



# Enhanced YOLOv5 algorithm for helmet wearing detection via combining bi-directional feature pyramid, attention mechanism and transfer learning

Yinfeng Fang<sup>1</sup> · Yuhang Ma<sup>1</sup> · Xuguang Zhang<sup>1</sup> · Yuxi Wang<sup>1</sup>

Received: 1 November 2022 / Revised: 31 December 2022 / Accepted: 12 January 2023 /  
Published online: 3 February 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

The complexity of infrastructure scenarios has led to a sustained increase in the global number of worksites related fatalities and injuries. Therefore, safety helmets play an essential role in protecting construction workers from accidents. It is essential to detect whether the helmets are correctly worn on the heads for smart construction site. However, due to the complex construction environments, it is challenging to precisely detect safety helmet wearing in real-time. This paper proposes an enhanced version of You Only Look Once version 5 (YOLOv5) to improve the detection accuracy, where bi-directional feature pyramid network (BiFPN), attention mechanism, and transfer learning are fully integrated. The BiFPN is taken to replace the original feature pyramid network (FPN) via adding additional cross layer edges with adaptive connecting weights. Attention mechanism is added after the end of backbone and neck network to let the network pay more attention on the interested region. Transfer learning is adopted for model training. The model is pre-trained by a head detection database and then fine-tuned by the helmet database. The proposed enhanced YOLOv5 is tested on a public GDU-HWD dataset, where both helmet and its color can be identified. This study achieves the accuracy at 93.3% , which is 4.8% higher than that of the original YOLOv5, but does not bring in much computing burden to the network. It is believed that the enhance version can also be successfully used in other similar detection tasks.

**Keywords** Helmet wearing detection · YOLO · Transfer learning · Attention mechanism · BiFPN

---

✉ Xuguang Zhang  
zhangxg@hdu.edu.cn

Yinfeng Fang  
yinfeng.fang@hdu.edu.cn

Yuhang Ma  
yuhang.ma@hdu.edu.cn

Yuxi Wang  
yxwang@hdu.edu.cn

<sup>1</sup> School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, 310018, China

# 1 Introduction

## 1.1 Research background

Construction is one of the most dangerous job sectors. The fatal injury rate for the construction industry is higher than other industries. The construction fatalities are always caused by the combination of different factors. Among them, traumatic brain injuries account for about 24% of all construction fatalities in the United States [1]. High-altitude falling objects occasionally injure the workers in areas of mining, electric power, chemical industries and other work areas [6]. Thus, to ensure the safety of the construction sites, an increasing number of helmet monitoring systems based on computer vision have been developed. It can not only reduce the workload of manual monitoring, but also highlight the unsafe operations and avoid the occurrence of accidents [9].

Monitoring safety equipment wearing before the operation, especially to detect whether helmets are properly worn, can effectively reduce the occurrence or cost of accidents. Traditional monitoring technologies largely rely on the observation and inspection by the experienced managers on site, which generally exists the problems of low automation level, large workload and limited inspection items. In addition, manual inspection is difficult to conduct continuous supervision during operation, which becomes one of the most potential safety hazards for construction sites.

With the background of internet of things and big data, smart construction site [2] has been proposed to improve the quality of safety monitoring via taking the advantage of artificial intelligence. The information can be transmitted to a big data management server for security analysis to fulfill automatic alarm or stop in real time. Object detection technology plays a critical role in these smart monitoring systems, where helmet detection becomes a hot research point. After years of development, helmet wearing detection have shifted from traditional machine learning methods to deep learning ones.

Object detection algorithms based on deep learning are generally divided into two categories: two-stage detection and one-stage detection method. The main idea of two-stage detection is to first generate a series of sample candidate boxes and classify these samples through a convolutional neural network. Some classic algorithms include the RCNN and Faster R-CNN [22]. Due to the RPN structure, two-stage method represented by the Faster R-CNN has a high detection accuracy but a slow speed, which makes it difficult to reach real-time processing requirements for construction scenes. Unlike the two-stage detection, one-stage detection does not necessarily generate sample candidate boxes, but directly converts the object location into regression problem, such as SSD [13], YOLO [20] and other algorithms. One-stage method can achieve the shared features of a single training, the speed can be significantly improved while keeping a certain accuracy. However, the SSD commonly fails to detect small-scale objects because of its inherent properties of weak features at the bottom layers of high resolution. Therefore, the YOLO series have a wider application in object detection tasks [4, 34], and have been continuously updated in recent years. YOLOv5 algorithm [29] is mostly appreciated due to its balanced performance among speed, accuracy and robustness. This paper proposes an enhanced YOLOv5 algorithm based on combining Bi-directional feature pyramid, attention mechanism and transfer learning.

## 1.2 Related works

From traditional methods to deep learning methods, helmet detection attracts great attentions [24]. Li et al. [9] proposed an innovative and practical safety helmet wearing detection

method based on image processing and machine learning. Firstly, Visual Background Extractor (ViBe) background modelling algorithm is exploited to detect motion region of interest, and then the Histogram of Oriented Gradient (HOG) feature is extracted to describe inner human. Secondly, with the input of HOG feature, the Support Vector Machine (SVM) is trained to identify pedestrians. Finally, the safety helmet is detected by its color. Mneymneh et al. [17] proposed a framework to monitor helmet wearing through detecting moving objects using standard deviation matrix (SDM) and then identifying human using the aggregate channel feature-based object detector. After that, a cascade object detector based on HOG features searched for helmets in the upper area of the identified personnel. Yue et al. [33] proposed a new Improved Boosted Random Ferns algorithm (IBRFs) for safety helmet wearing status detection. Firstly, based on HOG feature to construct random ferns, then weak classifiers are constructed. Finally, selected the most discriminative ones to build a strong classifier to detect the wearing status of the safety helmet. This method outperforms some of deep learning methods, including SSD, YOLOv3 and Faster R-CNN.

In the field of deep learning methods, Wu et al. [32] proposed an improved SSD algorithm to improve the efficiency of small target detection. This work developed a novel aggregation framework combined with the presented reverse progressive attention (RPA), which propagates the semantically strong features back to the bottom layers progressively. Deng et al. [3] proposed a lightweight YOLOv3 algorithm for safety helmet detection. This work integrated the CSPNet and GhostNet to design a more efficient residual network, and designed a new backbone network, the ML-Darknet. It solves the problem that YOLOv3 is expensive to calculate and difficult to deploy on mobile devices. Song et al. proposed [25] a novel object detection model based on anchor-free mechanism—Recurrent Bidirectional Feature Pyramid Detector (RBFPDet), it composed of multiple RBFP recursive units. This study regards helmet wearing detection as a strong semantic feature points detection task, and improves the accuracy of helmet wearing detection.

Although great effort has been made in helmet detection, it still suffers from the following problems: (1) unacceptable performance in varying environments, such as changeable weather and obstructions; (2) safety helmets are usually small targets in wide-angle monitoring system; (3) traditional detection methods are difficult to meet the real-time requirements on the premise of high accuracy, and thus a more lightweight detection model is needed.

This study takes YOLOv5 as the benchmark algorithm to tackle with the mentioned issues. On the basis of YOLOv3 [21], YOLOv5 integrates and innovates various advanced technologies, Cross Stage Partial Network(CSPNet) and FPN+PAN structure constitute the main detection network structure. This paper proposes an enhanced YOLOv5, for helmet wearing detection, where attention modules are added at the end of the backbone and neck network, and FPN is replaced by BiFPN, which improves the capability of network feature fusion. Transfer learning is adopted during the training process to improve the detection effect. Considering that different colors of helmets represent different types of work, this study also distinguish the helmets colors.

### 1.3 Our contributions

This paper contains the contributions in the following aspects:

- Mixed attention mechanisms are added after backbone and neck network to improve the capability of feature extraction, and better aggregate features from target region.

- BiFPN is employed in the neck network to enhance the capability of feature fusion. It better fuse the fine-grained features and utilize the contribution of different levels for feature map.
- Transfer learning is adopted to strengthen the model's perception of the head in the training phase. It brings the correlation between head and helmet to improve preception accuracy.
- Enhanced YOLOv5 model can achieve mAP of 93.3% on the GDUT-HWD dataset [32], which is better than the original YOLOv5 by 4.8%.

## 2 Material and methods

### 2.1 Benchmark YOLOv5

This study adopts YOLOv5 algorithm as the benchmark method for helmet wearing detection. YOLOv5 is the latest version of the YOLO architecture series. In comparison with previous versions, it has the most powerful performance without sacrificing the computational speed. The structure of YOLOv5 network consists of four components: input layer, backbone network, neck network and detect layer.

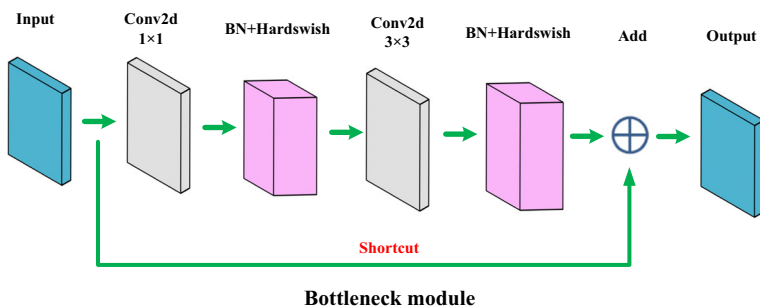
- (1) Input layer: It uses mosaic data enhancement, adaptive anchor box calculation and adaptive picture scaling technology to preprocess the image. As the core technology, mosaic data enhancement combines four images in random zoom, random crop and random arrangement, which can enrich the data set and improve the detection accuracy of small targets. An instance of Mosaic data enhancement is shown in Fig. 1.
- (2) Backbone network: It is a convolutional neural network that aggregates image features on different image granularities. Backbone has ten layers, which contains two main components, Focus structure and CSPNet. The Focus structure is used to generate more sufficient feature and reduce the calculation of the model. The input 3 channel image was segmented into four slices with the size of  $320 \times 320 \times 12$  per slice, and then through the convolutional layer composed of 32 convolution kernels, the output feature map with a size of  $32 \times 320 \times 320$  was generated.  
CSPNet network [30] is composed of BottleneckCSP(BCSP) module and convolution module. BottleneckCSP is mainly composed of a Bottleneck module, as shown in Fig. 2, which connects a convolutional layer whose convolution kernel size is  $1 \times 1$ , with a convolutional layer whose convolution kernel size is  $3 \times 3$ . The final output of the Bottleneck module is the addition of the output of this part and the initial input through the residual structure. Such a design can obtain richer gradient combination information.
- (3) Neck network: It contains a series of network layers that mix and combine image features, to pass the new features for the prediction layer. Neck network has 14 layers, which are composed of FPN and PAN structure. As shown in Fig. 3, the FPN [11] conveys strong semantic features from the top to the bottom, while the PAN [14] conveys strong positioning features from the bottom to the top, so that the output of different backbone layers are aggregated in the neck network. The feature maps of different scales are fused in pairs, so that the same target object with different sizes and scales can be accurately recognized.
- (4) Detect layer: It predicts categories based on image features and generates bounding boxes. The detection network is composed of three detect layers, whose input is a



**Fig. 1** Mosaic data enhancement. Four images are randomly cropped together and fed into the detection network

feature map with dimensions of  $80 \times 80$ ,  $40 \times 40$  and  $20 \times 20$  respectively, used to detect targets of different sizes.

With the same network components, YOLOv5 is divided into four sub versions by its size (YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x), where YOLOv5s is the smallest model with the fastest inference speed. The network structure of YOLOv5s model is shown in Fig. 4. Table 1 shows the configuration of four sub versions. The depth\_multiple and width\_multiple parameters are used to control the depth and width of different network structures. Depth\_multiple controls the number of BCSP in the network, and width\_multiple controls the number of convolution kernels in the network.



**Fig. 2** Structure of Bottleneck module

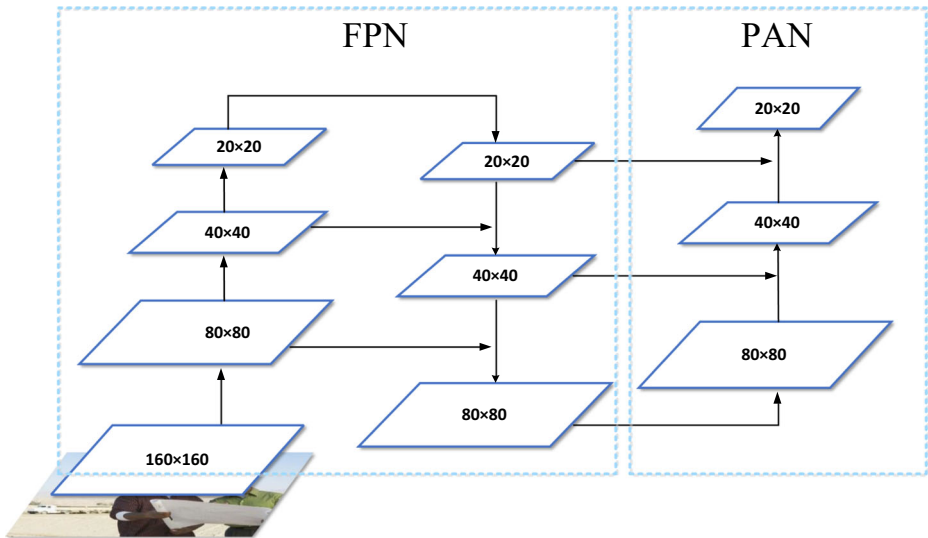


Fig. 3 Structure of FPN+PAN

## 2.2 Enhanced YOLOv5 for helmet detection

### 2.2.1 Enhanced YOLOv5 structure

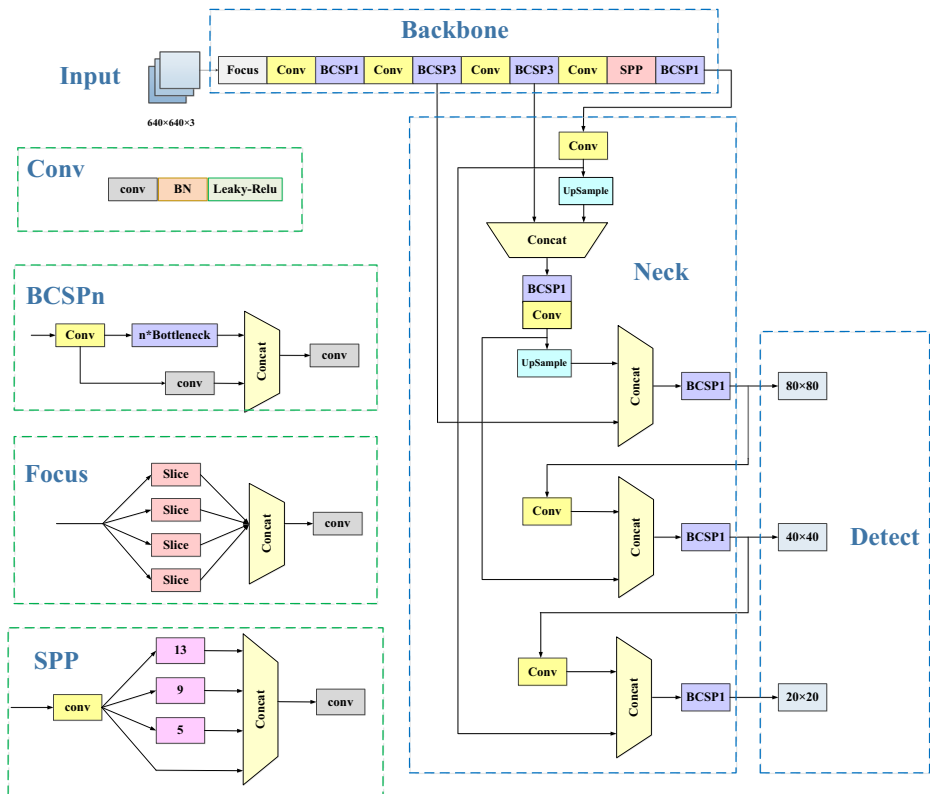
Figure 5 demonstrates the enhanced YOLOv5 structure proposed in this paper for helmet wearing detection, where BiFPN and CBAM are two new components. The enhanced YOLOv5 network has 28 network component layers, compared with the original YOLOv5 (Fig. 4), 4 layers are increased. CBAM modules are added after the 10th, 19th, 23th and 27th layers respectively. The utilised BiFPN structure includes four weighted-contact operation for feature fusion, and an additional edge that connects the third BCSP3 (7th layer) in backbone and third BCSP1 (22th layer) in neck network, which allows easy and fast multi-scale feature fusion. CBAM is integrated into the original YOLOv5 structure after several BCSPs, which can extract richer helmet related features from images.

For the detection task in this study, the detection network of enhanced has three detect layers, each detect layer outputs a 30-channel vector ((1 class probability + 4 surrounding box position coordinates + 5 helmet classes)  $\times$  3 anchor boxes), and then predicts the bounding boxes and categories of the helmet target.

### 2.2.2 BiFPN fusion structure

The fusion of feature maps for different scales is a significant way to improve the recognition performance of the detection network. YOLOv5 adopts the FPN+PAN structure, where all the feature images are changed into the same size for concat. However, different input features have different resolutions, and they usually contribute to the output feature unequally.

In order to better fuse fine-grained features and utilize the contribution of different levels for feature maps, BiFPN network is employed [27, 28]. The main ideas for BiFPN is efficient bidirectional cross-scale connections and weighted feature fusion.



**Fig. 4** Architecture of the original YOLOv5s network. The specific component of BottleneckCSP, Focus and SPP are described in the green box on the left

The enhanced YOLOv5 replaces the FPN+PAN structure by a customised BiFPN, an extra edge is added at the feature level of 40x40, and the connection of different feature layers is replaced by weighted feature fusion, as shown in Fig. 6. On one hand, at the feature level of 40x40, an additional edge is added to connect the first downsampling structure and the last downsampling structure. On the other hand, this study provides adaptive weights for different feature levels during fusion, so that the contributions of feature maps at different levels can be learned to describe the significance of each input feature. As described in the following equation,

$$O = \sum_i \bar{w}_i \cdot I_i = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot I_i, \tag{1}$$

**Table 1** Four different depths of YOLOv5 network

	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
depth_multiple	0.33	0.67	1.0	1.33
width_multiple	0.50	0.75	1.0	1.25

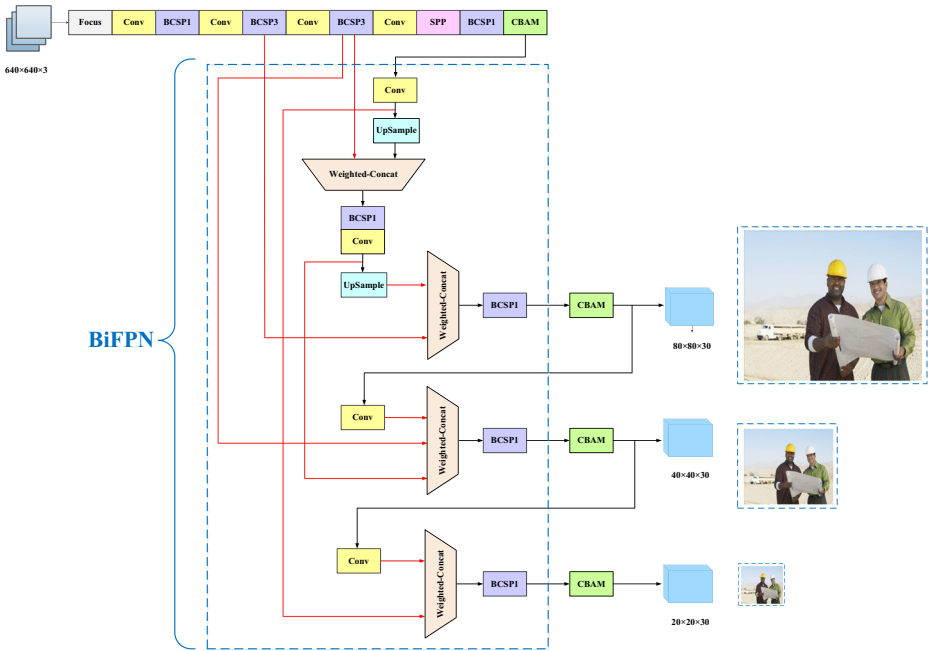


Fig. 5 The proposed enhanced YOLOv5 network structure, where CBAM are added to better aggregate features from target regions, and neck network is modified to improve the feature fusion ability

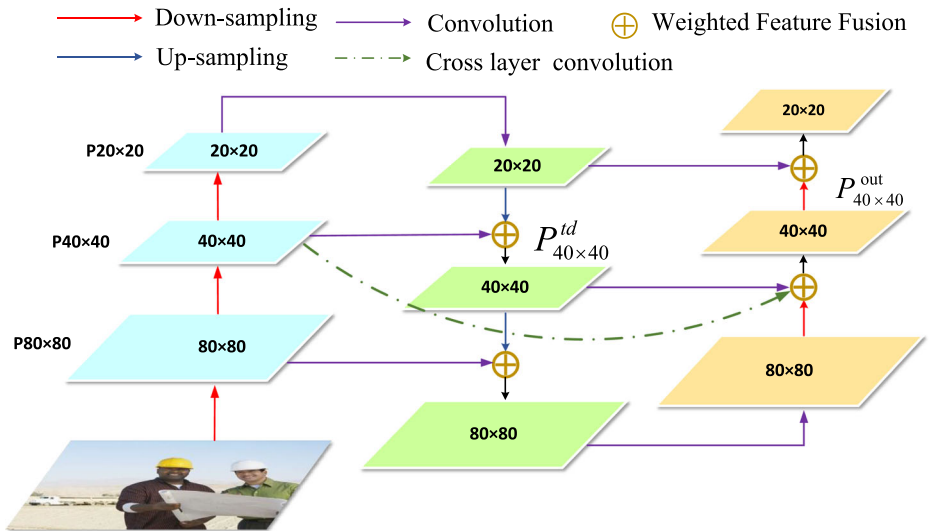


Fig. 6 The BiFPN fusion structure of the proposed enhanced YOLOv5, which is used to replace the original FPN structure



where  $w_i$  is a learnable weight for the  $i^{th}$  input feature;  $\sum_j w_j$  indicates the sum of all the weights;  $I_i$  is the feature that needs to be weighted, and  $\epsilon = 0.0001$  is a small value to avoid numerical instability. It is worth to be noticed that the final weights are normalised to make sure that the value of  $\bar{w}_i$  is limited to a range between 0 and 1, which could potentially avoid training instability.

Equations (2) and (3) describe two fused features at level  $40 \times 40$  in this study, which are consistent with Fig. 6.

$$P_{40 \times 40}^{td} = Conv \left( \frac{w_1 P_{40 \times 40}^{in} + w_2 Resize(P_{20 \times 20}^{in})}{w_1 + w_2 + \epsilon} \right), \quad (2)$$

$$P_{40 \times 40}^{out} = Conv \left( \frac{w'_1 P_{40 \times 40}^{in} + w'_2 P_{40 \times 40}^{td} + w'_3 Resize(P_{80 \times 80}^{out})}{w'_1 + w'_2 + w'_3 + \epsilon} \right), \quad (3)$$

where  $P_{40 \times 40}^{td}$  is the intermediate feature at level  $40 \times 40$  on the top-down pathway, and  $P_{40 \times 40}^{out}$  is the output feature at level  $40 \times 40$  on the bottom-up pathway.

### 2.2.3 CBAM attention module

Attention mechanism a simple but effective module for feedforward convolutional neural networks, it has great application prospects in convolutional neural networks [35]. Attention mechanism can focus on important features and suppress unnecessary ones. Common attention modules include Squeeze-and-Excitation (SE) block [7], Spatial Attention Mechanism (SAM) [37] and Convolutional Block Attention Modules (CBAM) [31]. SE block adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. SAM generates a spatial attention map by utilizing the inter-spatial relationship of features. CBAM exploits both spatial and channel-wise attention based on an efficient framework, and empirically verify that is superior to using only the channel-wise attention or spatial-wise attention. This work employs CBAM in YOLOv5. CBAM is a mixed attention mechanism combining channel-wise and spatial-wise attention, so that each branch can learn 'what' and 'where' to attend in the channel and spatial. Given an intermediate feature map, CBAM infers attention maps along two separate dimensions, channel and spatial. Then the attention maps are multiplied to the input feature map for adaptive feature refinement. The process can be summarized as:

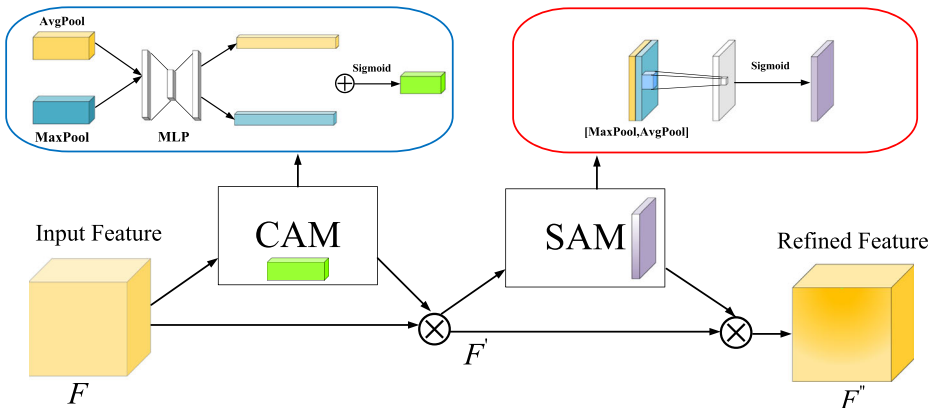
$$F' = M_c(F) \otimes F, \quad (4)$$

$$F'' = M_s(F') \otimes F', \quad (5)$$

where  $\otimes$  denotes element-wise multiplication;  $M_c(\cdot)$  extracts 1D channel attention maps;  $M_s(\cdot)$  extracts 2D spatial attention map;  $F$  is an input feature maps;  $F''$  is the final refined output.

The network structure of CBAM is shown in Fig. 7. It is divided into two parts: channel attention module (CAM) and spatial attention module (SAM). CBAM firstly performs channel weighting and then spatial weighting.

CAM aggregates spatial information of a feature map by using both average-pooling and max-pooling operations, generating two different spatial context descriptors: average-pooled features and max-pooled features. Then, they are forwarded to a shared network to produce channel attention map. The shared network is composed of multi-layer perceptron



**Fig. 7** The network structure of CBAM, where the input feature are processed by CBAM, containing CAM and SAM to obtain refined feature

(MLP). After the shared network is applied to each descriptor, CAM merges the output feature vectors using element-wise summation.

SAM aggregates channel information of a feature map by using two pooling operations to generate two 2D maps: average-pooled features and max-pooled features. Those are then concatenated and convolved by a standard convolution layer, producing 2D spatial attention maps.

Enhanced YOLOv5 adds CBAM at the end of the backbone network and neck network. Since CBAM can significantly improve the attention to the target and enhance the refining ability of detection, it can be employed before multi-scale prediction to improve detection performance.

#### 2.2.4 Transfer learning for training process

This study introduces transfer learning strategy for training. The model is firstly pre-trained by a head detection database to obtain a pre-trained model, and then transfers to the helmet detection model, via further training by a helmet database.

It is worth noting that the development of deep learning is accompanied by transfer learning [15, 16, 19]. Based on the similarity between source domain data and target domain data, source task and target task, transfer learning uses the knowledge learned in the source domain to solve the target domain task. In recent years, transfer learning has been widely used to promote object detection accuracy [10, 18]. Selecting a pretrained network model and using it as a starting point to learn a new task is the concept behind transfer learning. To be specific, transfer learning is represented by using a pretraining model, that is trained on a large benchmark dataset to solve similar problems. In this way, the training does not need to start from scratch, which can not only shorten the training time, but also improve its performance. The official YOLOv5 is also pre-trained by the COCO dataset [12] of 80 categories, which is an example of transfer learning.

Most of the helmet detection scenarios are of complicated environment and high crowd density, which becomes the main cause of missing detection. Besides, deep learning methods ignore the situations that the helmet is not properly worn on the head, but held in the hand. This is owing to the detection model does not have a favorable perception of human

heads in a complex scene. In terms of these problems, the head detection model is introduced as the pre-training model for the helmet detection, which can be regarded as a secondary transfer learning. As shown in Fig. 8, the head dataset is put into the original YOLOv5 network for pre-training. Then the helmet dataset is put into the enhanced YOLOv5 network for training and fine-tuning.

The role of transfer learning is reflected in three aspects. (1) alleviates the instability caused by the insufficient data amount of the helmet dataset; (2) improves the convergence speed; (3) brings the correlation between head and helmet to improve preception accuracy.

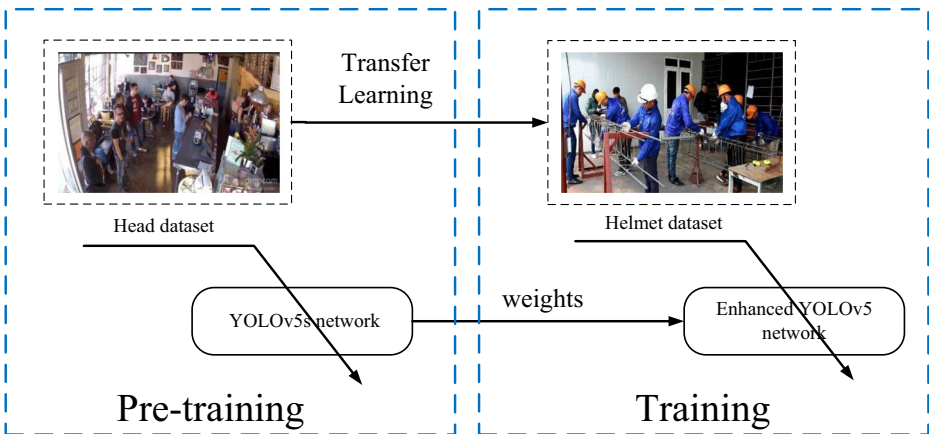
### 2.2.5 Loss function

The loss function includes: classification loss, confidence loss, and boundary loss (forecast the error between the Bounding box and the Ground truth). YOLOv5 uses the binary cross entropy loss function (BCELoss) to calculate the classification loss and confidence loss. In addition, YOLOv5 adopts CIOU Loss as the loss of bounding box regression. The loss function of YOLOv5 can be defined as:

$$Loss_{(object)} = Loss_{cls} + Loss_{obj} + CIOU\_loss, \tag{6}$$

YOLOv5 divides an image into  $S \times S$  cells and gets  $B \times S \times S$  prediction boxes. A mask matrix is created before the training phase to determine whether object appears in each prediction box, which is a  $B \times S \times S$  boolean matrix. If object appears in a prediction box, the corresponding position in the mask matrix is set to true, otherwise false. Based on this, classification loss and confidence loss are defined as:

$$Loss_{cls} = \frac{1}{N_t} \sum_{i=0}^{S \times S} \sum_{j=0}^B \mathbf{1}_{ij}^{obj} \sum_{t=0}^M loss_{BCE}(i, j, t), \tag{7}$$



**Fig. 8** Flow chart of transfer learning. The head detection model is transferred to the training of the helmet model

and

$$\begin{aligned}
 Loss_{obj} = & \frac{\alpha}{N_t} \sum_{i=0}^{S \times S} \sum_{j=0}^B \mathbf{1}_{ij}^{obj} loss_{BCE}(i, j) \\
 & + \frac{1 - \alpha}{N_f} \sum_{i=0}^{S \times S} \sum_{j=0}^B \mathbf{1}_{ij}^{noobj} loss_{BCE}(i, j),
 \end{aligned} \tag{8}$$

where  $M$  is the number of categories;  $N_t$  is the number of true elements in the mask matrix;  $\mathbf{1}_{ij}^{obj}$  denotes if object appears in cell  $i$  and the  $j$ th bounding box predictor is responsible for that prediction.  $N_f$  is the number of false elements in the mask matrix;  $\alpha$  is the weight coefficient when the mark matrix element is true.  $loss_{BCE}$  is BCE loss function.

In object detection tasks, bounding boxes are usually used to localize objects. Yu et al. [8] firstly introduced an intersection over union (IOU) loss function to evaluate the prediction results, which is defined as:

$$IOU\_loss = 1 - IOU = 1 - \frac{|A \cap B|}{|A \cup B|}, \tag{9}$$

where  $A$  is prediction box;  $B$  is ground truth. However, if the two boxes do not intersect, IOU becomes 0, and no gradient can be obtains. Therefore, after continuous optimization, GIOU [23], DIOU and CIOU [36] are proposed. CIOU takes the distance between the target frame and the prediction frame, the overlap rate, scale and penalty terms into account, making the target frame regression more stable.

The expressions of CIOU loss is as follows:

$$CIOU\_loss = 1 - CIOU = 1 - IOU + \frac{d^2}{c^2} + \alpha v, \tag{10}$$

where  $\alpha$  is a positive trade-off parameter, and  $v$  measures the consistency of aspect ratio; as shown in Fig. 9,  $c$  is the diagonal length of the smallest enclosing box covering two boxes; and  $d$  is the Euclidean distance of central points of two boxes.

$$\alpha = \frac{v}{(1 - IOU) + v}, \tag{11}$$

where

$$v = \frac{4}{\pi^2} \left( \arctan \frac{\omega^{gt}}{h^{gt}} - \arctan \frac{\omega}{h} \right)^2, \tag{12}$$

**Fig. 9** The normalized distance between the prediction frame and target frame. The upper left box represents the target frame, and the lower right block represents the prediction frame



where  $\omega^{st}$  and  $h^{st}$  represent the width and height of the target frame, and  $\omega$  and  $h$  represent the width and height of the prediction frame, respectively.

This study adopts Stochastic Gradient Descent (SGD) with momentum to optimize the training parameters. Momentum serves to accelerate the gradient in the right direction. In addition, during the training process, each batch of training is updated by means of Cosine annealing. Better convergence can be achieved by reducing the learning rate through cosine function.

## 2.3 Experimental setup

### 2.3.1 Platform and parameters

This study constructs an algorithm evaluation environment under Windows10( Microsoft, United States). Experiments are carried out on the Pytorch deep learning framework and programmed by Python3.8 . The training process uses GTX 1080Ti GPU processors with 11G memory.

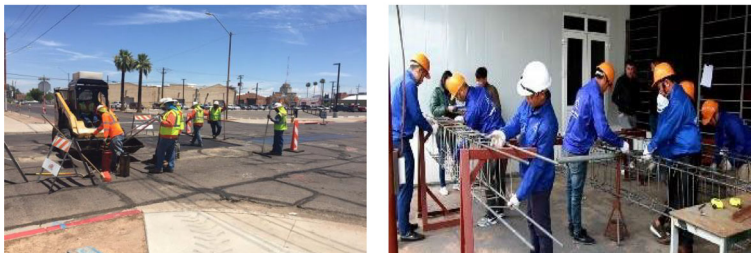
The parameter setting of the proposed network are as follows: the epoch (100), the batch-size (4), the momentum (0.95) and initial learning rate (0.001).

### 2.3.2 Datasets

Two datasets are selected to verify the effectiveness and feasibility of the algorithm. As shown in Fig. 10, BrainWash [26] is a dense dataset of human heads taken in convenience stores. In this study, 10,000 images from the dense head dataset BrainWash are used for pre-training, and 3174 images from helmet dataset GDUT-HWD with color label are used for



(a) BrainWash dataset



(b) GDUT-HWD dataset

**Fig. 10** Some image examples in BrainWah and GDUT-HWD datasets

the second stage of training. GDUT-HWD uses 2539 images for training and 635 images for testing. As shown in Table 2, five detection categories are marked, namely blue, yellow, white, red and none, which are used to classify the color of the helmet and test the multi-target detection effect of the trained model. The data set is divided in a 8:2 ratio, to form training set and testing set. Note that the label “none” indicates the person without wearing a helmet. By adding negative samples “none”, the trained model can have stronger robustness.

In addition, GDUT-HWD divides the helmets of different sizes into three categories: small (area < 322 pixels), middle (322 < area < 962 pixels) and large (area > 962 pixels). As shown in Table 3, the number of small instances is largest in the dataset, which increases the difficulty of helmet wearing detection.

### 2.3.3 Evaluation index

In object detection tasks, it not only needs to detect object in an image, but also to find out the position of the object. It is necessary to take both precision rate (P) and recall rate (R) into consideration. Therefore, a standard index average precision(AP) for judging the quality of the network model is introduced. Conceptually, AP is the area under the precision-recall curve. The expressions of precision and recall are as follows:

$$P = Precision = \frac{TP}{TP + FP}, \quad (13)$$

and

$$R = Recall = \frac{TP}{TP + FN}, \quad (14)$$

where  $TP$  is the number of helmets detected correctly;  $FP$  is the number of helmets misjudged;  $FN$  is the number of missed detection. After obtaining the values of Precision and Recall, AP is defined as:

$$AP = \int_0^1 P(R) dR, \quad (15)$$

It should be noticed that for multi-target detection, mean average precision (mAP) is the average value of AP for all categories, which is defined as:

$$mAP = \frac{1}{K} \sum_{i=1}^k AP_i, \quad (16)$$

where  $k$  represents the number of categories. In this paper, mAP is accessed as the index to test the detection accuracy of enhanced YOLOv5.

To evaluate the complexity of the proposed enhanced YOLOv5, the following evaluation indicators are used: the number of parameters (Params), Giga Floating-point Operations Per second (GFLOPs), Frames Per Second (FPS) and the final size of model (Weights). These indicators are used to analyze the number of parameters and reasoning speed of the model.

**Table 2** The category statistics of the GDUT-HWD data set

Label	Blue	Yellow	White	Red	None
Trainval	2060	3114	2980	3485	3407
Test	552	784	801	746	964
Total	2612	3898	3781	4231	4371

**Table 3** The number of instances in different sizes

Label	Small	Middle	Large
Trainval	4273	4098	984
Test	4713	3826	1035
Total	8950	7924	2019

### 2.3.4 Comparison and ablation study

To verify the validity of the enhanced YOLOv5, the detection accuracy was compared with Faster-RCNN and SSD algorithms. Faster R-CNN depends on region proposal algorithms to hypothesize object locations. It introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network. The RPN is trained end-to-end to generate high-quality region proposals. SSD is a kind of one-stage object detection algorithm. It is based on a feedforward convolutional network, which generates a fixed-size bounding box set and the corresponding score of the target category in the box, and then generates the final detection result according to the non-maximization suppression step. In addition, this study also compares the enhanced YOLOv5 with YOLOX. YOLOX is the latest achievement of the YOLO series [5]. It switches the YOLO detector to an anchor-free manner, and adopts a decoupled head and the leading label assignment strategy. YOLOX has demonstrated excellent performance on COCO data sets.

In order to clearly understand the contribution of each modification to accuracy, ablation experiments are carried out in this section. The different network structures under the ablation experiments are shown in Fig. 11.

Transfer learning+YOLOv5 indicates that the network structure is not changed and transfer learning is introduced in the training stage. YOLOv5+BiFPN is a changed YOLOv5 network, where the original neck model is replaced by BiFPN. YOLOv5+CBAM only adds CBAM to the backbone and neck of the original YOLOv5. YOLOv5+BiFPN+CBAM is the improved network that adds both CBAM and BiFPN. Transfer Learning+YOLOv5+BiFPN+CBAM is the enhanced YOLOv5. For the complexity problem, this study also conducts ablation experiments to verify the impact of every change on the algorithm complexity.

### 2.3.5 Visualization

YOLOv5 algorithm uses the entire image as the input of the CNN network, and directly returns to the position of the bounding box and the category. It is often hard to see which features can be learned in a particular part of the network. Therefore it is essential to visualize the feature maps. In a convolutional neural network, applying the filter to the resulting feature maps can provide insight into the internal representation of the input at a given point in the model. With the help of torch vision library of Pytorch framework, this study transforms the network layer to be visualized from tensor format to Python Image Library(PIL) image format. The image PIL object needs to be converted to a NumPy array of pixel data and expanded from a 3D array to a 4D array. Then the pixel values need to be scaled appropriately for the YOLOv5 model. Finally, the feature map of a certain layer can be obtained. This study performs visual operations on each layer of the network during training and detection.

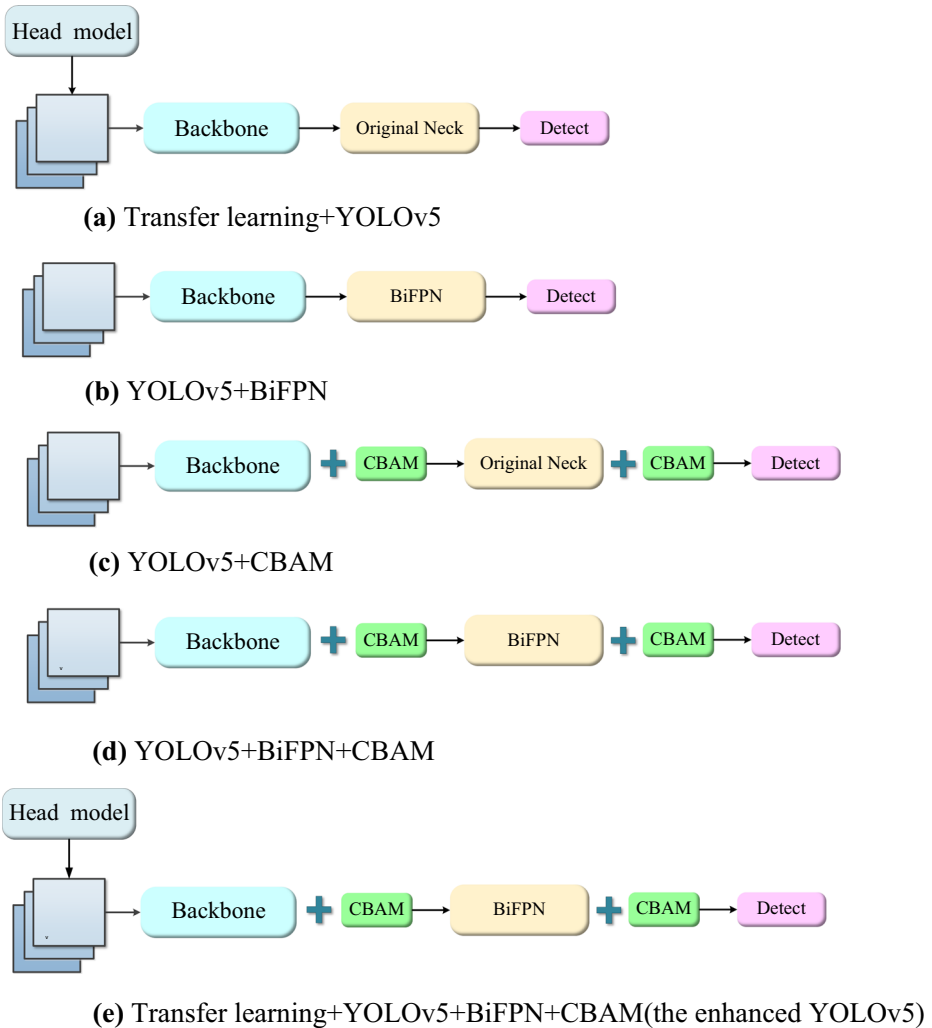


Fig. 11 Ablation experiment of the enhanced YOLOv5

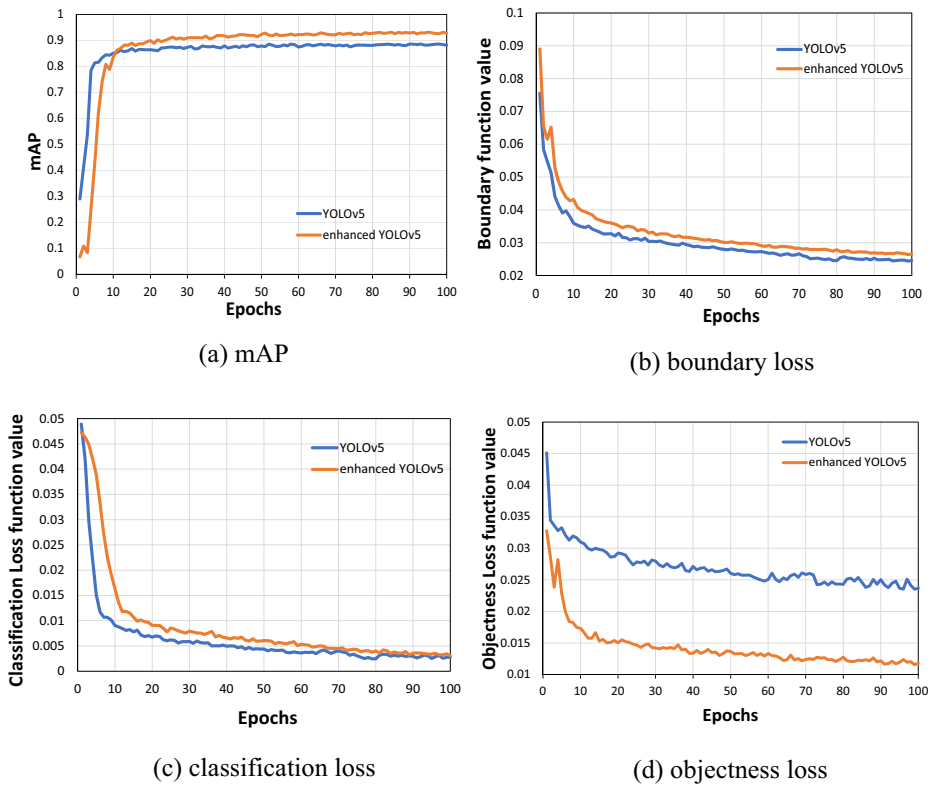
### 3 Experimental results

#### 3.1 Model convergence analysis

mAP and the loss of training are two important indicators to measure the quality of the object detection model. A well-performing model is supposed to have a high mAP value and low training losses. Figure 12 compares mAP and loss functions between YOLOv5 and enhanced YOLOv5 during training.

It can be seen that mAP of both can be stabilized to a high level after 20 epochs of training, while the decreasing speed of the three different loss functions are also gradually stable after about 20 epochs. It proves that YOLOv5 and its improved network model have a very





**Fig. 12** The changes of mAP, boundary loss, classification loss and objectness loss in training process, where YOLOv5 and the enhanced YOLOv5 are compared

fast convergence rate during training, and show the good performance on the experimental dataset.

It can be found that the enhanced YOLOv5 outperforms the original YOLOv5 in terms of accuracy. For the loss of training, both boundary loss and classification loss have a similar downward trend, and eventually stabilize at a very close value. In terms of the objectness loss, the enhanced YOLOv5 is significantly lower than the original YOLOv5, indicating that our modification improves the accuracy of prediction, resulting in a higher confidence score for model reasoning. In addition, the enhanced YOLOv5 does not increase the training time compared with the original YOLOv5. The training time for both is about 2.5 hours in the same training environment.

### 3.2 Accuracy

A total of 635 images are randomly selected from the safety helmet dataset GDUT-HWD as the verification set. Our modifications are tested for average accuracy respectively, and compare them with original YOLOv5.

As shown in Table 4, the first four lines are the comparison results of YOLOv5 and other three algorithms. The last five lines are the results of the ablation experiment of

**Table 4** Comparison of different algorithms by the indicator of AP

Method	Blue	White	Yellow	Red	None	mAP
Faster R-CNN	0.708	0.680	0.696	0.609	0.590	0.656
SSD	0.861	0.855	0.881	0.806	0.760	0.833
YOLOX-S	0.920	0.892	0.924	0.913	0.879	0.906
YOLOv5	<b>0.947</b>	0.945	0.868	0.850	0.817	0.885
Transfer learning+YOLOv5	0.942	0.894	0.918	0.873	0.856	0.896
YOLOv5+BiFPN	0.912	0.921	0.910	0.892	0.868	0.901
YOLOv5+CBAM	0.945	0.900	0.930	0.918	0.905	0.920
YOLOv5+BiFPN+CBAM	0.926	0.946	0.933	0.912	0.912	0.926
Enhanced YOLOv5	0.936	<b>0.949</b>	<b>0.939</b>	<b>0.921</b>	<b>0.920</b>	<b>0.933</b>

enhanced YOLOv5. The bold entries represent the highest precision of each category. Original YOLOv5 achieved the highest detection accuracy in the blue category, and enhanced YOLOv5 achieved the highest detection accuracies in the remaining categories. The mAP value of enhanced YOLOv5 is 93.3%. It is much higher than Faster R-CNN and SSD, which reflects its advantages. The mAP value of YOLOX is 2.1% higher than YOLOv5, but it is still 2.7% lower than the enhanced YOLOv5.

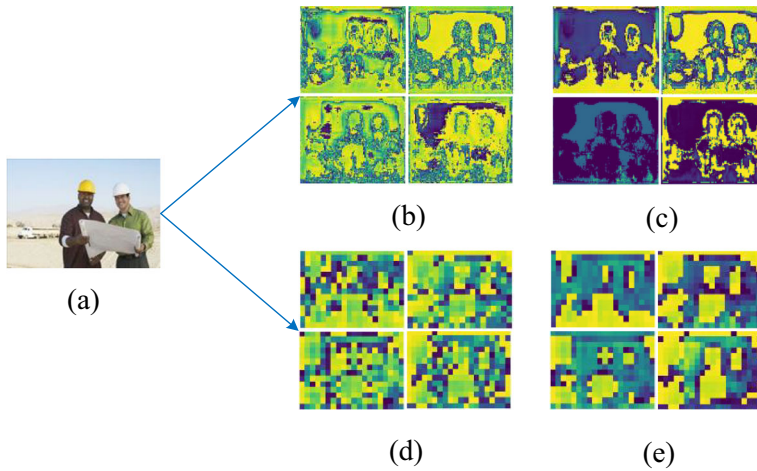
Ablation studies show that among all the improvements, CBAM brings the greatest improvement, increasing the mAP value by 3.5%. The training method of adopting transfer learning improves the accuracy by 1.1%, and the feature fusion based on BiFPN structure also brings 1.6% improvement. The enhanced YOLOv5 integrates the above improvements together. It brings a maximum increase of 4.8%. This proves that the modifications have a marked effect on the detection of safety helmet.

Experiments show that adding CBAM can significantly improve the accuracy. To further explain the contribution of CBAM, this work visualizes the feature maps of  $80 \times 80$  and  $20 \times 20$  scales, and outputs the visualization results before and after CBAM processing.

This study selects  $4 \times 4$  feature maps to observe and compare. As shown in Fig. 13, the feature information in the shallow layer are basically complete, while the features extracted from deeper layer are fuzzy. No matter the feature maps of  $80 \times 80$  or  $20 \times 20$  scale, there are significant changes after CBAM processing. The outline of the target characters and helmet area become more legible, and the distinction between foreground and background is more pronounced. It proves that the network that added CBAM is more interested in the helmets in an image. CBAM learns well to exploit information in target object regions and aggregate features from them.

### 3.3 Complexity

Table 5 shows the complexity of the original YOLOv5 and its variations, where params and GFLOPs represent the number of parameters and calculating speed of the algorithm; and FPS represents the detection speed of the algorithm; Weight is the final size of the model. The original YOLOv5 has the lowest complexity. The calculation cost brought by CBAM processing is negligible, while the BiFPN structure increases a certain amount of calculation cost and reasoning time. It is worth noting that the training method of transfer learning does not have any impact on the algorithm complexity. Overall, compared with the



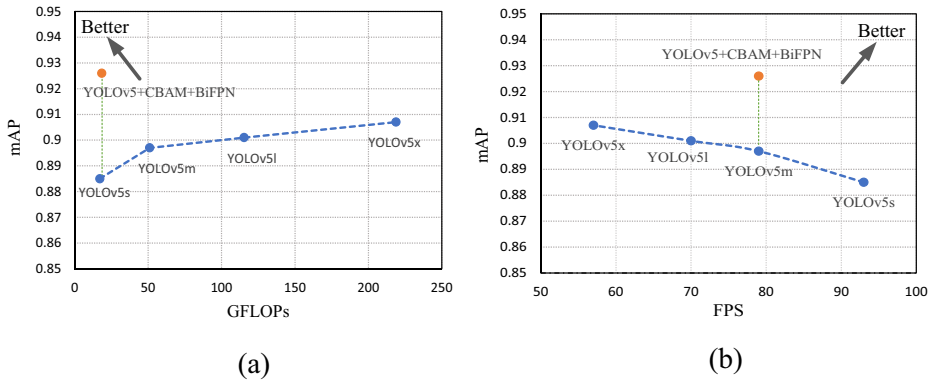
**Fig. 13** Feature map visualization. (a) shows the original image. (b) show the feature maps of  $80 \times 80$  scale before CBAM. (c) show the feature maps of  $80 \times 80$  scale after CBAM. (d) show the feature maps of  $20 \times 20$  scale before CBAM. (e) show the feature maps of  $20 \times 20$  scale after CBAM

improved accuracy, a small increase in algorithmic complexity for enhanced YOLOv5 is within acceptable bounds. In addition, the FPS of 79 indicates that the enhanced YOLOv5 fully meets the real-time requirements of the helmet detection scene. Its reasoning speed is several times faster than SSD and Faster-RCNN algorithms, and 30% faster than YOLOX-s.

As shown in Fig. 14, this study selects GFLOPs and FPS to measure the relationship between complexity and accuracy. From YOLOv5s model to YOLOv5x model, heavier networks with larger width and depth tend to obtain higher accuracy, but it leads to higher GFLOPs and lower FPS. In the comparison between YOLOv5s and YOLOv5x, the accuracy is increased only by about 2%, but the GFLOPs (about 220) of YOLOv5x is 11 times of that (about 20) of YOLOv5s, and the processing speed is also significantly reduced from 57 FPS from 93 FPS. In contrast, the proposed enhanced YOLOv5 structure (i.e. YOLOv5+CBAM+BiFPN) can obtain the accuracy at 92.6%, but the GFLOPs is almost the same as YOLOv5s, as seen in Fig. 14(a). Besides, it is also found that the computing speed of the enhance YOLOv5 is almost the same as YOLOv5m, but its accuracy is more than 2% higher.

**Table 5** Complexity comparison among YOLOv5 and three variations

Method	Params/M	GFLOPs	FPS	Weight/MB
YOLOv5	7.3	16.9	93	14.4
YOLOv5+CBAM	7.5	17.1	87	14.7
YOLOv5+BiFPN	8.0	18.3	81	16.8
Enhanced YOLOv5	8.1	18.4	79	17.1

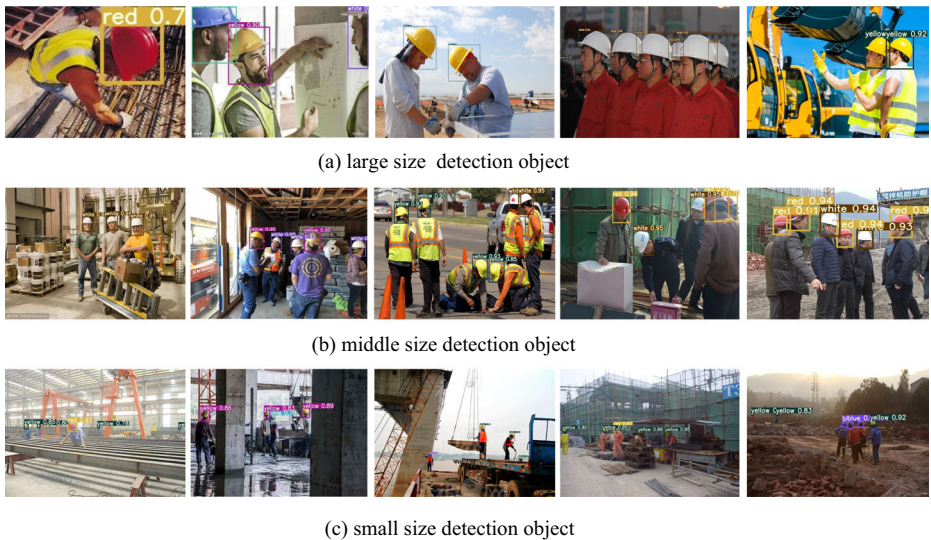


**Fig. 14** A comparison on the complexity among YOLOv5 series and the enhanced YOLOv5, where transfer learning is not applied during the training stage to obtain the accuracy. (a) and (b) show GFLOPs index and FPS index, respectively

### 3.4 Detection demonstration

#### 3.4.1 Different target scale

Figure 15 shows a few detection examples on GDUT-HWD dataset with the enhanced YOLOv5 model. The examples include helmet wearing detection for small, medium and large instances. These results show that the proposed model can well complete the helmet detection task and be applied to many construction sites.



**Fig. 15** Detection examples on GDUT-HWD test with enhanced YOLOv5

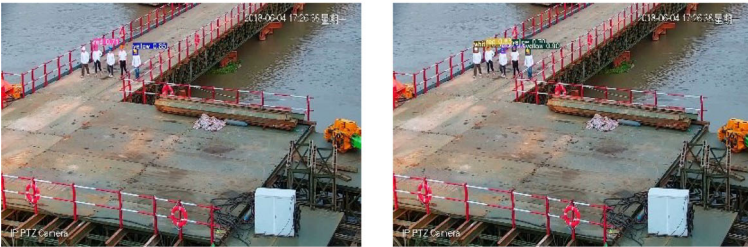
### 3.4.2 Comparison with YOLOv5 on different detection scenarios

In Fig. 16, three typical detection results are compared between YOLOv5 and enhanced YOLOv5. (a) It represents a situation where the background is dark and unclear. YOLOv5 failed to identify the person with the red helmet on the left. (b) It represents a small target detection task. The helmet occupies a tiny proportion of the whole image. YOLOv5 fails to detect a number of small targets. (c) It represents the case where the helmet is not worn on the head. YOLOv5 mistakenly marked a helmet placed on the table. The above problems of missed detection and wrong detection are effectively solved by the enhanced YOLOv5.

In sum, it can be found that enhanced YOLOv5 model has stronger inhibition to the interference of occlusions, better performance in complex environment with small targets.



(a) dark environment



(b) small target detection



(c) unworn safety helmet

**Fig. 16** The experimental comparison results of YOLOv5 and enhanced YOLOv5, where the left side are the detection results of original YOLOv5, and the right side are the detection results of enhanced YOLOv5

In addition, enhanced YOLOv5 has better head perception and a lower probability of error detection.

## 4 Discussions

This paper explores a safety helmet wearing detection method in real time. In order to meet both requirements on the accuracy and computing speed in industrial occasions, an enhanced YOLOv5 is proposed via BiFPN, CBAM and transfer learning. BiFPN strengthens the capability of network feature fusion. CBAM improves the ability of feature extraction. Transfer learning contributes to a better detection accuracy around head region. In general, the feasibility of this study is as follows:

- **Accuracy:** the detection accuracy of enhanced YOLOv5 on GDUT-HWD dataset is significantly higher than that of YOLOv5, which is due to its outstanding detection performance in challenging scenarios, such as dark environment, occlusion, and small targets etc. The achieved 93.3% mAP value demonstrates that the enhanced YOLOv5 can basically meet the detection requirement, which is much more superior than other traditional approaches.
- **Speed:** the proposed model can reach 79 FPS, which is fast enough to meet the real-time requirement. It is partly benefited from the advantage of the original YOLOv5 structure, but more importantly, it proves that our modification on the YOLOv5 does not result in heavier computing burden. Therefore, it can be directly used in industrial scenarios.
- **Applicability:** this study selects GDUT-HWD dataset to carry out the experiments. The dataset includes the information of helmet color, and a large number of negative samples. Therefore, it is believed that the model trained by such a dataset is robust. Besides, the adaptation of transfer learning allows the network to focus more on the head region, and neglect non-relevant region, which can avoid misrecognition.

In the experiments, some issues are identified and are worthy to be discussed. The first one is about whether to increase the detection layer in BiFPN. The original BiFPN structure contains five layers, whereas YOLOv5 has only three layers in FPN. Therefore, this study attempts to add additional layers and edges into YOLOv5. However, such changes significantly increase its complexity and reduce the reasoning speed of the model. Therefore, this study optimises the BiFPN structure by reducing the number of layers and edges. Although the accuracy is slightly reduced, it does not bring in much change on its complexity. The second one regards transfer learning. Experimental results show that the introduction of transfer learning brings 1% accuracy improvement, but it is far more than what we expected. The possible reasons are as follows: 1) the most suitable scenario of transfer learning is to transfer the source domain with a large number of samples to the target domain with a small number of samples, but in this study, the quantity of images in target domain may be enough to sufficiently train the model. 2) heads and safety helmets are visually similar. Although the introduction of the head detection model can improve the detection performance of the safety helmet, it may also lead to misrecognition.

In the future, it is recommended to further improve the detection speed of the model, so that it can be used in embedded systems for practical evaluation. Besides, transfer learning can be further studied to see how it influences the accuracy, and corresponding optimising strategies can be proposed.

## 5 Conclusion and further works

Detection of helmet wearing is of great importance to worker safety at construction sites. In this paper, an enhanced YOLOv5 model is proposed to improve the detection accuracy without sacrificing its computing speed. It integrates attention mechanism and BiFPN structure into the original YOLOv5 network, and takes transfer learning method for training. The experimental results show that the enhanced YOLOv5 model achieves a significant improvement in the accuracy. To make maximum use of its advantages in terms of small size and high accuracy, network downgauging will be focused in future research. It is to reduce the size of the model and improve its inference and operation speed without sacrificing its high accuracy. Eventually, the model could be deployed on mobile and hardware platforms, driving the intelligent and humanized development of smart construction sites.

**Data Availability** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

### Declarations

**Competing interests** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

1. Colantonio A, Mcvittie D, Lewko J, Yin J (2009) Traumatic brain injuries in the construction industry. *Brain injury* : BI 23(11):873–8. <https://doi.org/10.1080/02699050903036033>
2. Dakhli Z, Danel T, Lafhaj Z (2019) Smart construction site: ontology of information system architecture. *Modular Offsite Construct (MOC) Summit Proceed*:41–50. <https://doi.org/10.29173/mocs75>
3. Deng L, Li H, Liu H, Gu J (2022) A lightweight yolov3 algorithm used for safety helmet detection. *Sci Reports* 12(1):1–15. <https://doi.org/10.1038/s41598-022-15272-w>
4. Dewi C, Chen R, Jiang X, Yu H (2022) Deep convolutional neural network for enhancing traffic sign recognition developed on yolo v4. *Multimed Tools Appl*:1–25. <https://doi.org/10.1007/s11042-022-12962-5>
5. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) Yolox: exceeding yolo series in 2021. *arXiv*:2107.08430
6. Han G, Zhu M, Zhao X, Gao H (2021) Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection. *Comput Electr Eng* 95:107458. <https://doi.org/10.1016/j.compeleceng.2021.107458>
7. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7132–7141. <https://doi.org/10.1109/cvpr.2018.00745>
8. Jiang B, Luo R, Mao J, Xiao T, Jiang Y (2018) Acquisition of localization confidence for accurate object detection. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 784–799. [https://doi.org/10.1007/978-3-030-01264-9\\_48](https://doi.org/10.1007/978-3-030-01264-9_48)
9. Li J, Liu H, Wang T, Jiang M, Wang S, Li K, Zhao X (2017) Safety helmet wearing detection based on image processing and machine learning. In: *2017 Ninth international conference on advanced computational intelligence (ICACI)*. IEEE, pp 201–205. <https://doi.org/10.1109/icaci.2017.7974509>
10. Li G, Song Z, Fu Q (2018) A new method of image detection for small datasets under the framework of yolo network. In: *2018 IEEE 3rd advanced information technology, electronic and automation control conference (IAEAC)*. IEEE, pp 1031–1035. <https://doi.org/10.1109/iaeac.2018.8577214>
11. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2117–2125. <https://doi.org/10.1109/cvpr.2017.106>
12. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: *European conference on computer vision*. Springer, pp 740–755. [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48)

13. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37. [https://doi.org/10.1007/978-3-319-46448-0\\_2](https://doi.org/10.1007/978-3-319-46448-0_2)
14. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768. <https://doi.org/10.1109/cvpr.2018.00913>
15. Lu J, Behbood V, Hao P, Zuo H, Xue S, Zhang G (2015) Transfer learning using computational intelligence: a survey. *Knowl-Based Syst* 80:14–23. <https://doi.org/10.1016/j.knosys.2015.01.010>
16. Man CK, Quddus M, Theofilatos A (2022) Transfer learning for spatio-temporal transferability of real-time crash prediction models. *Accid Anal Prev* 165:106511. <https://doi.org/10.1016/j.aap.2021.106511>
17. Mneymneh BE, Abbas M, Khoury H (2019) Vision-based framework for intelligent monitoring of hardhat wearing on construction sites. *J Comput Civil Eng* 33(2):04018066. <https://doi.org/10.1109/icpr48806.2021.9412103>
18. Nie M, Wang K (2018) Pavement distress detection based on transfer learning. In: 2018 5th International conference on systems and informatics (ICSAI). IEEE, pp 435–439. <https://doi.org/10.1109/icsai.2018.8599473>
19. Pan SJ, Yang Q (2009) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
20. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788. <https://doi.org/10.1109/cvpr.2016.91>
21. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767
22. Ren S, He K, Girshick R, Sun J (2016) Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 39(6):1137–1149. <https://doi.org/10.1109/tpami.2016.2577031>
23. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 658–666. <https://doi.org/10.1109/cvpr.2019.00075>
24. Siebert FW, Lin H (2020) Detecting motorcycle helmet use with deep learning. *Accid Anal Prev* 134:105319. <https://doi.org/10.1016/j.aap.2019.105319>
25. Song R, Wang Z (2022) Rbfpdet: an anchor-free helmet wearing detection method. *Appl Intell*:1–16. <https://doi.org/10.1007/s10489-022-03664-4>
26. Stewart R, Andriluka M, Ng AY (2016) End-to-end people detection in crowded scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2325–2333. <https://doi.org/10.1109/cvpr.2016.255>
27. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, pp 6105–6114. arXiv:1905.11946
28. Tan M, Pang R, Le QV (2020) Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790. <https://doi.org/10.1109/cvpr42600.2020.01079>
29. (2020). Ultralytics.yolov5 online. <https://github.com/ultralytics/yolov5>
30. Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H (2020) Cspnet: a new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 390–391. <https://doi.org/10.1109/cvprw50498.2020.00203>
31. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19. [https://doi.org/10.1007/978-3-030-01234-2\\_1](https://doi.org/10.1007/978-3-030-01234-2_1)
32. Wu J, Cai N, Chen W, Wang H, Wang G (2019) Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. *Autom Constr* 106:102894. <https://doi.org/10.1016/j.autcon.2019.102894>
33. Yue S, Zhang Q, Shao D, Fan Y, Bai J (2022) Safety helmet wearing status detection based on improved boosted random ferns. *Multimed Tools Appl* 81(12):16783–16796. <https://doi.org/10.1007/s11042-022-12014-y>
34. Zhao J, Li C, Xu Z, Jiao L, Zhao Z, Wang Z (2022) Detection of passenger flow on and off buses based on video images and yolo algorithm. *Multimed Tools Appl* 81(4):4669–4692. <https://doi.org/10.1007/s11042-021-10747-w>
35. Zheng Y, Bao H, Meng C, Ma N (2021) A method of traffic police detection based on attention mechanism in natural scene. *Neurocomputing* 458:592–601. <https://doi.org/10.1016/j.neucom.2019.12.144>



36. Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, Zuo W (2021) Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans Cybern*:1–13
37. Zhu X, Cheng D, Zhang Z, Lin S, Dai J (2019) An empirical study of spatial attention mechanisms in deep networks. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 6688–6697. <https://doi.org/10.1109/iccv.2019.00679>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.