Helmet wearing detection based on YOLOv5

Chao WU[a], De-yong WANG[b,c], Wen-xi SHI[b,c], Jian FANG[b,c], Xue-yi Zhao[b,c], Yan-yun Fu[*,d]
[a.] School of Information Science and Engineering, Xinjiang University,Urumqi, 830000, China
[b.] XinJiang Lianhai INA-INT Information Technology Co.,Ltd,Urumqi, 830000, China
[c.] Key Laboratory of Big Data of Xinjiang Social Security Risk Prevention and Control,Urumqi, 830000, China
[d.] Beijing Academy of Science and Technology, Beijing,100035, China
[*] Corresponding author: fyyun163@163.com

ABSTRACT

Complex environments, such as dense personnel and background interference, affect the detection accuracy of whether personnel wear helmets. To solve this problem, a new detection algorithm for helmets based on YOLOv5 is studied in this paper. Firstly, RepVGGblock is used to replace the common convolution in the network to effectively utilize the computing power of GPU. Secondly, the efficient channel attention mechanism is incorporated into the C3 module of the backbone network to enhance the feature identification ability of the backbone network for the hard hat. Then, the boundary box regression function is changed to SIoU to redefine the distance loss and improve the regression accuracy. Finally, the detection performance of multi-scale targets is improved by adding detection layers. The test results on the self-made safety helmet data set show that the average accuracy of the improved YOLOv5 model reaches 94.6%, which is 4.1% higher than that of the original model, and can meet the requirements of target detection.

KEYWORDS: Helmet wearing test; Target detection; Attention mechanism; Deep learning

1. INTRODUCTION

Safety helmet is all walks of life safety production workers and high altitude personnel is an essential safety equipment, each worker should always remember not to wear a safety helmet, not into the construction site; Do not wear safety helmet, do not carry out construction operations. However, in the complex actual environment, workers often fail to wear safety helmets for some reasons, which not only causes hidden dangers to the personal and property safety of workers, but also poses a threat to the smooth development of construction projects. At present, the judgment of whether workers wear helmets according to regulations mostly relies on manual supervision, but this method is not suitable for the application scenarios with large construction scope and scattered personnel distribution, and is susceptible to subjective factors. Therefore, it is of great significance to detect whether workers wear helmets by intelligent means such as deep learning.

At present, the detection algorithms for wearing helmet mainly include traditional machine learning detection algorithm and deep learning-based detection algorithm. Traditional machine learning object detection algorithms, such as CAI Limei [1], realize the detection of hard hats by using the circular shape characteristics of hard hats and combining the histogram of the edge direction and orientation information as the feature map. Li Xiao [2] used Adaboost classifier and Har-like feature to detect whether to wear a helmet. Pathasu Doungmala[3] et al. used haar features [4] for face detection and circular Hough transform for helmet detection. Li Qirui [5] used moving object detection based on background subtraction to locate human body position, adopted HOG integral graph to reduce complexity, and finally realized helmet wearing detection by combining with SVM. Traditional detection methods use manual feature extraction, detection speed and recognition effect can not meet the actual needs. With the development of computers, object detection algorithms based on deep learning have emerged. Deng Benyang [6] et al used K-means algorithm to cluster prior boxes and adopted multi-scale training strategies to improve YOLOv4[7] algorithm to detect hard hats. Li Peng [8] proposed to replace the original backbone network of CascadeR-CNN[9] and add the strategy of variable convolution and void convolution. Liu Chuan [10] implemented algorithm detection by introducing attention mechanism in YOLOv3[11], adding space pooling pyramid module and optimizing loss function. These optimized algorithms have a good recognition effect.

3rd International Conference on Artificial Intelligence, Automation, and High-Performance Computing
(AIAHPC 2023), edited by Simon X. Yang, Dimitrios A. Karras, Proc. of SPIE Vol. 12717,
127170F · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2684745

In this paper, based on the YOLOv5[12] algorithm, RepVGG convolution [13] was first used to replace the common convolution operation with step size 2 in the network, multi-branch was used in training to improve the model accuracy, and the structure was reparameterized to single branch in reasoning to ensure the accuracy and improve the reasoning speed. Secondly, an efficient channel attention mechanism was introduced into the C3 module. Strengthen the ability of feature extraction, then modify the localization loss function of the algorithm, accelerate the convergence of the network, further improve the accuracy, and finally add a detection layer to improve the detection ability of the model for small targets. The experimental results show that the improved algorithm model improves the average accuracy and can better support the requirements of actual production environment detection.

2. Basic principles of YOLOv5 algorithm

YOLOv5 is a new-generation target detection network released by Ultralytics. Its network structure is shown in Figure 1, which is mainly composed of input, backbone, neck and prediction. The Input terminal is an input port, like Mosaic a series of data enhancement technologies. Backbone is mainly composed of Focus structure, Conv convolution, C3 module and SPP for feature extraction. Neck is composed of FPN (Feature Pyramid Network) +PAN (Path Aggregation Network) structure for feature fusion, and Prediction is used to output three kinds of feature maps without graph scale, predicting small, medium and large targets respectively.

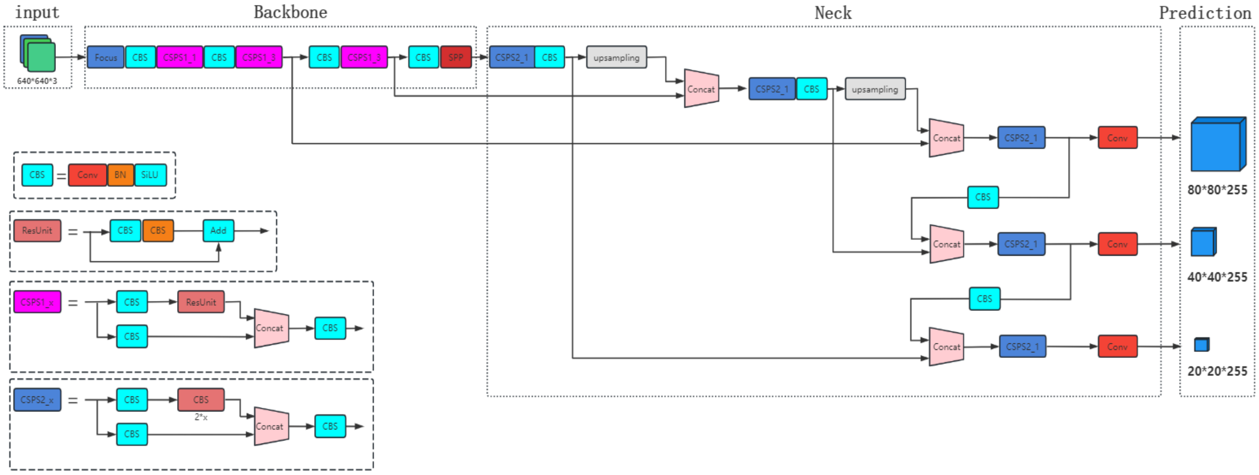

Fig.1 Schematic diagram of YOLOv5

3. Improved method of YOLOv5 algorithm

3.1. Introduce RepVGG convolution

RepVGG is a multi-branch architecture that can be reparameterized and fused to a single branch for reasoning, which can improve accuracy and speed of computation. The basic structure includes three branches: 3×3 convolution, 1×1 convolution and Identity residual. When the dimension of input and output is inconsistent, namely Stride=2, there are only two branches: 3×3 and 1×1, as shown in Figure 2 below:

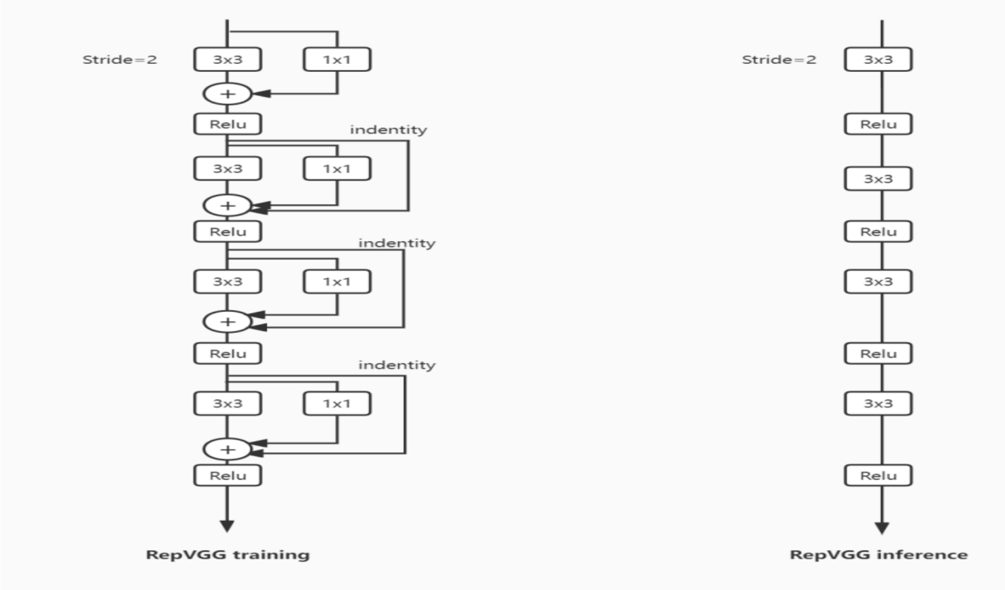

Fig.2 Schematic diagram of Repvgg

Reparameterization first combines the BN layer with the previous convolution layer and converts it into a convolution with a bias as shown in equations (1) and (2). $W, \mu, \sigma, \gamma, \beta$ are the parameters of the convolution layer, the mean value, variance, scaling factor and bias of BN.

$$\mathrm{W}'_{\mathrm{i},:,:,:} = \frac{\gamma_i}{\sigma_i}\mathrm{W}_{\mathrm{i},:,:,:} \tag{1}$$

$$\mathrm{b}'_i = -\frac{\mu_i \gamma_i}{\sigma_i} + \beta_i \tag{2}$$

Then convert the convolution on the three merged branches into 3×3 convolution. identity can be regarded as the 1×1 convolution whose convolution kernel is the identity matrix, and fill the two 1×1 convolution zeros into 3×3 convolution. Finally, add the three convolution and bias to get the final convolution term.

Due to the excellent performance of RepVGG network, it is introduced into the network model of YOLOv5 to replace the common convolution operation with step size 2. Strengthen the feature extraction ability of the network, so that the model can learn more feature information.

3.2. Introduce attention mechanisms

The attention mechanism focuses on local information and suppresses useless information. Efficient Channel Attention[14] can effectively avoid the impact of dimension reduction and achieve a local cross-channel interaction strategy without dimension reduction, which is realized through one-dimensional convolution. Only a small number of parameters is increased, but the ability of feature extraction is obviously improved. Efficient Channel Attention is shown in Figure 3:

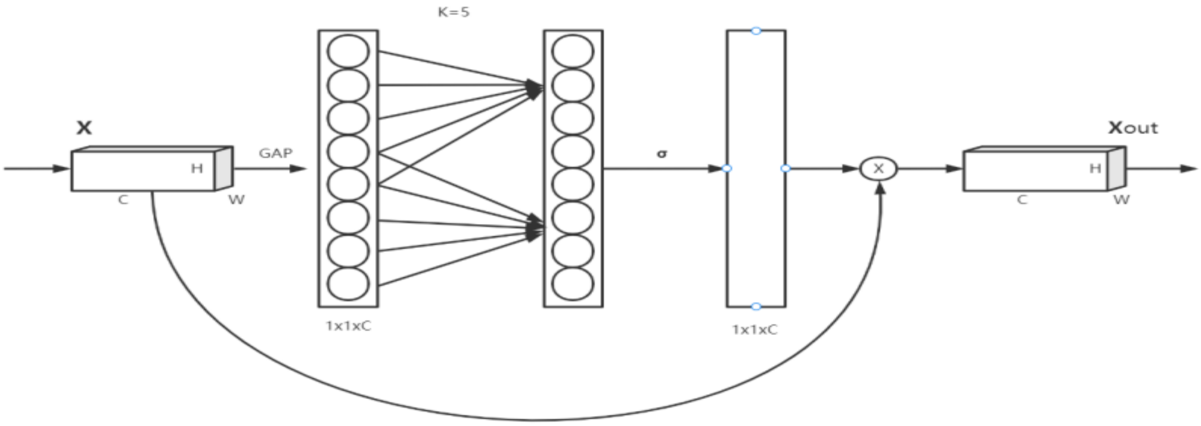

Fig.3 Schematic diagram of ECA attention mechanism

The principle is to obtain aggregation features through average pooling, and then perform one-dimensional convolution with the convolution kernel size of K to generate channel weights, where the value of K is determined by channel dimension C, and then generate the weight ratio of each channel through sigmoid function, and finally fuse with the original feature graph to form a new feature graph.

In this paper, the attention mechanism is added into the backbone network. The backbone network of YOLOv5s model is mainly composed of C3 modules. The attention mechanism is combined with it and improved into C3ECA module, which is added into the network.

3.3. Modified loss function

CIoU[15] regression loss function was adopted in YOLOv5s model, but this loss function did not take into account the direction between the real box and the predicted box, resulting in a slow convergence rate. In this regard, the vector Angle between the real box and the predicted box was introduced by SIoU[16] loss function to redefine the relevant loss function, reduce the degree of freedom of regression and accelerate the network convergence. Further improve the regression accuracy. SIoU consists of four parts:

(1) Angle cost is shown in Figure 4:

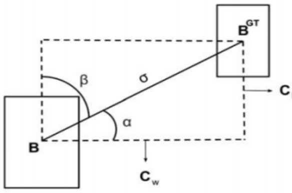

Fig.4 Calculation of Angle loss

Where ch is the height difference between the center point of the real frame and the predicted frame, σ is the distance between the center point of the real box and the predicted box,The Angle α is the arcsine of both.As shown in equations (3), (4) and (5), the center coordinate of the real box is ($b^{gt}_{c_x}, b^{gt}_{c_y}$),The center coordinate of the prediction box is(b_{c_x}, b_{c_y}).The Angle Λ loss function (6) is obtained.

$$\mathrm{c_h} = \max(b^{gt}_{c_y}, b_{c_y}) - \min(b^{gt}_{c_y}, b_{c_y}) \tag{3}$$

$$\sigma = \sqrt{\left(b^{gt}_{c_x} - b_{c_x}\right)^2 + \left(b^{gt}_{c_y} - b_{c_y}\right)^2} \tag{4}$$

$$\frac{c_h}{\sigma} = \sin(\alpha) \tag{5}$$

$$\Lambda = 1 - 2 * \sin^2(\arcsin(\frac{c_h}{\sigma}) - \frac{\pi}{4}) = \cos(2 * (\arcsin(\frac{c_h}{\sigma}) - \frac{\pi}{4})) \tag{6}$$

(2) Distance cost is shown in Figure 5:

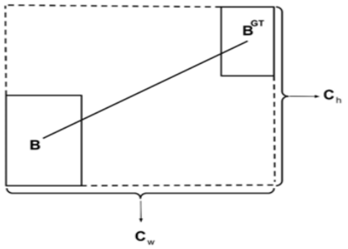

Fig.5 Calculation of distance loss

(c_w, c_h) Is the width and height of the minimum enclosing rectangle of the real box and the prediction box,Calculate ρ_x ρ_y γ and finally get the distance loss formula Δ ,See Formula (7) and (8).

$$\rho_x = (\frac{b_{c_x}^{gt} - b_{c_x}}{c_w})^2, \rho_y = (\frac{b_{c_y}^{gt} - b_{c_y}}{c_h})^2, \gamma = 2 - \Lambda \tag{7}$$

$$\Delta = \sum_{t=x,y}(1 - e^{-\gamma\rho_t}) = 2 - e^{-\gamma\rho_x} - e^{-\gamma\rho_y} \tag{8}$$

(3) Shape cost, as shown in Equation (9).

Where, (w, h) and (w^{gt}, h^{gt}) are the width and height of the prediction box and the real box respectively, as shown in Equation (10), θ controls the attention to shape loss, avoids excessive attention to shape loss and reduces the movement of the prediction box, θ is set between [2,6].

$$\Omega = \sum_{t=w,h}(1 - e^{-w_t})^\theta = (1 - e^{-w_w})^\theta + (1 - e^{-w_h})^\theta \tag{9}$$

$$w_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, w_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{10}$$

(4) IOU loss

Used to represent the intersection and union ratio of the prediction box and the real box. To sum up, the final SIoU loss function is defined as follows:

$$Loss_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{11}$$

3.4. Add detection layer for small targets

The original YOLOv5s model has 3 detection layers. 8,16,32 times downsampling was performed on the input image size of 640×640 to generate 3 kinds of feature graphs of 80×80, 40×40 and 20×20. The targets of 8×8, 16×16, 32×32 were predicted respectively, corresponding to small, medium and large size targets. In the actual construction site, because the distance between workers and detection equipment is different, the detection target of safety helmet is small and dense, and the detection accuracy is not high or even the problem of missing detection. To solve this problem, a small target detection layer is proposed. The improved model is shown in Figure 6. There are four detection layers of the model, and the corresponding feature map sizes are 160×160, 80×80, 40×40, 20×20. By adding a small target detection layer, the problem of insufficient detection accuracy of small targets can be effectively solved at the cost of increasing a small amount of calculation.

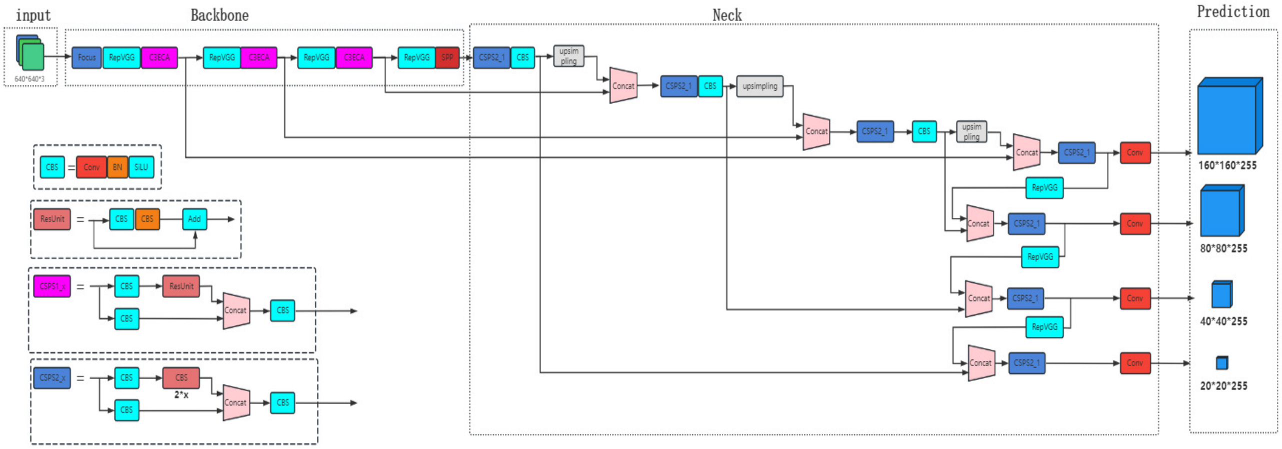

Fig.6 Improved model network structure

4. Experimental results and analysis

4.1. Data set and experimental environment

The existing open safety hat data set SafetyHelmetWearing, but part of the data does not meet the experimental conditions, so through online search, the crawler amplified the data set, a total of more than 7000 data sets, in which the label of wearing safety hat is hat, and the label of not wearing safety hat is person. It is divided into training set and test set according to 9:1. The improved algorithm is trained with training set, and the performance of the algorithm is verified with test set.

The experimental environment is based on WIN10 operating system, the CPU is Intel(R) Xeon(R) Silver 4210, the GPU is NVIDIA GeForce RTX 2080, and the deep learning framework Pytorch 1.11.0 is used. The experimental parameters are set as follows: The training rounds (batchsize) were 300 times, the training batchsize (batchsize) was 32, the initial learning rate was 0.01, and the stochastic gradient descent (SGD) method was used for optimization.

4.2. Evaluation indicators

In the field of deep learning object detection, the main evaluation indicators include Precision, Recall and detection speed, etc. In this paper, precision and recall are adopted. Average Precision and mean Average Precision indexes evaluate improved detection algorithms. The main parameters include: TP(True Positives) is an instance where the classifier considers a positive sample and it is,FP(False Positives) is an instance where the classifier considers a positive sample but it is not,TN(True Negatives) is the example that the classifier considers to be a negative sample and is indeed a negative sample, and FN(False Negatives) is the example that the classifier considers to be a negative sample but is not in fact a negative sample.

4.3. Experimental results and analysis

In order to verify the effectiveness of the algorithm, a comparative experiment was conducted. Experimental equipment and parameter Settings were consistent. The improved algorithm was compared with Faster-RCNN, SSD, YOLOv3, YOLOv4 and YOLOv5s, and the same experimental data set was used for training and testing. The experimental results mAP,Recall, weight and FPS are shown in Table 1 below:

Table 1 exprimental comparison results

model	mAP/%	Recall/%	Weight /MB	FPS
Faster-RCNN	91.8	92.3	160.1	5.3
SSD	75.2	78.0	101.2	10.2
YOLOv3	81.4	83.2	236	16.9
YOLOv4	88.5	90.3	244	24.4
YOLOv5s	90.5	87.2	14.1	40
ours	**94.6**	**90.1**	**15.1**	**41.7**

As can be seen from Table 1, the improved algorithm in this paper not only significantly improves the detection effect of whether workers wear safety hats, but also improves the detection speed. Compared with the original algorithm, when the weight of the improved model is increased by only 1MB, the mAP value is increased by 4.1%, and the detection speed is also improved to 41.7FPS. Compared with the typical two-stage target detection algorithm Faster-RCNN, although there is little difference between the average accuracy mean and recall rate, they have considerable advantages in weight and detection speed. In general, the improved algorithm has higher recognition accuracy and faster detection speed, which can meet the actual detection requirements.

In order to verify the improvement of each module, the ablation experiment was conducted, and the experimental results were shown in Table 2. It can be seen from Table 2 that after the modification of the backbone network, the mAP increased by 1.1% and the speed increased by 2ms. After the integration of attention mechanism, mAP increased by 0.4%, but the detection time increased. After modifying the loss function of the original network, mAP increases by 0.3%, and the detection speed does not change much. After adding a detection layer, mAP is increased by 1.3% and detection time is increased by 2ms. mAP of all the improved fusion algorithm model is increased by 4.1%, which improves the detection speed and can have a good detection effect on small and dense targets such as hard hats.

Table 2 Ablation experiments results

Model	RepVGG	Attentional mechanism	Modified loss function	Add detection layer	mAP/%	TTD/ms
YOLOv5s	×	×	×	×	90.5	25
YOLOv5s+Repvgg	√	×	×	×	91.6	23
YOLOv5s+Attention	×	√	×	×	90.9	26
YOLOv5s+loss function	×	×	√	×	90.8	24
YOLOv5s+detection layer	×	×	×	√	92.8	27
ours	√	√	√	√	**94.6**	**24**

5. Conclusion

In this paper, an improved model based on YOLOv5s model is proposed to improve the accuracy and detection efficiency of the model by modifying the convolutional operation in the network structure, increasing the attention mechanism, modifying the loss function and increasing the detection layer for small targets. The experimental results show that the improved model improves the detection accuracy of whether to wear a helmet, and is superior to other models in processing speed, which can better support the real-time monitoring of whether to wear a helmet in practical applications. In the follow-up work, the lightweight detection of the model will be further studied to achieve better detection effect.

ACKNOWLEDGMENT

Fund Project: Natural Science Foundation Project "Intelligent Perception and Real-time Simulation and Deduction Technology of Urban Emergency Management Events", Project No. U20B2060.

REFERENCES

[1] CAI Limei. Research on Personnel Target Detection and Tracking in Underground Coal Mine Based on Video [D]. China University of Mining and Technology,2010. https://www.cnki.net/

[2] Li Xiao. Research and Application of Video Recognition in HSE Monitoring Platform [D]. East China University of Science and Technology,2014.https://www.cnki.net/

[3] P. Doungmala and K. Klubsuwan, "Helmet Wearing Detection in Thailand Using Haar Like Feature and Circle Hough Transform on Image Processing," 2016 IEEE International Conference on Computer and Information Technology (CIT), 2016,

[4] Viola and Jones, "Rapid object detection using a boosted cascade of simple features", Computer Vision and Pattern Recognition, 2001.

[5] LI Qirui. Research and Implementation of Safety helmet Video Detection System Based on Human Body Recognition [D]. University of Electronic Science and Technology of China,2017.https://www.cnki.net/

[6] D. Benyang, L. Xiaochun and Y. Miao, "Safety helmet detection method based on YOLO v4," 2020 16th International Conference on Computational Intelligence and Security (CIS), 2020,

[7] Alexey Bochkovskiy et al. "YOLOv4: Optimal Speed and Accuracy of Object Detection" arXiv: Computer Vision and Pattern Recognition (2020).

[8] Li P. Research and implementation of Key technologies of construction site safety early warning based on target detection and depth estimation [D]. University of Electronic Science and Technology of China,2021.

[9] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018,pp.6154-6162,doi:10.1109/CVPR.2018.00644.

[10] Liu Chuan. Research on Helmet Wearing Detection Algorithm Based on Engineering Environment [J]. Henan Science and Technology,2022.https://www.cnki.net/

[11] Redmon, Joseph and Ali Farhadi. "YOLOv3: An Incremental Improvement." ArXiv abs/1804.02767 (2018).

[12] Ultralytics.YOLOv5[CP/OL].[2020-08-09]. https: //github.com/ultralytics/yolov5.

[13] Ding X , Zhang X , Ma N , et al. RepVGG: Making VGG-style ConvNets Great Again[C]// 2021.

[14] Wang Q , Wu B , Zhu P , et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[15] Zhaohui Zheng et al. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07) : 12993-13000.

[16] Zhora Gevorgyan https://arxiv.org/abs/2205.12740