IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 19, 2022, accepted June 2, 2022, date of publication June 8, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3180796

Investigation Into Recognition Algorithm of Helmet Violation Based on YOLOv5-CBAM-DCN

LIJUN WANG[1], YUNYU CAO[1], SONG WANG[1], XIAONA SONG[1], SHENFENG ZHANG[2], JIANYONG ZHANG[3], AND JINXING NIU[1]

[1]School of Mechanical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450011, China
[2]Suzhou Jiesheng Technology Company Ltd., Suzhou 215011, China
[3]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough TS1 3BA, U.K.

Corresponding author: Jinxing Niu (niujinxing@ncwu.edu.cn)

This work was supported in part by the Scientific and Technological Project of Henan Province under Grant 212102210069 and Grant 222102210043, in part by the ''ZHONGYUAN Talent Program'' of Henan Province under Grant ZYYCYU202012112, in part by the Zhengzhou Measurement and Control Technology and Instrument Key Laboratory under Grant 121PYFZX181, and in part by the Fund of Innovative Education Program for Graduate Students at North China University of Water Resources and Electric Power under Grant YK-2021-90.

ABSTRACT Recognition of safety helmets wearing by construction workers is a common target detection topic in applications of deep learning-based image processing. This paper provides a study of an enhanced YOLOv5-based method, in which the challenges caused by complicated construction environment backgrounds, dense targets, and the irregular shape of safety helmets are addressed. In a trunk network, feature extraction is more based on the target shape by using the Deformable Convolution Net instead of the conventional convolution; in the Neck, a Convolutional Block Attention Module is introduced to weaken feature extraction of complex backgrounds by giving weights to enhance the characterization ability of target features; and the original network's Generalized Intersection over Union Loss is replaced by Distance Intersection over Union Loss to overcome the problem of erroneous location when the population is dense. The dataset for the training network is created by mixing open-source datasets with autonomous collecting to evaluate the effectiveness of the algorithm. We observed that the improved model has a detection accuracy of 91.6%, up 2.3% over the original network model, and a detection speed of 29 frames per second, which is compliant with most security cameras' capture frame rate.

INDEX TERMS Safety helmet detection, attention mechanism, deformable convolution, YOLOv5.

I. INTRODUCTION

Construction is a high-risk business, and safety helmets are the most basic personal protective equipment and are critical to the safety of construction professionals. However, the importance of helmets is often disregarded due to lack of safety awareness. Statistical examination of pertinent data revealed that there were a total of 2133 construction safety production accidents in the country, with 2478 deaths, with high-altitude falls accounting for 52.41% of all accidents [1]. Helmet supervision is an essential part of the building operating environment. Manual supervision methods are commonly used in construction units, but with such excessive scope of supervision, it is impossible in practice to track and manage all workers in a timely manner. Therefore, helmet wearing detection under intelligent monitoring based on image processing is gradually becoming the main means for companies to implement management. The automatic supervision is faster in operation, more reliable with a larger coverage and low cost. As such it is more conducive to achieving the goal of on-site supervision than the traditional methods.

Computer vision detection algorithms can be classified into two types: traditional computer image detection algorithms and deep learning detection algorithms. In the traditional object detection algorithm, a window sliding strategy is adopted to achieve target positioning in a traversal manner with windows of various sizes and aspect ratios. Target regions are extracted according to the manually

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

designed features and then classified using machine learning algorithms. Clearly, such an exhaustive algorithm-based sliding window selection strategy can result in excessive number of redundant windows, which slow down the network model's detection process. Furthermore, the manually designed features have significant restrictions, which leads to a weak generalization ability that governs the entire algorithm [2]. As a result, academics across the world are shifting their focus from the traditional algorithms manual features-based algorithms to the deep learning-based target detection algorithms. The focus of deep learning-based target detection algorithms development is in two directions: 1) two-stage object detection algorithms such as the Region CNN (R-CNN) series, and 2) single-stage object detection algorithms, e.g., the YOLO series, Single Shot MultiBox Detector (SSD), and others [3], [4].

A common target detection challenge in helmet supervision is about 'whether the helmet is worn appropriately'. Many fellow researchers have undertaken extensive studies on the detection of safety helmets and their research outcomes are available. Yang Liqiong *et al.* combined the classical object detection and deep learning-based object detection approaches by using YOLOv3 to locate the face region, and the Support Vector Machines (SVM) classifier was chosen to classify the obtained features [5]. SenseNet was implemented in YOLOv3 by Daniel Zhang *et al.* to process the low-resolution feature layer, which could better pinpoint the position of the safety helmet using K-means mean clustering. An enhanced detection speed was achieved while ensuring accuracy [6]. Adding feature layers to the YOLOv3 network structure had increased the network's detection effect on small targets as suggested by Xu and Deng [7]. Guo and the research team replaced CSPdarknet53, the YOLOv4's backbone network, with MobileNetV3, and utilized the deep separable convolution instead of the regular convolution, a model that was both simpler and more accurate was realized [8].

Based on the aforementioned studies, this paper proposes a safety helmet detection method using YOLOv5-CBAM-DCN with an attention mechanism and deformable convolution, which addresses the problem of insufficient accuracy suffered by the traditional target algorithms due to the complex background of the site environment, uneven lighting, and irregular shape of the target. The following are the article's significant innovations and contributions:

(1) Replace traditional convolution with the Deformable Convolution Net (DCN) to improve the adaptability of the range of visual perception [9].

(2) Build YOLOv5 detection network and introduce the Convolutional Block Attention Mechanism (CBAM) in Neck to boost safety helmets' capacity to express themselves in complex working settings.

(3) The Distance Intersection over Union (DIoU) Loss is used to replace the Generalized Intersection over Union (GIoU) Loss, which solves the problem of non-regression caused by the dense distribution of workers by taking into consideration the distance between the center locations when calculating the border loss.

II. PRINCIPLES OF YOLOV5 ALGORITHM

YOLOv5 is a single-stage object detection model made up of four components: Input, Backbone, Neck, and Head. In comparison to the original YOLOv4 network model, this model adds the Focus module [10], at the same time, it also adopts the improved the Cross Stage Partial Network (CSPNet) as the backbone of the network to extract image features, a bottom-up Path Aggregation Network (PANet) layer based on the Feature Pyramid Structure (FPN) is also added in to strengthen the multi-scale feature fusion mechanism [11]. The structure of this model is shown in Fig. 1. These four elements are explained as follows:

Input: YOLOv5 uses the same Mosaic data enhancement method as YOLOv4 to scale and stitch each of the four images entered, enriching the background of the detected object and increasing batch size from the side. With this method, the number of GPUs can be reduced [12]. The network adapts the size of the input sample to 640*640 by adaptive picture scaling to improve detection outcomes. The adaptive anchor box calculation method is used to generate the preset anchor box during the model training process, and the network parameters are updated by backward transmission according to the deviation of the preset anchor box from the ground truth box.

Backbone: Backbone networks are convolutional neural networks that aggregate and form image features at different image fine-grains. The backbone of YOLOv5 is CSPDarknet, which is comprised of three parts: Focus, Cross Stage Partial Network (CSP), and Spatial Pyramid Pooling (SPP). By slicing the 640*640*3 (Width*Height*Channel) image entered in and using 32 convolutional kernel convolution operations with a size of 3*3*12, the Focus module generates four 320*320*3 feature layers, as seen in Fig. 2. And then connects the channels from depth though the Concat layer, as seen in Fig. 1, narrowing the data width.

The SPP structure uses pooling kernels of size 1*1, 5*5, 9*9, and 13*13 to perform the maximum pooling operation on the feature layer so that inputs of different sizes have feature vector outputs of the same size, increasing the range of visual perception while enhancing the nonlinear representation of the network [13].

In YOLOv5, there are two CSP structures, i.e., CSP1_X and CSP2_X, as shown in Fig. 3. In CSP1_X, the input splits into two branches, each performing convolution operations separately to halve the number of channels per branch. In one branch, the information passes through the CBL module (CBL = convolution + regularization + activation function) and then into several residual structures (Res units), in the other, information is directly convoluted. The two branches are then concatenated so that the model learns more features while keeping the size of BottleneckCSP output consistent with that of the input. CSP1_X is largely used in Backbone networks [14], [15], and its residual structure plays a key role

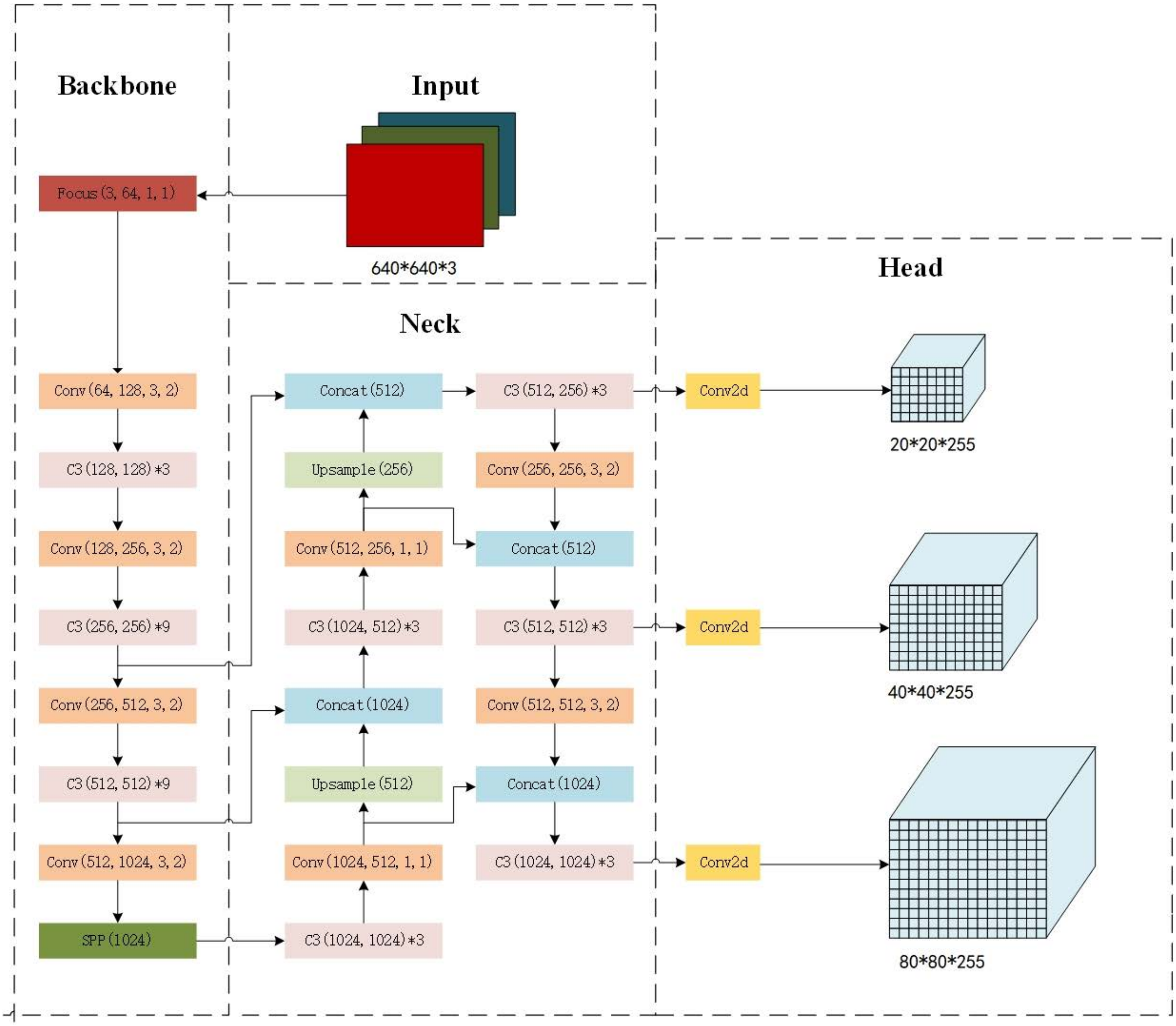


FIGURE 1. Schematic diagram of the YOLOv5 network structure.

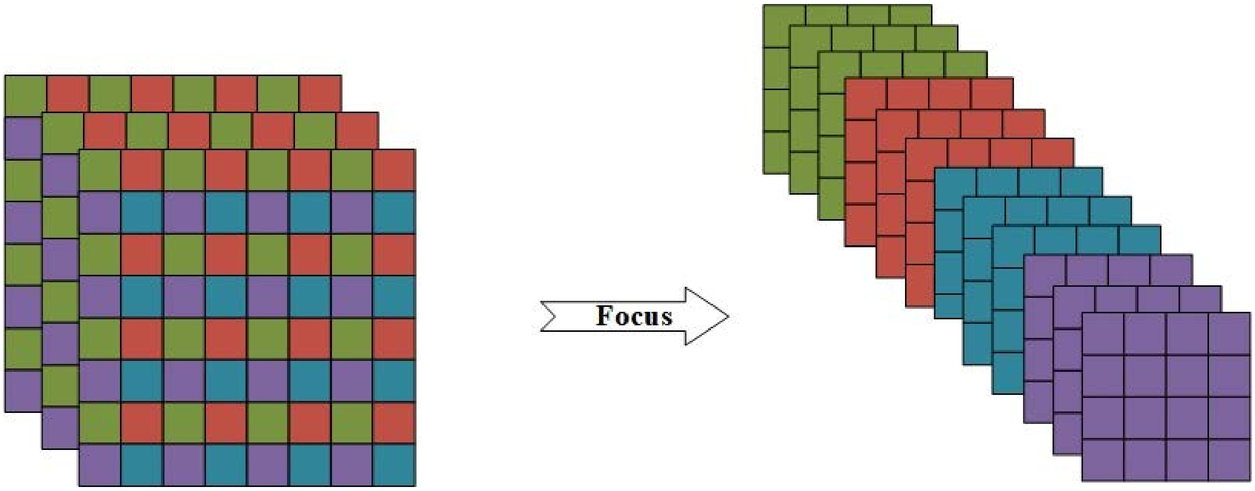


FIGURE 2. Schematic diagram of the focus module.

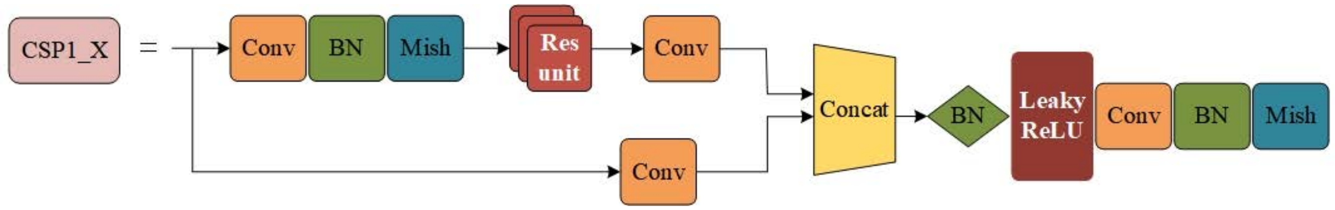


FIGURE 3. CSP1_X structure diagram.

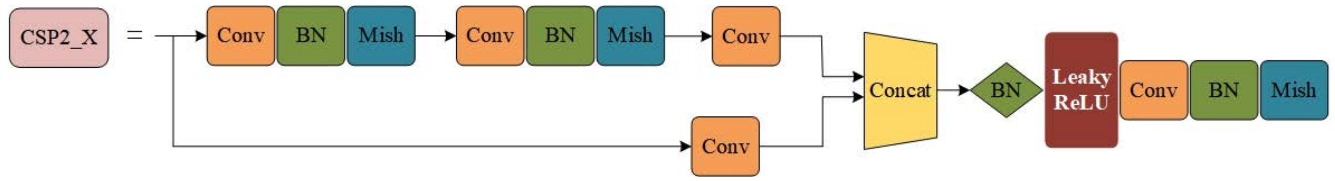


FIGURE 4. CSP2_X structure diagram.

in increasing backpropagation gradient values between layers and preventing gradient disappearance caused by network deepening. CPS2_X is primarily used in Neck networks. Compared to CSP1_X, in the CPS2_X, only the residual structure is replaced with numerous CBL modules on the structure, as seen in Fig. 4.

Neck: FPN is utilized to fuse shallow position information with deep semantic information, and the Neck part of YOLOv5 is used to fuse PANet on the FPN structure and add a bottom-up downward sampling process after top-down upsampling [16].

Head: YOLOv5 network model uses the GIoU Loss function to estimate the recognition loss of the rectangular box that detects the target.

III. IMPROVEMENTS BASED ON YOLOV5

A. CONVOLUTIONAL ATTENTION MECHANISM MODULE

The attention mechanism works by making the network adaptive through weight allocation and information filtering, with which, more valuable information for network training is extracted from the enormous amount of feature information and passed to the convolutional stack [17]. Two CBAM modules are added after the C3 module of the YOLOv5 network model [18], the construction of CBAM is depicted in Fig. 5. These modules are used to improve the characteristic expression of targets in the complex environment.

CBAM is an attention mechanism module that combines channel and spatial attention mechanisms. The Global Average Pooling (AvgPool) and Global Max Pooling (MaxPool) in the spatial dimension are first performed on the feature layers of the input to obtain rough global sensing fields, and then the shared full connection layer is used to construct the correlation between the channels based on the above two pooling results, and finally the results of the full connection layer processing are added and sent to the activation function Sigmoid [19]. The final result is a normalized weighted Mc with the same number of feature channels as the input [20]. The formula for the computation is shown in (1).

$$\begin{aligned} M_C(F) &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \end{aligned} \tag{1}$$

where W_0 and W_1 are the weights of the first and second layers, respectively. $\sigma(*)$ is the sigmoid activation function, and Multilayer Perceptron (MLP) represents two fully connected shared layers. MLP is used because with single layer of full connection, the nonlinear problem cannot be solved accurately. The first layer of full connectivity reduces the dimension to 1/16 of the original C, while the second layer restores it to C. The characteristic of AvgPool and MaxPool are represented by F^c_{avg} and F^c_{max}, respectively.

In CBAM a spatial attention mechanism is introduced to compensate the loss of information. Such loss is generated due to the channel attention mechanism that does not take location information into account [21]. Following the channel weighting process, the feature layer serves as an input to the spatial attention mechanism $\grave{F}$. First, global average pooling and global maximum pooling are performed on $\grave{F}$ in the channel dimension, and the maximum and average values of each feature point in the channel dimension are taken as the pooling results. The convolution operation with a convolution kernel size of 7*7 adjusts the two feature layers acquired by pooling to a feature layer, and this feature layer can be activated by the Sigmoid function to obtain a normalized weight Ms, which has the same number as the number of input features. Formula (2) is used for the computation. The size of the convolutional kernel is 7*7, as indicated by $f^{7\times 7}$ in the formula.

$$\begin{aligned} &M_S(F) \\ &\quad = \sigma\left(f^{7\times 7}\left(AvgPool\left(\grave{F}\right)\right) + f^{7\times 7}\left(MaxPool\left(\grave{F}\right)\right)\right) \\ &\quad = \sigma\left(f^{7\times 7}(\left[F^s_{avg}; F^s_{max}\right])\right) \end{aligned} \tag{2}$$

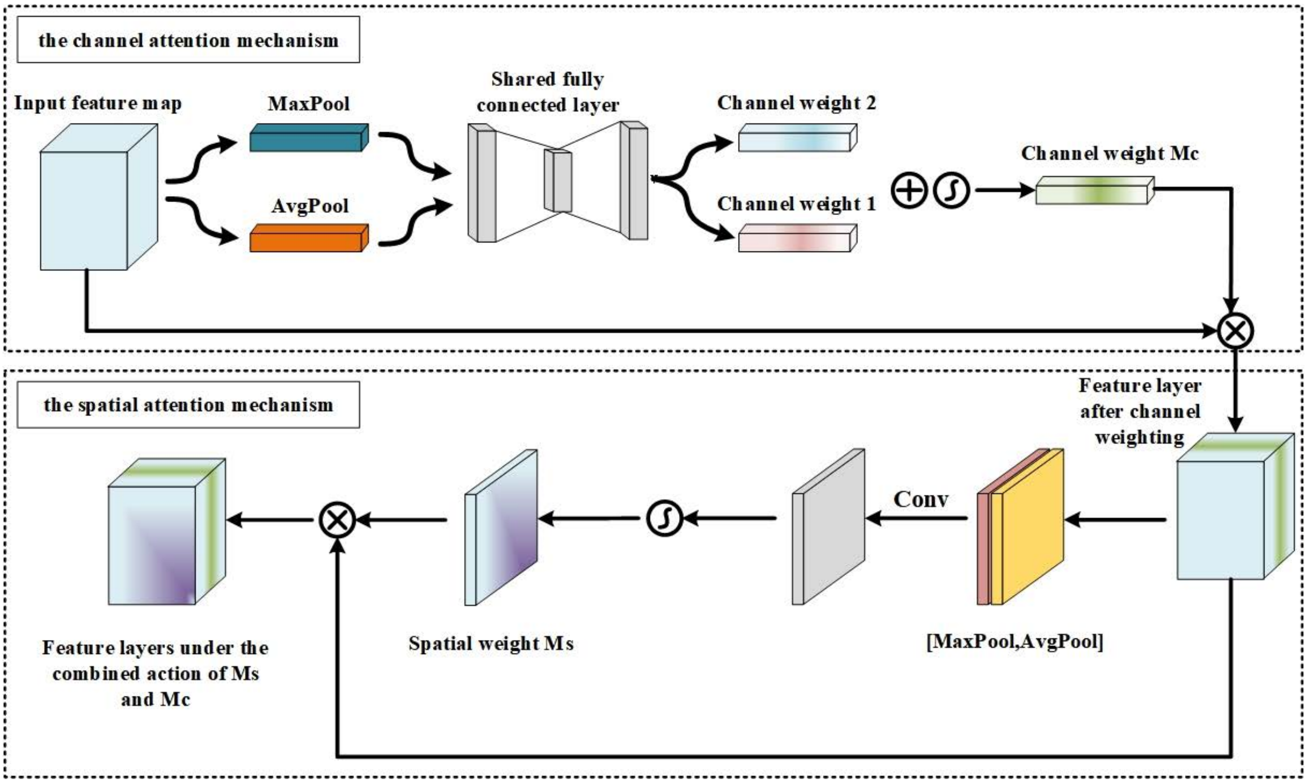

the channel attention mechanism
Input feature map
MaxPool
AvgPool
Shared fully connected layer
Channel weight 2
Channel weight 1
Channel weight Mc
the spatial attention mechanism
Feature layer after channel weighting
Conv
[MaxPool,AvgPool]
Spatial weight Ms
Feature layers under the combined action of Ms and Mc

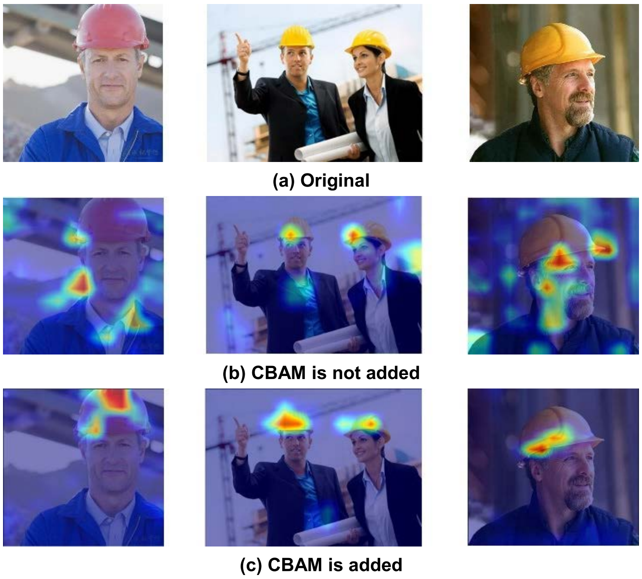
FIGURE 5. The CBAM modul.

The two CBAM modules are introduced in the Neck structure in this study to improve accuracy and detection time of the network detection. The model comparison results are displayed in Fig. 6, where the red regions indicate the regions that need high attention when the network extracts features, and the darker color indicates higher significance. From the experimental results, we observed that the incorporation of the CBAM module enhances the characteristic expression of targets in complicated situations, which is clearly reflected in Fig. 6 (c).

B. DEFORMABLE CONVOLUTIONAL

In typical traditional convolution operations, fixed-size convolutional kernels are employed, and the sensing field is always rectangular, regardless of network depth or width [22]. Because of this constraint, the generalizing capacity of traditional convolution is limited. Even if the CBAM module is added to the network, it remains a difficulty that feature extraction cannot cover the entire targeted item because a large number of background features may be extracted. To overcome this problem, DCN with geometric deformation capability is employed to improve target feature characterization in this study, so that the sample point is concentrated as much as possible on the target to be identified.

In contrast to the classic convolution, in the deformable convolution, an offset is added to each sample point, which alter the shape of the convolutional kernel as the offset varies.


(a) Original

(b) CBAM is not added

(c) CBAM is added

FIGURE 6. Comparison chart before and after CBAM joins YOLOv5.

This is depicted in Fig. 7. The position of the sampling points is modified during the training phase via the offset learning. In Fig. 7, the green dots represent typical convolutional sampling points, the blue represents the sample points after applying the deformation convolution, and the arrows indicate the offset direction.

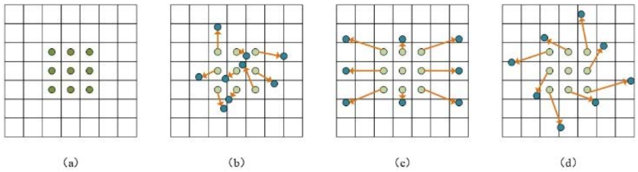


FIGURE 7. **a) Conventional convolution, convolutional kernel size 3 * 3; (b) Deformable convolution, increasing the offset on top of (a); (c) (d) Special forms of deformable convolution.**

Assuming the grid space in the image above is $R = \{(-1, 1), (-1, 0) \ldots (0, 1), (1, 1)\}$, the output representation of the standard convolution given by is shown in (3) and the output of the deformable convolution is displayed in (4). The value of x in (3) and (4) is calculated using bilinear in-line insertion as shown in (5) and (6), where g (a, b) is also defined because the offset Δp_n is usually a decimal number. p_{n} is the number of feature elements in the grid, Δp_n is the offset, $w(*)$ is the sampling weight, and $g(q_x, p_x) \cdot g(q_y, p_y)$ is a two-dimensional bilinear interpolated convolutional kernel, $x(q)$ is the value of the point at the integer position on the feature map. Bilinear interpolation uses four points in the original image to calculate the target pixel for interpolation, q_x and q_y represent the horizontal ordinate coordinate of the above four integer points, p_x and p_y represent the offset ordinate coordinate, and x(p) represents the value after bilinear interpolation.

$$y(p_0) = \sum\nolimits_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{3}$$

$$y(p_0) = \sum\nolimits_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + p_n) \tag{4}$$

$$x(p) = \sum\nolimits_{q} g(q_x, p_x) \cdot g(q_y, p_y) \cdot x(q) \tag{5}$$

$$g(a, b) = max(0, 1 - |\mathrm{a} - \mathrm{b}|) \tag{6}$$

The process of feature extraction using deformable convolution is shown in Fig. 8. One of change made in this study is that after the Focus module, the convolution layer is substituted with a deformable convolutional layer. A 3*3 convolution kernel is used to convolve the input feature map in the original structure. In the deformable convolution module, an additional 3*3 convolutional layer conv' is defined to learn about the offsets. The output dimension after convolution is consistent with the input feature map, and the number of channels has been changed from N to 2N, which is corresponding to the offset in the x and y directions, respectively. The offsets obtained from the convolutional layer conv' are bilinearly interpolated with the Input feature map to obtain the final Output feature map. A comparison of the convolution range of traditional and deformable convolution is shown in Fig. 9. Compared with traditional convolution, the sampling points of deformable convolution are more consistent with the location and shape of the detected target, which is helpful to enhance the saliency of target features.

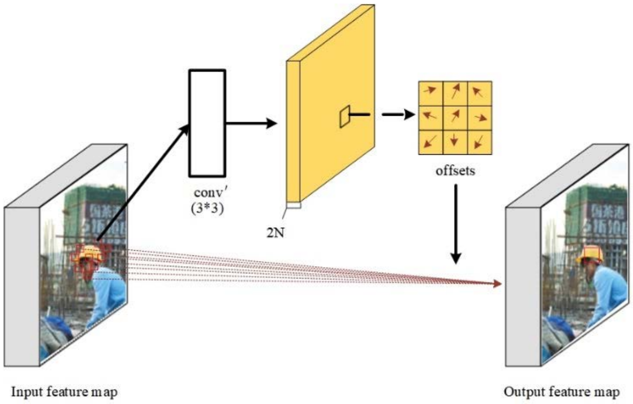


FIGURE 8. **Extraction of 3*3 deformable convolutional features.**

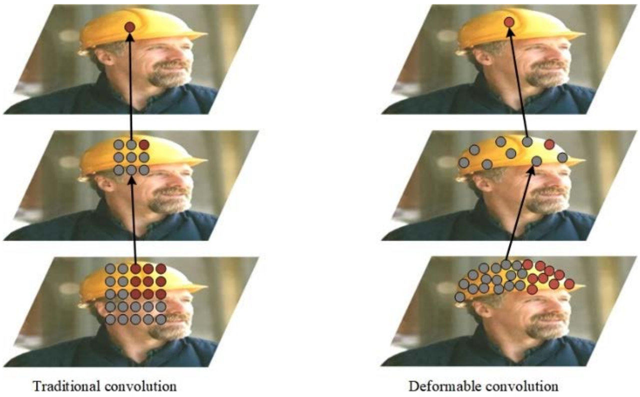


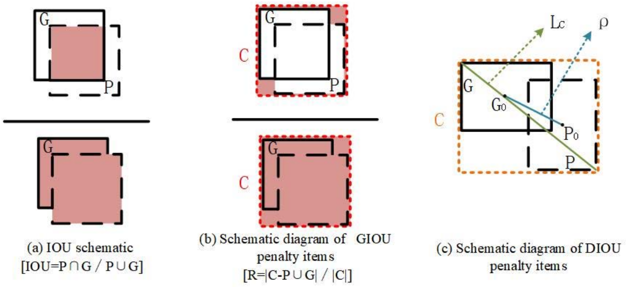
FIGURE 9. **Visualization of the range of sensations.**

C. OPTIMIZATION LOSS FUNCTIONO

The loss function is an operation function used to measure the degree of difference between prediction box (P) and the real box (G), in which L_n Loss is used to optimize the four parameters, i.e., the length, width, and center coordinates x and y of the candidate box as four independent variables.

Because the relationship between the variables is neglected, it leads to inaccurate regression box positioning. Based on this, the researchers opened a study based on Intersection over Union (IoU) Loss [23].

In the object detection task $IoU(P, G) = (P \cap G) / (P \cup G)$. In other words, IoU represents the ratio of the intersection and concatenation of P and G as shown in Fig. 10 (a). The IoU Loss function is illustrated in (7), and the analysis indicates that the gradient cannot be transferred back when $IoU(P, G) = 0$, that happens when the prediction box does not intersect with the real box. To address this flaw, in the YOLOv5, GIoU Loss is used as described in (8) to introduce a penalty term R, where $R = |C - P \cup G| / |C|$ [24]. Although GIoU resolves the problem of non-regression caused by disjoint between two boxes, GIoU Loss degenerates into IoU Loss when a box is wrapped inside another box. Hence GIoU does not reflect the quality of the regression state. As a result,



(a) IOU schematic [IOU=P∩G / P∪G]

(b) Schematic diagram of GIOU penalty items [R=|C-P∪G| / |C|]

(c) Schematic diagram of DIOU penalty items

FIGURE 10. Loss function schematic diagram.

TABLE 1. Experimental platform parameter configuration.

Configuration name	Version parameters
operating system	Ubuntu18.04
Graphics (GPU)	NVIDIA GeForce RTX 3080
Processor (CPU)	Intel Core i9-10900KF @ 3.70GHz
Deep learning framework	Pytorch

DIoU Loss illustrated in (9) is introduced in this article to include the center point distance in the consideration range, as well as the penalty term $R' = \rho^2 \left(P_0, C_0\right)/L_C^2$.

$$L_{IoU} = 1 - IoU\left(P, G\right) \tag{7}$$

$$L_{GIoU} = 1 - IoU\left(P, G\right) + |C - P \cup G|/|C| \tag{8}$$

$$L_{DIoU} = 1 - IoU\left(P, G\right) + \rho^2\left(P_0, G_0\right)/L_C^2 \tag{9}$$

where P and G stand for prediction and real boxes, respectively. P_0 and G_0 are for center points of the P and G boxes, C for the area of the two boxes of the closure area, $\rho\left(*\right)$ for the Euclidean distance between the center points of the two boxes, and L_C for the diagonal distance between the two boxes of the closure area.

IV. EXPERIMENTS AND ANALYSIS

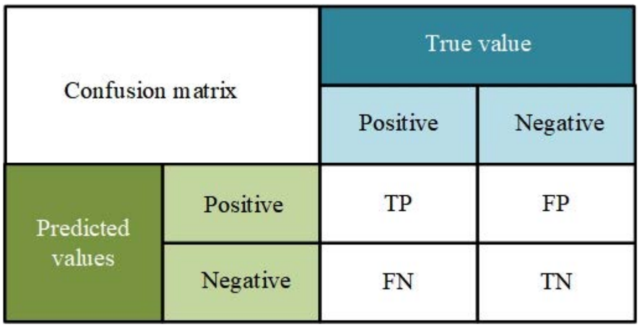
A. EXPERIMENTAL PLATFORM SETTINGS AND DATA ACQUISITION

The hardware and software platform configurations used in the experiment are detailed in Tab. 1.

Through autonomous collection, real shooting, reorganization of open-source datasets, and web crawlers, a total of 7450 photos were acquired as a dataset for model training in this study. LabelImg is used to annotate images into two categories, and they are saved in VOC format to improve dataset readability. The images are then transformed to YOLO format using script code and divided into 8:2 training and validation sets.

B. EXPERIMENTAL PARAMETER SETTING AND EVALUATION INDEX

In the experiments, the training set and validation set are uniformly scaled to 640*640 size to improve the reliability of the data in the model training process. The data enhancement method Mosaic is used to increase the diversity of the data through randomly cutting and stitching any four images to increase the diversity of data; and the asynchronous random gradient descent (SGD) method is adopted to update gradients. In the training process of the model, 32 sample images are fed into the network as a batch in 8 times in each training round. We are consistent with the original network architecture of YOLOv5 for the settings of hyperparameters during model training. The momentum is set to 0.937 so as to avoid the model training to fall into local optimum; the weight decay coefficient is set to 0.0005 to prevent overfitting; the starting learning rate is 0.01 and the termination learning rate is 0.2, and the ideal weights of the model are obtained after 200 rounds of training.



Confusion matrix		True value	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

FIGURE 11. Binary confusion matrix.

Model evaluation is a crucial task in the field of deep learning, and evaluation indicators commonly include precision, recall, and Mean Average Precision (mAP), among others. The precision rate, also known as the accuracy rate, is used to represent the likelihood that an algorithmically predicted positive class sample is correct. The number of correctly classified positive class samples as a percentage of the actual number of positive class samples is described as the recall rate. The most widely used assessment index in object detection is the mAP, which is determined by taking the accuracy rate of each single category when IoU = 0.5 [25], [26]. Each indicator's calculating formula is given in (10) to (12):

$$Precision = TP/(TP + FP) \tag{10}$$

$$Recall = TP/(TP + FN) \tag{11}$$

$$mAP = \sum\nolimits_{i=1}^{N} \int_0^1 P_i\left(r\right) dr/N \tag{12}$$

The accuracy of each category is represented by $P_i\left(r\right)$ in the formula, and i is the category ordinal number, TP, FP, and FN in the formula are illustrated by the binary confusion matrix, as shown in Fig. 11. The TP indicates the number of samples in which positive class samples are correctly predicted as positive, the FP indicates the number of samples in which positive class samples are incorrectly predicted as positive, and the FN indicates the number of samples in which positive class samples are incorrectly predicted as negative.

C. TRAINING RESULTS

The red and black curves shown in Fig. 12 are the average mean accuracy curves of the network designed in this

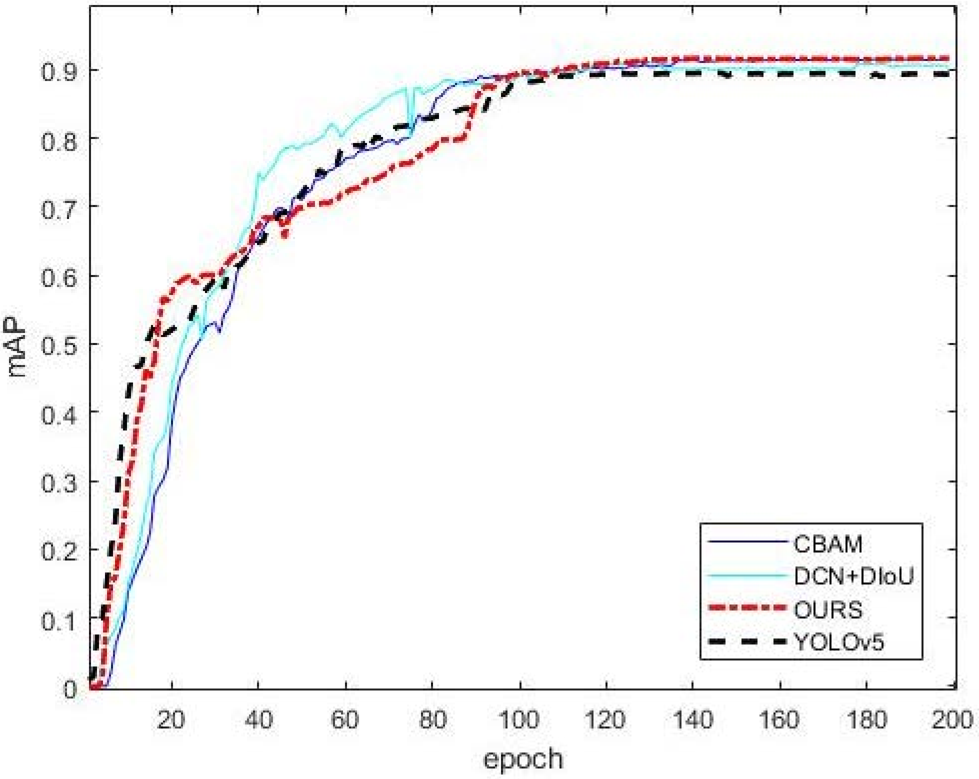


FIGURE 12. Comparison curves of different improved model experiments.

article and the original YOLOv5 network were trained for 200 rounds under the same configuration, respectively. The results are shown in Fig. 12, where the abscissa is the training time epoch, and the ordinate coordinate represents the mAP. According to experimental results, we observed that both the algorithm designed in this paper and the original network structure of YOLOv5 converge rapidly in the first 50, and gradually stabilize after 100 rounds, without overfitting during the training process. The average detection accuracy of the improved network model in this paper rises to 0.7 after the first 50 rounds of iteration and finally stabilizes at 0.916; the original YOLOv5 network model slowly rises to 0.7 after 50 rounds of iteration and eventually stabilizes at 0.893. The improved model has a 2.3% improvement in accuracy over the original YOLOv5 network model, which verifies the feasibility of the improved strategy.

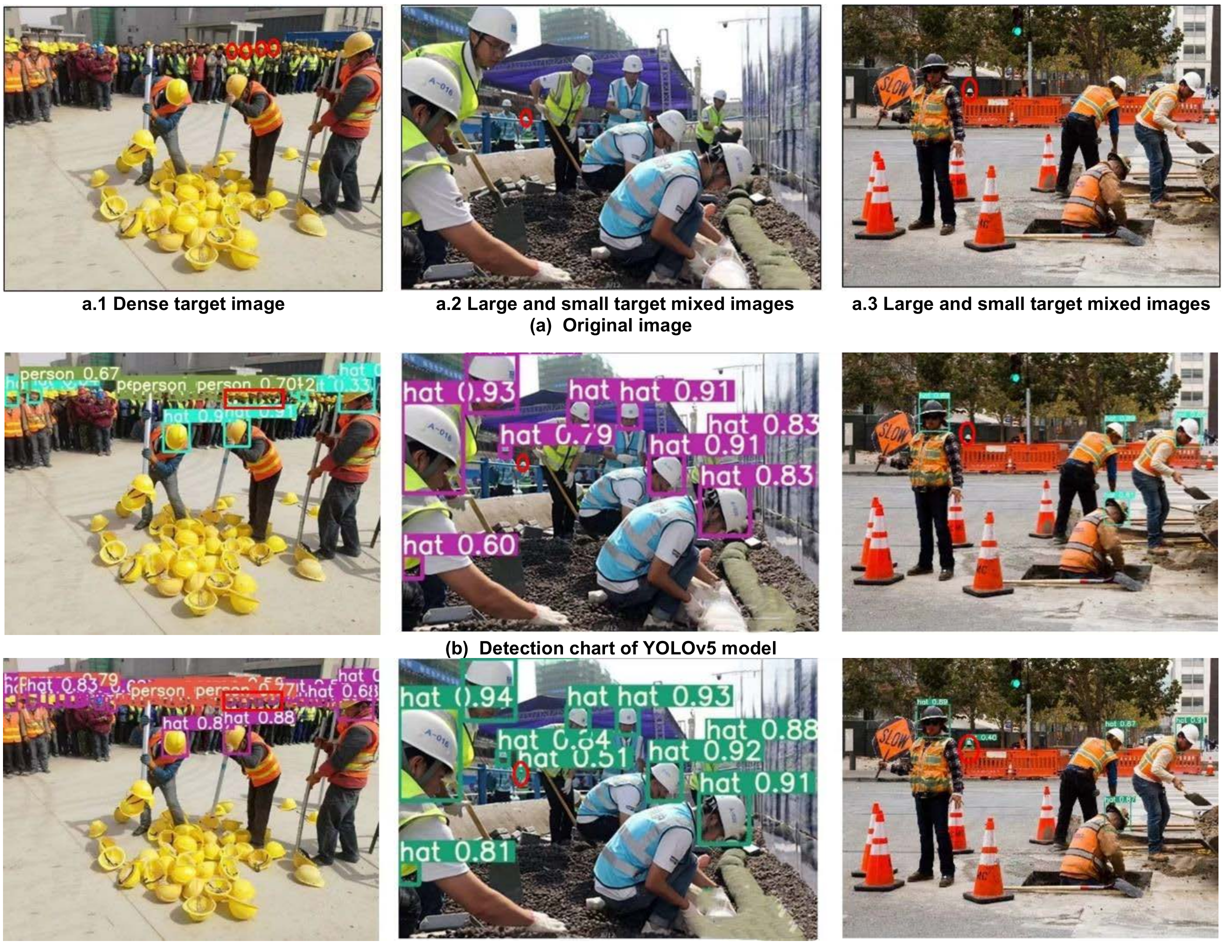
D. COMPARISON OF DETECTION RESULTS

Some of the detection findings are presented in Fig. 13. The targets that overlooked by the original model and detected by the upgraded model are circled in red in the diagram. Figure (a) shows the original image captured, Figure (b) displays the original YOLOv5 network detection result, and Figure (c) show the enhanced network detection result. The picture of a dense target helmet is a.1, whereas the photos of helmets with mixed size targets are a.2 and a.3.

When comparing (b) and (c), we observed that the upgraded network model can discover targets that the original model misses due to their small size or overlap, and the improved model has a higher confidence score for target categorization. Additionally, the detection results in Figure a.1 indicate that during the detection of helmets, the trained model only recognizes helmets that are appropriately worn on the head and does not identify helmets that are placed in other locations. This shows that the trained network model can prevent false detection when the worker does not correctly wear the helmet, such as holding it in his hand or hanging it on other parts of his body.

E. BLATION EXPERIMENTS

In this paper, an ablation experiment was also performed to verify the effect of each module on the model performance, and the results are contained in Tab. 2. The first row in the table are the results based on the unimproved YOLOv5 model detection, whilst the subsequent rows indicate the results with the addition of CBAM modules, DIoU loss functions, and DCN modules to the network, respectively. By comparing the experimental results under the above different methods, we observed that by adding two CBAM modules in Neck the network's feature extraction ability is improved in a complex working environment, and 2.1% improvement in detection accuracy over the original network. For the situation where the loss function is changed from GIoU to DIoU, and the



a.1 Dense target image

a.2 Large and small target mixed images

a.3 Large and small target mixed images

(a) Original image

(b) Detection chart of YOLOv5 model

(c) Detection diagram of the improved model

FIGURE 13. Comparison chart of detection effect before and after network improvement.

TABLE 2. Ablation experiment comparison.

CBAM	DIOU+DCN	mAP@0.5/%	Recognition speed /(frames/s)
		89.3	32
√		91.4	27
	√	90.1	31
√	√	91.6	29

deformable convolution is added to the network, a 0.8% improvement in detection accuracy over the original network is achieved. The network created by combining the three modifications has the best detection effect, which further verifies the feasibility of the improved model.

F. COMPARATIVE EXPERIMENTS ON OTHER NETWORKS

A comparative test is done using the existing target identification algorithm with higher comprehensive performance in order to objectively evaluate the superiority of the enhanced algorithm in this work, and the identical samples and parameter settings are utilized in the experiment. The experimental results are shown in Tab. 3. In comparison to other algorithms, we observed that the average detection accuracy of the upgraded algorithm is 3.2% higher than that of PP-YOLO and 2.3% higher than that of YOLOv5. The detection speed of the improved algorithms is 19frames per second, which is 2 frames per second higher than the detection speed of PP-YOLO and basically meets the detection requirements (the frame rate of surveillance camera acquisition is normally 25 frames/s to 30 frames/s) [27]. The comparative experiment once again verifies the superiority and feasibility of the model.

TABLE 3. Experimental comparison of different algorithms.

Model	mAP@0.5/%	Recognition speed /(frames/s)
Faster RCNN	88.4	13
YOLOv3	82.5	24
YOLOv4	85.1	20
YOLOv5	89.3	32
PP-YOLO	90.5	27
OURS	91.6	29

V. CONCLUSION

From this study, an improved model based on YOLOv5-CBAM-DCN is provided to tackle the low detection accuracy problem suffered by the traditional target algorithm due to the complex background of the site environment, uneven lighting, and irregular shape of the targe. The average detection accuracy of the improved model has been improved, the missed detection rate of the safety helmet has been reduced, and the detection speed also meets the standard frame rate of general monitoring cameras. According to experimental verification on the self-made dataset. The article algorithm's detection impact for the safety helmet in wet situations is currently not optimal, therefore consider boosting object detection in such background environments by increasing pre-processing of picture de-fogging and other factors.

REFERENCES

[1] Z. Ma, "Detection on wearing behavior of safety helmet based on machine learning method," *Urban rural Stud.*, vol. 7, no. 3, pp. 52–55, Feb. 2022.

[2] Y.-Z. Xiao, Z.-Q. Tian, J.-C. Yu, Y.-S. Zhang, S. Liu, S.-Y. Du, and X.-G. Lan, "A review of object detection based on deep learning," *Multimedia Tools Appl.*, vol. 79, no. 33, pp. 23729–23791, Sep. 2020.

[3] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[4] H. Pan, Y. Li, and D. Zhao, "Recognizing human behaviors from surveillance videos using the SSD algorithm," *J. Supercomput.*, vol. 77, no. 7, pp. 6852–6870, Jan. 2021.

[5] L. Q. Yang, L. Q. Cai, and S. Gu, "Detection on wearing behavior of safety helmet based on machine learning method," *J. Saf. Sci. Technol.*, vol. 15, no. 10, pp. 152–157, Oct. 2019.

[6] Y. Zhang, K. P. Wu, K. Gao, and X. Yang, "Helmet detection based on modified YOLOV3," *Comput. Simul.*, vol. 38, no. 5, pp. 413–417, May 2021.

[7] K. Xu and C. Deng, "Research on helmet wear identification based on improved YOLOv3," *Laser Optoelectronics Prog.*, vol. 58, no. 6, pp. 300–307, Mar. 2021.

[8] S. H. Guo, J. R. Jing, X. D. Zhang, and X. H. Qin, "Research on detection of safety helmet wearing based on improved YOLOv4," *J. Saf. Sci. Technol.*, vol. 17, no. 12, pp. 135–141, Dec. 2021.

[9] Z. Liu, B. Yang, G. Duan, and J. Tan, "Visual defect inspection of metal part surface via deformable convolution and concatenate feature pyramid neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9681–9694, Dec. 2020.

[10] F. Jubayer, J. A. Soeb, A. N. Mojumder, M. K. Paul, P. Barua, S. Kayshar, S. S. Akter, M. Rahman, and A. Islam, "Detection of mold on the food surface using YOLOv5," *Current Res. Food Sci.*, vol. 4, pp. 724–728, Oct. 2021.

[11] Z.-Z. Wang, K. Xie, X.-Y. Zhang, H.-Q. Chen, C. Wen, and J.-B. He, "Small-object detection based on YOLO and dense block via image super-resolution," *IEEE Access*, vol. 9, pp. 56416–56429, 2021.

[12] L. Chen, M. Zheng, S. Duan, W. Luo, and L. Yao, "Underwater target recognition based on improved YOLOv4 neural network," *Electronics*, vol. 10, no. 14, p. 1634, Jul. 2021.

[13] Y. S. Tan, K. M. Lim, C. Tee, C. P. Lee, and C. Y. Low, "Convolutional neural network with spatial pyramid pooling for hand gesture recognition," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 5339–5351, Sep. 2020.

[14] B. Yan, P. Fan, X. Lei, Z. Liu, and F. Yang, "A real-time apple targets detection method for picking robot based on improved YOLOv5," *Remote Sens.*, vol. 13, no. 9, p. 1619, Apr. 2021.

[15] W. Li, W. Sun, Y. Zhao, Z. Yuan, and Y. Liu, "Deep image compression with residual learning," *Appl. Sci.*, vol. 10, no. 11, p. 4023, Jun. 2020.

[16] S. Chen, Z. Zhang, R. Zhong, L. Zhang, H. Ma, and L. Liu, "A dense feature pyramid network-based deep learning model for road marking instance segmentation using MLS point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 784–800, Jan. 2021.

[17] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[18] Y.-L. Li and S. Wang, "HAR-Net: Joint learning of hybrid attention for single-stage object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3092–3103, 2020.

[19] W. Cao, Z. Feng, D. Zhang, and Y. Huang, "Facial expression recognition via a CBAM embedded network," *Proc. Comput. Sci.*, vol. 174, pp. 463–477, Jan. 2020.

[20] M. Xue, M. Chen, D. Peng, Y. Guo, and H. Chen, "One spatio-temporal sharpening attention mechanism for light-weight YOLO models based on sharpening spatial attention," *Sensors*, vol. 21, no. 23, p. 7949, Nov. 2021.

[21] R. Ranjbarzadeh, A. B. Kasgari, S. J. Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–17, May 2021.

[22] Z.-D. Zhang, M.-L. Tan, Z.-C. Lan, H.-C. Liu, L. Pei, and W.-X. Yu, "CDNet: A real-time and robust crosswalk detection network on Jetson nano based on YOLOv5," *Neural Comput. Appl.*, vol. 1, pp. 1–12, Feb. 2022.

[23] J. Shen, X. Xiong, Y. Li, W. He, P. Li, and X. Zheng, "Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 36, no. 2, pp. 180–196, Jun. 2020.

[24] X. Liu, J. Hu, H. Wang, Z. Zhang, X. Lu, C. Sheng, S. Song, and J. Nie, "Gaussian-IoU loss: Better learning for bounding box regression on PCB component detection," *Expert Syst. Appl.*, vol. 190, no. 15, Mar. 2022, Art. no. 116178.

[25] M. H. Yap, R. Hachiuma, A. Alavi, R. Brüngel, and E. Frank, "Deep learning in diabetic foot ulcers detection: A comprehensive evaluation," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104596.

[26] I. Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104519.

[27] M.-Z. Yuan, L. Gao, H. Fu, and S. Xia, "Temporal upsampling of depth maps using a hybrid camera," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 3, pp. 1591–1602, Mar. 2019.

LIJUN WANG received the B.S. degree from the North University of China, Taiyuan, China, in 1993, the M.S. degree from Henan Agricultural University, Zhengzhou, China, in 2003, and the Ph.D. degree from PLA Information Engineering University, Zhengzhou, in 2010.

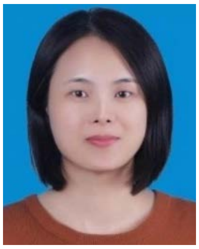
Since 2010, she has been a Professor with the School of Mechanical Engineering, North China University of Water Resources and Electric Power, Zhengzhou, where she is currently the Dean of the School of Mechanical Engineering and the Director of the Zhengzhou Key Laboratory of Measurement and Control Technology and Instrument. She has led approximately 20 important scientific research projects. She has authored over 60 scientific articles in international journals and conferences. Her current research interests include signal processing, image processing and machine learning, fault diagnosis, and intelligent control.

YUNYU CAO received the B.S. degree in electrical and information engineering from the Anhui University of Technology, Anhui, China, in 2020. She is currently pursuing the M.S. degree with the School of Mechanical Engineering, North China University of Water Resources and Electric Power, Zhengzhou, China.

Her current research interests include deep learning, object detection, and image processing.

SONG WANG received the B.S. degree in mechanical engineering from the North China University of Water Resources and Electric Power, Zhengzhou, China, in 2020, where he is currently pursuing the M.S. degree with the School of Mechanical Engineering.

His current research interests include deep learning, image processing, and pattern recognition.



XIAONA SONG received the B.S. degree in automation from the Nanjing University of Science and Technology, in 2005, and the Ph.D. degree in mechanical engineering from the Army Engineering University of PLA, in 2020.

She is currently a Lecturer with the North China University of Water Resources and Electric Power. Her current research interests include deep learning, object recognition, computer vision, and machine learning.

SHENFENG ZHANG received the B.S. and M.S. degrees in mechanical engineering from the North China University of Water Resources and Electric Power, Zhengzhou, China, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in business administration with the University of Chinese Academy of Social Sciences, Beijing, China.

He is currently working as the Manager of the Research and Development Department and the Vice President of the Research Institute, Suzhou Jiesheng Technology Company Ltd. He has been engaged in machine vision, artificial intelligence, optical measurement, signal acquisition, and fault diagnosis. He has published 12 papers and 17 patents. His current research interests include machine vision, artificial intelligence, image processing, signal acquisition, and troubleshooting.

JIANYONG ZHANG received the B.S. degree from Northeastern University, Liaoning, China, in 1983, the M.S. degree from the University of Science and Technology, Beijing, China, and the Ph.D. degree from Teesside University, U.K., in 2002.

He worked as an Associate Professor at USTB. Prior to joining Teesside University. Since joining Teesside University, he has worked on several EU and British government funded projects, and led several international, U.K., government and industry funded projects as a Principal Investigator. He is currently an Associate Professor in instrumentation and control with Teesside University. Over 80 papers, including book chapters, peer-reviewed journals and conference papers have been published. His current research interests include measurement and control, energy saving, industrial applications, signal conditioning and processing, and sensing technologies.

Dr. Zhang has been a Reviewer of several international journals, including *Electronics* journal, IEEE SENSORS, and IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. He is a member of the Institute of Measurement and Control and a Contact Person of EU RFCS.

JINXING NIU received the B.S. and M.S. degrees from Henan University, Kaifeng, China, and the Ph.D. degree from the Xi'an Institute of Optics and Precision Mechanics of CAS, Xi'an, China, in 2010.

He is currently a Teacher with the North China University of Water Resources and Electric Power. His current research interests include machine vision and machine learning.