algorithm takes the images collected at the industrial site as input, and first performs preprocessing operations such as size scaling, normalization and data enhancement to meet the input requirements of model training and inference. The preprocessed images are fed into the object detection model to complete feature extraction and PPE target prediction. In this paper, five models are selected to output the target category, confidence level and bounding box position. In the processing stage of the detection results, non-maxima suppression and confidence threshold screening are used to reduce redundant detection frames and improve the recognition reliability. Finally, the behavior of the operator is judged according to the PPE category and the "not worn" state, and the early warning mechanism is triggered when a violation is detected, so as to realize a complete closed loop from image input to safety risk warning.

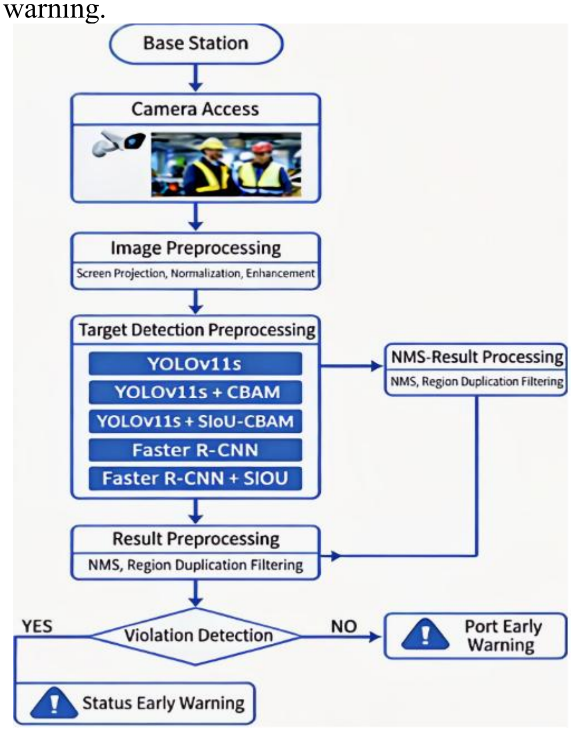


**Figure 2. Flow Chart of PPE Detection Algorithm in Industrial Scenarios**

In order to solve the problems of large target scale difference, severe occlusion and complex background in PPE detection in industrial scenarios, a variety of data enhancement strategies are introduced in the model training stage to improve the generalization ability and robustness. Specifically, it includes random rotation, horizontal flipping, color jitter, scaling, cropping, and Gaussian noise injection to simulate different lighting conditions, working postures, and occlusion conditions, and combines normalization and batch normalization to accelerate model convergence and enhance feature expression capabilities. In terms of loss function design, the YOLOv11s series models adopt weighted comprehensive loss in category prediction, confidence and bounding box regression, in which the bounding box regression introduces the SIoU loss function, and considers the center point distance, scale ratio and angle information to improve the accuracy of target positioning. FasterR-CNN uses SIoU to optimize candidate box generation and regression processes in both RPN and detection head stages. Experimental results show that the joint optimization of data enhancement and improved loss function effectively improves the detection performance of the model in complex industrial environments.

## 3. Experimental Setup and Results

### 3.1 Introduction to Datasets

This paper uses the Construction-PPE dataset (*e.g.*, Figure 3) released by Ultralytics for industrial construction scenarios, including 1132 training drawings, 143 verification diagrams, and 141 test drawings, including helmet, gloves, vest, boots, goggles, none, Person, no_helmet, no_goggle, no_gloves, no_boots A total of 11 types of label PPE With personnel targets, the image background is complex, the target scale changes largely, and the occlusion is widespread, which can truly reflect the requirements of industrial on-site wearing state detection, and provide a reliable basis for model training and performance evaluation. Although the dataset used in this paper is relatively limited, the data is derived from real-world industrial operation scenarios, covering a wide range of human postures, lighting conditions, and occlusion conditions, and is complex and representative to reflect the main challenges in PPE inspection tasks.

**Figure 3. Partial Display of the Dataset**

### 3.2 Model Training

In order to ensure the fairness and reproducibility of the experimental results, five object detection models are trained and tested under a unified software and hardware environment, namely the original YOLOv11s, YOLOv11s + CBAM, YOLOv11s + SIoU + CBAM, FasterR-CNN and FasterR-CNN + SIoU. The experimental platform uses the PyTorch framework to implement model construction, training, and inference. The YOLO series models rely on the implementation provided by Ultralytics, and the Faster R-CNN model is optimized for loss functions. All experiments were performed under the same dataset partition to ensure consistent performance comparisons between different models.

In the training stage, the Stochastic Gradient Descent (SGD) optimizer is used to improve the convergence speed and stability of the model by combining momentum and weight decay. The input image size is uniformly 640×640, the batch size is set to 16, the initial learning rate is 0.01, and the Cosine Annealing strategy is used to gradually decay during training to balance convergence efficiency and stability. The number of training rounds is fixed at 100, and the optimal model parameters on the validation set are selected for the final test combined with the early stop strategy. To enhance the generalization capabilities of the model and reduce overfitting, various data augmentation strategies are enabled during the training process, including image flipping, scaling, and color transformation.

### 3.3 Experimental Results and Analysis

In order to comprehensively evaluate the performance of different detection models in PPE recognition tasks in industrial scenarios, mAP, Precision, Recall and FPS are selected as the main indicators, and five models are compared: original YOLOv11s, YOLOv11s+CBAM, YOLOv11s+SIoU+CBAM, FasterR-CNN and FasterR-CNN+SIoU. All models are trained and tested in a unified dataset and experimental environment. Experimental results show that after the introduction of CBAM attention mechanism and SIoU bounding box regression in YOLOv11s, mAP, Precision and Recall are improved, which enhances the expression of key PPE regions and improves the accuracy of bounding box regression. FasterR-CNN is slightly better in Precision and Recall, but the detection speed is slower. The YOLOv11s series has obvious advantages in FPS and can meet the needs of real-time monitoring. Combined with accuracy and speed performance, YOLOv11s+SIoU+CBAM strikes the best balance in industrial PPE inspection. Table 1 shows the detection accuracy and inference speed of the five models on the Construction-PPE dataset.

**Table 1. Performance Comparison Results of the Five Models**

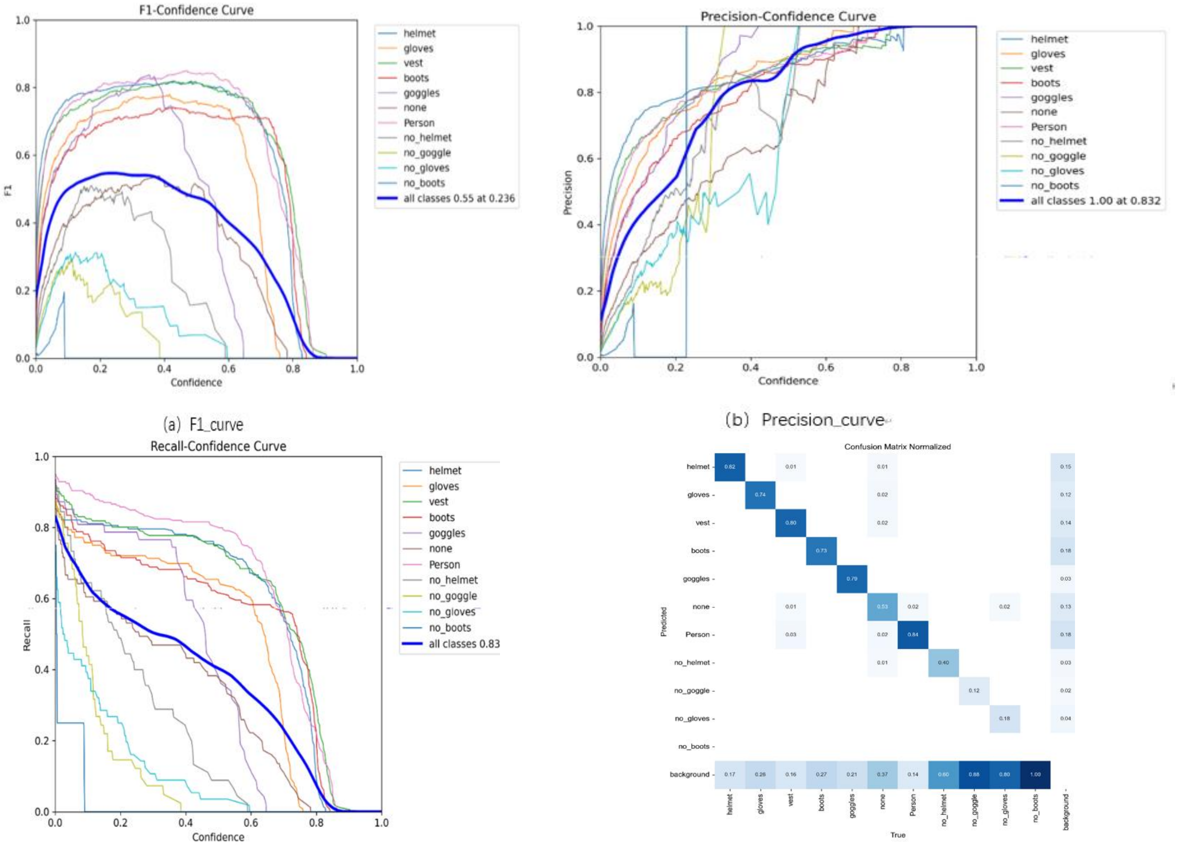
| Model name | Precision | Recall | F1-score | mAP@0.5 |
|---|---|---|---|---|
| YOLOv11s | 0.6467 | 0.5163 | 0.5742 | 0.5777 |
| YOLOv11s + CBAM | 0.6706 | 0.5373 | 0.5966 | 0.6213 |
| YOLOv11s + SIoU + CBAM | 0.7078 | 0.5368 | 0.6106 | 0.6481 |
| Faster R-CNN | 0.0083 | 0.0398 | 0.0138 | 0.0079 |
| Faster R-CNN + SIoU | 0.1469 | 0.7014 | 0.2430 | 0.1829 |

The comprehensive experimental results show that after the introduction of CBAM attention mechanism and SIoU loss function, the YOLOv11s model has achieved significant improvement in PPE detection accuracy while maintaining a high detection speed. Compared with the traditional two-stage detection model, the improved YOLOv11s has better comprehensive performance while meeting real-time requirements, making it more suitable for deployment in actual industrial safety monitoring and early warning scenarios.

In order to further verify the influence of attention mechanism on model performance, a separate experimental analysis was conducted on the YOLOv11s model with CBAM. Experimental results show that YOLOv11s+CBAM shows relatively stable and balanced detection ability in PPE multi-category object detection tasks. The model reaches a Precision of 0.6706 on the test set, a Recall of 0.5373, and a mAP@0.5 of 06213, the overall performance is better than the basic model without the introduction of attention mechanisms.

From the F1–Confidence curve in Table 2(a), when the confidence threshold is about 0.236, the maximum F1 value is about 0.55, which achieves the optimal balance between Precision and Recall. In categories with obvious characteristics such as hard hats and reflective vests, the F1 value is higher and stable, while the "no wearing" category has a low F1 value due to small differences in appearance and difficult to distinguish samples, which lowers the overall performance. Judging from the accuracy curve in Table 2(b) and the recall curve in (c), the CBAM module adaptively assigns weights in the channel and spatial dimensions, making the model pay more attention to the key areas of PPE and reduce background interference, thereby improving precision. However, there is still room for improvement in Recall, indicating that it is still possible to miss detections in complex occlusion, small targets, or low-confidence samples. The normalized confusion matrix curve in Table 2(d) shows that the model has a high accuracy in recognizing the main categories such as helmet, vest, goggles and Person, and the diagonal values of helmet and Person are 0.82 and 0.84, respectively, which verifies the advantages of CBAM in enhancing the expression of key features. However, the identification of illegal targets such as no_helmet, no_goggles, and no_gloves is relatively weak, and there is a background phenomenon of misjudgment, which is reflected in the similar category characteristics and unbalanced samples, and the ability to distinguish violations still needs to be improved. The Recall–Confidence curve further shows that the recall rate is high when the confidence threshold is low, but the decrease of non-compliance recalls is more obvious with the increase of the threshold, indicating that YOLOv11s+CBAM is more conservative under high confidence prediction, improving detection reliability but sacrificing part of the recall. In summary, YOLOv11s+CBAM effectively focuses on the key areas of PPE while maintaining high precision, which verifies the effectiveness of the attention mechanism in industrial scene detection, but there are still shortcomings in the detection of complex backgrounds, small targets and illegal objects, and the performance can be further improved through sample balancing, feature fusion optimization and special loss design.

**Table 2. Performance of the YOLOv11s + CBAM model**



| Predicted \ True | helmet | gloves | vest | boots | goggles | none | Person | no_helmet | no_goggle | no_gloves | no_boots | background |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| helmet | 0.82 | | 0.01 | | | 0.01 | | | | | | 0.15 |
| gloves | | 0.74 | | | | 0.02 | | | | | | 0.12 |
| vest | | | 0.80 | | | 0.02 | | | | | | 0.14 |
| boots | | | | 0.73 | | | | | | | | 0.18 |
| goggles | | | | | 0.79 | | | | | | | 0.03 |
| none | | | 0.01 | | | 0.53 | 0.02 | | | 0.02 | | 0.13 |
| Person | | | 0.03 | | | 0.02 | 0.84 | | | | | 0.18 |
| no_helmet | | | | | | 0.01 | | 0.40 | | | | 0.03 |
| no_goggle | | | | | | | | | 0.12 | | | 0.02 |
| no_gloves | | | | | | | | | | 0.18 | | 0.04 |
| no_boots | | | | | | | | | | | | |
| background | 0.17 | 0.28 | 0.16 | 0.27 | 0.21 | 0.37 | 0.14 | 0.60 | 0.88 | 0.80 | 1.00 | |

(a) F1_curve

(b) Precision_curve

(c) Recall_curve

(d) Confusion_Matrix_curve

In order to evaluate the impact of the improvement strategy on the performance of PPE detection in industrial scenarios, the effects of CBAM attention mechanism and SIoU bounding box regression loss were systematically analyzed. Firstly, the original YOLOv11s is used as the baseline model to evaluate the performance on the Construction-PPE dataset. Subsequently, the CBAM module is introduced to enhance the network's modeling ability of key PPE regional features, and the model is significantly improved in mAP, Precision and Recall indicators. After replacing the bounding box regression loss with SIoU, the target positioning accuracy is significantly improved, especially in small target and occlusion scenarios. Experimental results show that CBAM and SIoU are complementary in feature expression and localization optimization, and the combination of the two can effectively improve the detection accuracy and robustness of the model.

## 4. Summary

Based on the Construction-PPE dataset, five models, namely YOLOv11s, YOLOv11s+CBAM, YOLOv11s+SIoU+CBAM, FasterR-CNN and FasterR-CNN+SIoU, are systematically compared and analyzed. The experimental results show that the first-stage model has obvious advantages in inference speed, which is more suitable for real-time monitoring of industrial scenarios. The detection accuracy of the two-stage model is high, but the real-time performance is limited. The introduction of CBAM attention mechanism can enhance the feature modeling ability of YOLOv11s on key PPE regions, and the target positioning accuracy is further improved after SIoU loss. Based on the accuracy and speed indicators, YOLOv11s+SIoU+CBAM performed the best in industrial PPE inspection tasks. Despite the results, this paper still has problems such as limited data scale, insufficient generalization ability under complex working conditions, and insufficient use of video timing information. In the future, the practicability and robustness of industrial PPE intelligent detection methods can be further improved by expanding multi-scenario datasets, fusing detection and tracking methods, and exploring lightweight and edge deployment strategies, so as to improve the generalization ability and engineering practicability of the model.

## References

[1] Wang Lei, Zhang Qiang, Li Ming. Research on Detection Method of Wearing Personal Protective Equipment on Construction Site Based on Deep Learning. Computer Engineering and Application, 2022, 58(18): 215-221

[2] Zhou Kai, Sun Lixin. Research on safety helmet wearing detection based on YOLO. Computer Application Research, 2020, 37(12): 3658-3662

[3] Li Peng, Huang Zhiqiang. Review of the application of deep learning in industrial safety monitoring. Automation Technology and Application, 2023, 42(6): 1-7

[4] Liu Yang, Zhao Xin, Chen Zhigang. Improvement and application of object detection algorithm for industrial safety. Computer Engineering, 2021, 47(10): 268-274

[5] Liu Z, Mao H, Wu C Y, et al. A ConvNet for the 2020s: ConvNeXt//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 11976–11986.

[6] Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212–3232.

[7] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv9: Learning what you want to learn using programmable gradient information//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Seattle: IEEE, 2024: 12345–12354.

[8] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.

[9] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3–19.

[10]Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 12993–13000.

[11]Zhao Y, Lv J, Wang S, et al. RT-DETR: Real-Time Detection Transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Vancouver: IEEE, 2024: 1234–1243.

[12]Liu H, Zhang Y, Chen X, et al. Enhanced Feature Pyramid Networks for Small Object Detection. IEEE Transactions on Image Processing, 2023, 32(6): 4123–4135.

[13]Park H J, Kim S, Lee J, et al. ssFPN: Scale-Sequence Feature Pyramid Network for Small Object Detection. Sensors, 2023, 23(9): 4215.