Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.043671

ARTICLE

Detection of Safety Helmet-Wearing Based on the YOLO_CA Model

Xiaoqin Wu, Songrong Qian* and Ming Yang

State Key Laboratory of Public Big Data, Guizhou University, Guiyang, 550025, China

*Corresponding Author: Songrong Qian. Email: qiansongr@163.com

Received: 09 July 2023 Accepted: 10 October 2023 Published: 26 December 2023

ABSTRACT

Safety helmets can reduce head injuries from object impacts and lower the probability of safety accidents, as well as being of great significance to construction safety. However, for a variety of reasons, construction workers nowadays may not strictly enforce the rules of wearing safety helmets. In order to strengthen the safety of construction site, the traditional practice is to manage it through methods such as regular inspections by safety officers, but the cost is high and the effect is poor. With the popularization and application of construction site video monitoring, manual video monitoring has been realized for management, but the monitors need to be on duty at all times, and thus are prone to negligence. Therefore, this study establishes a lightweight model YOLO_CA based on YOLOv5 for the automatic detection of construction workers' helmet wearing, which overcomes the shortcomings of the current manual monitoring methods that are inefficient and expensive. The coordinate attention (CA) addition to the YOLOv5 backbone strengthens detection accuracy in complex scenes by extracting critical information and suppressing non-critical information. Further parameter compression with deeply separable convolution (DWConv). In addition, to improve the feature representation speed, we swap out C3 with a Ghost module, which decreases the floating-point operations needed for feature channel fusion, and CIOU_Loss was substituted with EIOU_Loss to enhance the algorithm's localization accuracy. Therefore, the original model needs to be improved so as to enhance the detection of safety helmets. The experimental results show that the YOLO_CA model achieves good results in all indicators compared with the mainstream model. Compared with the original model, the mAP value of the optimized model increased by 1.13%, GFLOPs cut down by 17.5%, and there is a 6.84% decrease in the total model parameters, furthermore, the weight size cuts down by 4.26%, FPS increased by 39.58%, and the detection effect and model size of this model can meet the requirements of lightweight embedding.

KEYWORDS

Safety helmet; CA; YOLOv5; ghost module

1 Introduction

Since construction sites [1] and other construction and building sites are generally exposed to the outdoors, therefore, the risk variables are going to be much greater than in other industry sectors, resulting in a higher accident rate. As a result, wearing personal protection equipment [2] can keep workers from safety accidents and decrease injuries and even deaths [3]. However, in reality, there are many examples of helmets not being worn or not being worn right, examples like sultry weather and

lack of safety awareness among employees. Moreover, it is sometimes difficult for safety managers on construction sites to keep track of whether or not employees are wearing helmets, resulting in several incidents involving manufacturing safety. In China, based on data from the Ministry of Housing and Urban-Rural Development, there were 689 production security mishaps in housing and municipal construction in 2020, and 794 workers died during production activities. Among them, there were 407 accidents of falling from a height, accounting for 59.07% of the total; and 83 accidents involving object strikes, accounting for 12.05% of the total [4]. Similarly, 123 people suffered fatal injuries on the job in the United Kingdom in 2021/2022, as reported by the Health and Safety Executive (HSE), with falls from height being the most serious fatal accident accounting for 23.6% [5] and being struck by an object also accounted for 14.6%. Helmets protect workers by absorbing the impact of objects hitting their heads directly, and studies have shown that helmets are an effective way for construction workers to lessen the risk of skull fractures, neck sprains, and concussions when falling from heights [6]. At the same time, helmets can also minimize the risk of serious brain injury from impacts [7].

Helmet-wearing supervision is an important part of creating a safe environment for construction operations. Usually, construction units use manual supervision, but due to the excessive range of workers' activities, they cannot be managed promptly in real scenarios. Therefore, helmet-wearing detection based on intelligent technology is gradually becoming a vital management measure for companies. Safety helmet-wearing detection may be broken down into two groups: those that rely on sensors and those that rely on vision-based detection methods. Remote location and tracking technologies are the primary focus of emphasis for sensor-based detection approaches. Cyber-Physical Systems (CPS) were suggested for real-time monitoring and detection by Zhang et al. [8]. The Internet of Things (IoT) architecture was used by Zhang et al. to design a smart helmet system [9], which identifies whether the helmet is being used based on whether or not both the infrared light detector and the thermal infrared sensor in the helmet are activated. Nevertheless, detector tracking technology can be limited by the need to wear physical tags or sensors, and requires a large investment, by adding a significant number of extra devices (like physical tags and sensors), resulting in low scalability. Besides, with the present radio-frequency identification (RFID) solution, connecting to the network requires workers to wear a terminal device, which is inconvenient for workers' work [10].

Contrasted with sensor-based detection techniques, visual technology detection methods are gaining attention [11]. By collecting rich pictures of building sites to get a faster and more comprehensive grasp of complex scenes of construction sites [12]. Fang et al. suggested an improved target detection method for the case of workers not wearing helmets, but the method is inefficient and does not match the real-time requirement [13]. In contrast, Wu et al. improved the performance of Single Shot MultiBox Detector (SSD) for helmet-wearing detection by using a reverse progressive attention mechanism [14]. K-nearest neighbor (KNN) was used by Wu et al. in order to detect moving objects from video to classify pedestrians, heads, and helmets [15]. Xie et al. examined the performance of several detection techniques using that same dataset, and You Only Look Once (YOLO) had the best average accuracy and the speediest detection compared to SSD and Faster region based convolutional neural network (Faster R-CNN) [16]. Wang et al. enhanced the representation ability of target features by introducing convolutional block attention modules in the neck to assign weights and weaken feature extraction from complex backgrounds [17]. Wen et al. used the soft-NMS algorithm to optimize the YOLOv3 model, and the improved YOLOv3 algorithm was able to effectively detect occluded targets, but the target detection was not satisfactory when the occlusion rate exceeded 60% [18]. Wang et al. proposed an improved helmet wear detection algorithm, YOLOv4-P, which improves the accuracy of helmet wear detection by increasing the mAP value by 2.15% compared to the YOLOv4 algorithm [19]. Proposed an improved lightweight YOLOv5 vehicle detection method that improves

the model's performance by inserting C3Ghost and Ghost modules in the YOLOv5 neck network, adding a convolutional block attention module (CBAM) attention mechanism in the backbone, etc. Headcounting is gradually becoming an emerging research hotspot in the field of video surveillance, Khan et al. generated scale-aware head suggestions based on scale graphs, thus proposing a method for counting the number of people in a sports video by detecting their heads, which solves the problem of different scales, and which is clearly superior to the state-of-the-art (SoA) method [20]. In addition, an end-to-end scale-invariant head detection framework is proposed, which can handle a wide range of scales and is important for high-density crowd counting and crowd safety management [21].

The following is the major organizing framework for the remainder of this article: Section 2 describes the relevant materials and methods. In Section 3, conduct experiments and analyses. Introduce a detailed discussion in Section 4. Section 5 is the final section, which is the conclusion of this article.

2 Materials and Methods

2.1 Environment and Data for Experimentation

The hardware and software settings of the experimental platform are as follows: the operating system is Linux, the GPU is NVIDIA Tesla A10 GPU 24 GB, and the deep learning framework is Python.

For some existing datasets mostly collected from advertisements some of these images have backgrounds other than construction sites [14]. Considering the robustness that the model should have in practical applications, this research creates a new dataset that can detect construction workers wearing helmets in terms of building background, angle, and category. A total of 10,700 pictures were gathered as the dataset for model training through open-source dataset restructuring and were randomly divided into an 1:1:8 a set for validation, a set for testing, and a set for training. Using the graphical image annotation application LabelImg [22] identify the photographs as being two classes, then save them in the YOLO format. To increase reliability in the experimental data, both the training and validation sets were equally scaled to a 640 × 640 size before being used to train the models. Using a technique for improving data Mosaic cuts and stitches together any four photos at random, increasing data variety; gradients are updated using the asynchronous stochastic gradient descent (SGD) approach. Following training, the model performance is assessed using the test dataset.

2.2 Performance Evaluation Metrics

To validate the YOLO_CA model proposed in this study, Evaluation criteria like accuracy, recall, and mAP are applied in order to assess the performance of the network model. mAP is the most extensively used evaluation statistic in target identification algorithms, as the area under the precision-recall curve (P-R), is the average accuracy achieved at varying recall percentages [23]. The greater the Mean Average Precision (mAP) number, the more effective the present target detection model is for this dataset [24]. It is calculated by setting all categories' precision to IOU = 0.5 [25]. Eqs. (1) to (4) demonstrate how to calculate accuracy, recall, and mAP:

$$Precison = \frac{TP}{TP + FP} \tag{1}$$

The proportion of all targets that can be successfully forecasted is referred to as the recall. The following is a definition of it:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where TP is the quantity of properly categorized positive instances, FP indicates the number of negative examples that are incorrectly classified, FN is the number of instances of positives that were misclassified, and TN is the number of properly categorized negative instances.

$$AP = \int_0^1 P(R)\, dR \tag{3}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{4}$$

The amount of time spent processing detections on average includes both the amount of time spent processing by the NMS and the amount of time spent processing by the network. The size of the model is determined by the model's size that was stored after the completion of all training.

2.3 Building a Deep Learning Network

2.3.1 YOLO_CA Network Structure

YOLOv5 is broken down into four distinct network architectures (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) based on the required depth and breadth of the network [26]. The four model structures each see a progressive rise in their overall size as well as the number of parameters they need [27]. Due to the fact that the helmet-wearing algorithm for detecting in building construction scenarios has to meet real-time requirements, in this research, we use the smallest version of YOLOv5s with the smallest size and fewest parameters as the basic network for detection, Fig. 1 presents the proposed organizational framework for the YOLO_CA model. To compress the network parameters, minimize the network computation, and enhance the model inference performance, introducing the Ghost module into the backbone of the model to reduce the number of model parameters and GFLOPs [28]. In the bottleneck module, deep separable convolution (DWConv) is used in place of the original model's CONV to achieve parameter compression. Finally, the backbone module adds the Coordinate Attention (CA) technique to improve the characteristics [29], thereby enhancing helmet detection's precision in such complicated circumstances.

2.3.2 Coordinate Attention Mechanism

The idea of an attention mechanism resembles the nerve system of the human body. Distinguishing between useful and useless data, and focusing on the essentials of the objective to be found as a result [30]. To enable more accurate localization and detection of target regions in complex environments, a CA structure [29] was introduced, which has the following benefits. First of all, it may collect information that is position- and orientation-aware in addition to cross-channel information, enabling the model to more accurately recognize and identify the object of interest. Second, CA is light and flexible, and easy to insert into classical modules. Eventually, an already-trained model, the CA technique could prove very helpful for downstream activities on lightweight networks, particularly those requiring extensive prediction, including techniques like semantic segmentation. There are a pair of steps to capturing channel relations and long-range dependencies with precise location

information using coordinate attention: coordinated attention creation and information embedding. Fig. 2 illustrates the specific principle.

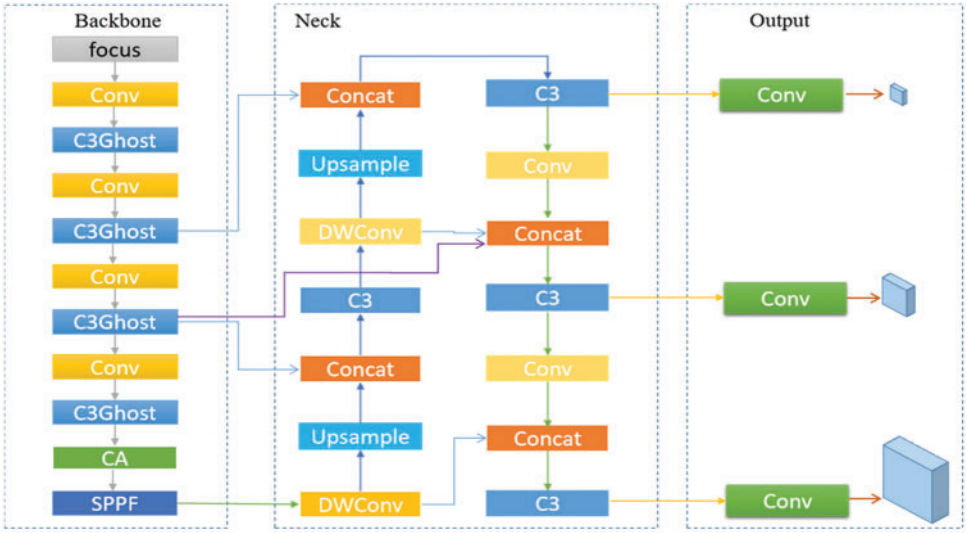


Figure 1: Structure of the proposed YOLO_CA model

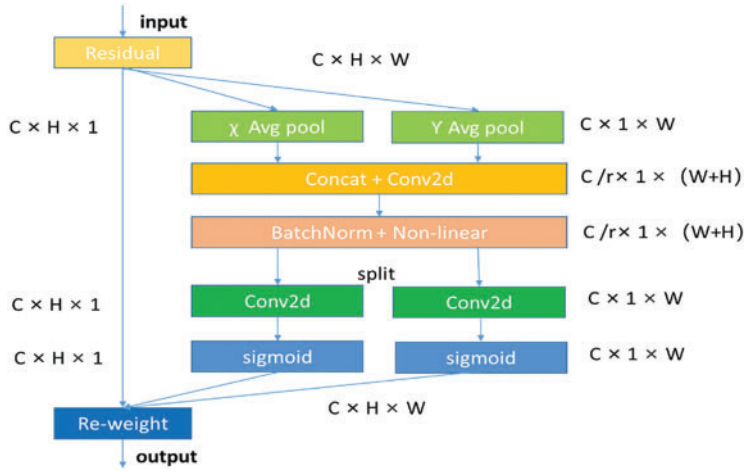


Figure 2: CA module structure

Step 1: Information embedding that is coordinated. For the attention module to accurately capture spatial long-range dependencies, channel attention often uses global pooling to globally encode spatial information into channel descriptors. Two one-dimensional feature encoding procedures are then constructed from the initial global pooling:

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

For input X, the information of different channels is encoded along the horizontal and vertical directions using pooling kernels of size (H, 1) and (1, W). Consequently, the c channel's output at

height h can be as:

$$z_c^h(h) = \frac{1}{w}\sum_{0\le i<w} x_c(h,i) \tag{6}$$

Similarly, the output of channel c with width w is shown as follows:

$$z_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j,w) \tag{7}$$

A pair of direction-aware feature maps are generated by applying these two transformations to feature aggregations along two spatial directions [31]. Furthermore, because capturing the long-range dependencies along one spatial orientation and retaining the precise location details along the other spatial orientation, strengthens the network's capacity to pinpoint the location of a desired target [32].

Step 2: Coordination of attention generation. The feature maps of the global receptive field's width and height are stitched together and passed to the 1 × 1 convolution module with dimensionality reduced to the original C/r. A feature map f of form 1 × (W + H) × C/r is obtained by feeding the batch-normalized feature map F1 into the sigmoid activation function, as seen by Eq. (8).

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{8}$$

δ represents the nonlinear activation function, where $f \in R^{(C/r)\times(W+H)}$ is the intermediate horizontal and vertical feature map of spatial information, and the downsampling rate is r.

Using the original height and width, multiply the feature map f by 1 × 1 to get the feature maps F_h and F_w, having exactly as many channels as the input x. Making use of the sigmoid activation function, determine the attention weights g^h along the feature maps' height axis and g^w along their width axis.

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{9}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{10}$$

Eventually, as shown in Eq. (11), the initial weighted feature map is obtained by multiplication for the purpose of obtaining a feature map with width and height attention weights.

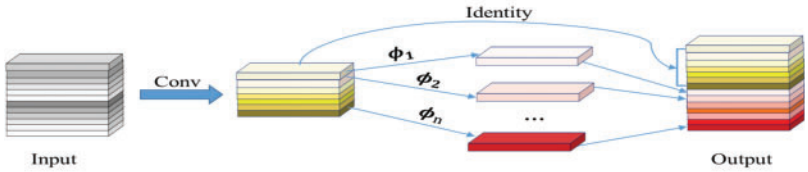
$$y_c(i,j) = x_c(i,j) \times g_C^h(i) \times g_C^w(j) \tag{11}$$

CA mechanism is a new mechanism that embeds location information into channel attention. It has been proved that embedding it into the backbone network is a lightweight structure, and in subsequent experiments, it improves the performance of helmet detection in complex backgrounds.

2.3.3 Ghost Module

The YOLOv5 model uses the C3 structure for backbone feature extraction, but the structure's vast array of parameters, as well as sluggish detection speed, can lead to limited applications, making it difficult to apply in some practical scenarios, such as mobile or embedded devices [33]. Therefore, this paper uses Han et al. [34] to propose a new Ghost module for creating effective neural networks to replace the original C3 structure to achieve a lightweight network model that balances speed and accuracy. The fundamental Ghost module divides the initial convolutional layer into two sections and generates many intrinsic feature maps with fewer filters. After that, in order to efficiently construct the reimage feature maps, there will be a certain amount of transformations performed. The underlying concept is shown in Fig. 3.

(a) The convolutional layer



(b) The Ghost module

Figure 3: Conventional convolution and Ghost module

Assume that the input feature map size is $h \times w \times c$, and $h' \times w' \times c'$ is the output feature map size. The input feature map's height and width are h and w, respectively, whereas the output feature map's height and width are h′ and w′. There is one constant mapping and $m \times (s-1) = n \times s \times (s-1)$ linear operations, each of which has a convolution kernel size of $d \times d$ and a regular convolution kernel size of $k \times k$. Linear operations of the same size are used in a Ghost module, and after s transformations, d and k are of similar size when $s \ll c$. The theoretical ratio between the Ghost module and the standard convolution is:

$$\begin{aligned} r_s &= \frac{c' * h' * \omega' * c * k * k}{\frac{c'}{s} * h' * \omega' * c * k * k + (s-1) * \frac{c'}{s} * h' * \omega' * d * d} \\ &= \frac{c * k * k}{\frac{1}{s} * c * k * k + \frac{s-1}{s} * d * d} \\ &\approx \frac{s * c}{s + c - 1} \approx s \end{aligned} \tag{12}$$

In the same way, the number of parameters is:

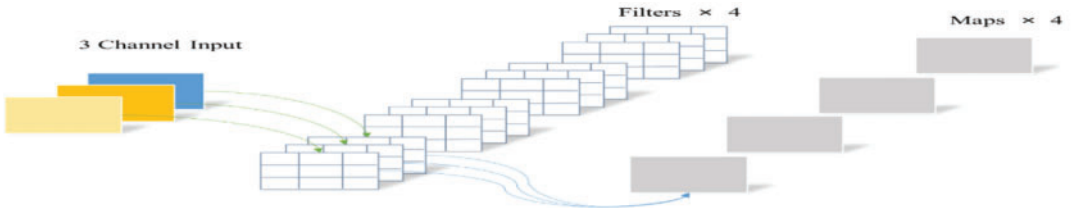
$$\begin{aligned} r_c &= \frac{c' * c * k * k}{\frac{c'}{s} * c * k * k + (s-1) * \frac{c'}{s} * d * d} \\ &\approx \frac{s * c}{s + c - 1} \approx s \end{aligned} \tag{13}$$

From the above, it is obvious that Ghost modules are superior in terms of computational cost. The Ghost Bottleneck can be formed by stacking two Ghost modules, replacement of the bottleneck module in the C3 module with the Ghost Bottleneck generates a new C3Ghost, this may reduce computing costs and reduce model size.

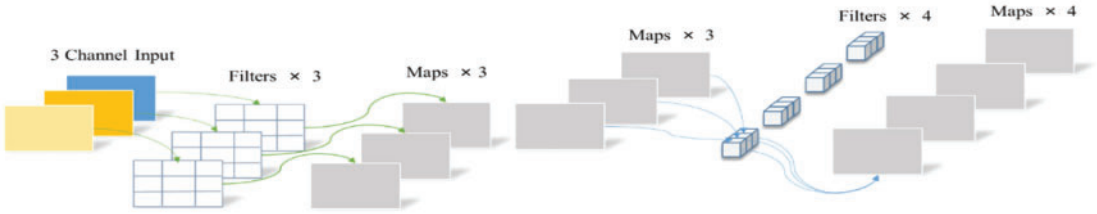
2.3.4 Deeply Separable Convolution Module

To further reduce the parameters in a network and to create a model that is lightweight, the YOLOv5s' original Neck layer has been updated. The Conv in the original PANet module is replaced by DWConv [35]. Unlike traditional convolutional operations, Deep Convolutional Xception [36] and MobileNet [37] the core idea of DWConv Xception and MobileNet is to split a convolution process into two separate sections: Depthwise Convolution (DW) layer and Pointwise Convolution (PW) layer, which is a network [38]. DW is a convolution of separate channels, i.e., each convolution

kernel corresponds to each channel of input, thus the features of each layer are separated, and the effective information of different layers at the same spatial location cannot be effectively utilized, then a second part (PW) is needed to produce a fresh feature map by combining the separate feature maps of the first part (DW), as shown in Fig. 4b. To prevent gradients from disappearing and the establishment of complex parameters, the BN algorithm adjusts the distribution of the data [39].



(a) Standard convolution



(I) Depthwise convolution (II) Pointwise convolution

(b) Deeply separable convolution

Figure 4: Standard convolution and depth-separable convolution

2.3.5 EIOU Loss Function

1. Limits of CIOU

Three important geometrical elements are taken into account by Complete Intersection over Union (CIOU) Loss: overlap area, centroid distance, and aspect ratio [40]. Following is the definition of the CIOU loss given the prediction frame B and the target frame B^{gt}:

$$L_{CIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{14}$$

where the centroids of B and B^{gt} are indicated by b and b^{gt}, and Intersection over Union (IOU) stands for the overlap ratio between the real frame and the detected frame. $IOU = \frac{|b \cap b^{gt}|}{|b \cup b^{gt}|}$, $\rho(\cdot) = \|b - b^{gt}\|_2$ stands for the Euclidean distance between them, and the tiniest closed box that covers the two boxes has a diagonal length of c, and $v = \frac{4}{\pi^2}(arc\tan\frac{w^{gt}}{h^{gt}} - arc\tan\frac{w}{h})^2$ and $\alpha = \frac{v}{(1 - IOU) + v}$ measures the difference in aspect ratio. Although CIOU loss's convergence speed and detection precision have considerably increased, the confidence levels for width and height are not taken into account in the equation for v; just the aspect ratio difference is reflected, so it sometimes slows down the convergence speed of CIOU, decreases the real differences between width and height and the confidence levels, this makes it more difficult for the model to successfully optimize the similarity.

2. Suggested Approach

To deal with the above problem, the more efficient Efficient Intersection over Union(EIOU) Loss is used instead of the network model's CIOU Loss [41]. EIOU is based on CIOU Using the aspect ratio's influencing factor to determine the target box's and anchor box's individual lengths and widths, the width-height true difference, the center distance, and the overlap area are the components of this loss function, The approach used in EIOU in the first two portions is continued in CIOU. However, the loss of width-to-height resolves the ambiguous aspect ratio definition based on CIOU, it immediately reduces the height and breadth differences between the target box and anchor box, hastening convergence. It is defined as follows:

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp}$$

$$= 1 - IOU + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \tag{15}$$

where w^c and h^c are the smallest closed box's width and height that encloses these two boxes. That is, three sections make up the loss function: the IOU loss L_{IOU}, the distance loss L_{dis} and the aspect ratio loss L_{asp}. In the above case, we may keep the CIOU loss's profitability features. Meanwhile, the disparity between the target and anchor frames' width and height is immediately minimized by the EIOU loss. Having a quicker convergence rate as well as better localization results.

3 Experiment and Analysis

3.1 The Outcomes of Model Training

During training, the YOLO_CA network model uses the Adam optimizer with a batch size of 16, a learning rate momentum of 0.93, a weight decay of 0.0005, with a 0.01 starting learning rate. Each model has the same parameters for a total of 100 iterations, and the best weights obtained are used as the weights for helmet-wearing detection. According to Fig. 5, the model's training precision gradually increases with the number of iterations, and the loss value gradually decreases with an increase in the number of iterations. The model learning efficiency is higher in the initial model training phase, and the training loss curve converges faster. In the first 20 periods of training, mAP, recall, and precision exhibit a tendency for rapid growth as the network rapidly converges. Around 50 cycles later, in terms of precision, recall, and mAP, the model achieves a stable and then stabilizes at around 80. Finally, mAP, Precision, and Recall stabilized at 96.73%, 95.87%, and 95.31%, respectively. Precision-Recall value distributions with training are shown in Fig. 5. During the testing period, the thresholds for non-maximal suppression (NMS), confidence (C), and the intersection of sets (IoU) were set at 0.45, 0.25, and 0.5, respectively.

To test the recognition performance of the model, first, input the images of the divided test set and validation set into the trained network for detection. The validation results of the YOLO_CA model are shown in Table 1, the recognition accuracy in the validation set is 93.6%, the recall rate is 90.1%, the mAP is 94.8%, and the recognition speed is 134 FPS. Fig. 6 displays the recognition outcomes of a few test sets.

In order to validate the generality of the model proposed in this study, 600 and 402 images from the public data SHWD [42] and CHV [43] were used for validation, respectively, and the validation results are shown in Table 1. As shown in Table 1, the mAP on the two publicly available datasets is 93.0% and 93.6%, the accuracy is 96.0% and 95.7%, and the FPS on GPU is 117 and 119, respectively.

Fig. 7 displays some of the test results from publicly available datasets. In summary, it can be seen that this model has good robustness and generality.

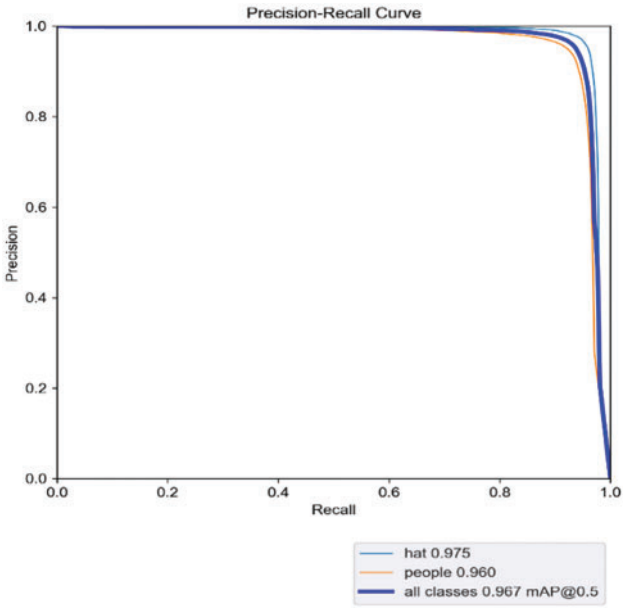


Figure 5: Precision recall rate

Table 1: YOLO_CA model validation

Data set	mAP (%)	Accuracy (%)	Recall (%)	Number of pictures	Inference time (ms)	FPS
Ours	94.8	93.6	90.1	1070	6.5	134
CHV	93.0	96.0	90.3	600	7.7	113
SWHD	93.6	95.7	89.1	402	7.1	119

Figure 6: Partial test set identification results

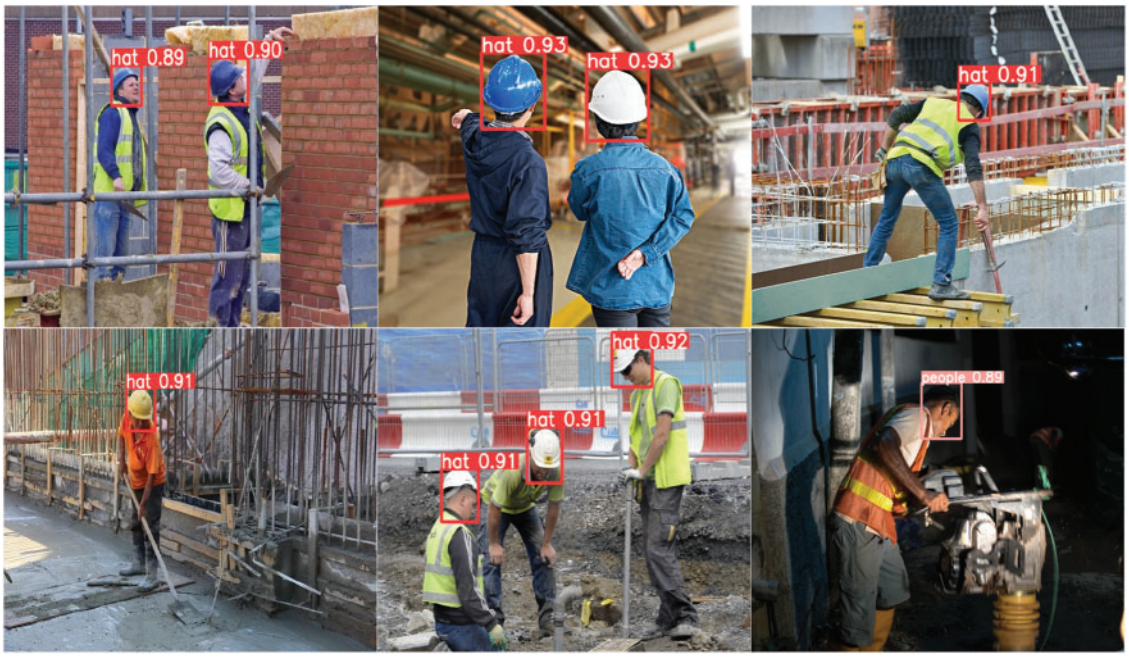
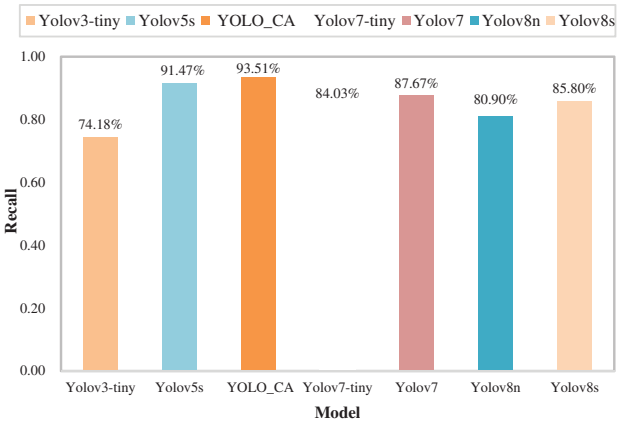


Figure 7: Partial public dataset test results

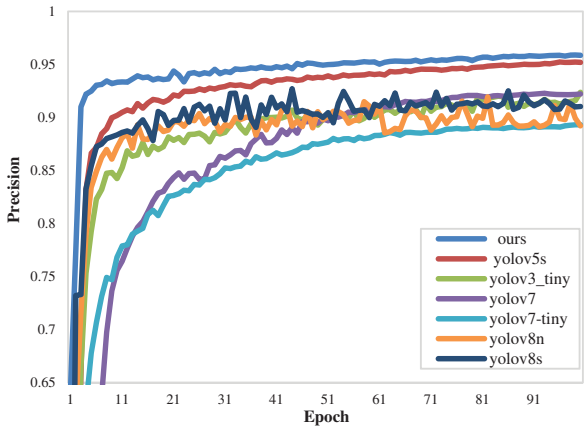
3.2 Analysis of Model Improvement Performance

3.2.1 Comparative Analysis of Similar Methods

To further verify the performance of the put-forward method, it was contrasted with network models like the YOLOv3_tiny, Yolov5s, Yolov7, and the newly proposed Yolov8. Fig. 8 shows the comparison of Recall, Precision, and detection speed on GPU, and mAP for each model, the weight parameters and other evaluation indicators are shown in Table 2, and Fig. 9 shows the detection rate FPS of each model.

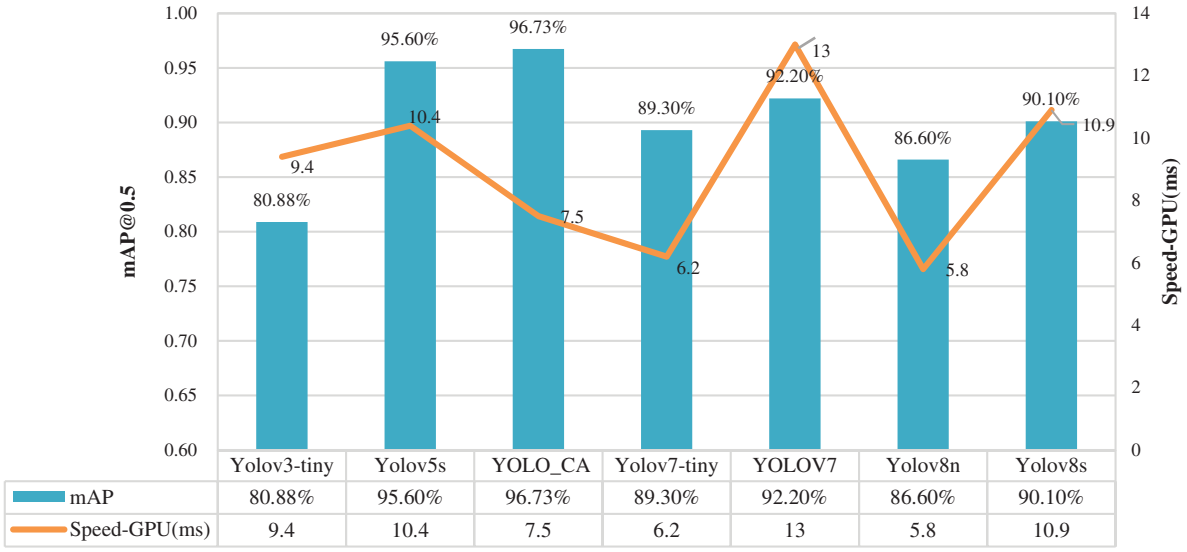


(a) Recall



(b) Precision

Figure 8: (Continued)



(c) Speed-GPU and mAP

Figure 8: Comparing the evaluation indexes of different models

From Table 2, it can be seen that compared to other models in the YOLO series, YOLO_ CA has the best mAP value, accuracy, and recall. In terms of model weights and parameters, both Yolov7-tiny and Yolov8n have smaller weights and parameters than the YOLO_CA model, with Yolov7-tiny's weights and parameters reduced by 9.7% and 8.7%, respectively, while it is of interest that Yolov8n's weights and parameters are both reduced by about 54%, however, this model's FPS and mAP values are also reduced 59.5% and 7.63% from YOLO_CA, respectively. In terms of the detection rate of the model, as shown in Table 2 and Fig. 9, the FPS of Yolov7-tiny and Yolov8n are 20% and 28% faster than YOLO_CA, respectively, but the Recall of YOLO_CA is 11.31% and 14.41% higher than Yolov7-tiny and Yolov8n, respectively. As a result of the above study, it is obvious that the proposed method in this paper offers benefits over existing typical network models.

Table 2: Comparison of evaluation indexes of different models

Model	mAP (%)	Precision (%)	Recall (%)	Weight (M)	Parameter size	Speed-GPU per image (ms)	FPS
YOLOv3_tiny	80.88	92.38	74.18	16.6	8672186	9.4	106
Yolov5s	95.60	95.20	91.47	14.1	7025023	10.4	96
Yolov7-tiny	89.30	91.20	84.00	12.3	6017694	6.2	161
Yolov7	92.20	92.60	87.70	71.3	37201950	13	77
Yolov8n	86.60	89.90	80.90	6.2	3011238	5.8	172
Yolov8s	90.10	91.10	85.80	22.5	11136374	10.9	92
Yolov5_CA	96.73	95.87	95.31	13.5	6544487	7.5	134

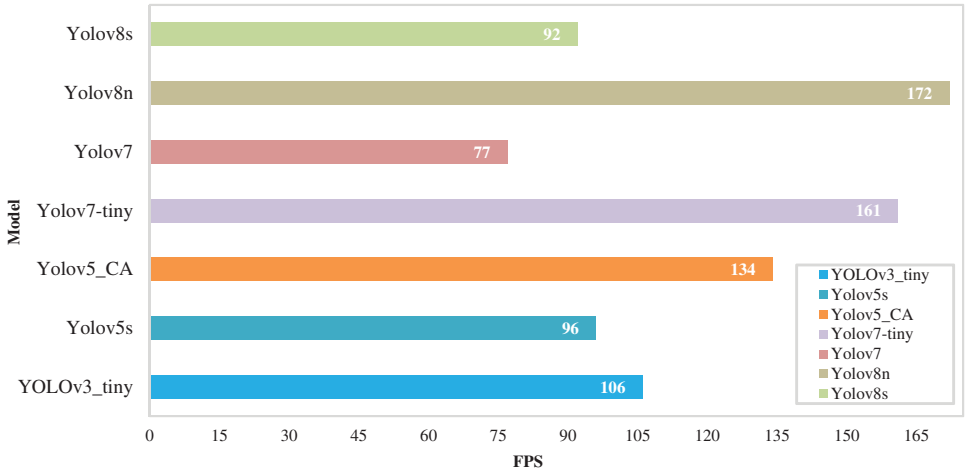


Figure 9: PFS for each model

3.2.2 Ablation Experiments

Ablation tests were carried out to validate the performance of various components so as to assess the validity and feasibility of the proposed model. Because the proposed model in this study depends on Yolov5s, the Yolov5s model is used as the benchmark for the ablation experiments to conduct a comparative analysis of 2 different parts with CA attention mechanisms at different locations, and different lightweight models. They are the following:

1. Comparative analysis of CA attention mechanisms in different locations

To further evaluate the impact of the location of CA mechanism addition on algorithm performance, attention mechanisms (CA) were added to the backbone and neck of the model, respectively. The experimental results are shown in Table 3. Fig. 10's a–c and a + c are the CA addition locations 1–4 in Table 3, respectively. According to the results of the experiment, the CA addition position used in this study has the most effective detection effect, moreover, its accuracy has been substantially enhanced.

Table 3: Comparison of the results of CA attention mechanism in different locations

No.	Location of CA	Precision (%)	Recall (%)	mAP (%)
1	Backbone (a)	96.36	93.98	97.10
2	Backbone (b)	95.50	92.88	96.38
3	Neck (c)	95.64	93.11	96.50
4	Backbone (a) + Neck (c)	95.73	93.30	96.65

2. Comparative analysis of different lightweight models

In order to make the model more convenient to be applied in actual production practice, the model is subjected to lightweight processing. To achieve optimal experimental results, Yolov5s was used as the experimental benchmark, and the other parameters of the model were kept consistent.

The experimental results are shown in Table 4, and the final model has good recognition accuracy and can balance the model weight size and running speed. From Table 4, it can be concluded that the 4th model has the best experimental results, having the fastest detection speed and the best mAP values. Where the FPS was 39.58%, 10.74%, and 14.53% faster than the other three models, respectively, and the mAP value for this model was 96.73%, which is 1.13%, 0.62%, and 0.41% higher than the other 3. Compared with the original model, the number of parameters is decreasing by 20.54%, 3.86%, and 6.84%, while the GFLOPs are lowered by 4.7, 1.2, and 2.8, as compared to the model, weight sizes were reduced by 2.5, 0.9, and 0.6, respectively, compared to yolov5s. Therefore, the model lightweight method used in this study is effective. In summary, choose the fourth model in the table as the final model.

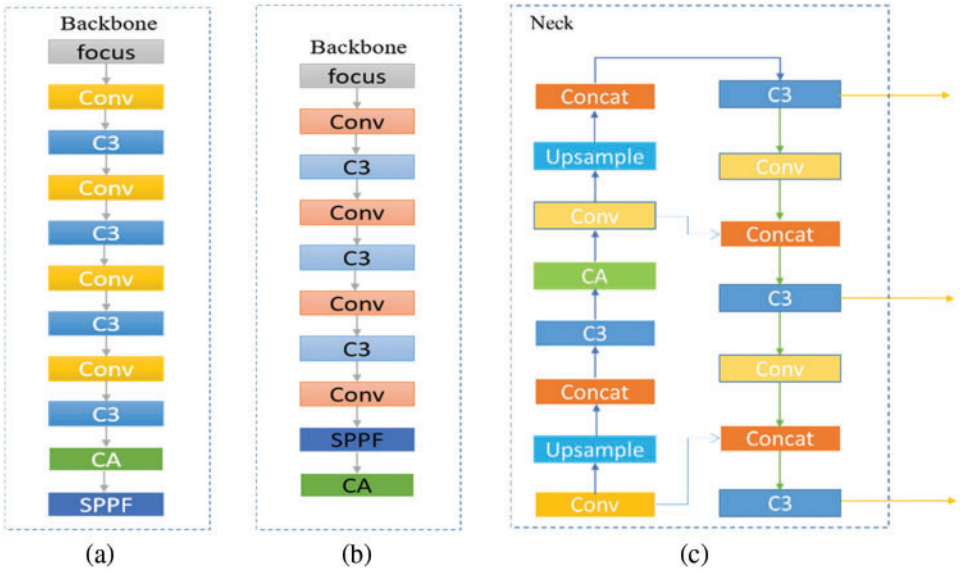


Figure 10: CA attention mechanism in different locations

Table 4: Comparison of different lightweight models

No.	Models	mAP (%)	Parameter size	GFLOPs	Weight (M)	Speed-GPU per image (ms)	FPS
1	YOLOv5s	95.60	7025023	16	14.1	10.4	96
2	YOLOv5s_Ghost_all	96.11	5582199	11.3	11.6	8.2	121
3	YOLOv5s_Ghost_neck	96.32	6753903	14.8	13.2	8.5	117
4	YOLOv5s_Ghost_backbone	96.73	6544487	13.2	13.5	7.5	134

4 Discussion

In this paper, an automatic detection method for real-time detection of helmet-wearing is discussed. To satisfy the requirements of construction enterprises to know the helmet-wearing status of construction workers at any time, this study uses the YOLO_CA model for a more in-depth study. The detection performance in complex scenes is improved by adding an attention mechanism. EIoU is used

as the loss function, it results in faster convergence and better localization. The method's feasibility is as follows:

1. When it comes to model weight size and detection precision, the dataset scenes used in this study are all based on images of construction sites with relatively complex background information, which can meet the application in actual building construction scenarios. In order to be better deployed in the device, the model was optimized to decrease the number of parameters and the size of the model's weight. After optimization, the weight size and the number of parameters are both reduced by 4.26% and 6.8%, respectively, but the accuracy and mAP value are only reduced by 0.37% and 0.49%, respectively, during the lightweight of the model. As a consequence, the model suggested in this study may be applied in practice and produces good results.

2. As for the speed of detection, to satisfy the real-time requirements of supervision and management, as well as to detect models across numerous objects, our proposed YOLO_CA model, which serves as a useful example of a sophisticated single-stage object identification approach, is considered. In the same hardware environment, the one-stage model outperforms the two-stage method in processing speed. When compared with other one-stage methods (Such as YOLOv3_tiny, Yolov7, and Yolov8 in the above experiments), this model has a higher mAP value and accuracy, as well as faster detection speed. Although the updated network structure is still complicated when compared to the other basic model, the improved model has improved detection speed and accuracy, which can meet real-time requirements.

3. Regarding the capacity to generalize models, the method of using public datasets for validation improves the generalization ability and robustness of the model.

Following the preceding discussion, it is believed that the proposed method in this study is an effective method for construction personnel helmet-wearing detection, which might boost firm safety monitoring management's effectiveness and reduce the corresponding manual supervision cost, as well as promote the intelligent development of construction production safety management.

5 Conclusion

In this essay, we put forth and build a lightweight model YOLO_CA for real-time helmet detection in construction scenarios, based on the CA mechanism and the Yolov5 target identification algorithm. A specific dataset under the building construction scene is constructed for training the network model to overcome the issue of poor detection precision due to complex background and uneven illumination of the site environment. The addition of the Ghost module, CA attention mechanism, and DWConv decreases the model's overall size while also enabling the model to interpret redundant information better and faster, as well as saving model parameters and operating costs, thus resulting in a lighter model. The results of the comparison experiments showed that the model proposed in this study achieved good results in all indicators, with Precision and Recall values of 95.87% and 95.31%, mAP value of 96.73%, and FPS of 134. Compared with Yolov3_tiny, Yolov5s, Yolov7-tiny, Yolov7, Yolov8n and Yolov8s, Precision increased by 3.49%, 0.67%, 4.67%, 3.27%, 5.97% and 4.77%, respectively, and mAP was up by 15.85%, 1.13%, 7.43%, 4.53%, 10.13% and 6.63%, respectively. It can be seen that the improved method has good superiority and effectiveness, and the proposed YOLO_CA model's higher detection accuracy and recognition speed. Therefore, the YOLO_CA model can be better applied to mobile and embedded devices, so that the supervision and inspection of helmet-wearing can be gradually shifted from manual to artificial intelligence, and the efficiency of safety supervision can be improved.

In future studies, to expand the uses of algorithms and be able to expand to more detection devices, the following aspects can be addressed. Firstly, data sources from different domains need to be expanded to further extend the application in helmet detection. Because more scenes of helmet-wearing samples can not only test the method's generalizability but also look into other types of complex scene-wearing features, which will assist in increasing the detection level. Secondly, the enhanced dataset is included in the subsequent training model to provide a more accurate and comprehensive helmet detection algorithm, therefore that the proposed YOLO_CA model in this paper can be used not only for helmet-wearing detection in building construction scenes but also in other industrial scenes as well as traffic scenes. In addition, two-stage target detection methods or traditional deep learning methods can be included in the follow-up work to compare with the methods used to make the model more convincing; it is also necessary to pay attention to the cutting edge of the technology, and the latest models or techniques can be considered in the follow-up to replace the previous methods, so as to better enhance the effectiveness of the application. Finally, the proposed model can be applied to embedded devices with limited computational power and real-time computing requirements, such as UAVs and smartphones.

Acknowledgement: We thank the funders of this project Guizhou Optoelectronic Information and Intelligent Application International Joint Research Center, and all the teams and individuals who supported this research.

Funding Statement: This research was funded by Guizhou Optoelectronic Information and Intelligent Application International Joint Research Center (Qiankehe Platform Talents No. 5802 [2019]).

Author Contributions: Conceptualization, X.Q.W. and M.Y.; methodology, X.Q.W.; software, X.Q.W. and M.Y.; validation, M.Y.; formal analysis, X.Q.W. and M.Y.; investigation, S.R.Q.; resources, S.R.Q.; data curation, X.Q.W; writing—original draft preparation, X.Q.W.; writing—review and editing, M.Y. and S.R.Q.; visualization, X.Q.W. and M.Y.; supervision, S.R.Q.; project administration, M.Y. and S.R.Q.; funding acquisition, S.R.Q. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: Two open datasets SHWD and CHV. SHWD [42] at https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset (accessed on 05/12/2022). CHV dataset [43] at https://github.com/ZijianWang-ZW/PPE_detection (accessed on 06/01/2023).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. H. Kim and H. S. Kong, "The effects of apartment facility maintenance on the residential satisfaction of residents," *The Journal of Convergence on Culture Technology*, vol. 6, no. 3, pp. 175–183, 2020.

[2] N. J. Kwak and D. J. Kim, "A study on detecting the safety helmet wearing using YOLOv5-S model and transfer learning," *The International Journal of Advanced Culture Technology*, vol. 10, no. 1, pp. 302–309, 2022.

[3] U.S. Department of Labor. "Occupational Safety and Health Act of 1970," [Online]. Available: https://www.osha.gov/laws-regs/oshact/completeoshact (accessed on 10/03/2023)

[4] Ministry of Housing and Urban-Rural Development, "Announcement on the production safety accidents of housing municipal engineering in 2020," 2020. [Online]. Available: https://www.mohurd.gov.cn/gongkai/zhengce/zhengcefilelib/202210/20221026_768565.html (accessed on 10/03/2023)

[5] Fatal Injuries in Great Britain, "Technical report," 2017. [Online]. Available: https://www.hse.gov.uk/statistics/fatals.htm (accessed on 10/03/2023)
[6] B. Michael, D. G. Gina, T. Stanley and M. Steven, "Effect of helmet design on impact performance of industrial safety helmets," *Heliyon*, vol. 8, no. 8, pp. e09962, 2022.
[7] B. L. Suderman, R. W. Hoover, R. P. Ching and I. S. Scher, "The effect of hardhats on head and neck response to vertical impacts from large construction objects," *Accident Analysis and Prevention*, vol. 73, pp. 116–124, 2014.
[8] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan and A. Alamri, "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, vol. 11, no. 1, pp. 88–95, 2017.
[9] H. Zhang, X. Yan, H. Li, R. Jin and H. F. Fu, "Real-time alarming, monitoring, and locating for non-harhat use in construction," *Journal of Construction Engineering and Management*, vol. 145, no. 3, pp. 1943–7862, 2019.
[10] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M. J. Khazaee *et al.,* "Mobile passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites," *Automation in Construction*, vol. 36, pp. 38–52, 2013.
[11] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison *et al.,* "Fast personal protective equipment detection for real construction sites using deep learning approaches," *Sensors*, vol. 21, no. 10, pp. 3478, 2021.
[12] J. O. Seo, S. U. Han, S. H. Lee and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 239–251, 2015.
[13] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo *et al.,* "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, vol. 85, pp. 1–9, 2018.
[14] J. Wu, N. Cai, W. Chen, H. Wang and G. Wang, "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset," *Automation in Construction*, vol. 106, pp. 102894, 2019.
[15] H. Wu and J. Zhao, "Automated visual helmet identification based on deep convolutional neural networks," *International Symposium on Process Systems Engineering*, vol. 44, pp. 2299–2304, 2018.
[16] Z. Xie, H. Liu, Z. Li and Y. He, "A convolutional neural network based approach towards real-time hard hat detection," in *Proc. of the 2018 IEEE Int. Conf. on Progress in Informatics and Computing*, Suzhou, China, pp. 430–434, 2018.
[17] L. Wang, Y. Cao, S. Wang, X. Song, S. Zhang *et al.,* "Investigation into recognition algorithm of Helmet violation based on YOLOv5-CBAM-DCN," *IEEE Access*, vol. 10, pp. 60622–60632, 2022.
[18] P. Wen, M. Tong, Z. Deng, Q. Qin and R. Lan, "Improved helmet wearing detection method based on YOLOv3," in *Int. Conf. on Artificial Intelligence and Security*, Hohhot, China, pp. 670–681, 2020.
[19] B. Wang, H. Xiong and L. Liu, "Safety helmet wearing recognition based on improved YOLOv4 algorithm," in *2022 IEEE 6th Information Technology and Mechatronics Engineering Conf. (ITOEC)*, Chongqing, China, pp. 1732–1736, 2022.
[20] S. D. Khan, H. Ullah, M. Ullah, N. Conci, F. A. Cheikh *et al.,* "Person head detection based deep model for people counting in sports videos," in *2019 16th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, pp. 1–8, 2019.
[21] S. D. Khan and S. Basalamah, "Scale and density invariant head detection deep model for crowd counting in pedestrian crowds," *The Visual Computer*, vol. 37, no. 8, pp. 2127–2137, 2021.
[22] Tzutalin, "LabelImg," 2015. [Online]. Available: https://github.com/tzutalin/labelImg (accessed on 20/05/2020).
[23] Q. Wang, M. Cheng, S. Huang, Z. Cai, J. Zhang *et al.,* "A deep learning approach incorporating YOLO v5 and attention mechanisms for field real-time detection of the invasive weed Solanum rostratum Dunal seedlings," *Computers and Electronics in Agriculture*, vol. 199, pp. 107194, 2022.
[24] M. H. Yap, R. Hachiuma, A. Alavi, R. Brüngel, E. Frank *et al.,* "Deep learning in diabetic foot ulcers detection: A comprehensive evaluation," *Computers in Biology and Medicine*, vol. 135, pp. 104596, 2021.

[25] I. Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Computers in Biology and Medicine*, vol. 134, pp. 104519, 2021.
[26] M. O. Lawal, "Tomato detection based on modified YOLOv3 framework," *Scientific Reports*, vol. 11, no. 1, pp. 1447, 2021.
[27] S. W. Li, X. Y. Gu, X. R. Xu, D. W. Xu, T. J. Zhang *et al.,* "Detection of concealed cracks from ground penetrating radar images based on deep learning algorithm," *Construction and Building Materials*, vol. 273, pp. 121949, 2021.
[28] Z. Y. Zhao, X. X. Yang, Y. C. Zhou, Q. Q. Sun, Z. D. Ge *et al.,* "Real-time detection of particleboard surface defects based on improved YOLOv5 target detection," *Scientific Reports*, vol. 11, no. 1, pp. 21777, 2021.
[29] Q. Hou, D. Zhou and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 13708–13717, 2021.
[30] S. Y. Guo, L. L. Li, T. Y. Guo and Y. L. Li, "Research on mask-wearing detection algorithm based on improved YOLOv5," *Sensors*, vol. 22, no. 13, pp. 4933, 2022.
[31] S. F. Li, K. Y. Li, Y. Qiao and L. X. Zhang, "A multi-scale cucumber disease detection method in natural scenes based on YOLOv5," *Computers and Electronics in Agriculture*, vol. 202, pp. 107363, 2022.
[32] Q. B. Hou, D. Q. Zhou and J. S. Feng, "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, pp. 13708–13717, 2021.
[33] X. Dong, S. Yan and C. Duan, "A lightweight vehicles detection network model based on YOLOv5," *Engineering Applications of Artificial Intelligence*, vol. 113, pp. 104914, 2022.
[34] K. Han, Y. H. Wang, Q. Tian, J. Y. Guo, C. J. Xu *et al.,* "GhostNet: More features from cheap operations," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New York, USA, pp. 1580–1589, 2020.
[35] S. J. Ma, W. K. Liu, W. Cai, Z. W. Shang and G. Liu, "Lightweight deep residual CNN for fault diagnosis of rotating machinery based on depthwise separable convolutions," *IEEE Access*, vol. 7, pp. 57023–57036, 2019.
[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii State, USA, pp. 1251–1258, 2017.
[37] A. G. Howard, M. L. Zhu, B. Chen, D. Kalenichenko, W. J. Wang *et al.,* "MobileNets: Efficient convolutional neural networks for mobile vision applications," *Remote Sensing*, vol. 14, no. 12, pp. 2763, 2017.
[38] Z. Xiao, Z. Zhang, K. W. Hung and S. Lui, "Real-time video super-resolution using lightweight depth-wise separable group convolutions with channel shuffling," *Journal of Visual Communication and Image Representation*, vol. 75, pp. 103038, 2021.
[39] C. Liu, H. Sui, J. Wang, Z. Ni and L. Ge, "Real-time ground-level building damage detection based on lightweight and accurate YOLOv5 using terrestrial images," *Remote Sensing*, vol. 14, no. 12, pp. 2763, 2022.
[40] Z. Zheng, P. Wang, W. Liu, L. Jinze, R. Ye *et al.,* "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, vol. 34, no. 7, pp. 12993–13000, 2020.
[41] Y. F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang *et al.,* "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, 2022.
[42] njvisionpower, "Safety-helmet-wearing-dataset," 2019. [Online]. Available: https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset (accessed on 05/12/2022)
[43] ZijianWang-ZW/PPE_detection, "CHV dataset," 2020. [Online]. Available: https://github.com/ZijianWang-ZW/PPE_detection (accessed on 06/01/2023)