Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Safety Helmet Wearing Detection Based on YOLOv5 of Attention Mechanism

To cite this article: Z P Xu *et al* 2022 *J. Phys.: Conf. Ser.* **2213** 012038

Safety Helmet Wearing Detection Based on YOLOv5 of Attention Mechanism

Z P Xu, Y Zhang, J Cheng and G Ge

University of Jinan, Jinan, 250022, China

Correspondence: cse_zhangy@ujn.edu.cn

Abstract. Aiming at problems of low accuracy and strong detection interference of the existing safety helmet wearing detection algorithms, an object detection algorithm by adding the squeeze-and-excitation block based on the YOLOv5 algorithm is proposed in this paper. The proposed method can not only obtain the weight of picture channel, but also accurately separate the foreground and background of the picture. Keeping all parameters unchanged, the proposed method and the YOLOv5 algorithm are applied to detect the safety helmet data set in the experiment. The result shows that the YOLOv5 algorithm with the squeeze-and-excitation block has an average detection accuracy of 94.5% for safety helmets and an average detection accuracy of 92.7% for human heads. The mAP value detected by the proposed method is 2% ~2.5% higher than using YOLOv5 algorithm directly.

1. Introduction

During the period of vigorous development of economic construction and modernisation in my country, construction was carried out in many places. However, the environment of these construction sites is complex and there are many life-threatening factors. The head is the most critical part of the human body, and it is also the most vulnerable part. Once injured, it is easy to be fatal. However, some workers lack safety awareness and often do not wear safety helmets, so safety helmet detection has become an important technology to ensure construction safety.

The earliest detection of whether workers wear safety helmets adopted manual supervision, but the scope of work of construction workers is very broad. So this method cannot supervise all workers in time. Then, many researchers began to use machine learning and image processing technology to detect safety helmets. For example, Li Qirui uses a hog algorithm to realize feature extraction, and then combines support vector machine to achieve safety helmet wearing detection. This traditional method has poor stability and generalization ability [1].

The current mainstream deep learning object detection algorithms are mainly divided into two categories. The first category of object detection algorithm is a two-stage and region-based series of RCNN algorithms, such as Fast-RCNN and Faster-RCNN [2]. This variety of algorithms has high accuracy, however the speed is slow. The second category of object detection algorithm is a one-stage and converts the detection into the You Only Look Once (YOLO) [3] series algorithm and Single Shot MultiBox Detector(SSD)[4] algorithm to solve the regression problem. Bin Dai and others aimed at the problem of strong interference such as different light intensity and various weather changes. On the foundation of SSD, they introduced multi-layer feature fusion, multi-layer deconvolution structure and light weight network structure. They constructed MD- SSD network, and the mAP of the network detects safety helmets can reach 86.7% [5].

YOLO is a fast and efficient object detection algorithm based on feature extraction by convolution neural networks. At present, based on the YOLOv3 algorithm [6], Peizhi Wen and others first used the K-Means algorithm to cluster the object, and then also used multi-scale feature extraction and non-maximum value suppression. They used this method to detect safety helmets with a detection accuracy of 90.7% [7]. Based on YOLOv4, Deng Benyang and others first used the K-means algorithm to perform clustering to obtain some appropriate prior boxs. Subsequently, a multi-scale training strategy is used in the network to detect objects of different sizes. They used this technique to detect the mAP of safety helmets up to 92% [8]. A few months later, YOLOv5 came out, and the principal detection algorithm used in this article is YOLOv5.

2. Detection Algorithm Network

YOLOv5 mainly adds some new improvement methods, and its speed and accuracy have been greatly improved, as follows. Input: in the model training stage, a series of improved methods such as mosaic data enhancement and adaptive image scaling are proposed. Backbone network: some new detection methods are used for reference, mainly including Focus structure and CSP structure [9]. Neck Network: YOLOv5 mainly uses FPN and PAN structures in this part. Output layer: the localization loss function is CIoU loss.

2.1. Backbone network

The first thing the backbone network performs is the Focus operation. Focus operation is to slice the picture. The slicing operation is to take a value every other pixel in a picture, and finally 4 pictures are obtained. Then, backbone network performs the convolution operation on the obtained new picture, and finally obtains the double down-sampling feature map without information loss. Focus structure is shown in figure 1.

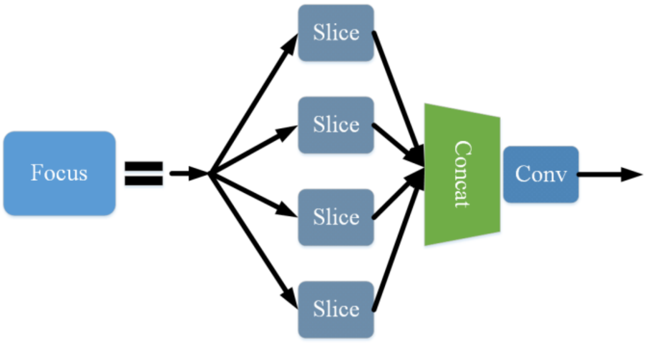


Figure 1. Focus structure.

The main function of the Focus structure is to reduce the number of network layers, minimize calculation parameters and decrease CUDA memory consumption.

The Focus structure is followed by the CSP structure. Two CSP structures are designed in YOLOv5. Since backbone is a relatively deep network, adding the residual structure can not only avoid the disappearance of the gradient, but also extract more fine-grained features. So backbone chooses CSP1_X with a residual structure. Its network structure is shown in figure 2. Neck uses CSP without residual structure, and its network structure is shown in figure 3.

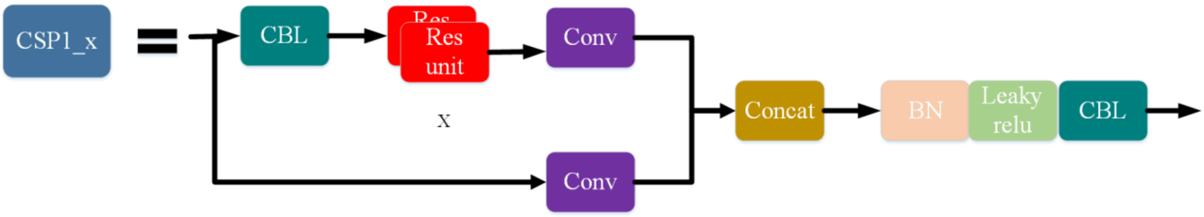


Figure 2. CSP structure with residuals.

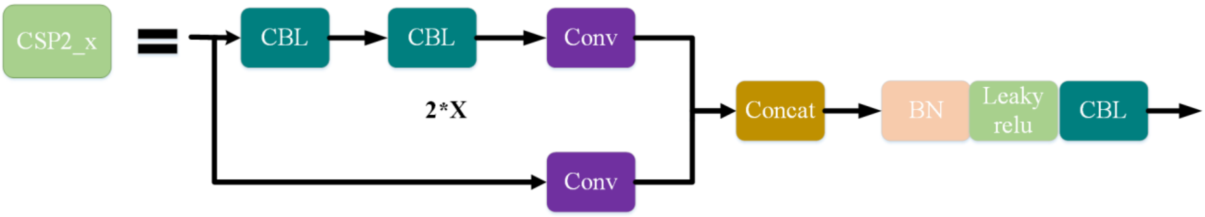


Figure 3. CSP structure without residuals.

Compared with the structure of Leaky Relu followed by ordinary convolution, the CSP with two branches can extract richer feature information.

2.2. Neck Network

Generally, the shallow feature maps have strong location information but weak semantic features, and deep feature maps have strong semantic features but weak location information. FPN is a top-down structure. The purpose of FPN is to fuse high-level features and low-level features through upsampling, and then predicts on multiple scales. PAN adds a bottom-up feature pyramid based on FPN [10]. Therefore, PAN has both strong semantic features and strong location features.

2.3. Loss function

The loss function of the YOLOv5 algorithm is divided into 3 parts. The first part is the classification loss; the second part is the confidence loss; the third part is the localization loss. Considering categories are mutually exclusive, the classification loss uses the binary cross-entropy loss function. The confidence loss is the same as the classification loss.

The formula of binary classification cross entropy loss function [11]:

$$L=-\{y\log\hat{y}+(1-y)\log(1-\hat{y})\} \tag{1}$$

IoU is the intersection of the ground truth box and the predicted box divided by the union of the ground truth box and the predicted box. But IoU has the following two shortcomings. First, according to the definition, when the two boxes have no intersection, IoU=0. IoU=0 cannot reflect the confidence of the two boxes at all. Second, equal IoU does not mean that the confidence is the same.

Because IoU has these two shortcomings, the YOLOv5 algorithm chooses CIoU. CIoU not only considers the aspect ratio, but also considers the center point distance [12]. The complete CIoU loss function is defined as follows:

$$I_{CIoU}=1-IoU+\frac{\rho^2(b,b^{gt})}{c^2}+\alpha v \tag{2}$$

Where α is the weight parameter, and v represents the similarity of the aspect ratio, the formula is as follows:

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h}\right) \tag{3}$$

3. Squeeze-and-Excitation Block

The current attention mechanisms are summarized into the following three categories. The first category is the channel attention mechanism. The representative network is squeeze-and-excitation networks [13]. The second category is the spatial attention mechanism. The third category is the hybrid domain attention mechanism. The representative network is convolutional block attention module [14].

In the traditional convolution network, the weight of each channel of the default feature map of the convolution network is equal. In the problem of safety helmet detection, because the interference is different, the proportion of different channel is not equal. The detection algorithm introduces the squeeze-and-excitation block to solve the problem of different proportion of different channel and minimize the impacts of interferences. The network structure of squeeze-and-excitation block is shown in figure 4.

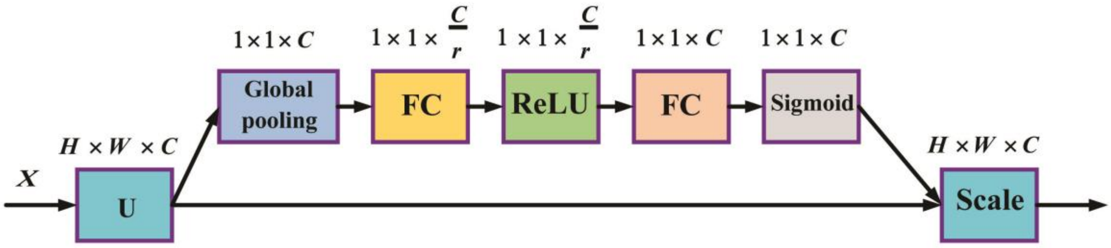


Figure 4. Squeeze-and-Excitation block.

The squeeze-and-excitation block first turns the input $X \in R^{H'\times W'\times C'}$ into $U \in R^{H\times W\times C}$ by convolution. Convolution operation uses $V = [v_1, v_{2,} \ldots, v_c]$ as filter kernels, where v_c represents the parameters of the c-th filter. The convolution formula is as follows:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{4}$$

Where * means convolution. Since the output is generated by summing all channels, channel dependencies are implicitly embedded in v_c. In order to solve the problem of channel dependencies, this article uses global average pooling to generate $1\times1\times c$ channel descriptor when performing squeeze operation. The calculation formula for global average pooling is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{5}$$

Where $z \in R^C$, z_c is the c-th element of z . To make use of the information characteristics aggregated by squeeze operation, the excitation operation must fully capture the channel-wise dependencies. To achieve this function, it must be able to learn non-linear interaction between channels. Second, this article wants to ensure that multiple channels are emphasized, so it must learn non-mutually exclusive relationships. In order to meet these two criteria, this article chooses to use a simple gating mechanism with a sigmoid activation:

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)) \tag{6}$$

Where δ is ReLU activation function, $W_1 \in R^{\frac{C}{r}\times C}$ and $W_2 \in R^{C\times\frac{C}{r}}$. In order to limit the complexity of the model, this article parameterizes the gating mechanism with two fully connected (FC) layers around the non-linearity. One is the dimensionality-reduction fully connected layer with the reduction ratio r. And the output of the information feature after a dimensionality-reduction fully connected layer becomes $1\times1\times\frac{c}{r}$. Through the ReLU activation function, the output of the information feature after a dimensionality-increasing fully connected layer becomes $1\times1\times c$. The final output of the block is obtained by rescaling U with the activations s:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c \tag{7}$$

Where $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_c]$. $F_{scale}(u_c, s_c)$ represents the channel-wise multiplication between the scalar s_c and the feature map $u_c \in R^{H\times W}$.

4. Experiment

Experimental environment configuration: The GPU is GeForce RTX 3080, the memory is 64GB, CUDA is version 11.1, python is version 3.6, and the Pytroch learning framework is used for the experiment.

4.1. Evaluation Method

This experiment is evaluated by mean Average Precision（mAP）. The mAP is the average value of each category of AP. First, for each different Recall value (including 0 and 1), the experiment selects the maximum values of Precision that is greater than or equal to these Recall values, and then calculates the area under the PR curve. The area is the AP value. Therefore, we must first calculate the precision and recall rate. The calculation formula is as follows:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Re}\,\text{call} = \frac{TP}{TP + FN} \tag{9}$$

The meaning of TP is the positive sample predicted as a positive value by the model; the meaning of FP is the positive sample predicted as a negative value by the model; and the meaning of FN is the negative sample predicted as a negative value by the model [15].

4.2. Network Training

Before the start of the experiment, 2400 pictures containing workers wear safety helmets were divided in a ratio of 0.8. Finally, the training set contains 1920 pictures, and the validation set contains 480 pictures. In the experiment, the optimizer chooses Adam. The iteration epochs are 30, and the batch size is 8. The PR curve obtained by formula 8 and formula 9 is shown in figure 5.

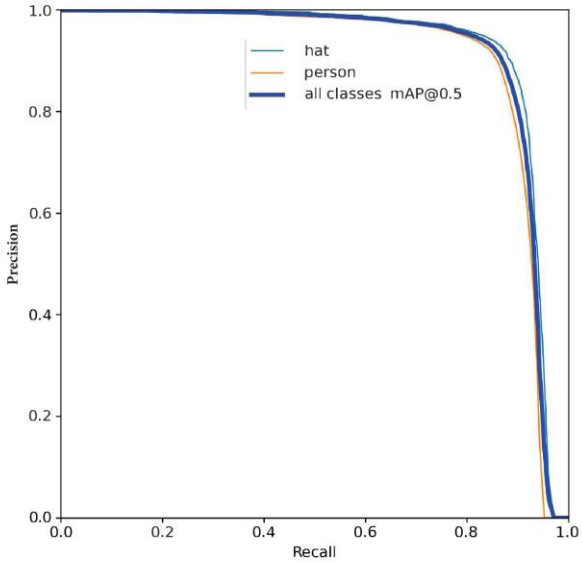


Figure 5. PR curve

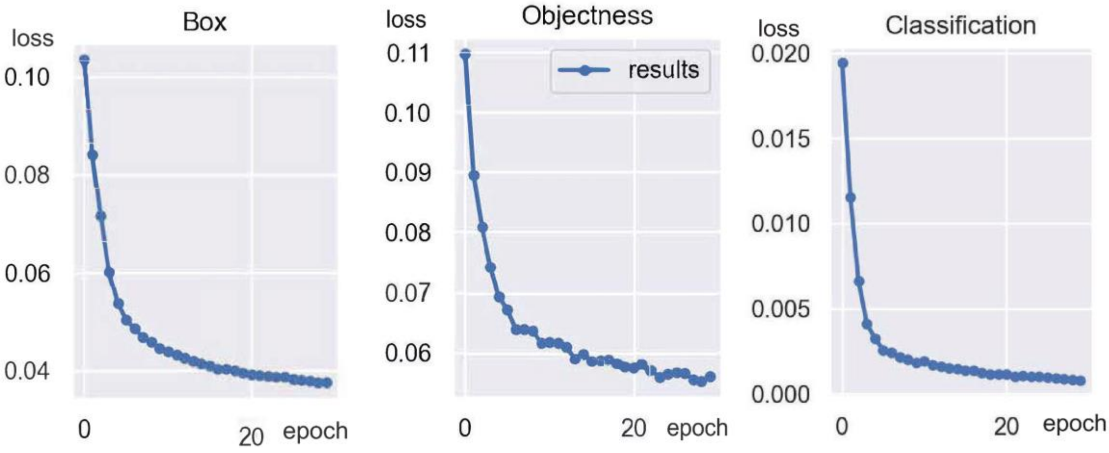
Figure 5 shows the PR curve of the original YOLOv5 algorithm for detecting safety helmets, human heads and all classes. The area enclosed by the PR curve, the horizontal axis and the vertical axis represents the mAP of the algorithm’s detection.

Table 1. mAP of YOLOv5 algorithm.

detection site	mAP
hat	92.3%
human head	90.2%
all classes	91.2%

The calculated area is the number in table 1. The data in table 1 shows that when IoU is 0.5, the original YOLOv5 algorithm detects the mAP of the safety helmets can reach 92.3%, the mAP of the human heads can reach 90.2%, and the average mAP of all classes can reach 91.2%.

This article keeps other parameters unchanged, and adds the class squeeze-and-excitation block to the last layer of the backbone network. The loss obtained from the experiment is shown in figure 6.



(a) Localization loss. (b) Confidence loss. (c) Classification loss.

Figure 6. Loss function.

Generally, in the training process of the model, the smaller the value of the loss function, the better the model is, and the expected value is 0. From figure 6, the horizontal axis represents the iteration epoch, and the vertical axis represents the loss value. As iteration epochs increase, the localization loss, confidence loss and classification loss are basically close to 0.

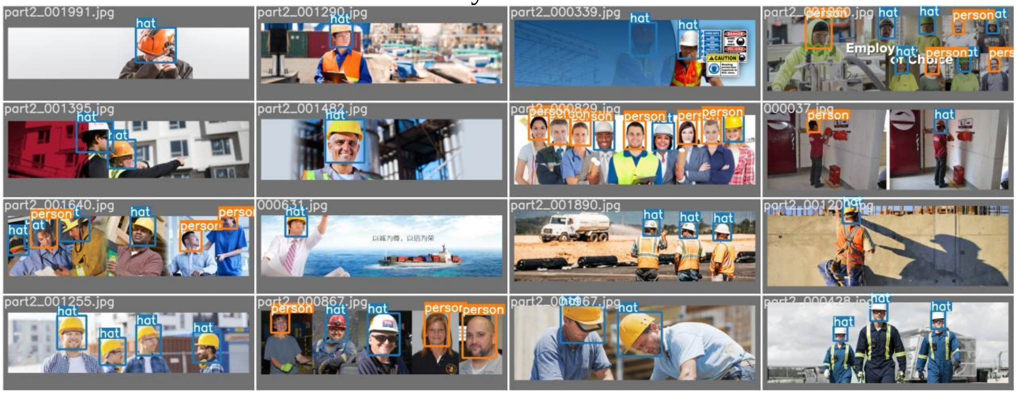


Figure 7. Ground truth box.

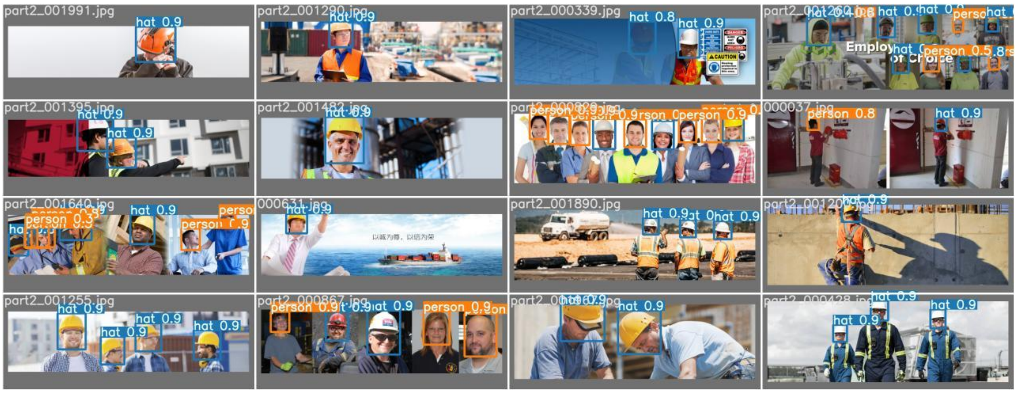


Figure 8. Prediction box.

Figure 7 is a manually labeled box, and figure 8 is a box predicted by the model. The category name is on the box, and the number next to the name represents the confidence. As can be seen from the numbers in the figure, the confidence is relatively high. Comparing the two pictures, the box predicted by the model is basically the same as ground truth box.

The PR curve of YOLOv5 algorithm with the squeeze-and-excitation block obtained by formula 8 and formula 9 is shown in figure 9. Figure 9 shows the PR curve of YOLOv5 algorithm with the squeeze-and-excitation block for detecting safety helmets, human heads and all classes.

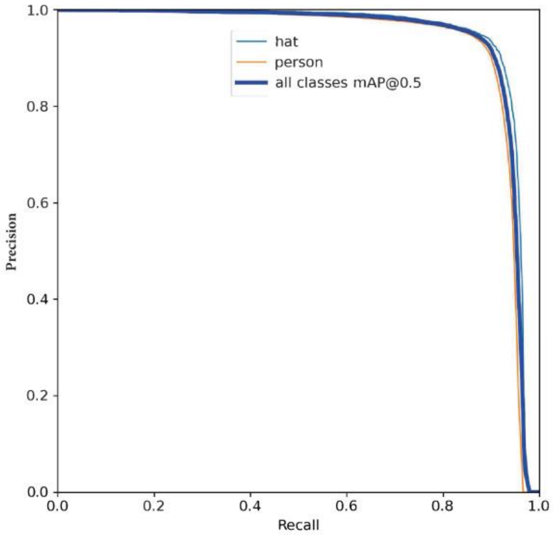


Figure 9. PR curve

Table 2. mAP of proposed method

detection site	mAP
hat	94.5%
human head	92.7%
all classes	93.6%

The data in table 2 shows that when IoU is 0.5, the mAP of the safety helmets detected by the YOLOv5 algorithm with the squeeze-and-excitation block can reach 94.5%, the mAP of the human heads can reach 92.7%, and the average mAP of all classes can reach 93.6%. Comparing table 1 and table 2, whether it is to detect safety helmets or human heads, the mAP detected by the YOLOv5 algorithm with the squeeze-and-excitation block is 2%~2.5% higher than using the YOLOv5 algorithm directly.

5. Conclusion and Future Work

The squeeze-and-excitation block is a kind of attention mechanism module. The squeeze-and-excitation block can obtain the weight of the picture channel. The YOLOv5 algorithm with the squeeze-and-excitation block can accurately separate the foreground and background of the picture. This article compares in detail the YOLOv5 algorithm with the squeeze-and-excitation block and directly use YOLOv5 algorithm to detect the safety helmets. The results prove that the YOLOv5 algorithm with the squeeze-and-excitation block detects safety helmets is 2%~2.5% higher than using YOLOv5 algorithm directly. This basically meets the accuracy requirement of safety helmet wearing detection in construction scenarios.

Although the feature fusion effect of PAN is very good, it will also increase the amount of calculation. In the future, this article may explore whether BiFPN can reduce the number of network model parameters and improve the speed of network model detection [16].

6. References

[1] Jie L, Liu H and Wang T 2017 Safety helmet wearing detection based on image processing and machine learning *Ninth International Conference on Advanced Computational Intelligence (ICACI).*

[2] Ren S, He K and Girshick R 2016 Faster R-CNN: towards real-time object detection with region proposal networks *IEEE transactions on pattern analysis and machine intelligence.*

[3] Redmon J, Divvala S and Girshick R 2016 You Only Look Once: Unified, Real-Time Object Detection *Computer Vision & Pattern Recognition.*

[4] Liu W, Anguelov D and Erhan D 2016 SSD: Single Shot MultiBox Detector *European Conference on Computer Vision.*

[5] Dai B, Nie Y and Cui W 2020 Real-time Safety Helmet Detection System based on Improved SSD *Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture.*

[6] Redmon J and Farhadi A 2018 Yolov3: An incremental improvement *arXiv preprint.*

[7] Wen P, Tong M and Deng Z 2020 Improved Helmet Wearing Detection Method Based on YOLOv3 *International Conference on Artificial Intelligence and Security.*

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

[9] Tan S, Lu G and Jiang Z 2021 Improved YOLOv5 Network Model and Application in Safety Helmet Detection *IEEE International Conference on Intelligence and Safety for Robotics (ISR).*

[10] Wang S 2021 Substation Personnel Safety Detection Network Based on YOLOv4 *IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE).*

[11] Bo Y, Huan Q and Huan X 2019 Helmet detection under the power construction scene based on image analysis *IEEE 7th international conference on computer science and network technology (ICCSNT).*

[12] Bochkovskiy A and Liao H 2020 YOLOv4: Optimal Speed and Accuracy of Object Detection

[13] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proceedings of the IEEE conference on computer vision and pattern recognition.*

[14] Woo S, Park J and Lee J Y 2018 Cbam: Convolutional block attention module *Proceedings of the European conference on computer vision (ECCV).*

[15] Long X, Cui W and Zheng Z 2019 Safety helmet wearing detection based on deep learning *Information technology, networking, electronic and automation control conference (ITNEC).*

[16] Tan M, Pang R and Le Q V 2020 Efficientdet: Scalable and efficient object detection *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.*