Lightweight Worker-Safety Equipment Detection Algorithm Based on YOLOv7

Changbin Wang

Abstract—**With the development of artificial intelligence technology, target detection technology has been widely used in security monitoring and early warning systems. The purpose is to improve the efficiency of safe production supervision. However, the existing target detection frameworks require a large amount of computational resources to perform. To address this issue, lightweight deep learning models have become a research hotspot. This article proposes a lightweight target detection model. By adjusting the neural network structure, it significantly reduces computational complexity while maintaining high-performance detection. Firstly, we improve the Ferry_Network structure using depthwise separable convolution. Replace the CBS module in the Ferry_Network structure with a depthwise separable module. Then we modify the branch structure of MCN structure. Introduce depthwise separable convolution into some branches of the MCN structure. Perform lightweight operations on the previously improved model to achieve improvements in parameter reduction. Furthermore, comparative experiments were conducted. Comparing the lightweight network model YOLOv7-DSE with the baseline network model and other excellent object detection models, demonstrating the effectiveness of the improvements. Finally, ablation experiments were performed to verify the effects of each improved module. Proving that the model proposed in this paper not only has good accuracy and speed but also exhibits advantages in model size due to its lightweight nature. While also demonstrating strong robustness.**

Index Terms—**target detection, safety equipment testing, lightweight model, Ferry-Network structure, MCN structure.**

I. Introduction

With the rapid development of deep learning technology, target detection technology has been widely used in safety monitoring and early warning systems. Target detection technology plays a crucial role in multiple practical applications. In highly dangerous live working sites, workers must perform important operations such as checking the electricity and disconnecting it. Wearing appropriate work safety equipment is a key means of achieving self-protection in production work. However, due to the widespread lack of safety awareness among workers. The complexity of the working environment and the current safety supervision work not yet meeting standards. The wearing rate of work safety equipment is low, forming a vicious cycle. These factors lead to major safety accidents and cause significant losses to enterprise property. To improve the effectiveness of safety supervision, using target detection technology to assist in achieving safety supervision has become a low-cost and efficient means.

Object detection is an important field in computer vision. It identifies and locates specific objects in images or videos. Traditional machine learning methods have laid the foundation for object detection. In this period from 2010 to 2014, common methods included the use of feature-based approaches such as Haar features [1] and HOG features [2]. These were combined with machine learning classifiers such as Adaboost [3] and SVM [4]. These methods achieved certain results in specific scenarios. These algorithms have strong understandability and good interpretability. They have good generalization performance for tasks with small amounts of data. However, they require manual design of appropriate features and may encounter difficulties or insufficient feature extraction. For complex scenes and large variations in targets. they have insufficient generalization ability and weak robustness. For example, Sri-Kaushik Pavani et al. proposed an extended Haar feature for examining the human frontal and cardiac regions. The feature rectangle with the best weight is used for target detection to maximize its ability to distinguish objects from clutter. These features maintain the simplicity of traditional formula evaluation while being more discriminative [5]. Yanwei Pang et al. proposed a cell-based CHOG algorithm [6]. It only extracts sub-feature vectors. Calculates the sub-decision values of partial linear classifiers. Reduces the dimension of traditional block-based gradient direction histogram (BHOG) feature vectors. Using CHOG as a feature descriptor, it detects hands, faces, and pedestrians in videos. Experimental results demonstrate the superiority of this method [7]. Christos Kyrkou successfully designed and implemented a flexible parallel hardware architecture based on AdaBoost algorithm for real-time target detection. This architecture can efficiently and accurately detect targets, and has flexibility to handle different input image sizes and training set formats [7].

Since 2015, with the development of deep learning technology, target detection methods based on deep learning have emerged. Target detection methods based on convolutional neural networks CNN（Convolutional Neural Network [8] have achieved breakthrough progress. Alex Krizhevsky trained a large-scale deep convolutional neural network called AlexNet algorithm [9]. AlexNet performed brilliantly in the ImageNet LSVRC-2012 competition, achieving a test error rate of 15.3% and surpassing the second place by 10.9%. His performance won first place. This neural network has 60 million parameters and 650,000 neurons. It consists of five convolutional layers, some of which are followed by max pooling layers [10] and three

Manuscript received Mar 28, 2024; revised Jan 19, 2025. This work was supported by Ministry of Education industry-university cooperative education project (231105181075653), and the Education Department of of Liaoning Province ,Youth Project (LJKQZ20222440).

Changbin Wang is an associate Professor of School of Artificial Intelligence (Big Data Industry College), Anshan Normal University, Anshan, China (Corresponding author to provide phone: +86-152-4220-3316; e-mail: 15242203316@163.com).

fully connected layers [11]. Finally, there is a 1000-way softmax normalization exponential function [12].

In recent years, target detection technology based on deep learning has been more widely used in the field of engineering construction. This has important application significance in scenarios such as industrial hazardous environments and construction sites. It can ensure the safety of worker operations. More and more experts and scholars at home and abroad have begun to study algorithms for detecting the wearing of work safety equipment. Currently, there are generally two types of commonly used methods for detecting safety helmet algorithms.one based on machine learning and the other based on deep learning. When studying machine learning-based methods, firstly, researchers usually use traditional image feature extraction methods (such as Haar features, HOG features, and color features). Then they use classifiers (such as SVM, AdaBoost, etc.) to judge whether it is a safety helmet or not. Then they use classifiers such as support vector machines (SVM) and AdaBoost to train and predict. Pathasu Doungmala et al. combined Haar feature and Hough transform detection, two kinds of helmet detection methods. To detect the phenomenon of not wearing helmets and wearing helmets, so as to achieve better detection results [13]. A fast color image safety helmet detection algorithm based on Haar-like features is proposed to detect helmet regions. A method for detecting facial features using circular Hough transform is proposed to determine the type of helmet worn by the wearer. Miao Jin used SVM, HOG features, and color features to propose an algorithm for deformable part models. Gradient histograms are used for feature training. Support vector machines are used to detect helmets and ultimately determine whether workers wear helmets or not [14]. Fan Min et al. used the Vibe algorithm [15] to detect moving target areas. They combined Haar features with HSV color space features to extract helmet features. Input them into the Adaboost algorithm for classification to achieve helmet recognition.

In addition, various efficient object detection frameworks such as Faster R-CNN, YOLO, and SSD have been proposed in recent years. However, these models have excessive parameters and require substantial computational resources for execution. This challenge is particularly pronounced on resource-constrained devices. Such as mobile devices and embedded systems, which typically have limited computational capabilities and battery life. In response to this challenge, lightweight deep learning models have become an active research area. Aiming to develop models that are computationally efficient, energy-efficient, and do not significantly compromise performance. However, reducing model parameters and computational load often sacrifices some degree of detection accuracy. Therefore, how to balance the efficiency and accuracy of lightweight models becomes a core issue. Given the diversity of safety equipment for workers and the complexity of industrial scenarios. Traditional machine learning methods for detecting safety equipment have limitations in feature design and lack robustness in complex scenarios. While they perform well in tasks with small amounts of data and have good generalization capabilities. They require manual design of appropriate features, which may lead to difficulties or inadequacies in feature extraction. In complex scenarios and situations with significant target variations. They suffer from insufficient generalization capabilities and weak robustness. Therefore, this research chooses to adopt a deep learning-based object detection method to achieve the detection of workers' safety equipment.

This article improves the baseline network model YOLOv7 by reducing the model's parameter count through optimized depthwise separable convolution. First, the MCN module in the main network is improved by replacing ordinary convolution with depthwise separable convolution modules in the branch of MCN modules. Then, the FN structure is improved to ensure feature extraction capability while reducing the model's parameter count and optimizing the main network structure. Finally, this article describes the experimental part of this study. This experimental part compares Faster RCNN algorithms horizontally and compares YOLOv7, CEAM-YOLOv7, YOLOv7-RAR vertically. The effectiveness of algorithm improvement is verified by observing model size, mAP, precision curve, recall curve, PR curve and other evaluation indicators. Finally, ablation experiments are conducted by sequentially freezing improved modules FN-DSC, MCN-DSC, and MCN-SPD. The purpose is to observe model size and mAP to prove that the improvement of the algorithm is effective. It is proved that the model proposed in this paper not only has good accuracy and speed, but also has the advantage of lightweight in terms of the number of parameters. it also has strong robustness. This paper addresses the issue of the large parameter count in the baseline network model YOLOv7 and proposes improvements to achieve model lightweighting. The depthwise separable convolution is optimized. First, the MCN module of the backbone network is improved by replacing the regular convolutions on the MCN module branches with depthwise separable convolution modules. Second, the FN structure is optimized to ensure feature extraction capability while reducing the model's parameter count, thereby optimizing the backbone network structure. Third, the experimental section of this paper is described, which initially compares the Faster R-CNN algorithm horizontally. Followed by a vertical comparison with the baseline models YOLOv7, CEAM-YOLOv7, and YOLOv7-RAR. The effectiveness of the algorithm improvements is verified by observing evaluation metrics. Such as model size, mAP, precision curves, recall curves, and PR curves. Finally, ablation experiments were conducted. The improved modules FN-DSC, MCN-DSC, and MCN-SPD were frozen in sequence. The effectiveness of the algorithm improvements was demonstrated by observing the model size and mAP. It was proven that the model proposed in this paper not only exhibits good accuracy and speed but also has the advantage of being lightweight in terms of parameter count. Additionally, it possesses strong robustness.

II. YOLOv7 MODEL

A. YOLOv7 model structure

The algorithm framework of YOLOv7 is composed of three parts: input, backbone feature layer and head prediction layer network. The entire working process of

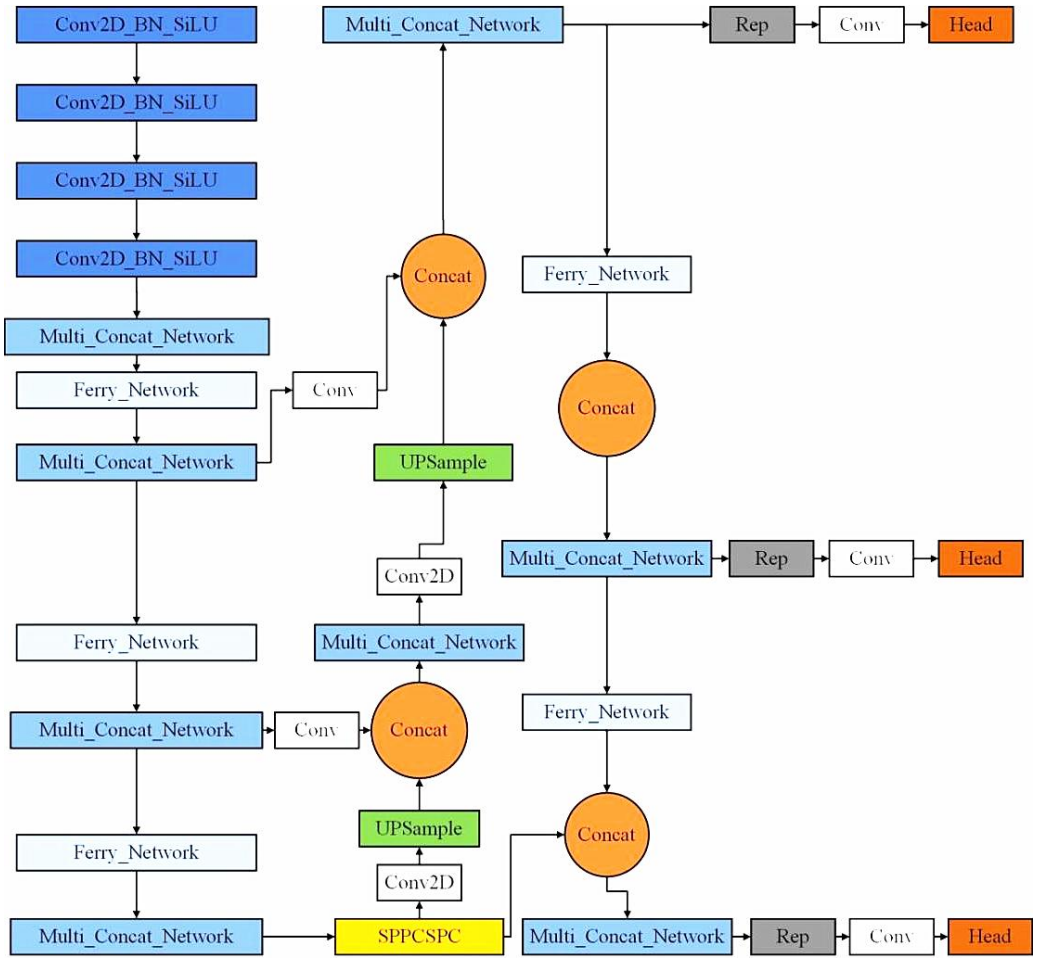


Fig.1. Structure of YOLOv7

YOLOv7 can be simplified as follows. Feature extraction, feature enhancement and prediction of different prior frames correspond to different scales. The structure of YOLOv7 model is shown in Fig.1.

III. YOLOv7 Backbone Network Architecture

The main part of YOLOv7 utilizes a convolutional neural network composed of a large number of CBS (Conv2D_BN_Silu) modules. This module is constantly reused in the network structure of YOLOv7. It is constructed by a convolutional layer, a BN layer, and an activation function Silu layer.

The MCN structure comprises four branches, as shown in Fig.2. The leftmost branch is a CBS module. The second branch from the left is a CBS module. The rightmost branch is a series of five CBS modules connected in series. The second branch from the right is a series of three CBS modules connected in series. These four branches are then stacked and undergo feature fusion using another CBS module. By controlling the gradient path, more features are learned, enhancing the robustness of the network.

The main part of YOLOv7 also constructs a FN (Ferry Network) structure for downsampling. The FN structure comprises two branches, as shown in Fig.3. The upper branch consists of a stride-2 max pooling layer and a CBS module connected in series. The lower branch consists of two CBS modules connected in series. Their filter sizes are 1×1 and 3×3, respectively. The main network constructs a more dense residual structure. A more dense residual network can also increase depth to improve accuracy. The residual network modules inside it use skip connections to alleviate the problem of gradient vanishing caused by deeper neural networks.

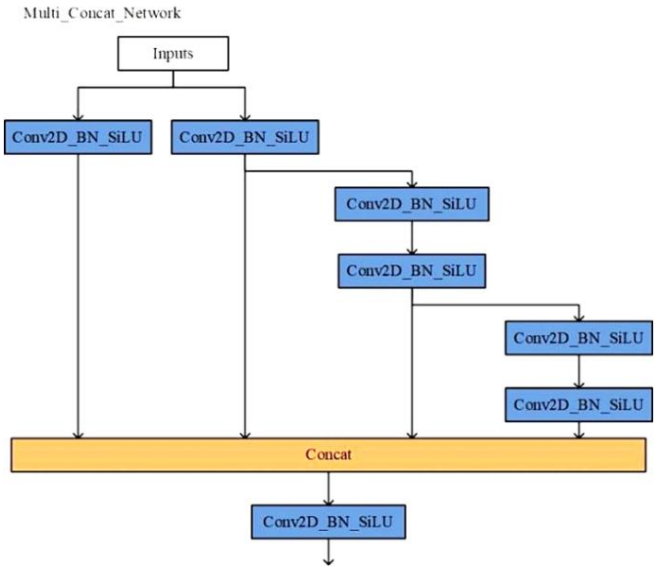


Fig. 2. Structure of MCN

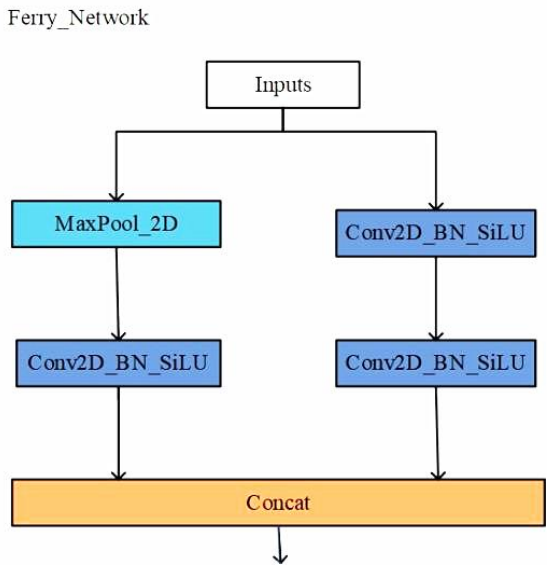


Fig. 3. Structure of FN

IV. The Improved Network Model YOLOv7-DSE

Given the complexity of live-work scenarios. It is important to improve the detection accuracy of safety equipment. While also reducing the false positive and false negative rates of network models. Although enhancing feature extraction capabilities is necessary. It can also increase the number of parameters and computational complexity. In the baseline model YOLOv7 algorithm, there are a large number of 3×3 convolutions. However, as the network hierarchy deepens, it leads to more parameter calculations. To improve the feature extraction capabilities of the network without increasing the number of parameters and computational complexity. Depthwise separable convolution is adopted to improve the FN structure and MCN structure in the backbone part.

A. Depth Separable Convolution

Deep separable convolution and spatial separable convolution are different.and it often referred to as "separable convolution" in deep learning frameworks such as TensorFlow and Keras. This includes channelwise convolution, which performs spatial convolution independently on each channel of the input.and point wise convolution, which projects the output channels of the depthwise convolution to a new channel space. Depthwise separable convolution is typically implemented in nonlinear situations. Due to its characteristic of two-dimensional convolution operations. The mapping of cross-channel correlation and spatial correlation in the feature maps of convolutional neural networks can be fully decoupled.

The steps of depthwise separable convolution are as follows:

(1) Perform channelwise convolution (depthwise Convolution). It uses filters to perform separate convolution operations on each channel of the input feature map. Generating a set of depthwise convolutional feature maps. This step only performs convolution operations on each input channel without expanding the filters. The depthwise separable step is illustrated in Fig.4.

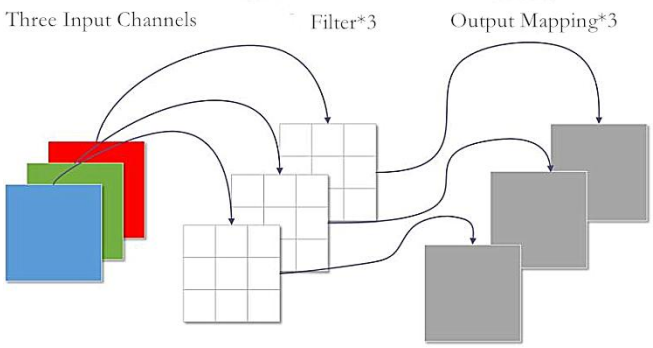


Fig. 4. Schematic Diagram of Depthwise Convolution

(2) Point wise Convolution: It uses 1x1 filters to perform convolution operations on the depthwise convolutional feature maps. The features between channels are linearly combined to generate the final output feature map. This step can be viewed as a traditional fully connected layer operation. However, only the convolution operation is performed on the channels, and no convolution operation is performed on the positions. The Point wise Convolution has two functions. First, it allows the depthwise separable convolution to freely change the number of channels. Second, it fuses the feature maps output by the channelwise convolution. The steps of Point wise Convolution are illustrated in Fig. 5.

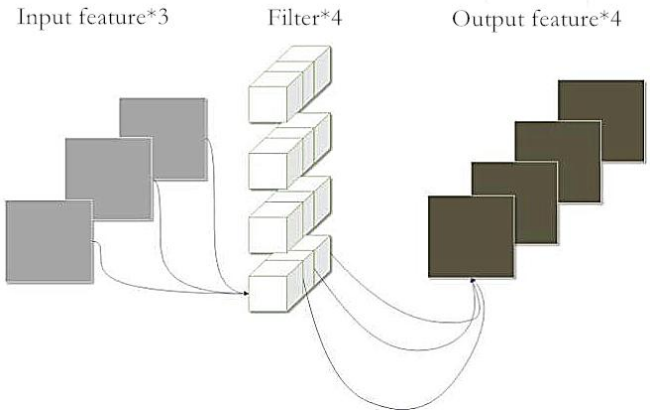


Fig. 5. Schematic Diagram of Pointwise Convolution

To demonstrate the superiority of the improvement. It is necessary to calculate and compare the parameter count and computational cost of ordinary convolution and depthwise separable convolution. We can draw clear conclusions. W represents the width of the filter, H represents its height. C_{in} represents the number of input channels, and C_{out} represents the number of output channels. The parameter P can be expressed as formula (1):

$$P = W \times H \times C_{in} \times C_{out} \qquad (1)$$

W' represents the width of the input image. H' represents the height of the input image. and the computational cost C can be expressed as formula (2):

$$C = W \times H \times (W' - W + 1) \times (H' - H + 1) \times C_{in} \times C_{out} \qquad (2)$$

When taking an input image with three channels and a size of 5×5. Aiming to get a feature map of size $3\times3\times4$.

The ordinary convolution requires a filter of size 3×3×3×4. The number of parameters *Pcon* in its convolution layer can be calculated as 108. The computational cost *Ccon* in its convolution layer can be calculated as 972. If we want to get the same feature map of size $3 \times 3 \times 4$. The depthwise separable convolution performs channelwise convolution on the input image with three channels through a 3×3 filter. And then uses a 1×1 filter to combine different channels for pointwise convolution, getting a new set of output feature maps. The number of parameters PDSC used by depthwise separable convolution can be calculated using formula (1), and the result is 39. The computational cost CDSC used by its convolution layer can be calculated using formula (2), and the result is 351.

Through the calculation of parameters and computational cost. we find that when getting feature maps of the same size. The number of parameters and computational cost used by depthwise separable convolution are about one-third of those used by ordinary convolution. Therefore, under the condition of ensuring unchanged detection effect. Using depthwise separable convolution can reduce the computational cost of neural networks. The experimental part of this chapter will analyze the specific improvement brought by improving depthwise separable convolution for the model.

B. Improve the Ferry_ Network Structure

The FN structure is mainly used for downsampling features in the main trunk. With two branches: one is the pooling branch, and the other is the convolution branch. By applying depthwise separable convolution to the second CBS's 3×3 convolution in the convolution branch of the FN structure. We can reduce the network depth and structural parameters, thus reducing the computational cost of the model and achieving lightweight. The improved structure of the FN-DSC module is shown in Fig.6.

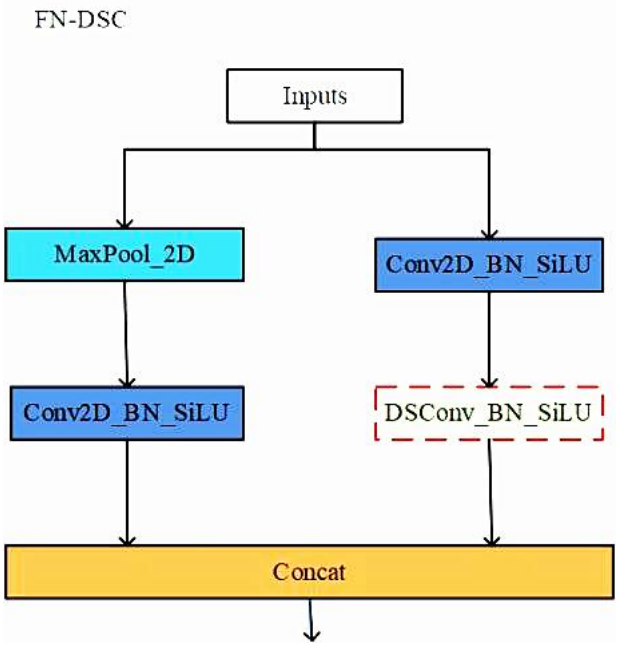


Fig. 6. Structure of FN-DSC Module

C. Improved the Multi _Concat_ Network Structure

(1) Improved network model MCN-SPD

In the MCN structure used in the head network of YOLOv7, 3x3 convolutions are used. This can cause the loss of some target features when the network model focuses on target features. An improvement is made to the space-to-depth non-stride convolution. First, the Multi_Concat_Network structure of the head network is improved. A space-to-depth non-stride convolution is proposed. And it is applied to the MCN structure. The 3x3 convolution after the Concat module is replaced. An additional space-to-depth non-stride convolution is added. Finally, the MCN-SPD module is obtained. Because YOLOv7 uses 5 convolutional layers of step size 2 in the backbone to downsample the feature map with a factor of 25. And two convolutional layers of step size 2 are used in the neck. There is a cascade layer after each step convolution in the neck of YOLOv7, so this does not affect the SPD's approach. It only keeps the space-to-depth non-stride convolution between convolutions. This method used in the paper retains all the image features. The structure of the improved MCN-SPD is shown in Fig.7.

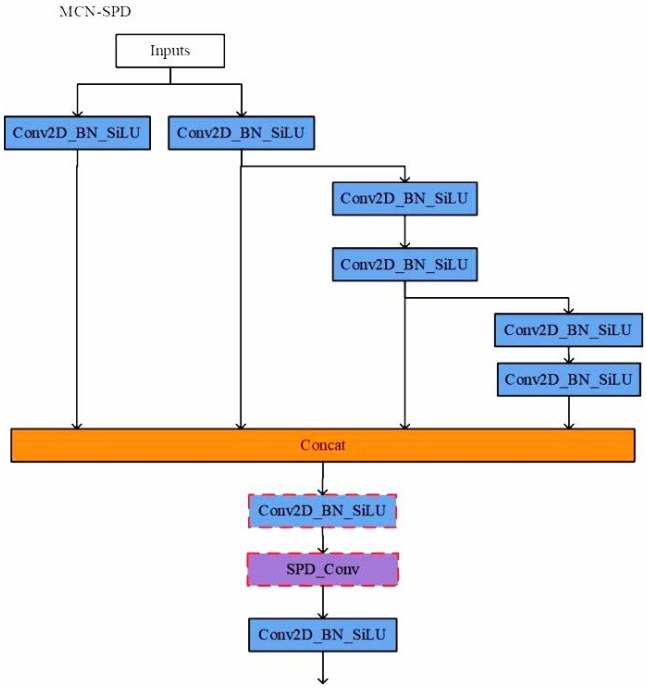


Fig.7.Improved Module MCN-SPD

(2) Improved MCN-DSC

In Part II, the MCN structure is mainly composed of four branches connected in parallel. Each branch uses a CBS module composed of convolution, batch normalization function, and activation function. The number of CBS modules on each branch is different, which is used for feature extraction of different sizes. After stacking, the four branches pass through a CBS module for feature fusion, which is used for downsampling features in the main trunk. This part improves depthwise separable convolution and applies it to the MCN structure in the backbone network. As shown in Fig.8. In the right two branches, multiple stacked CBS modules cause an increase in the number of parameters and computational cost. In the improvement, the structure of convolution-batch normalization function-activation function is not changed. The optimized depthwise separable convolution is replaced with ordinary convolution. The improved MCN-DSC module structure is shown in Fig.8.

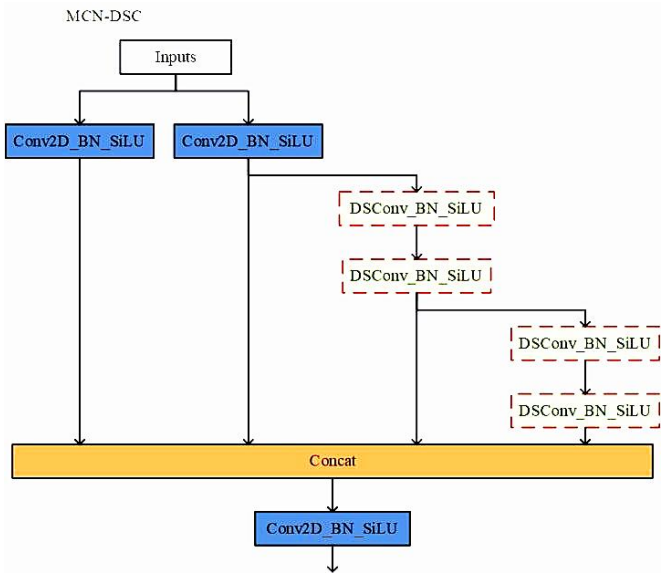


Fig.8. Struceture of MCN-DSC Module

In the baseline model YOLOv7, the role of MCN is to extract and fuse features. This structure is continuously reused in the model. By applying MCN-DSC to the backbone network, the model achieves the effect of optimizing the network structure.

D. Improved network model YOLOv7-DSE (Depthwise Space-to-depth Efficiency)

The overall improved and optimized YOLOv7-DSE is used to detect worker safety equipment. The backbone network part has optimized the MCN-DSC module and the FN-DSC module. Improving depthwise separable convolution has made it partially lightweight, reducing the number of parameters and the computational cost of the network model. In the head network part, the MCN-SPD has been optimized. Improving spatial-depth non-stride convolution has allowed the capture of image information without losing any image features. It has improved the detection effect of small targets and optimized the loss function. A new regression box loss calculation method, EIoU, has been used. A new penalty term has been added to limit the target anchor box and localization anchor box. This makes the model's prediction box positioning more accurate in complex scenes, increasing the robustness of the network model. The improved YOLOv7-DSE structure is shown in Fig.9.

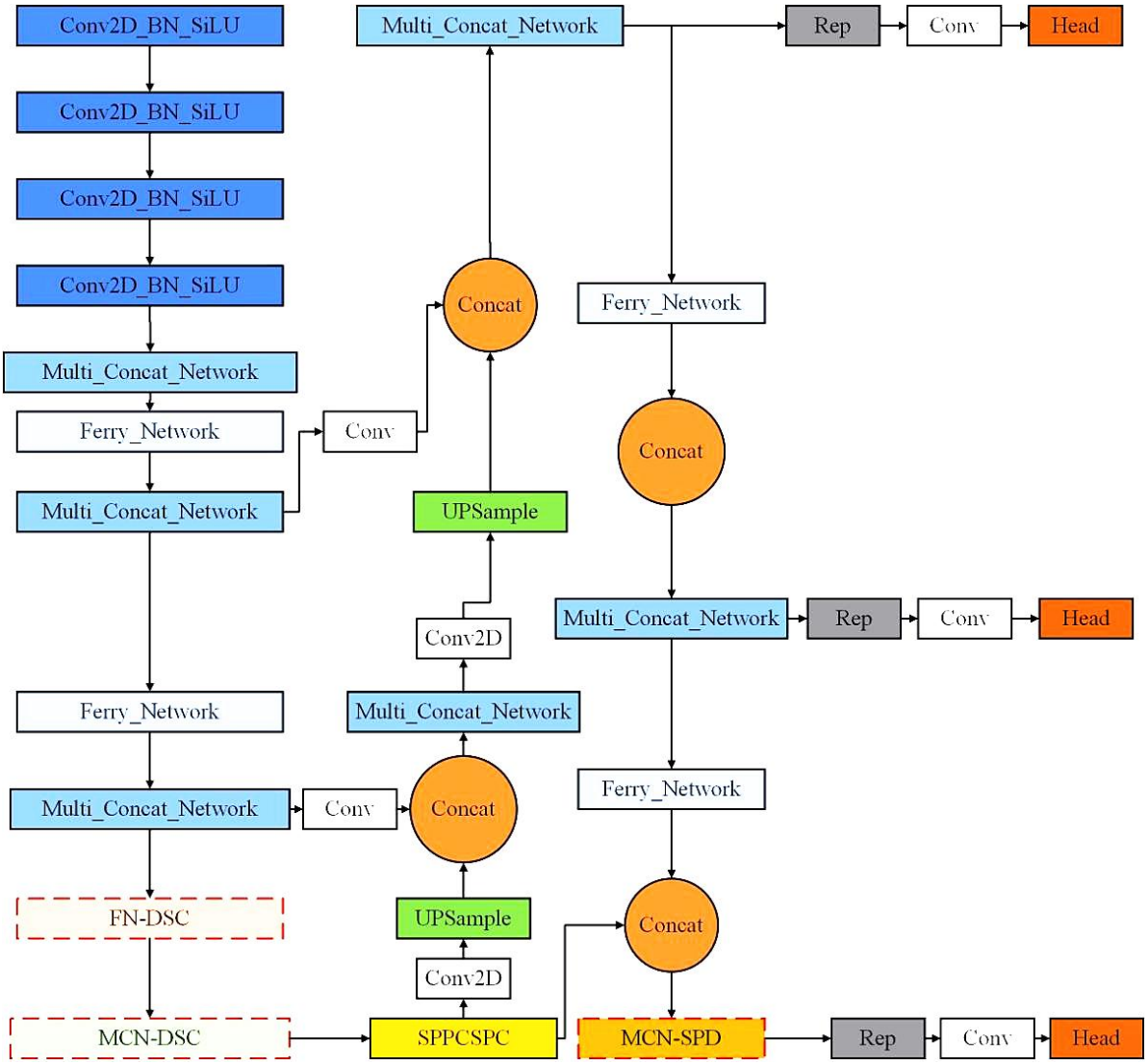


Fig.9. Struceture of YOLOv7-DSE

TABLE I
Comparison Experiment Results

Models	F1	Miss rate	mAP	Model size
YOLOv4	0.74	0.06%	72.94%	103.2MB
YOLOv5	0.74	0.06%	75.75%	101.6MB
YOLOv7	0.80	0.05%	79.75%	94.1MB
YOLOv7-DSE	0.84	0.04%	82.39%	72.9MB
CEAM-YOLOv7	0.81	0.04%	80.44%	96.4MB
YOLOv7-RAR	0.79	0.05%	79.86%	84.3MB

V. Experimental Design and Result Analysis

A. Experimental Environment

The experimental equipment includes a NVIDIA GeForce GTX3070 graphics card with 8GB memory. An Intel Core i7-10700F CPU, and a Windows 10 Professional operating system with 16.0GB RAM. These are used for comparative experiments of object detection algorithms. Multiple object detection algorithms use the PyTorch 1.8.0 framework based on Python 3.6 in this experimental environment. The language compiler is PyCharm 2022.3.2, and the GPU acceleration libraries used for training are CUDA 11.1 and cudnn 8.0.4.

B. Experimental Design

In this paper, the equipment_dataset is used as the training and validation dataset. The training iteration is set to 100 epochs, and the ratio of the training set to the validation set is 8:2. A horizontal comparative experiment is conducted by comparing the improved YOLOv7-DSE with YOLOv4 and YOLOv5. Three vertical comparative experiments are designed by comparing the improved YOLOv7-DSE with YOLOv7, CEAM-YOLOv7, and YOLOv7-RAR, which are three object detection algorithms that perform well in industrial settings. Finally, to verify the effectiveness of improvements made to each part of the network model. A set of ablation experiments is designed to demonstrate the effectiveness of each optimization module.

C. Experimental Results and Analysis

The comparative experimental results are shown in Table I.

(1) Model size comparison

The improved YOLOV7-DSE was compared with the YOLOv7 algorithm, YOLOv4 algorithm, YOLOv5 algorithm, CEAM-YOLOv7 algorithm and YOLOV7-RAR algorithm in terms of model size.

By optimizing and improving depthwise separable convolution, the model size of the improved model has been reduced. The experimental results are shown in Fig.10. The model size of the baseline network model YOLOv7 algorithm is 94.1MB. While the model sizes of YOLOv4, YOLOv5, CEAM-YOLOv7, and YOLOv7-RAR are 103.2MB, 101.6MB, 96.4MB, and 84.3MB, respectively. The model size of the optimized YOLOv7-DSE algorithm is only 72.9MB, which is a reduction of 22.4% compared to the original baseline network model YOLOv7. Demonstrating that the model has achieved lightweight results.

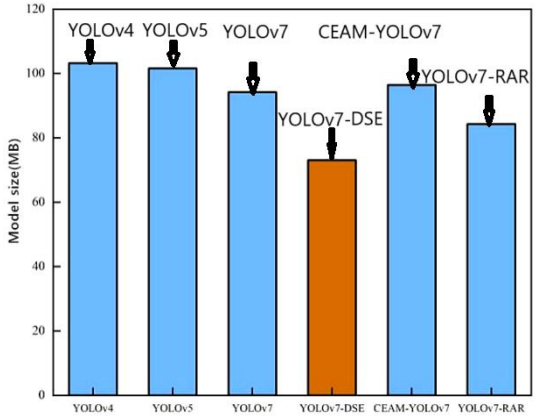


Fig.10. Model Size Comparison Diagram

(2) Precision ratio comparison

Compare the accuracy of the improved YOLOV7-DSE algorithm with that of YOLOv7 algorithm. Fig.11. shows the comparison of precision of YOLOv7 object detection algorithm and the proposed YOLOv7-DSE worker operation safety equipment detection algorithm with confidence. Confidence is a decimal between 0 and 1. Representing the accuracy of all detection results that are less than the confidence level when all equipment under this confidence level are considered operation safety equipment. At the same time, connecting the precision of all confidence levels can clearly show the change of precision with confidence. The black line chart in Figure 10 shows that when the confidence threshold is 0.5, the accuracy rate of YOLOv7 algorithm to detect all categories is 86.23%. The red line chart shows that when the confidence threshold is 0.5, the accuracy of the improved YOLOv7-DSE algorithm to detect all categories is 90.41%.

From the experimental results in Fig.11. It can be observed that as the confidence threshold increases, the overall accuracy of the YOLOv7 algorithm continues to improve. This means that as the criteria for identifying workers' safety equipment become stricter. The algorithm is more capable of locating areas where safety equipment is present. When the confidence threshold is set to 0.5. The accuracies for the two scenarios are 86.23% and 90.41% respectively. This indicates that when the algorithm determines there is a 50% probability of safety equipment being present in a region of the image. It classifies that region as containing safety equipment. The improved YOLOv7-DSE algorithm in this paper can correctly identify 4.18 times per 100 frames compared to the baseline model.

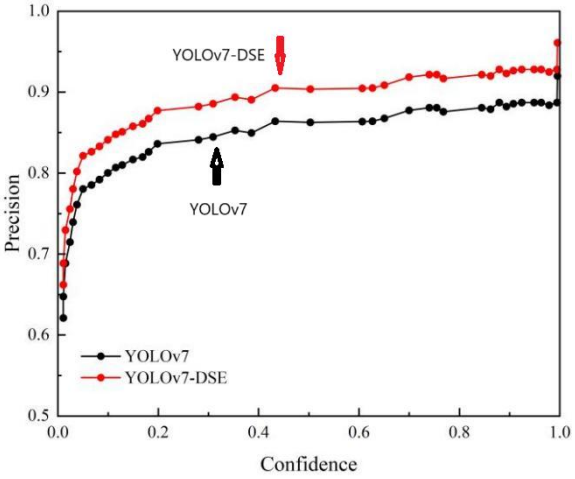


Fig.11. Comparison of Precision Experiment

(3) Recall rate comparison

The recall rate of the improved YOLOV7-DSE and YOLOv7 algorithm was compared. Fig.12. shows the Recall rate of YOLOv7 object detection algorithm and the YOLOV7-DSE worker safety equipment detection algorithm proposed in this study. Image contrast that changes as confidence increases. Changes in recall rates can be observed. The black line chart in Fig.12. shows that when the confidence threshold is 0.5, the recall rate of all categories detected by YOLOv7 algorithm is 90.38%. The red line chart in Fig.12. shows that when the confidence threshold is 0.5, the recall rate of all categories detected by the improved YOLOv7-DSE algorithm is 93.72%.

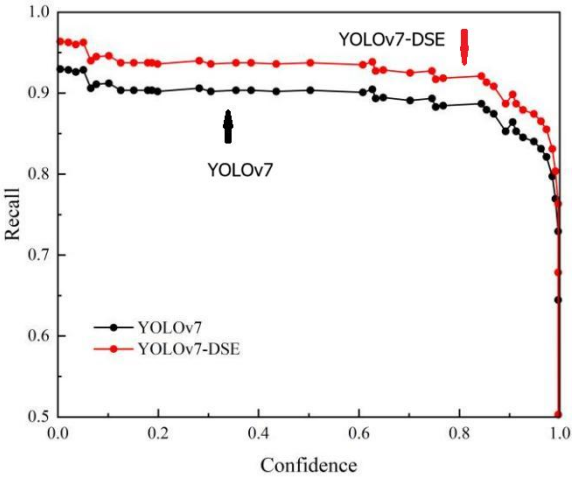
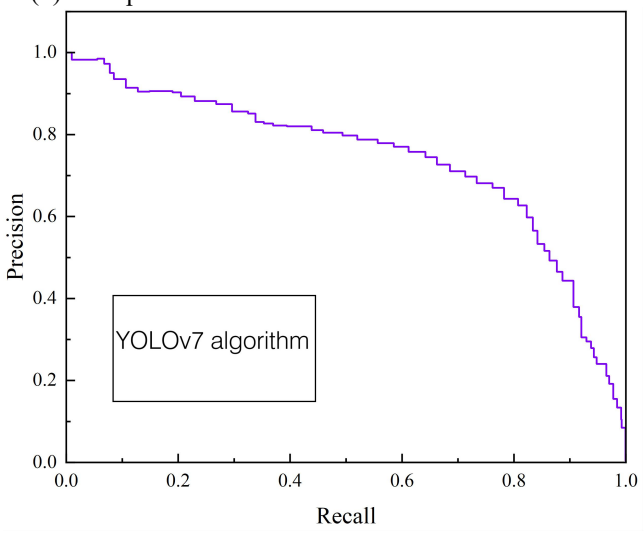


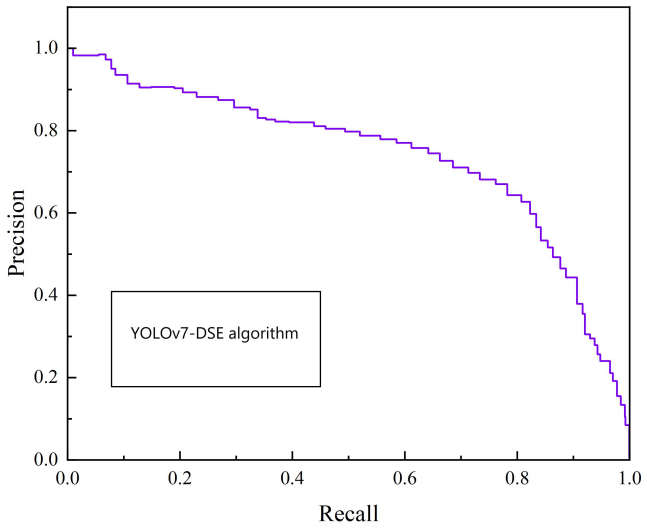
Fig.12. Comparison of Recall Experiment

From the experimental results in Fig.12. With the increase of the confidence threshold, the recall rate of YOLOv7 algorithm continues to decrease on the whole. That is, with the stricter the identification requirements for workers' work safety equipment, the more difficult it is to detect whether there is workers' work safety equipment in the area. When the confidence threshold is set to 0.5, the recall rates of both are 90.38% and 93.72%, respectively. It means that when the algorithm judges that there is 50% probability of workers' operating safety equipment in the area of the image. It will judge that there is operating safety equipment in the area. The improved YOLOv7-DSE algorithm in this paper can recognize 3.34 times more than the baseline model per 100 frames on average.

(4) Comparison with mAP



(a) YOLOv7 algorithm



(b) YOLOv7-DSE algorithm

Fig. 13. PR Curve of Comparison

Compare the improved YOLOV7-DSE algorithm with YOLOv7 algorithm on mAP. It can be seen from fig.11. and fig.12. As confidence grows, so does accuracy, and at the same time, the corresponding recall rate decreases. The PR curve in Fig.13. shows the correspondence between recall rate and accuracy. Connect each corresponding value with a curve. The area of the curve surrounded by the horizontal and vertical positive coordinate axes is used as a comprehensive evaluation index, that is, all kinds of mAP. In Fig.13(a). indicates that the mAP of all categories of workers' work safety equipment detected by YOLOv7 algorithm is 79.75%. In Fig.13(b). indicates that the mAP of all categories of workers' work safety equipment detected by the YOLOv7-DSE algorithm proposed in this study is 82.39%.

The improved YOLOv7-DSE was compared with YOLOv4 algorithm and YOLOv5 algorithm on mAP. Longitudinal comparison with CEAM-YOLOv7 and YOLOv7-RAR on mAP. In this chapter, four other algorithms were tested using the same method. The experimental results are shown in Fig.14. The experimental results show that the mAP of the baseline network model YOLOv7 algorithm is 79.75%. The mAP of YOLOv4 is

72.94%. The mAP of YOLOv5 is 75.75%. The mAP of CEAM-YOLOv7 is 80.44%. The mAP of YOLOv7-RAR is 79.86%. The mAP of the proposed YOLOv7-DSE is 82.39%. Compared with other network models, the improvement is 2.64% compared with the baseline model.

(5) Training loss comparisons

The improved YOLOV7-DSE algorithm is compared with YOLOv7 algorithm in terms of training loss. The changes in training loss of YOLOv7 and YOLOv7-DSE are shown in Fig.15. As can be seen from Fig.15. The training loss of the optimized object detection model YOLOv7-DSE decreases more. The loss reduction during training is also very stable.

(6) Ablation experiments

Table II shows the results of the ablation experiments. This study proposes three optimization modules: MCN-SPD, FN-DSC and MCN-DSC. A checkmark in the table indicates that this module was used in the experiment, while the frozen part was frozen during the experiment. In terms of evaluation indicators. The evaluation indicators with upward arrows indicate that the higher the value of this indicator, the better the algorithm performs. Conversely, evaluation indicators with downward arrows indicate that the lower the value of this indicator, the better the algorithm performs. According to the data in the table. It can be seen that the optimized parts have a specific impact on mAP and model size. When the optimized depthwise separable convolution is frozen. The number of model parameters increases by 22.5%. When the spatial-to-depth non-strided convolution is frozen, mAP decreases by 1.85%, which demonstrates that each improved module is effective.

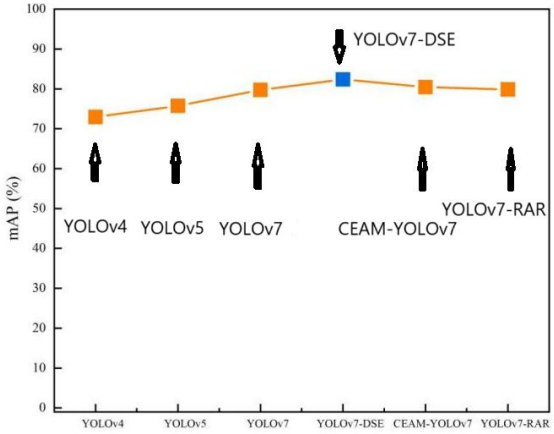


Fig. 14 mAP of Comparison Experiment

D. Model visualization

The baseline network model YOLOv7 algorithm and the improved and optimized YOLOV7-DSE algorithm were used to predict the same group of images respectively. Feel the visual representation ability of the test model. The test results are shown in Fig.16. and 17. To the left of the picture is the original. In the middle is the test result output by YOLOv7. On the right is the test result output by YOLOv7-DSE.

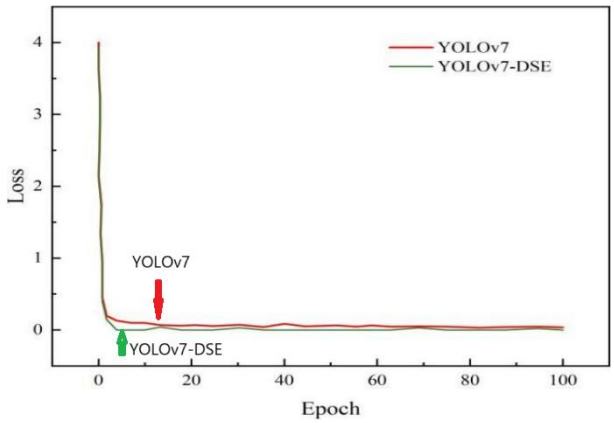


Fig.15. Comparison of training losses

It can be intuitively seen from the figure that the original baseline network model YOLOv7 has poor performance in detecting workers' operating safety equipment. There are a large number of error detection and leakage detection cases, and the accuracy is very low. The improved and optimized YOLOv7-DSE can detect the safety equipment of workers. It not only greatly reduces the probability of error detection in complex scenes. The detection ability of small targets has also been improved.

VI. Conclusion

Through the analysis of the above experimental results. It can be clearly seen that the optimized YOLOv7-DSE algorithm is compared with the original baseline network model YOLOv7. A significant lightweight effect was achieved in the model size, with a reduction of 22.4%. At the same time, under the condition that the confidence threshold is 0.5. The YOLOv7-DSE algorithm achieved 90.41% accuracy in detecting all categories. Compared with 86.23% of YOLOv7, it is significantly improved. On average, every 100 frames of images were correctly identified 4.18 times more than the baseline model. In addition, the recall rate of YOLOv7-DSE algorithm also increased to 93.72%. Compared with 90.38% of YOLOv7, the average recognition rate is 3.34 times per 100 frames. The comparison of training losses also shows that the training losses of YOLOv7-DSE decreased more and more steadily. The three optimization modules proposed in this paper, MCN-SPD, FN-DSC and MCN-DSC, have been proved effective by ablation experiments. In summary, the experimental results fully prove the effectiveness of the model improvement. It shows the excellent generalization ability of the model. Compared with the baseline network model, the error detection rate and missing detection rate are significantly reduced.

TABLE II
Ablation Experiment Results

Baseline YOLOv7	FN-DSC	MCN-DSC	MCN-SPD	Epoch	Batchsize	mAP↑	Model size↓
√				100	16	79.75%	94.1MB
	√			100	16	80.22%	81.9MB
	√	√		100	16	80.54%	70.5MB
	√	√	√	100	16	82.39%	72.9MB

Fig. 16. Model Test Effect Comparison One（person, badge, wrongglove, operatingbar）

Fig. 17. Model Test Comparison Two(person, wrongglove)

REFERENCES

[1] R. Padilla, and F. Costa, “Evaluation of haar cascade classifiers designed for face detection,” Engineering and Technology, vol. 64, no.35, pp. 62-371, 2012.

[2] H. S. Dadi and G. K. M. Pillutla, “Improved face recognition rate using HOG features and SVM classifier,” Journal of Electronics and Communication Engineering, vol. 11, no.4, pp. 34-44, 2016.

[3] C. Ying, G. Qi , and L. Jia , “Advance and prospects of AdaBoost algorithm,” Acta Automatica Sinica, vol. 39, no.6, pp. 745-758, 2013.

[4] H. Yu , S. Kim SVM, “Tutorial-Classification”, Regression and Ranking, Handbook of Natural computing, vol. 1, no.10, pp. 479-506, 2012.

[5] S. K. Pavani, D. Delgado and A. F. Frangi, “Haar-like features with optimally weighted rectangles for rapid object detection,” Pattern Recognition, vol. 43, no.1, pp. 160-172, 2010.

[6] V. Chandrasekhar, G. Takacs and D. Chen, “Compressed histogram of gradients a low bit-rate feature descriptor,” Conference on Computer Vision and Pattern Recognition, Miami, 2009, pp. 2504-2511.

[7] C. Kyrkou, T. Theocharides, “A flexible parallel hardware architecture for AdaBoost-based real-time object detection,” Transactions on Very Large Scale Integration Systems, vol. 19, no.6, pp. 1034-1047, 2010.

[8] Z. Li, F. Liu, and W. Yang, “A survey of convolutional neural networks: analysis, applications, and prospects,” Transactions on Neural Networks and Learning Systems, vol. 32, no.14, pp215-223, 2021.

[9] Z. W. Yuan and J. Zhang, “Feature extraction and image retrieval based on AlexNet,” Eighth International Conference on Digital Image Processing, ChengDu, 2016, pp. 65-69.

[10] A. Giusti, D. C. Cireşan and J. Masci, “Fast image scanning with deep max-pooling convolutional neural networks,” 2013 IEEE International Conference on Image Processing, Melbourne, 2013, pp. 4034-4038.

[11] J. Tang, H. Xia and J. Zhang, “Deep forest regression based on cross-layer full connection,” Neural Computing and Applications, vol. 3, no.15, pp. 9307-9328, 2021.

[12] A. Krizhevsky , I. Sutskever and G E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol. 25, no.3, pp. 112-119, 2012.

[13] P. Doungmala and K. Klubsuwan, “Helmet wearing detection in Thailand using Haar like feature and circle hough transform on image processing,” International Conference on Computer and Information Technology, Yanuca Island, 2016, pp. 611-614.

[14] M.Jin , J.Zhang and X.Chen, “Safety helmet detection algorithm based on color and hog features,” 19th International Conference on

Cognitive Informatics & Cognitive Computing, Beijing, 2020, pp. 215-219.
[15] M.Kocabas , N.Athanasiou and M J.Black “Vibe: Video inference for human body pose and shape estimation,” Conference on Computer Vision and Pattern Recognition, Virtual, 2020, pp. 5253-5263.