International Core Journal of Engineering Volume 8 Issue 3, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

A Computer Vision-based Approach for Detecting Safety Harnesses and Helmets

Zhijing Xu[a], Jiajing Huang[b]

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

[a]zjxu@shmtu.edu.cn, [b]jjhuang96@163.com

Abstract

Wearing safety helmets and safety harnesses when working at heights on construction sites is an effective means to protect workers from accidents. An improved YOLOv5 detection method for safety harnesses and helmets is proposed to improve the detection performance of small objects in complex backgrounds. Firstly, a mixed attention mechanism is introduced to design the backbone network, which can effectively suppress the negative impact of complex backgrounds and improve the model detection performance. Secondly, a cross-layer complementary feature fusion network is constructed to strengthen the fusion between high and low-level features. It can improve the model's ability to adapt to small and medium-sized objects. Finally, DIoU-NMS is used to reduce the over-suppression of bounding boxes caused by the proximity of objects. A large number of experiments are carried out on a self-built dataset. The results show that the mean Average Precision (mAP) of the proposed method reaches 94.5%, which is better than that of mainstream detection methods, and the frame rate per second is 48.

Keywords

YOLOv5; Attention Mechanism; Feature Fusion; Safety Harnesses and Helmets.

1. Introduction

Accidents caused by construction workers not wearing helmets during construction and accidents caused by falling due to not wearing safety harnesses when working at height are the more frequent types of accidents at construction sites[1,2]. The environment of construction sites is very complex and the number of construction workers is large. The way of safety supervision by safety manager is less efficient.

The current rapid development of computer vision technology and target detection technology has provided a more efficient and intelligent method for construction site safety supervision. Xu et al.[3] introduced Focal Loss to the loss function based on the YOLOv3 algorithm and used GIoU Loss as the border loss to improve the helmet recognition accuracy. Gu et al.[4] used the improved OpenPose to estimate the posture of construction workers, determined the head region of workers according to different postures, and used YOLOv4 to The helmet wearing situation was detected. Wang et al.[5] combined YOLOv3 with FaceNet to identify the violating workers while detecting the helmet wearing situation. Chu et al.[6] introduced a self-attention module based on SSD target detection algorithm for helmet recognition. Gu et al.[7] optimized Faster R-CNN by multi-scale training, increasing the number of anchor points and introducing OHEM (Online Hard Example Mining) to improve the detection accuracy. Li et al.[8] improved the Faster R-CNN by using Focal loss instead of the original loss function and using ROI Align instead of ROI Pooling to construct a helmet recognition network for complex operational scenarios. Pang et al.[9] simplified the MTCNN network structure by replacing the maximum pooling layer using a common convolutional layer and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

introducing MobileNet to improve the accuracy and detection speed. Due to the lack of public datasets, there are fewer related studies on safety harness detection. Zhang et al.[10] used safety harness model images as matching templates to automate detection by OpenCV template matching functions. Fang et al.[1] used two convolutional neural network (CNN) models to determine whether workers were wearing safety harness while working at heights.

An analysis of previous studies shows that the complex background of construction sites plays a negative role in the target detection process, which can easily lead to false detection. Safety harnesses and helmets belong to small targets, and the wearing of safety equipment by workers far from the camera is prone to missed detection. To address the problems of existing methods, this paper proposes a safety harness and helmet detection method based on the improved YOLOV5. The backbone network is designed by introducing the CBAM block structure[11] in the residual structure. This can effectively enhance the focus of the model on the target to be detected and reduce the influence of the complex background. The proposed cross-layer complementary feature fusion network can enrich the feature semantic information and enhance the small target localization detection capability. The DIoU-NMS[12] (Distance-IoU Non Maximum Suppression, NMS) is chosen to replace the original NMS to reduce excessive suppression caused by the proximity of the target.

2. Proposed Method

Considering the dual requirements of accuracy and real-time for target detection algorithms in practical engineering applications, this paper designs a model structure based on YOLOv5s[13] as shown in Figure 1.

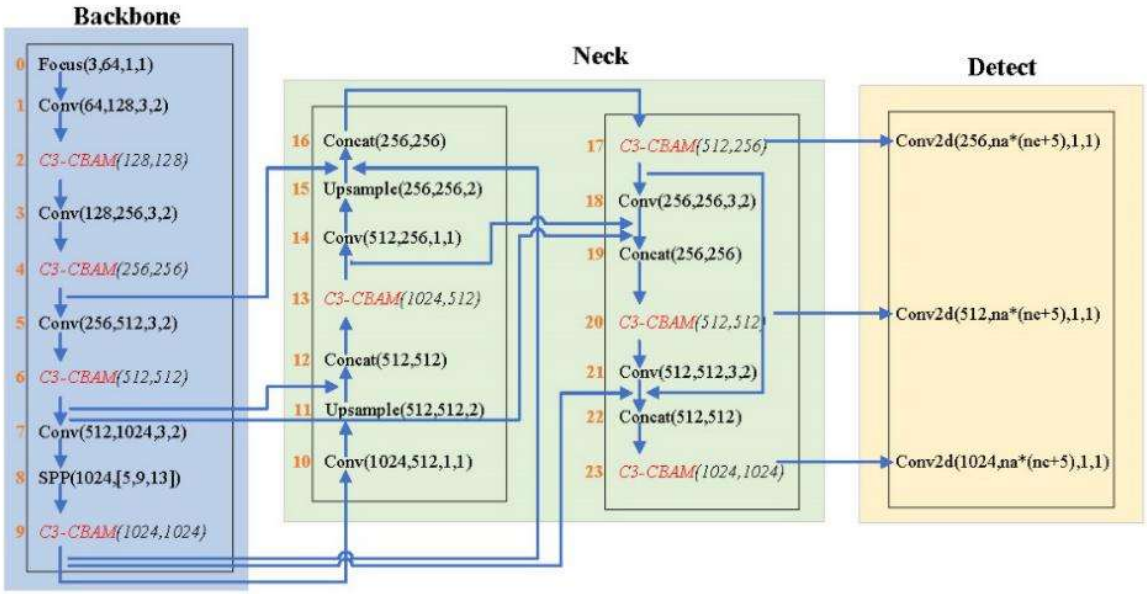

Figure 1. The model structure proposed in this paper

The input image is firstly pre-processed at the input side by scaling, rotating and color space adjustment of the input image, and the Mosaic data enhancement method is used. Multiple data enhancement methods are used to effectively enrich the data samples and improve the robustness of the network model, so that the model can be adapted to detect safety harnesses and helmets in more construction scenarios. Due to the different sources and sizes of the images in the self-built dataset, it is also necessary to uniformly scale the original images input to the network to 640×640, which helps to improve the inference speed of the model. Second, the scaled images are passed into the backbone network proposed in this paper for preliminary feature extraction. Third, the cross-layer complementary feature fusion network constructed is used to deeply fuse the obtained feature maps

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

at each scale to further enhance the fusion of high-level semantic information and the underlying features to realize the advantages of complementarity between high and low-level features. It can improve the detection performance of small targets in the model. Finally, the fully fused feature maps of three different scales are used for target prediction, and DIoU-NMS is used to screen the generated detection frame to obtain the final detection result.

2.1 Backbone Network Design

The backbone network is a key part of the object detection network. It is responsible for extracting the features of the object to be detected. Facing the complex environment, the backbone network cannot fully focus on the object to be detected. By introducing an attention mechanism, the model can pay more attention to the object to be detected, and improve the performance of the model.

Currently, attention mechanisms are mainly divided into three categories: channel attention, spatial attention, and mixed attention. This paper introduces mixed attention in the backbone network. Figure 2 shows the structure of a CBAM block[11], which consists of channel attention and spatial attention. The CBAM block is a serial structure, and its execution process can be represented by the following formula:

$$F' = M_c(F) \otimes F, \tag{1}$$

$$F'' = M_s(F') \otimes F', \tag{2}$$

Among them, F represents the feature map input into the module, and F' represents the feature map after the input is weighted by the attention of the M_c channel. Taking F' as new input, M_s is the weight of the spatial attention, and the final feature map F'' is obtained. The details of M_c and M_s are shown in the formulas:

$$\begin{aligned} M_c(F) &= \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F^c avg\right)\right) + W_1\left(W_0\left(F^c max\right)\right)\right), \end{aligned} \tag{3}$$

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}\left(\left[AvgPool(F);\ MaxPool(F)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s avg;\ F^s max\right]\right)\right), \end{aligned} \tag{4}$$

where σ represents the sigmoid activation function. $AvgPool$ and $MaxPool$ represent average pooling and max pooling. MLP represents a multi-layer perceptron.

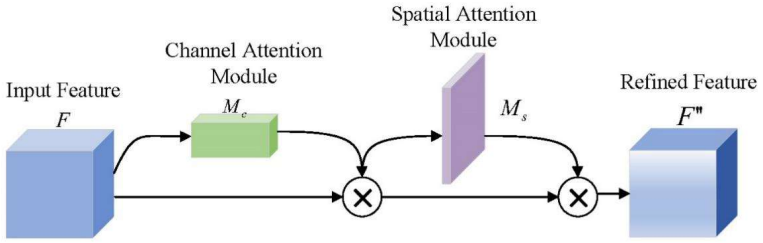

Figure 2. The overview of CBAM

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

The CBAM block can be used as a plug-and-play module. The C3 structure is the basic module of the yolov5 backbone network[13]. In order to weaken the negative impact of complex background in the process of feature extraction, and at the same time do not increase too much computation. We combined the CBAM block with the residual structure and designed the C3-CBAM structure, the details of which are shown in Figure 3.

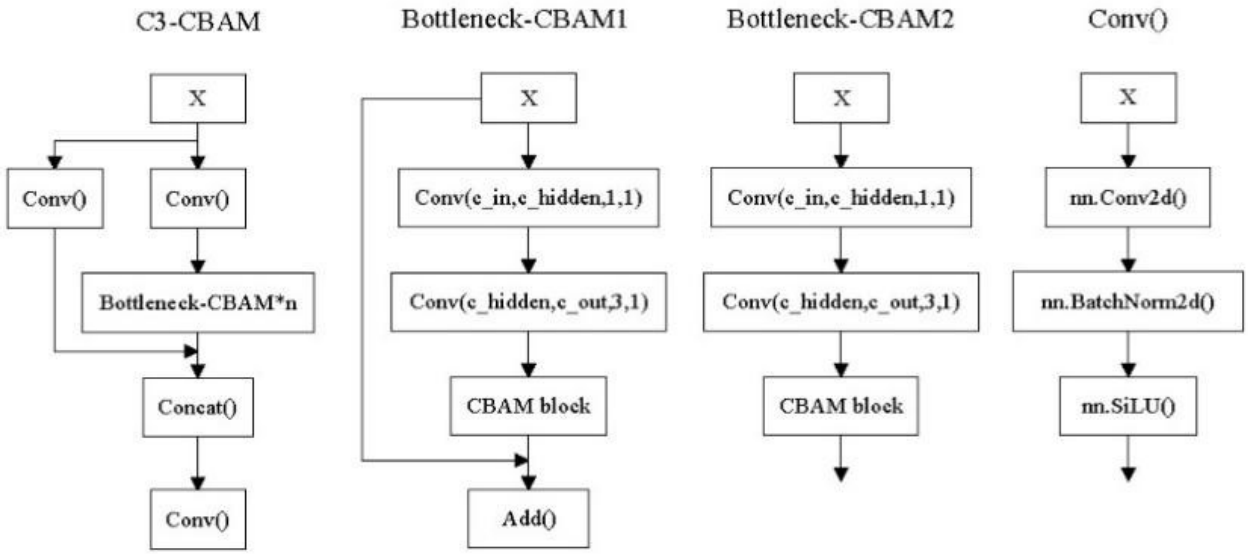

Figure 3. The overview of C3-CBAM

2.2 Cross-layer Complementary Feature Fusion Network

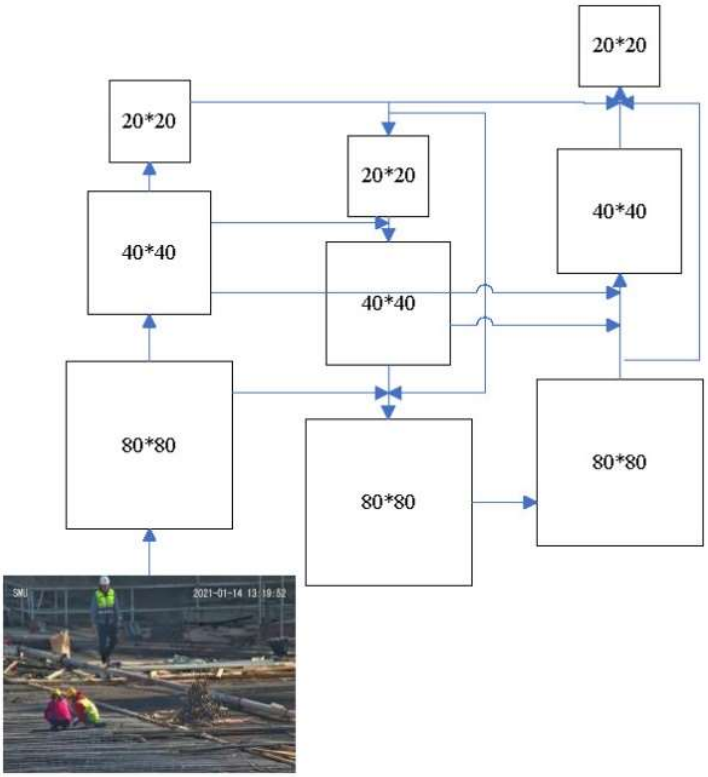

Figure 4. The cross-layer complementary feature fusion network

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

According to previous researches, low-level features contain more location information, and high-level features contain more semantic information. With the deepening of the network, the resolution of the feature map gradually becomes smaller, and the high-level semantic information extracted by the model becomes more and more abundant. However, the spatial location information is continuously lost, and some small objects are easily lost in the high layers of the network, which is not conducive to the detection of the model[14,15].

The cross-layer complementary feature fusion network we proposed in this paper makes full use of the advantages of high and low layer features. Commix high-level semantic information and localization information to break the imbalance between high-level features and low-level features. It can improve model performance.

The structure shown in Figure 4 is the cross-layer complementary feature fusion network proposed in this paper. The top-down structure transfers the rich semantic information of the high-level to the lower-level through the up-sampling operation, and the bottom-up structure transfers the low-level positioning information upwards to improve the utilization of low-level features. Considering that the distance between high-level and low-level is far, and high-level feature information is more likely to be lost, cross-level connections in the horizontal and vertical directions are designed. The horizontal connection in the figure fuses the middle and high-level feature maps at the same level through the Concat operation so that the overall features are more abundant, and the vertical connection is used to quickly transfer the high and low-level features to improve the detection accuracy.

2.3 DIoU-NMS

In the inference and prediction link, the detection end needs to screen a large number of overlapping bounding boxes. NMS uses the Intersection over Union (IoU) of the predicted box and the real box as the only evaluation index. In the actual construction site, the activities of workers are relatively frequent. When the distance between the two objects to be measured is very close, the IoU value is usually larger, and the bounding boxes of the adjacent objects are easily eliminated. Therefore, this paper used DIoU-NMS to suppress overlapping bounding boxes, and while considering the IoU index, the distance between the center points of the two prediction boxes is also judged[12]. The formula for DIoU-NMS is as follows:

$$R_{DIoU}(M,B_i)=\frac{\rho^2(M,B_i)}{c^2}, \tag{5}$$

$$s_i=\begin{cases} s_i & IoU-R_{DIoU}(M,B_i)<\varepsilon \\ 0 & IoU-R_{DIoU}(M,B_i)\geq\varepsilon \end{cases}, \tag{6}$$

$\rho(M,B_i)$ represents the Euclidean distance between the center points of the two prediction boxes. c represents the diagonal length of the minimum circumscribed rectangle of the two overlapping bounding boxes, and ε is the set threshold size.

3. Experiments and Results

3.1 Dataset

There are currently no publicly available data sets of safety harnesses for workers working at heights. For this, we built a dataset with 3150 images for training and testing. The dataset comes from three main sources: field video recordings, adaptations of the hard hat dataset published by Wang et al.[16], and web crawlers. We use LabelImg to annotate the collected images in four categories: safety-harness, dangerous, helmet, head. Working at heights without a safety lanyard is an extremely dangerous act, and we have marked it as dangerous. Safety-harness and dangerous mark the location

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

of the worker's chest (where safety harnesses often appear), and helmet and head mark the worker's entire head. We divided the dataset into training set, validation set, and test set according to 16:4:5.

Figure 5 shows the base case of the dataset. The dataset includes common safety harness styles and colors on construction sites. In order to alleviate the problem that reflective clothing is easily mistakenly identified as a safety harness, we specifically added reflective clothing of various styles and colors to the data set. The construction scenes contained in the dataset are also relatively rich.

Figure 5. Dataset

3.2 Ablation Experiments

In order to fully verify the effectiveness of each module proposed in this paper, based on YOLOv5s, ablation experiments are used for comparative analysis. The effectiveness of the redesigned backbone network, the cross-layer complementary feature fusion network, and DOU-NMS is verified. The experimental results are shown in Table 1. Using DIoU-NMS for post-processing can slightly improve the model performance and reduce the excessive suppression of the bounding boxes caused by the proximity of workers. After using the backbone network constructed by C3-CBAM, the mAP of the model increased from 93.4% to 93.9%. This shows that the CBAM block can make the model pay more attention to the objects under detection. It can reduce the negative impact caused by the complex background. The cross-layer complementary feature fusion network proposed in this paper increases mAP from 93.4% to 94.2%. This shows that the feature fusion structure designed in this paper can effectively integrate the information of high and low layers so that the advantages of high and low layers are fully complementary. When the three modules are introduced at the same time, the mAP of the model is improved by 1.1%.

Table 1. Ablation experiments based on YOLOv5s

DIoU-NMS	C3-CBAM	Our FPN	mAP/%
-	-	-	93.4
√	-	-	93.6
-	√	-	93.9
-	-	√	94.2
√	√	√	94.5

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

3.3 Comparison of Different Network Architectures

In order to further verify the effectiveness of the method proposed in this paper, this paper compares the current more popular target detection algorithms. All experiments are done in the same environment and on the same training set, validation set, and test set, and the confidence level is uniformly set to 0.7. Table 2 shows the comparison results of different detection methods. As can be seen from the table, the method proposed in this paper is superior to the comparison methods in detection performance and speed, and the model size is only 19.1MB, which is easy to deploy. Figure 6 shows the comparison of experimental methods and different algorithms.

Table 2. Comparison results with other object detection methods

Algorithm	mAP/%	FPS	Weights file/MB
Faster-RCNN	89.05	17	521
SSD	85.98	39	92.1
YOLOv4	89.14	25	244
Ours	94.5	48	19.1

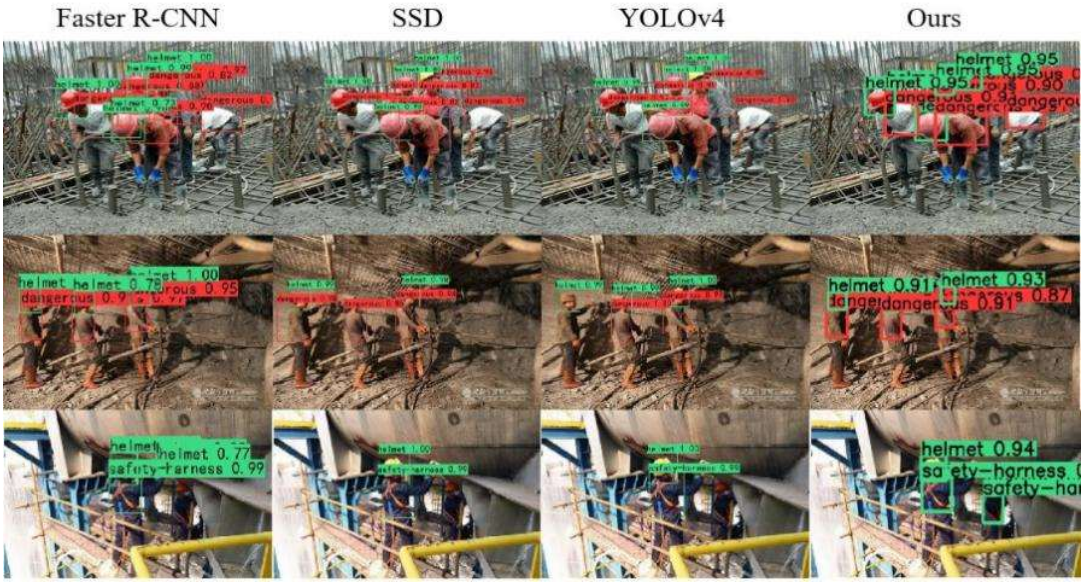

Figure 6. Comparison of test results

4. Conclusion

This paper proposes an improved YOLOv5 model for the detection of safety harnesses and helmets during building construction. By introducing the CBAM module to improve the backbone network, so as to improve the feature expression ability of the model. A cross-layer complementary feature fusion network is designed. It makes full use of the advantages of high-level and low-level features, breaks the imbalance between them, and achieves complementary advantages. The detection performance of the model for small and medium-sized objects is improved. In the post-processing stage, DIoU-NMS is used to reduce the excessive suppression of bounding boxes caused by the proximity of the objects. Through the comparison experiment with mainstream methods, it is confirmed that the method proposed in this paper can accurately identify the personal safety protection equipment in the complex construction environment. Detection speed meets real-time application requirements.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0075

References

[1] W. Fang, L. Ding, H. Luo, and P.E.D. Love. Falls from heights: A computer vision-based approach for safety harness detection [J]. Automation in Construction, 2018, 91: 53-61.

[2] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, and W. An. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos [J]. Automation in Construction, 2018, 85: 1-9.

[3] K. Xu and C. Deng. Research on Helmet Wear Identification Based on Improved YOLOv3 [J]. Laser & Optoelectronics Progress, 2021, 58(06): 300-307.

[4] Y. Gu, Y. Wang, L. Shi, N. Li, and S. Xu. Automatic detection of safety helmet wearing based on head region location [J]. IET Image Processing, 2021, 15(11): 2441-2453.

[5] H. Wang, Z. Hu, Y. Guo, Y. Ou, and Z. Yang, A Combined Method for Face and Helmet Detection in Intelligent Construction Site Application. 2020: Recent Featured Applications of Artificial Intelligence Methods. LSMS 2020 and ICSEE 2020 Workshops.

[6] Y. Chu, Y. Huang, X. Zhang, and H. Liu. SSD image target detection algorithm based on self-attention [J]. J.Huazhong Univ.of Sci.& Tech.(Natural Science Edition) 2020, 48(09): 70-75.

[7] Y. Gu, S. Xu, Y. Wang, and L. Shi. An Advanced Deep Learning Approach for Safety Helmet Wearing Detection[C]. in 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). pp.669-674.

[8] H. Li, Y. Wang, P. Yi, T. Wang, and C. Wang. Research on recognition of safety helmets under complex operation scenes based on deep learning [J]. Journal of Safety Science and Technology, 2021, 17(01): 175-181.

[9] S. Pang and S. Lu. Multi-scale safety helmet detection based on improved MTCNN [J]. Application Research of Computers, 2021, 38(06): 1907-1912+1916.

[10] J. Zhang, Y. Han, J. Yao, and S. You. Design and Implementation of Automatic Inspection System for Safety Equipment of Construction Workers [J]. Construction Technology, 2017, 46(24): 83-86.

[11] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon. CBAM: Convolutional Block Attention Module. in Proc. Eur. Conf. Comput. Vis. (ECCV). 2018: Munich, Germany. pp.3-19.

[12] Z. Zheng, P. Wang, W. Liu, J. Li, and D. Ren. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression[C]. in AAAI Conference on Artificial Intelligence.Conference, Year.

[13] Ultralytics. YOLOv5. 2021; Available from: https://github.com/ultralytics/yolov5.

[14] T.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection[C]. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Conference, Year.

[15] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path Aggregation Network for Instance Segmentation[C]. in IEEE. Conference, Year.

[16] L. Wang, L. Xie, P. Yang, Q. Deng, and L.J.S. Xu. Hardhat-Wearing Detection Based on a Lightweight Convolutional Neural Network with Multi-Scale Features and a Top-Down Module [J]. 2020, 20(7): 186 8.