Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(5):8685–8707.
DOI: 10.3934/mbe.2023381
Received: 14 December 2022
Revised: 15 February 2023
Accepted: 27 February 2023
Published: 06 March 2023

Research article

Research on application of helmet wearing detection improved by YOLOv4 algorithm

Haoyang Yu[1], Ye Tao[1,*], Wenhua Cui[1], Bing Liu[2] and Tianwei Shi[1]

[1] School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Liaoning, China

[2] School of Electronic and Information Engineering, University of Science and Technology Liaoning, Liaoning, China

* **Correspondence:** Email: taibeijack@163.com; Tel: +8613304224928; Fax: +8604125929818.

Abstract: Aiming at the problem that the model of YOLOv4 algorithm has too many parameters and the detection effect of small targets is poor, this paper proposes an improved helmet fitting detection model based on YOLOv4 algorithm. Firstly, this model improves the detection accuracy of small targets by adding multi-scale prediction and improving the structure of PANet network. Then, the improved depth-separable convolution was used to replace the standard 3 × 3 convolution, which greatly reduced the model parameters without reducing the detection ability of the model. Finally, the k_means clustering algorithm is used to optimize the prior box. The model was tested on the self-made helmet dataset helmet_dataset. Experimental results show that compared with the safety helmet detection model based on Faster RCNN algorithm, the improved YOLOv4 algorithm has faster detection speed, higher detection accuracy and smaller number of model parameters. Compared with the original YOLOv4 model, the mAP of the improved YOLOv4 algorithm is increased by 0.49%, reaching 93.05%. The number of model parameters was reduced by about 58%, to about 105 MB. The model reasoning speed is 35 FPS. The improved YOLOv4 algorithm can meet the requirements of helmet wearing detection in multiple scenarios.

Keywords: safety helmet detection; depthwise separable convolution; multiscale prediction; YOLOv4

1. Introduction

With the rapid progress of society, science and technology get rapid development. Artificial intelligence technology, in particular, has not only changed the way we live, but also improved our life experience. As the environment of traditional industrial manufacturing industry is complex and changeable, the scene is varied, the intensity of light is uneven, and the size of the target to be detected is different, making the application of target detection technology very difficult [1]. Therefore, the development of artificial intelligence technology in industrial scenarios is particularly slow. The wearing of safety helmet is an important measure to effectively protect the head of front-line employees engaged in infrastructure industry and industrial production industry. Because of the lack of safety awareness of employees, wearing safety helmet and other safety measures are not considered. Therefore, there will always be dangerous accidents at industrial production sites and infrastructure construction sites [1]. At present, many work sites have installed cameras for manual monitoring. However, the cost of manual monitoring is too high, and the monitoring results are easily affected by the subjective consciousness and emotions of the staff. Unable to carefully and responsibly monitor the personal safety of front-line employees. Therefore, in this case, helmet wear real-time detection is very important. Because the monitoring process is not affected by the subjective consciousness and emotions of the staff, it can realize real-time monitoring and timely warning.

At present, there are two kinds of detection algorithms for helmet wearing. One is traditional machine learning algorithms. This algorithm uses scale invariant feature transform (SIFT) [2], directional gradient histogram (HOG) [3] and other methods to extract local features of the target. Canny, Prewitt [4] and other methods were used to extract edge texture features for target detection. Li [5] firstly used Vibe algorithm to locate the human body. Then the convex word algorithm is used to locate the header. Finally, the detection of helmet wearing is realized by combining HOG algorithm and Support Vector Machine (SVM). Traditional machine learning algorithms rely on strong prior knowledge and require sufficient understanding of the characteristics of specific objects. It also takes a lot of time to design features. The algorithm model design is not only complex, but also the model detection accuracy is low, the detection speed is slow, especially the robustness and generalization of the poor, can not meet the requirements of real-time detection.

The other is the helmet wearing detection based on convolutional neural network. Because of the strong feature extraction capability of convolutional neural network, the detection algorithm of helmet wearing based on convolutional neural network has been developed. The safety helmet detection algorithm based on convolutional neural network can also be divided into two kinds: one is the two-stage target detection method based on regional suggestion mechanism, the other is the single-stage target detection algorithm based on regression model

- Two-stage target detection method based on region suggestion mechanism

For example, based on region convolutional neural network (RCNN) series algorithm improved helmet wear detection algorithm. The representative method is Faster RCNN algorithm for helmet wearing detection. Its characteristic is relatively high detection accuracy, but the detection speed is very slow. Xu et al. [6] introduced the online difficult sample mining strategy to improve the algorithm Faster RCNN, and combined it with the multi-component algorithm for helmet wearing detection. The detection rate is improved by 7% compared with the original Faster RCNN algorithm. Wu et al. [7] improved Faster RCNN algorithm through multi-feature layer information fusion and multi-scale prediction for helmet wearing detection.

- A single - stage object detection algorithm based on regression model

Such as the use of YOLO series algorithm and SSD algorithm to improve the helmet wear detection algorithm. Wang et al. [8] improved YOLOv3 algorithm for helmet wearing detection by using CSP structure and adding SPP structure. The improved algorithm achieved a mAP of 90% and an FPS of 20. Zhang et al. [9] adopted DenseNet (Dense convolutional network) method to process low resolution feature layer and improved YOLOv3 algorithm for helmet detection. This results in a mAP of 96.5% on a homemade small hardhat dataset. Gu et al. [10] used YOLOv4 algorithm combined with attitude estimation to carry out helmet wearing detection. There are 96.60 percent AP(helmet) tests, but they are slow. Cong et al. [11] adopted the lightweight YOLOv4-Tiny network and increased the attention mechanism for helmet detection. Wang et al. [12] used MobileNet [13] as the feature extraction network to modify SSD network for human body detection and then safety helmet detection. Not only improve the detection accuracy, but also improve the detection speed. Xiao et al. [14] used a lightweight convolutional neural network MobileNetV3-small [15] and multi-scale feature fusion to improve SSD for helmet wearing detection, making the detection speed reach 108 FPS.

At present, helmet wear detection algorithms are verified on self-built small data sets. The detection effect of these algorithms on small targets is not ideal. In addition, most of the improved helmet wear detection algorithms have large number of model parameters. Therefore, the real-time detection algorithm of helmet wearing is still in progress. At present, YOLOv4 algorithm has a very significant target detection effect in other scenarios. Therefore, the wear detection algorithm model of helmet detection in this paper is improved based on YOLOv4. The algorithm model can not only improve the detection accuracy of the model to the greatest extent, but also reduce the number of parameters of the model.

The main work of this paper is as follows. Firstly, the 104 × 104 feature layer is added to form the multi-scale prediction, and the large-scale feature layer is used to predict the small target and strengthen the detection effect of the small target. At the same time, a 4x up-sampling and a 4x down-sampling were added respectively to carry out feature fusion across the feature layer. The PANet structure was improved to enhance the feature fusion capability. Second, using the idea of residual network to improve the depth separable convolution instead of the standard 3 × 3 convolution, to reduce the number of model parameters. Meanwhile, the feature extraction capability of depth-separable convolution is enhanced. Third, the helmet wearing detection model based on the improved YOLOv4 algorithm and Faster RCNN algorithm is compared vertically. Horizontal comparison of helmet wearing detection model based on YOLOv4 algorithm and improved YOLOv4 algorithm. Compare and analyze the experimental results.

2. Introduction of YOLOv4 algorithm

2.1. YOLOv4 network structure

The YOLOv4 algorithm framework has three parts. The backbone feature extraction network CSPDarknet53, the depth feature extraction network PANet structure and the prediction layer network YOLO head. The YOLOv4 algorithm framework is shown in Figure 1.

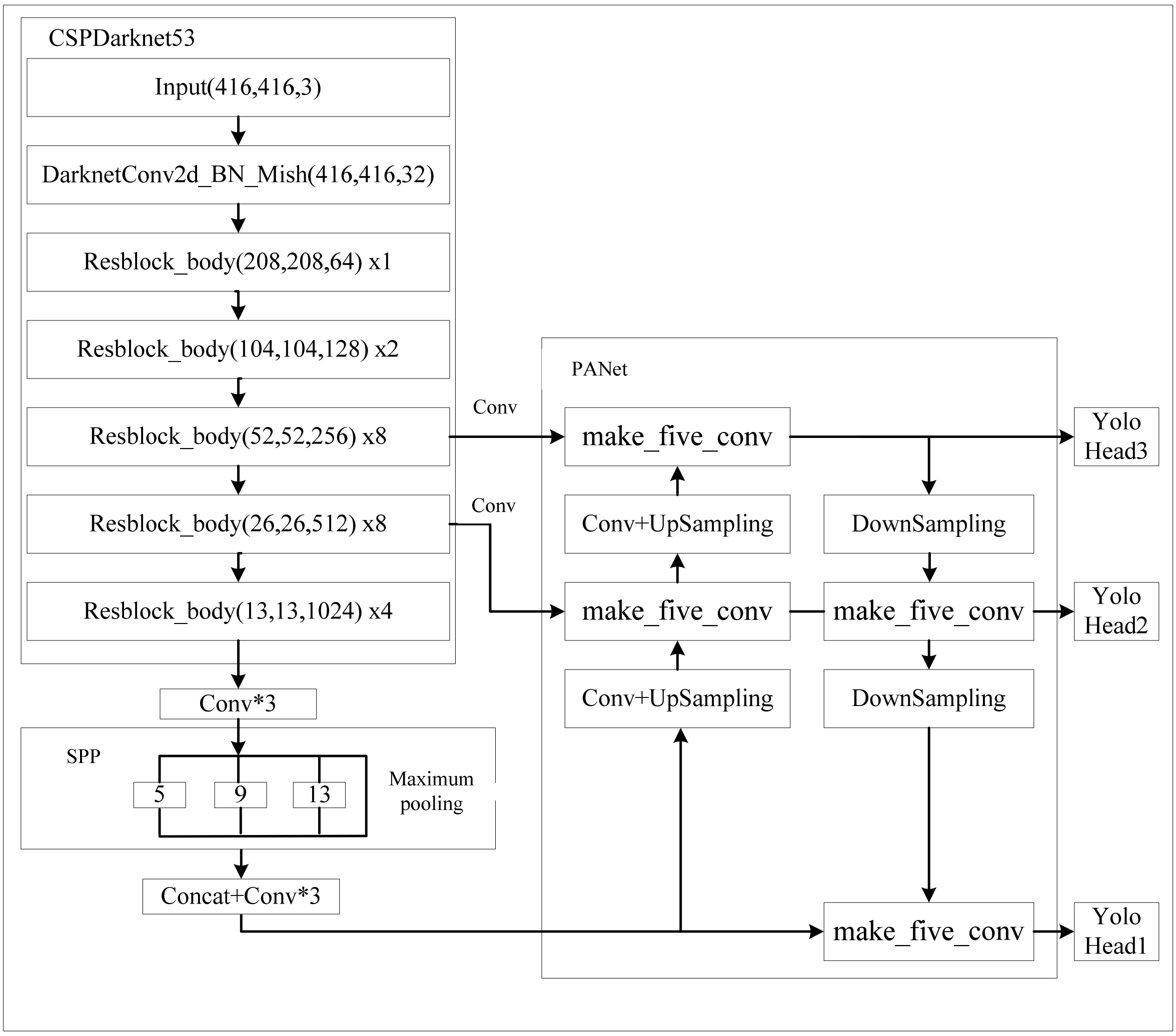


Figure 1. Structure diagram of YOLOv4 algorithm.

2.2. YOLOv4 model prediction

The YOLOv4 model prediction steps are as follows. First, the input image was divided into $S\times S$ grids, and each grid was pre-allocated with B anchor boxes. Then, each grid cell needs to do model forward reasoning by training model weights. Calculate the adjustment parameters of the center coordinates of each prediction box and the adjustment parameters of the width and height of each prediction box (t_x, t_y, t_w, t_h). At the same time, it is necessary to predict whether the adjusted anchor box contains targets and which targets. Therefore, the output dimension of the model prediction network in the training stage is $S \times S \times B\times (4 + 1 + C)$. Where, 4 represents the center coordinates of the prediction box and the four adjustment parameters (t_x, t_y, t_w, t_h). 1 indicates whether the prediction box contains the confidence box_score of the target. The confidence calculation formula of whether the target is included is shown in Eq (5). C represents the number of prediction target categories, and each number represents the probability of which target the prediction box belongs to. In this paper, there are only two helmet and head targets, so C is 2. Therefore, the output dimension of the model in the training stage is $S \times S \times B \times (4 + 1 + 2)$.

The generation process of the model prediction box is shown in Figure 2. Assuming that the input image is divided into 3 × 3 grid cells, if the center of a category target happens to fall in the middle grid cell. Then the target is predicted by the *B* prior box anchor box in the upper left corner of the grid cell. The dotted black rectangle box represents anchor box, while the solid red rectangle box represents the model prediction box obtained after the adjustment parameters predicted by the model were scaled and shifted through anchor box. The solid green rectangle represents the actual tag box, that is, it contains the tag target. The prediction accuracy of the model was judged by comparing the overlap between the predicted rectangle and the real rectangle. The adjustment formulas of the model prediction box are (1)–(4), where c_x, c_y, c_w and c_h respectively represent the central coordinates and width and height of the prior frame anchor box. b_x, b_y, b_w, b_h represent the center coordinates and width and height of the target prediction box respectively. t_x, t_y, t_w and t_h are respectively the adjustment parameters of the prediction box obtained by the model through prediction. $\sigma\left(t_x\right)$ indicates the Sigmoid activation function, which fixes the adjustment parameters between 0 and 1. This is why the center of the prediction target falls in the grid cell, and the prior box in the upper left corner of the grid cell is the anchor box responsible for the prediction.

$$b_x = \sigma\left(t_x\right) + c_x \tag{1}$$

$$b_y = \sigma\left(t_y\right) + c_y \tag{2}$$

$$b_w = p_w e^{t_w} \tag{3}$$

$$b_h = p_w e^{t_h} \tag{4}$$

$$\text{box\_score} = Pr(\text{project}) \times IOU(\text{b,object}) \tag{5}$$

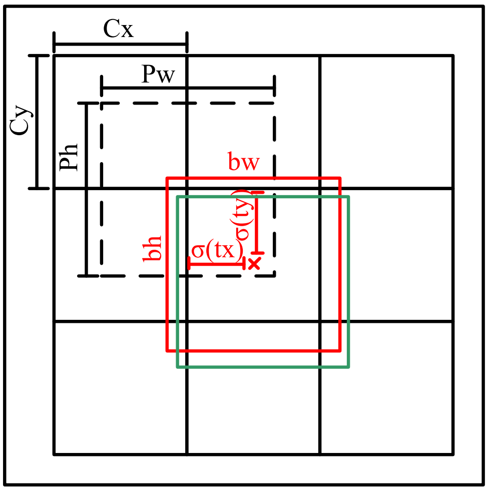


Figure 2. Schematic diagram of prediction box regression.

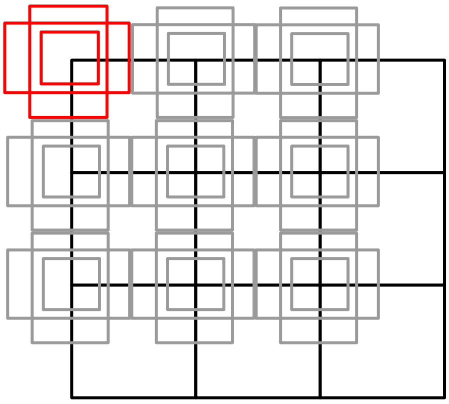


Figure 3. Anchor box distribution diagram of feature maps.

If three prior frames anchor box are allocated to each grid cell, the prior box template distribution of the 3 × 3 scale feature graph is shown in Figure 3. Setting too much anchor box will affect the detection speed of the model, while setting too little anchor box will affect the recall rate of the model. Feature maps of different scales assign anchor boxes of different sizes. Large-scale feature map allocates small-scale anchor box, and small-scale feature layer allocates large-scale anchor box. In this paper, each grid cell is allocated three anchor boxes. Therefore, the output dimension of each feature layer of the algorithm in this paper is $S \times S \times 3\times (4 + 1 + 2)$.

3. Improved YOLOv4 algorithm

YOLOv4 target detection algorithm is one of the most excellent target detection algorithms at present, but the application of YOLOv4 algorithm in real-time helmet wearing detection has the following two serious problems. On the one hand, the recall rate and detection accuracy of helmet with small target is low. On the other hand, the algorithm model carries a large number of parameters, which is not conducive to the deployment of front-end devices. Therefore, in order to solve the above two problems, this paper firstly makes small target prediction by adding 104 × 104 large-scale feature layer. At the same time, a 4-fold up-sampling and a 4-fold down-sampling were added respectively for cross-feature layer information fusion to improve the PANet structure. Secondly, an improved depth-separable convolution is proposed to improve the PANet structure, and the CSPDarknet53 network structure is improved by depth-separable convolution. On the premise of ensuring the feature extraction ability of the model as much as possible, the model parameters were reduced and finally the anchor box was optimized by clustering algorithm.

3.1. Improve PANet network structure

There are minimal targets for helmet wear detection in metallurgical workshops and construction sites. In order to enhance the detection effect of small targets in this scene, improve the recall rate and detection accuracy of the overall hard hat. In this paper, a layer of 3 × 3 convolution is added after the fourth convolution layer of CSPDarknet53, the backbone feature extraction network of YOLOv4 algorithm, which is used to output 104 × 104 feature layer and detect minimal hard hat targets. For the trunk feature extraction network with the input image size of 416 × 416, the final output of 4 feature layers. The dimensions of the characteristic layers are 13 × 13, 26 × 26, 52 × 52 and 104 × 104, respectively. However, with the addition of another 104×104 large-scale feature layer, the PANet structure of feature fusion network should be deepened accordingly.

PANet structure is composed of FPN network and PAN network. FPN network is up-sampled by the feature layer, and then spliced with the feature layer of corresponding size output by the main feature extraction network to achieve the purpose of feature fusion. The FPN structure is designed to enhance the semantic information missing from the shallow feature layer. In PAN network, feature layer is subsampled by convolution and spliced with feature layer of corresponding scale output of FPN network to achieve the purpose of feature fusion. The PAN structure is designed to enhance positional information that is lacking in deep feature maps.

In this paper, the YOLOv4 algorithm was improved into four layers of feature layers for model prediction, resulting in four layers of feature information for feature fusion in PANet network, and up-sampling and down-sampling were carried out three times respectively. In order to prevent excessive up-sampling and down-sampling times from damaging feature information and preventing feature loss, a 4-fold up-sampling was added to the FPN structure and a 4-fold down-sampling was added to the PAN structure. Feature fusion across feature layers is carried out respectively to ensure the richness of feature information fusion. The improved PANet network structure is shown in Figure 4. The YOLOv4 algorithm improved by this step is named YOLOv4_4out.

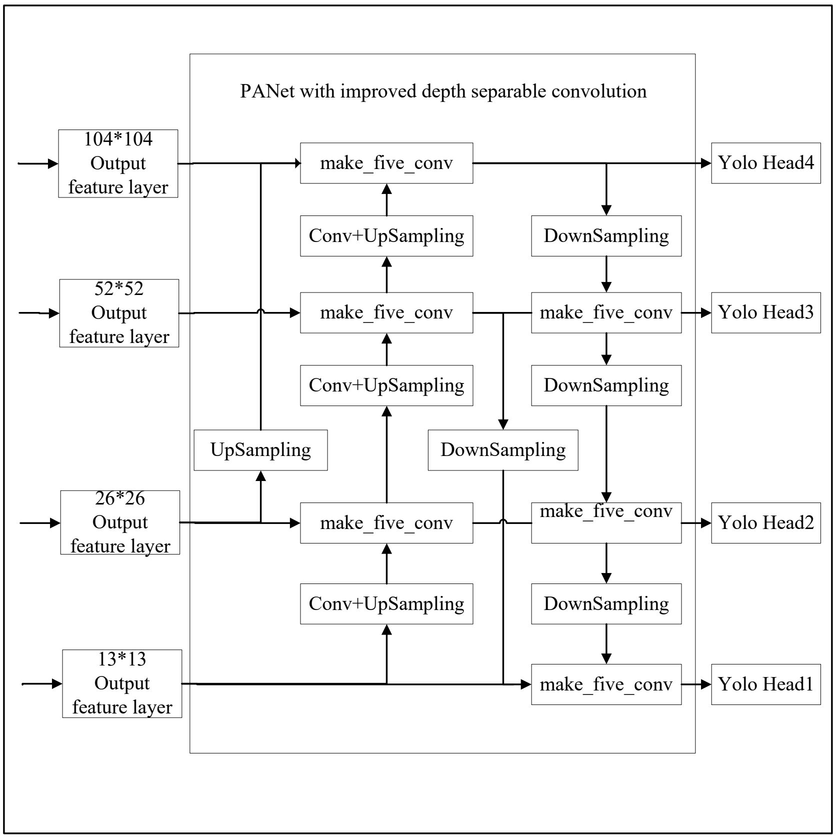

PANet with improved depth separable convolution
104*104 Output feature layer
52*52 Output feature layer
26*26 Output feature layer
13*13 Output feature layer
make_five_conv
Conv+UpSampling
UpSampling
DownSampling
Yolo Head4
Yolo Head3
Yolo Head2
Yolo Head1

Figure 4. PANet with improved depth separable convolution.

3.2. Improved depthwise separable convolution

In Section 3.1, in order to increase the recall rate and detection accuracy of small targets, a large-scale feature layer is added for output by increasing the width of prediction model structure. At the same time, the feature fusion structure PANet network is improved for more sufficient feature fusion. Although these improvement measures can enhance the feature fusion ability and the detection effect of small targets, they will inevitably lead to the increase of the number of model parameters. At the same time, there are a large number of standard 3 × 3 convolution cores in YOLOv4 algorithm, and using standard 3 × 3 convolution for feature extraction will lead to a large number of parameters carried. In order to reduce the number of parameters in the model, the memory space occupied by the model is reduced. Therefore, in this paper, deep separable convolution will be used to improve the structure of CSPDarknet53 network, and the improved deep separable convolution will be used to improve the structure of PANet network. In this way, the model parameters can be reduced and the detection ability of the model can be maintained.

3.2.1. Depth separable convolution

Depth separable convolution is composed of one channel convolution and one point convolution. Channel by channel convolution uses 3 × 3 convolution kernel and does not change the number of channels after convolution. It carries out convolution operation in a two-dimensional plane and cannot fuse the feature information of different channels in the same position. Point by point convolution is a 1 × 1 convolution operation, after convolution does not change the size of the feature layer, but changes the number of channels. Its essence is to fuse the information of different channels at the same location. The depth-separable convolution process has 4 channels for input and 2 channels for output, as shown in Figure 5. The YOLOv4 algorithm uses standard 3 × 3 convolution, with 4 channels for input and 2 channels for output, as shown in Figure 6.

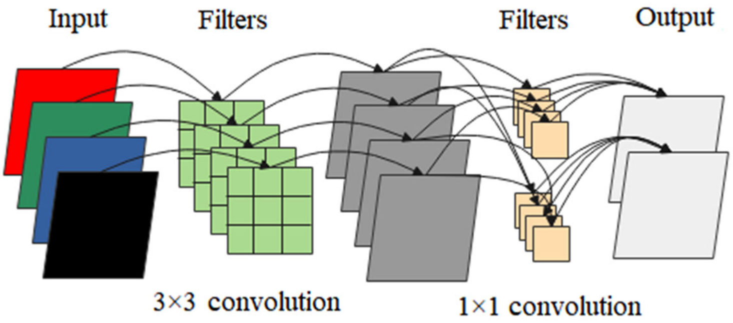


Figure 5. Schematic diagram of depthwise separable convolution.

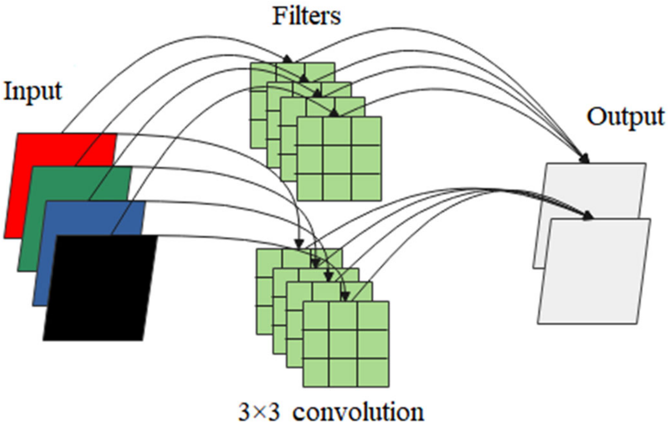


Figure 6. Schematic diagram of standard 3 × 3 convolution.

The parameters of the convolutional neural network model are mainly concentrated in the convolution kernel. According to the diagram of standard 3 × 3 convolution and depth-separable convolution, calculate the number of parameters involved in two different convolution modes. Formula (6) is the number of parameters to calculate the standard 3 × 3 convolution, and formulas (7) is the number of parameters to calculate the depth separable convolution. Where $D_k \times D_k$ is the size of the convolution kernel, M is the number of input channels, and N is the number of output channels. The ratio of the number of depth-separable convolution parameters to the number of conventional convolution parameters is shown in formulas (8), indicating that the number of model parameters can be compressed to $\frac{1}{\mathrm{N}}+\frac{1}{\mathrm{D}_k^2}$ times of the standard 3 × 3 convolution kernel of the original model by using depth-separable convolution. In the deep learning object detection model, the convolution kernel size D_k used is usually equal to 3, and the number of channels N in the convolution process is usually [32, 64, 128, 256, 512, 1024]. It can be seen that the depth separable convolution, instead of the standard 3 × 3 convolution, can reduce the number of parameters of the model by about 88% and greatly reduce the memory storage space of model parameters.

$$F_1 = D_k \times D_k \times M \times N \tag{6}$$

$$F_2 = D_k \times D_k \times M \times M \times N \times 1 \times 1 \tag{7}$$

$$\frac{F_2}{F_1} = \frac{D_k \times D_k \times M \times M \times N \times 1 \times 1}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{1}{D_k{}^2} \tag{8}$$

3.2.2. Improved depth separable convolution

According to the analysis in Section 3.2.1, it can be seen that the replacement of standard 3 × 3 convolution by deep separable convolution can significantly reduce the number of model parameters. However, according to the structure of depth-separable convolution, it can be seen that depth-separable convolution does not fuse the feature information between channels in real time, because depth-separable convolution cannot connect the feature time in the depth direction with the feature time in the width direction. Depth separable convolution The first 3×3 convolution only operates on a two-dimensional plane, meaning that information between channels cannot be fused in real time. The second 1 × 1 convolution will only fuse the characteristic information of different channels at the same position. This makes the feature fusion in the depth direction and the feature fusion in the width direction out of sync. This convolution method reduces the feature extraction capability of depth-separable convolution.

In order to solve the problem of insufficient feature extraction ability caused by the depth separable convolution structure, an improved depth separable convolution structure was proposed combined with the idea of residual. This can enhance the feature extraction capability of depth-separable convolution. The structure of the improved depth-separable convolution is shown in Figure 7. By analyzing the improved depth-separable convolution structure, it is clear that the structure brings more parameters than the depth-separable convolution, but less than the standard 3 × 3 convolution. The improved formula for calculating the depth separable convolution parameters is shown in Formula (9), which is about twice the depth separable convolution, and the formula for calculating their ratio is shown in Formula (10).

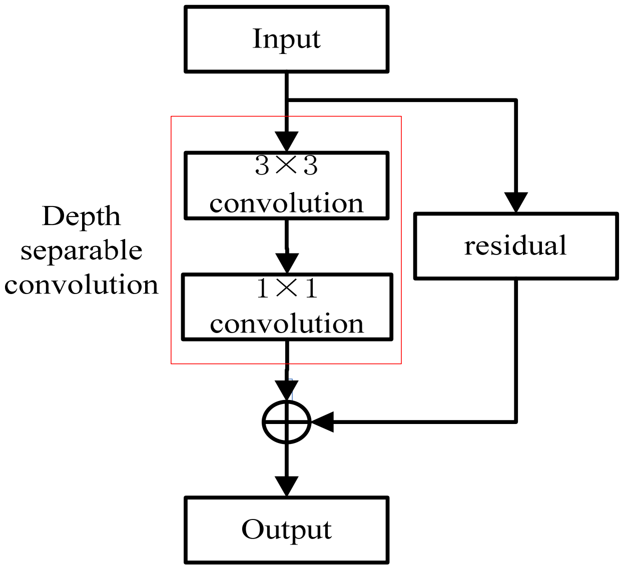


Figure 7. Improved depth separable convolution.

$$F_3 = D_k \times D_k \times M + M \times N \times 1 \times 1 + M \times N \times 1 \times 1 \tag{9}$$

$$\frac{F_3}{F_1} = \frac{D_k \times D_k \times M + M \times N \times 1 \times 1 + M \times N \times 1 \times 1}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{2}{D_k{}^2} \tag{10}$$

The CSPDarknet53 network of YOLOv4 algorithm already contains a large number of residual structures. Therefore, the backbone feature extraction network only uses standard depth-separable convolution instead of standard 3 × 3 convolution to reduce the parameters of the model. However, the structure of PANet feature fusion network is at the back end of the whole model, and the construction of PANet network is mainly serial stack standard 3 × 3 convolution. If depth-separable convolution is only used to replace standard 3 × 3 convolution, the feature extraction capability of depth-separable convolution is inferior to that of standard 3 × 3 convolution, and the model will lose feature information in the process of deepening convolution. In order to enhance the feature extraction capability of PANet network, the improved depth separable convolution was used to replace the standard 3 × 3 convolution, so as to improve the structure of PANet network and enhance the feature fusion capability of PANet network. The overall structure of the improved YOLOv4 algorithm is shown in Figure 8.

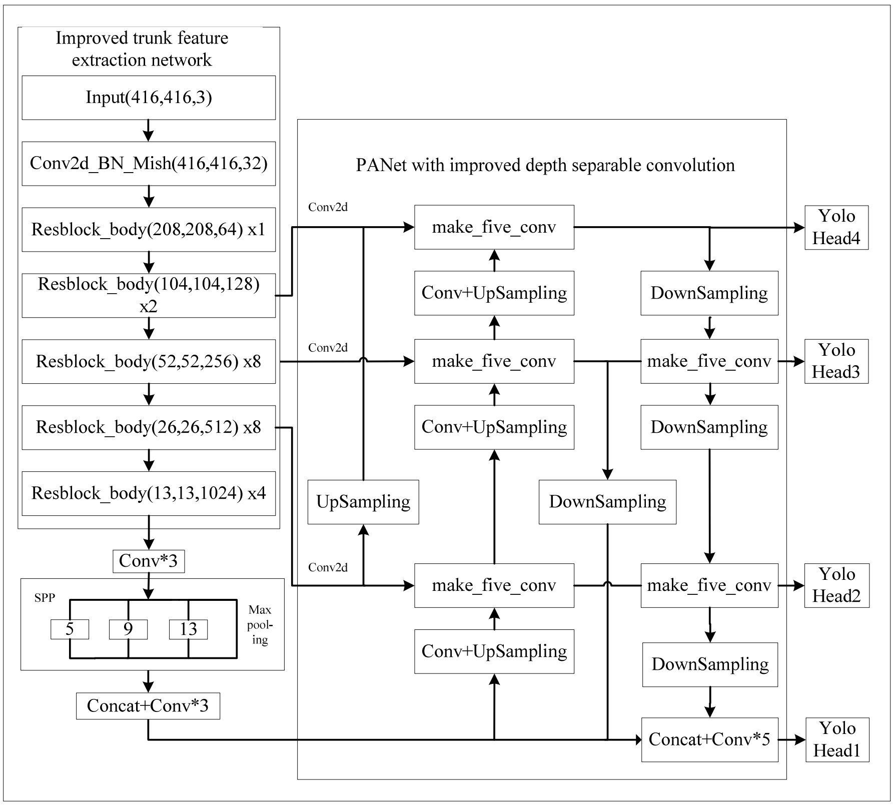


Figure 8. Improved YOLOv4 algorithm framework.

3.3. Optimize anchor box

The regression of the maximum pooled prediction box was completed based on the scale of anchor box. If the scale of anchor box does not conform to the scale of the real target, the scale of the prediction box obtained based on the adjustment of anchor box must not match the real marked box well. This will affect the loss calculation of the model and then affect the convergence of the model during training. Therefore, the scale size of anchor box can be obtained from the real box labeled by the data set itself and the clustering algorithm. The size of anchor box will determine the rationality of the model prediction box to some extent. In this paper, K-means clustering algorithm, an unsupervised learning algorithm, is used to obtain the prior frame anchor box. The steps to obtain a prior box are as follows.

1) K real boxes are randomly selected from all real boxes as the initial clustering center prior boxes.

2) Calculate the distance between each real box and K initial center prior boxes in 1), and use

IOU to express the distance between the two boxes instead of Euclidean distance *d*. The distance formula is (11).

$$d = 1 - IOU \tag{11}$$

3) The distance *d* between each real box and all initial center prior boxes is calculated by 2). The actual boxes that are less than a certain threshold are left to indicate that they belong to the same category.

4) After 3) to determine the category of each real box. Then the real boxes in the same category are sorted by their width-height ratio. Take the real box corresponding to the intermediate value as the new initial center prior box. Follow this approach to update all real boxes.

5) Repeat steps 2), 3) and 4) until the initial center prior box size of K categories no longer changes, the clustering algorithm ends. Output the width and height of K prior boxes.

When self-built helmet data set is used to study helmet wearing detection algorithm, categories 9, 8, 12 and 16 can be tried according to the number of clustering categories. The clustering accuracy table is shown in Table 1. The calculation amount and clustering accuracy of the algorithm are combined, and the prior box with 12 types of clustering results is selected. The accuracy reached 79.13%. The 12 categories of prior boxes obtained by clustering are evenly distributed to the four feature layers of the improved algorithm, and the results are shown in Table 2.

Table 1. Clustering accuracy table.

number of categories	9 categories	8 categories	12 categories	16 categories
match with the original data set	75.90%	76.97%	79.13%	81.21%

Table 2. Distribution table of prior boxes on feature layers.

characteristic layer size	104 × 104	52 × 52	26 × 26	13 × 13
HM_DT	7,12	13,16	24,28	42,46
dataset	10,19	18,22	28,39	53,74
anchor box	13,26	19,34	32,60	82,118
detection object	very small target	small target	middle target	big target

4. Comparative experiment of helmet wearing detection algorithm

4.1. Helmet data set

At present, there is no open source large data set to support the algorithm research of real-time helmet wear detection. Therefore, in order to support the application research of the helmet wearing detection algorithm in this paper, this paper integrates 5000 open source small data sets, 2662 video collection data under the scene, 1000 network collection data, a total of 8662 image data. In this paper, the self-built helmet dataset was named helmet_dataset. There were 20,684 targets wearing helmets and 27,506 targets without helmets. Part of the data set of the training model is shown in Figures 9, 10 and 11. The data sets were divided into training sets (80%) and test sets (20%).

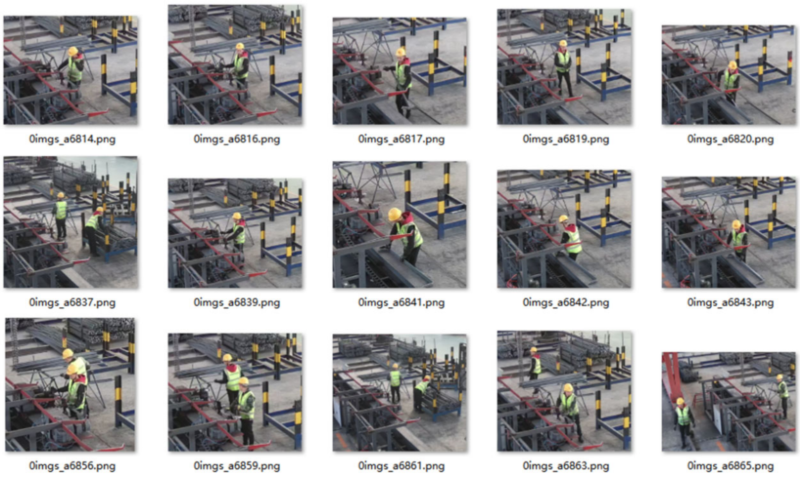


Figure 9. Part I of the data set.

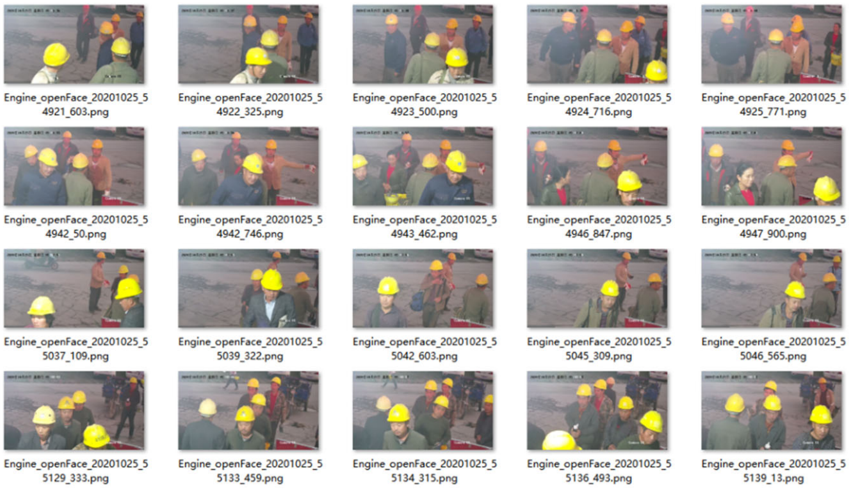

Engine_openFace_20201025_54921_603.png
Engine_openFace_20201025_54922_325.png
Engine_openFace_20201025_54923_500.png
Engine_openFace_20201025_54924_716.png
Engine_openFace_20201025_54925_771.png
Engine_openFace_20201025_54942_50.png
Engine_openFace_20201025_54942_746.png
Engine_openFace_20201025_54943_462.png
Engine_openFace_20201025_54946_847.png
Engine_openFace_20201025_54947_900.png
Engine_openFace_20201025_55037_109.png
Engine_openFace_20201025_55039_322.png
Engine_openFace_20201025_55042_603.png
Engine_openFace_20201025_55045_309.png
Engine_openFace_20201025_55046_565.png
Engine_openFace_20201025_55129_333.png
Engine_openFace_20201025_55133_459.png
Engine_openFace_20201025_55134_315.png
Engine_openFace_20201025_55136_493.png
Engine_openFace_20201025_55139_13.png

Figure 10. Part II of the data set.

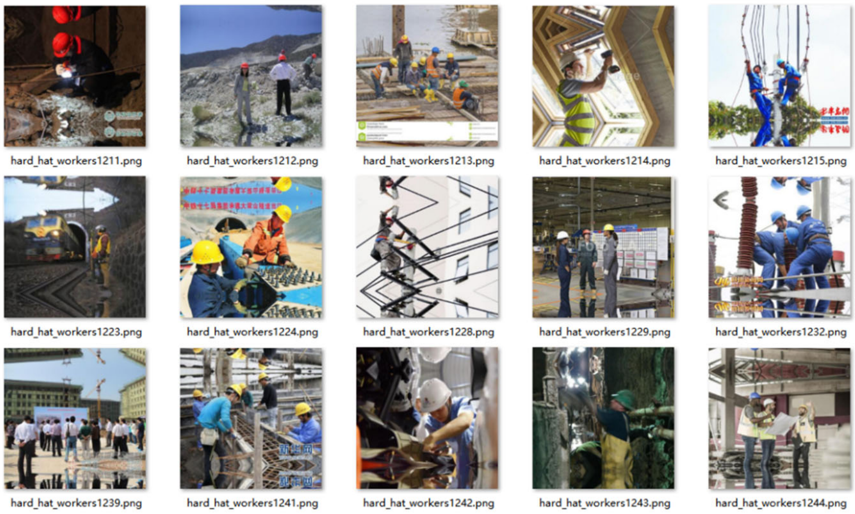


Figure 11. Part III of the data set.

4.2. Experimental process and analysis

4.2.1. Experimental environment

The experimental environment configuration of this paper is shown in Table 3.

Table 3. Experimental environment.

Experimental environment	Configuration
Operating System	Windows10
CPU	I7 10700F
GPU	NVIDIA RTX3070 8G
GPU Library of acceleration	CUDA11.1，cuDNN8.0.4
Language	Python3.6
Compiler	Pycharm2017.2.7
Deep learning framework	Pytorch1.8.0

A total of 4 groups of experiments were designed in this paper. It includes the classic two-stage Faster RCNN algorithm, the classic single-stage YOLOv4 algorithm, YOLOv4_4out algorithm and the improved YOLOv4 algorithm for helmet wearing detection. The four experiments were conducted model training, verification and testing on the helmet_dataset. The YOLOv4_4out algorithm represents the model after adding 104 × 104 large-scale feature layer on the basis of the original YOLOv4 algorithm for model prediction, and adding one up-sampling and one down-sampling respectively to improve the PANet network. The improved YOLOv4 algorithm represents a progressive improved model based on the YOLOv4_4out algorithm by using depth-separable convolution to reconstruct make_five_conv structure. And the Faster RCNN algorithm and the YOLOv4 series algorithm used the transfer learning method in order to accelerate the convergence speed of the model in the training stage. Its pre-training model is obtained from training in the COCO data set.

4.2.2. Experimental main process and result analysis

In the experimental design process of the whole algorithm, the Faster RCNN algorithm was used for model training on the self-built safety helmet dataset helmet_dataset. In 0-10 epochs, the learning rate of the model is set to 10-4, the Batch size is set to 8, and Adam is selected as the optimizer. Freeze the trunk feature extraction network and RPN network part of the model, and only update the unfrozen part of the network parameters. In 10-110 epochs, the learning rate of the model is set to 10-5, the Batch size is set to 4, and all parameters of the whole network model are involved in training and updating. The experimental results show that the change curves of model training loss and model validation loss tend to be stable and converge around the 100th epoch of the model. The changes of model training loss and validation loss function are shown in Figure 12.

When the best model obtained after training is tested on the test set, the mAP of the algorithm only reaches 84.53%. The AP values of helmet and head classes are shown in Figure 13. Through testing, the detection speed of the Faster RCNN model is only 25 frames per second, but the number of parameters of the model is as high as 530.37 MB. Therefore, it can be concluded that the two-stage Faster RCNN algorithm based on the safety helmet detection accuracy is poor, the number of parameters in the model is large, the detection speed is slow, and it is not suitable for real-time detection on the mobile side.

Therefore, this paper does not use the two-stage target detection algorithm to carry out the application research of helmet wearing detection. The YOLOv4 algorithm of single-stage target detection with better detection speed and accuracy is used for application research.

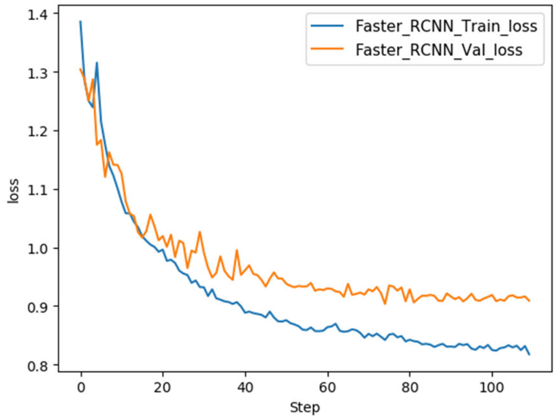


Figure 12. Comparison diagram of training loss and verification loss of safety helmet detection algorithm based on Faster RCNN.

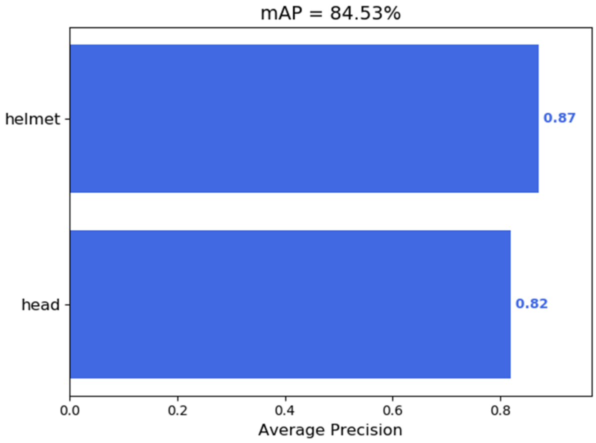


Figure 13. mAP of helmet detection algorithm based on Faster RCNN.

YOLOv4, YOLOv4_4out and the improved YOLOv4 algorithm all freeze the model parameters of the CSPDarknet53 structure at the epoch of 0-10. batch_size is set to 8, and the initial learning rate is set to 10-4. At 10-60 epochs, batch_size is set to 2 (the limit of the hardware device), the initial learning rate is set to 10^{-4}, and the ownership of the training model is heavy parameter. Set batch_size to 2, initial learning rate to 10^{-5} and ownership heavy parameters of the training model at 60-120 epochs. The learning rate was dynamically adjusted by cosine annealing algorithm in the training process, and Adam optimizer was selected to optimize the model parameters. The image data size of the input model is set to 416 × 416. Through a large number of experiments, the training loss change curve of YOLOv4 algorithm, YOLOv4_4out algorithm and the improved YOLOv4 algorithm in the training process is shown in Figure 14. The change curve of verification loss of these three models in the training process is shown in Figure 15.

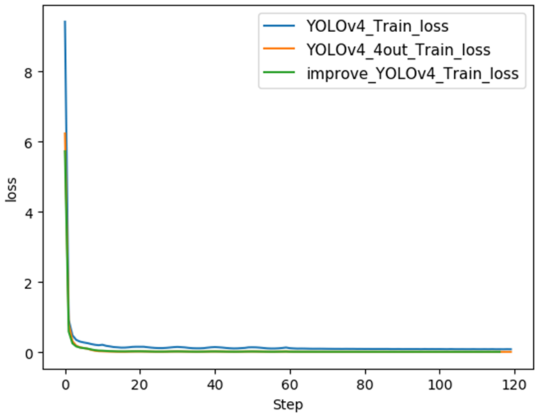


Figure 14. Comparison of training losses.

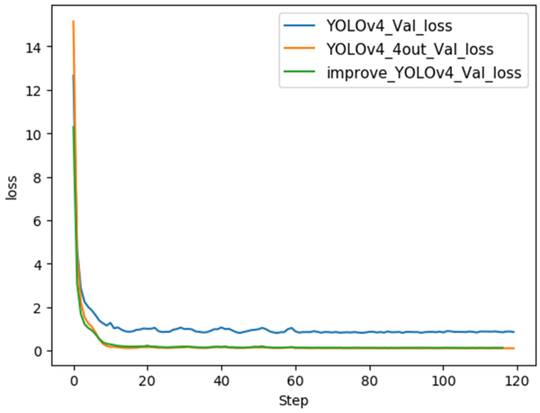


Figure 15. Verify the loss comparison diagram.

It can be analyzed from Figure 14 that in the training stage, the loss of YOLOv4_4out algorithm decreases more than that of YOLOv4, and the loss convergence is faster and more stable. It can be analyzed from Figure 15 that the validation loss of YOLOv4_4out algorithm decreases rapidly and steadily, and the loss value is significantly smaller after stabilization. It shows that adding 104×104 large-scale feature layer for model prediction and improving PANet network structure is conducive to fast convergence of the model and can reduce the loss value of the model. Thus, it is indirectly proved that the improvement in this paper can increase the recognition effect of small targets.

It can also be analyzed from Figure 14 and Figure 15 that the loss curve of the improved YOLOv4 algorithm remains basically unchanged compared with that of YOLOv4_out, no matter in terms of training loss or validation loss. It means that by improving the depth separable convolution to improve the YOLOv4_4out algorithm, the number of model parameters can be reduced, but the loss changes of the model in the training stage and the convergence of the model in the training will not be affected. It shows that the function of the improved depth separable convolution structure reaches the expected goal. In other words, make_five_conv structure was improved by improving depth-separable convolution, so that model parameters could be greatly reduced, but the feature extraction capability of the model was not reduced.

The above is a comparison of the performance changes shown by the improved YOLOv4 algorithm, YOLOv4 algorithm and YOLOv4_4out algorithm in the training stage. Through the analysis of the model in the training stage and the verification stage of the loss change magnitude and the trend of loss change to analyze the advantages and disadvantages of the model.

The following is an analysis of the performance of the best model obtained by each algorithm after training on the test set. The AP curves obtained by the model on the same test set were compared and analyzed. The AP curves of helmet and head of the YOLOv4 algorithm are shown in Figure 17, the AP curves of helmet and head of the YOLOv4_4out algorithm are shown in Figure 16, and the AP curves of helmet and head of the improved YOLOv4 algorithm are shown in Figure 18.

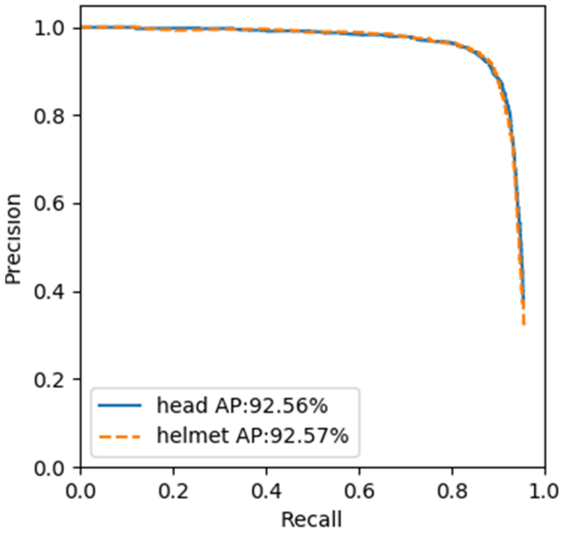


Figure 16. AP curve of YOLOv4 algorithm.

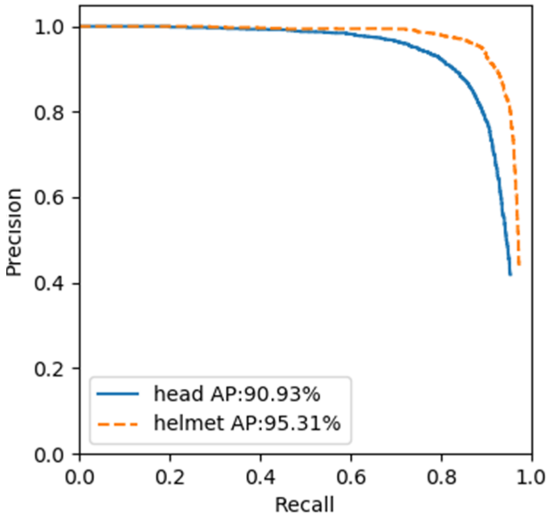


Figure 17. AP curve of YOLOv4_4out algorithm.

According to the analysis in Figures 16, 17 and 18, the mAP of the improved YOLOv4 algorithm model is about 93.05%, the mAP of the YOLOv4_4out algorithm model is 93.12%, and the mAP of the YOLOv4 algorithm model is about 92.56%. From the mAP indexes of the three models on the test set, it is obvious that the improved YOLOv4 algorithm and YOLOv4_4out algorithm both perform better in mAP than the YOLOv4 algorithm. Meanwhile, the AP values of the improved YOLOv4 algorithm model are closer than those of the two classes of the YOLOv4_4out algorithm. This indicates that the average prediction effect of the model is better in the two categories. The detailed index pairs of 4 groups of algorithms are shown in Table 4.

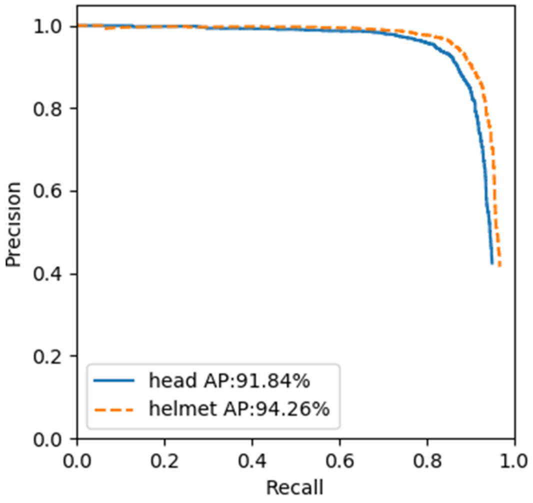


Figure 18. AP curve of improved YOLOv4 algorithm.

According to the analysis in Figure 16, 17 and 18, the mAP of the improved YOLOv4 algorithm model is about 93.05%, the mAP of the YOLOv4_4out algorithm model is 93.12%, and the mAP of the YOLOv4 algorithm model is about 92.56%. From the mAP indexes of the three models on the test set, it is obvious that the improved YOLOv4 algorithm and YOLOv4_4out algorithm both perform better in mAP than the YOLOv4 algorithm. Meanwhile, the AP values of the improved YOLOv4 algorithm model are closer than those of the two classes of the YOLOv4_4out algorithm. This indicates that the average prediction effect of the model is better in the two categories. The detailed index pairs of 4 groups of algorithms are shown in Table 4.

Table 4. Comparison table of model test performance index.

Experiment	Model	mAP(%)	AP(%)	FPS	Model size (MB)
1	Faster RCNN	84.53	87.08	24	530.37
2	YOLOv4	92.56	92.56	**45**	244.42
3	YOLOv4_4out	**93.12**	**95.31**	39	248.39
4	Improved YOLOv4	93.05	94.26	36	**104.32**

It can be clearly seen from Table 4 that the performance indexes of experiment 2, experiment 3 and experiment 4 are all better than those of experiment 1. In conclusion, the single-stage target detection algorithm YOLOv4 is far better than the two-stage Faster RCNN algorithm for helmet wearing detection. Therefore, the single-stage target detection algorithm YOLOv4 is adopted in this paper for improvement. The improved YOLOv4 algorithm also achieves the expected goal. The improved YOLOv4 algorithm also achieves the expected goal. In the case that the indicators do not lag behind, greatly reduce the storage space.

Table 5 shows the detailed index comparison of the recognition accuracy and recall rate of the improved YOLOv4 algorithm, YOLOv4_4out algorithm and YOLOv4 algorithm on targets of different sizes. COCO data sets are defined differently for different size targets. The predicted target area less than 32 × 32 is identified as small target, the area between 32 × 32 and 96 × 96 as medium target, and the area larger than 96 × 96 as large target. However, in the actual application scenario, it is usually more inclined to use the ratio of the original diagram to define. That is, the product of the length and width of the object label box, divided by the product of the length and width of the whole image, and then take the square root. If the result is less than 5%, call it a small goal. If the result is between 5% and 15%, call it a medium target. If the result is greater than 15%, call it a big goal.

Table 5. The comparison table of average accuracy and average recall rate of the three models for targets of different sizes.

Model evaluation index	YOLOv4 (%)	YOLOv4_4out (%)	Improved YOLOv4 (%)
Average Precision (area = all)	92.56	**93.12**	93.05
Average Precision (area = small)	80.4	**82.88**	82.80
Average Precision (area = medium)	87.45	87.46	**87.47**
Average Precision (area = large)	**96.56**	96.50	96.52
Average Recall (area = all)	88.41	**88.58**	88.50
Average Recall (area = small)	75.90	**77.82**	77.63
Average Recall (area = medium)	83.92	83.89	**84.01**
Average Recall (area = large)	**91.07**	91.05	90.97

It can be seen from Table 5 that the YOLOv4_4out algorithm is significantly better than the YOLOv4 algorithm in the prediction effect of small targets. For small targets, AP value increased by 2.48% and AR value increased by 1.92%. This also directly proves that after adding 104×104 feature layers and improving the PANet network with 4x up-sampling and 4x down-sampling, the recognition effect of the improved YOLOv4_4out algorithm on small targets can be enhanced.

By comparing the experimental results between YOLOv4_4out algorithm and YOLOv4 algorithm, the mAP of YOLOv4_4out algorithm is increased by 0.56%, and the number of model parameters is increased by 3.97 MB. Therefore, it can be concluded that adding a large scale feature layer output and improving the feature fusion PANet structure can improve the detection effect of small targets in the model. However, as the parameter size of the model increases, the memory storage space increases. By comparing the experimental results of the improved YOLOv4 algorithm and the YOLOv4_4out algorithm, the number of parameters of the improved YOLOv4 algorithm is reduced by about 144 MB, but the model mAP is only reduced by 0.07%. It can be concluded that the improved depth-separable convolution and the improved depth-separable convolution with the improved make_five_conv structure and the improved Resblock structure can greatly reduce the model parameters. However, because the feature extraction capability of depth-separable convolution is slightly insufficient, the mAP of the model has a very small decrease. Overall, compared with the YOLOv4 algorithm, the improved depth-separable convolution enables the model to significantly reduce model parameters while still ensuring the detection effect of the model. mAP index of model value is increased by 0.49%, and the number of model parameters is reduced by 140.1 MB. The improvement of the YOLOv4 model in this paper has also reached the expected goal on the whole at the algorithm level.

4.2.3. Model generalization test

The above experiments have proved the effectiveness of the improved algorithm from the model training stage, model verification stage and model testing stage respectively. Therefore, we used the improved YOLOv4 algorithm and YOLOv4 algorithm to predict the same group of untrained data images of other scenes respectively. Through the test, we can feel the generalization performance of the model directly. The image test results of the five scenarios are shown in Table 6.

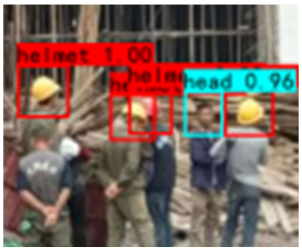
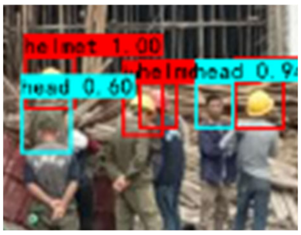
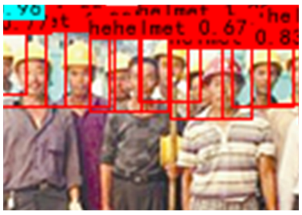
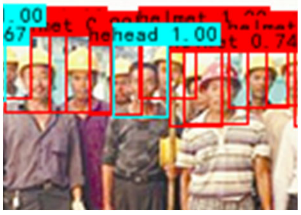
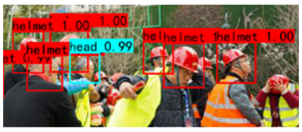
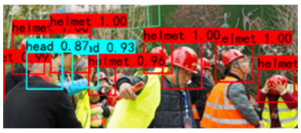
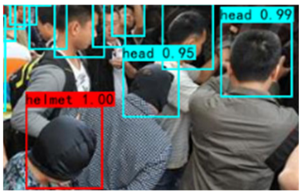
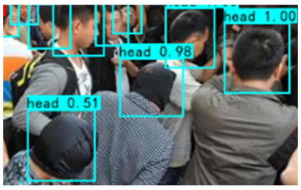
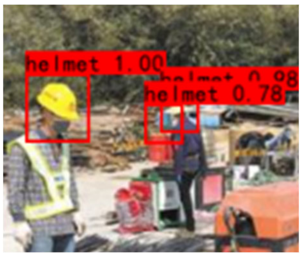
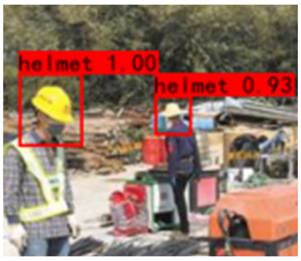
As can be seen from the table, in the prediction results of scenario 1 and scenario 2, the original version of YOLOv4 algorithm missed an obvious head target respectively. However, the improved YOLOv4 algorithm model in this paper can be successfully detected. The local enlarged image of the predicted results of the original YOLOv4 algorithm and the proposed algorithm can be clearly compared, and the proposed algorithm is more excellent. In the test of scenario 3, it can be found that one head and one helmet target were missed respectively by the original YOLOv4 algorithm, and the algorithm model in this paper was also successfully detected. In the test of scenario 4 and Scenario 5, the original YOLOv4 algorithm has error detection targets, and the algorithm model in this paper has correct detection. From the data of the five scenes tested, it can be shown from the side that the algorithm model in this paper has stronger feature extraction ability. It can still correctly identify small targets at a distance relatively dense targets and targets with too much background interference. However, the data prediction effect is still not ideal for scenarios where the hard hat targets are highly overlapped together and the targets are small. For example, there are still difficult targets not detected in scenarios 3 and 4.

5. Conclusions

The research of safety helmet wearing detection algorithm is very important to the safety operation of workers in construction site and traditional metallurgical manufacturing workshop. This paper mainly studies the two problems of small target detection and large number of model parameters in the current safety helmet wear detection algorithm. The main work of the algorithm in this paper is summarized as follows. First, self-built helmet wearing data set. Second, increase large-scale output and improve PANet network. Third, the depth separable convolution is improved to replace the standard 3 × 3 convolution. The experimental results show that the loss change convergence of the improved YOLOv4 algorithm in the training stage and the verification stage is faster and more stable, no matter in the training set or the verification set. The number of references was reduced by about 58%, the mAP on the test set was increased by 0.49% to 3.05%, and the model detection speed reached 35 FPS.

This paper has achieved the goal of reducing the number of model parameters without reducing the detection ability of the model. However, there are more missed detection cases where the helmet target is highly overlapped together. In order to make helmet wearing detection more accurate and rapid model reasoning in front-end equipment, future research needs to be carried out from four aspects. First, increase the amount of data in the hard hat data set, which shall include multiple backgrounds, multiple target sizes and multiple scene environments. Second, model pruning or build a better feature extraction network. Third, because the YOLOv4 algorithm is relatively far away and dense, the prediction probability of the prediction results is often relatively low. Therefore, another classification algorithm can be added to the YOLOv4 algorithm as a supplement. Forth, the model size can be reduced, the model detection accuracy can be maintained, the model detection speed can be improved, and the detection effect of the model can be guaranteed in various complex environments such as the occlusion of the target and the high overlap of the target.

Table 6. Model test effect comparison diagram.

scenario	original images	YOLOv4 prognostic images	improved YOLOv4 prognostic images
1			
2			
3			
4			
5			

Acknowledgments

This work was supported by Joint Fund Project of the National Natural Science Foundation of China (U1908218), the Natural Science Foundation project of Liaoning Province (2021-KF-12-06), and the Department of Education of Liaoning Province (LJKFZ20220197).

References

1. H. Fan, Application of machine vision technology in Industrial inspection, *Digital Commun. World*, **12** (2020), 156–157. https://doi.org/10.3969/J.ISSN.1672-7274.2020.12.068
2. D. G. Lowe, Distinctive image features from scaleinvariant keypoints, *Int. J. Comput. Vision*, **60** (2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
3. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, (2005), 886–893. https://doi.org/10.1109/CVPR.2005.177
4. J. Canny, A computational approch to edge detection, *IEEE Trans. Pattern Anal. Mach. Intell*, **8** (1986), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851
5. Q. Li, *A Research and Implementation of Safety-helmetVideo Detection System Based on Human Body Recognition*, University of Electronic Science and Technology in Chengdu, M. S. thesis, 2017.
6. S. Xu, Y. Wang, Y. Gu, N. Li, L. Zhuang, L. Shi, Safety helmet wearing detection study based on improved Faster RCNN, *Appl. Res. Comput.*, **37** (2020), 267–271. https://doi.org/10.19734/j.issn.1001-3695.2018.07.0667
7. D. Wu, H. Wang, J Li, Safety helmet detection and identification based on improved faster RCNN, *Inf. Technol. Informatization*, **1** (2020), 17–20. https://doi.org/10.3969/j.issn.1672-9528.2020.01.003
8. H. Wang, Z .Hu, Y. Guo, Z. Yang, F. Zhou, P. Xu, A real-time safety helmet wearing detection approach based on CSYOLOv3, *Appl. Sci.*, **10** (2020), 6732. https://doi.org/10.3390/app10196732
9. Y. Zhang, K. Wu, K. Gao, X. Yang, Helmet detection based on modified yolov3, *Comput. Simul.*, **38** (2021), 5–10.
10. Y. Gu, Y. Wang, L. Shi, N. Li, L. Zhang, S. Xu, Automatic detection of safety helmet wearing based on head region location, IET *Image Process.*, **15** (2021), 2441–2453. https://doi.org/10.1049/ipr2.12231
11. Y. Cong, X. He, H. Zhu, X. Zhu, Helmet Monitoring System Based on Improved Yolov4-Tiny Network, *Electron. Technol. Software Eng.*, **19** (2021), 121–124.
12. F. Wang, L. Chen, L. Jiao, Research on the algorithm of helmet detection based on SSD-MobileNet, *Inf. Res.*, **3** (2020), 34–39.
13. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017. Available from: https://www.semanticscholar.org/reader/3647d6d0f151dc05626449ee09cc7bce55be497e
14. T. Xiao, L. Cai, K. Tang, X. Gao, C. Zhang, Improved SSD's Helmet wearing detection method, *J. Sichuan Univ. Light Chem. Technol.: Nat. Sci. Ed.*, **33** (2020), 9–15. https://doi.org/10.11863/j.suse.2020.04.10

15. A. Howard, M. Sandler, G. Chu, W. Wang, L. Chen, M. Tan, et al, Searching for MobileNetV3, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140