Hindawi
Journal of Sensors
Volume 2021, Article ID 4746516, 12 pages
https://doi.org/10.1155/2021/4746516

Research Article

DWCA-YOLOv5: An Improve Single Shot Detector for Safety Helmet Detection

Zhang Jin,[1,2,3] **Peiqi Qu**,[1] **Cheng Sun**,[4] **Meng Luo**,[4] **Yan Gui**,[2] **Jianming Zhang**,[2] **and Hong Liu**[1]

[1]*School of Information Science and Engineering, Hunan Normal University, Changsha 410081, China*
[2]*School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China*
[3]*Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310058, China*
[4]*School of Mathematics and Statistics, Hunan Normal University, Changsha 410081, China*

Correspondence should be addressed to Hong Liu; qpeggy@hunnu.edu.cn

Received 24 July 2021; Accepted 9 September 2021; Published 7 October 2021

Academic Editor: Ruizhen Yang

Aiming at solving the problem that the detection methods used in the existing helmet detection research has low detection efficiency and the cumulative error influences accuracy, a new algorithm for improving YOLOv5 helmet wearing detection is proposed. First of all, we use the *K*-means++ algorithm to improve the size matching degree of the a priori anchor box; secondly, integrate the Depthwise Coordinate Attention (DWCA) mechanism in the backbone network, so that the network can learn the weight of each channel independently and enhance the information dissemination between features, thereby strengthening the network's ability to distinguish foreground and background. The experimental results show as follows: in the self-made safety helmet wearing detection dataset, the average accuracy rate reached 95.9%, the average accuracy of the helmet detection reached 96.5%, and the average accuracy of the worker's head detection reached 95.2%. Making a comparison with the YOLOv5 algorithm, our model has a 3% increase in the average accuracy of helmet detection, which is in line with the accuracy requirements of helmet wearing detection in complex construction scenarios.

1. Introduction

According to a series of statistical reports issued by the Ministry of Housing and Urban-Rural Development, compared with 934 accidents and 840 deaths in 2018, there were a total of 773 construction production safety accidents and 904 deaths across the country in 2019. The number of accidents and deaths increased by 5.31% and 7.62%. In general, the number of accidents in the construction industry is showing a gradual increase. In the literature [1], when studying the relationship between the use of safety protection equipment and the number of deaths in construction sites, it was found that 67.95% of the victims had not used or used safety protection (such as safety helmets and safety belts). Due to the weak awareness of safety protection of construction workers, the importance of wearing safety helmets is often ignored. At the construction site, manual supervision is usually used to monitor whether workers wear safety helmets [2], which makes it impossible to manage all construction workers promptly on the construction site and to know the movement tracks of all construction workers. The use of automatic monitoring methods helps to monitor the construction personnel and confirm the specific conditions of all construction workers wearing helmets at the construction site, especially when the traditional monitoring methods are time-consuming and expensive, easy to detect errors, and are not enough to meet the safety of modern building construction management requirements. The use of automatic supervision of deep learning methods is conducive to supervising all construction personnel onsite.

Traditional object detection often uses an artificial selection of features and design and training classifiers based on specific detection objects. This method is highly subjective, complex in the design process, has poor generalization

ability, and has great limitations in engineering applications. In recent years, due to the fact that convolutional neural networks (CNN) do not use an artificial selection of features, they have gradually been sought after by scholars in the field of deep learning. The deep convolutional neural network has good comprehensive performance in the field of object detection. In 2014, Girshick et al. successfully proposed R-CNN [3], fast R-CNN [4], and faster R-CNN [5], which were verified in the PASCAL VOC2007dataset, respectively, and gradually improved the experimental effect. The method of extracting feature frames by these models gradually changes from selective search to regional proposal network (RPN), thus getting rid of the traditional manual feature extraction method. In 2015, Redmon and others proposed a one-stage object detection model YOLO [6], which abstracted the detection task as a regression problem for the first time, avoiding the cumbersome operation of dividing the detection task into two steps in the R-CNN series. In 2016, Liu et al. proposed the SSD [7] detection algorithm, which introduced a multiscale detection method, which can effectively detect groups of small targets. In 2017, Lin et al. proposed the RetinaNet [8] dense detector, which solves the problem of extreme foreground and background imbalance encountered during training by reshaping the standard entropy loss. In 2017, Redmon and others proposed the YOLOv2 [9] detection model, which selected a new basic model Darknet-19 to achieve end-to-end training. In 2018, Redmon et al. proposed YOLOV3 [10] based onYOLOV1 and YOLOV2. In this model, the FPN method was adopted to integrate three different sizes feature maps to accomplish detection tasks, which significantly improved the detection effect of small-size targets. In April 2020, Bochkovskiy proposed YOLOv4 [11], which uses PANet instead of FPN used in YOLOv3 as the path aggregation method; at the same time, the backbone network uses CSP Darknet53, which significantly enhances the detection accuracy of the network. In June 2020, Glenn proposed YOLov5 [12], which designed a new focus structure and added it to the backbone network to achieve a new benchmark for the perfect combination of speed and accuracy.

Because of the rapid rise of computer vision in the direction of object detection, more and more researchers are focusing on combining deep learning with practical application scenarios. For example, Chen et al. [13] improved the SSD model by adding an inception module before the prediction layer to achieve rapid and accurate detection of small vehicles. Tian et al. [14] used DenseNet to optimize the low-resolution layer in the feature layer of the YOLOv3 network and applied the improved YOLOv3 to the detection of anthrax lesions on the surface of orchard apples to achieve real-time detection. Dashun et al. [15] applied the improved RetinanNet network to the field of pedestrian detection and realized the rapid detection of multispectral pedestrians. Zhong et al. [16] used the LocNet positioning module to replace the boundary regression module to improve the faster R-CNN model and applied it to multidirectional text instance detection. Zhang et al. [17, 18] used the residual network (reset) in the prediction part to encode the input features of the image and chose to increase the deconvolution layer to change the MMDetection network model in the process of feature information decoding, to achieve a higher crowd in dense scenes. And it can be seen that deep learning has become a popular research direction, and it has become the mainstream field in combination with actual application scenarios.

Safety helmet detection is one of the application areas of object detection. So far, many researchers at home and abroad have conducted several related investigations on safety helmet detection. In 2013, Kelm [19] and others designed a mobile radio frequency identification (RFID) portal to check the compliance of construction workers wearing safety protective equipment. However, the recognition area of the radio frequency identification reader is limited. It is only recommended that the helmet be close to the worker, but it cannot be confirmed whether the helmet is worn correctly. In 2014, Liu [20] and others used a combination of support vector machines and skin color detection to achieve helmet detection. In 2016, Rubaiya [21] and Silva [22] and others combined the histogram of gradient (HOG) algorithm with the frequency domain-related information in the image for human detection and then used the circular Hough transform (CHT) to detect the helmet. In 2017, Li [23] and others used the vibe algorithm to locate the human body position, followed by the embossing algorithm to detect the worker's head and finally combined the HOG algorithm and SVM to realize the helmet wearing detection. In 2018, Wu et al. [24] used Hu moment invariant (HMI), color histogram (CH), and local binary pattern (LBP) to extract the characteristics of different color helmets and then constructed a hierarchical support vector machine (H-SVM) for safety cap wearing detection. Due to the complex environment, the detection accuracy of helmet wearing detection is low at this stage, which is quite different from the management requirements in actual building construction.

In this paper, two types of targets for construction workers wearing helmets and those not wearing helmets are the detection tasks, and more than 7,000 pictures are collected from the Internet for preprocessing to construct a helmet detection dataset. Select the YOLOv5 network model as the main body, and first, use the k-means++ algorithm to cluster the target anchor box to obtain a bounding box suitable for the target, so that the model can converge faster. Secondly, a new DWCA module is designed and integrated with the features of the backbone network to strengthen the attention to enhance the attention of the detection target and improve the ability to resist background interference. According to the final experimental results, the average detection accuracy (mAP) of the DWCA-YOLOv5 detection model has been significantly improved, and it can effectively detect the unsafe behavior of workers on the construction site not wearing helmets.

2. Related Work

2.1. YOLOv5 Algorithm Principle. YOLOv5 is a new-generation target detection network of the YOLO series. It is a product of continuous integration and innovation based on YOLOv3 and YOLOv4. Secondly, YOLOv5 has achieved

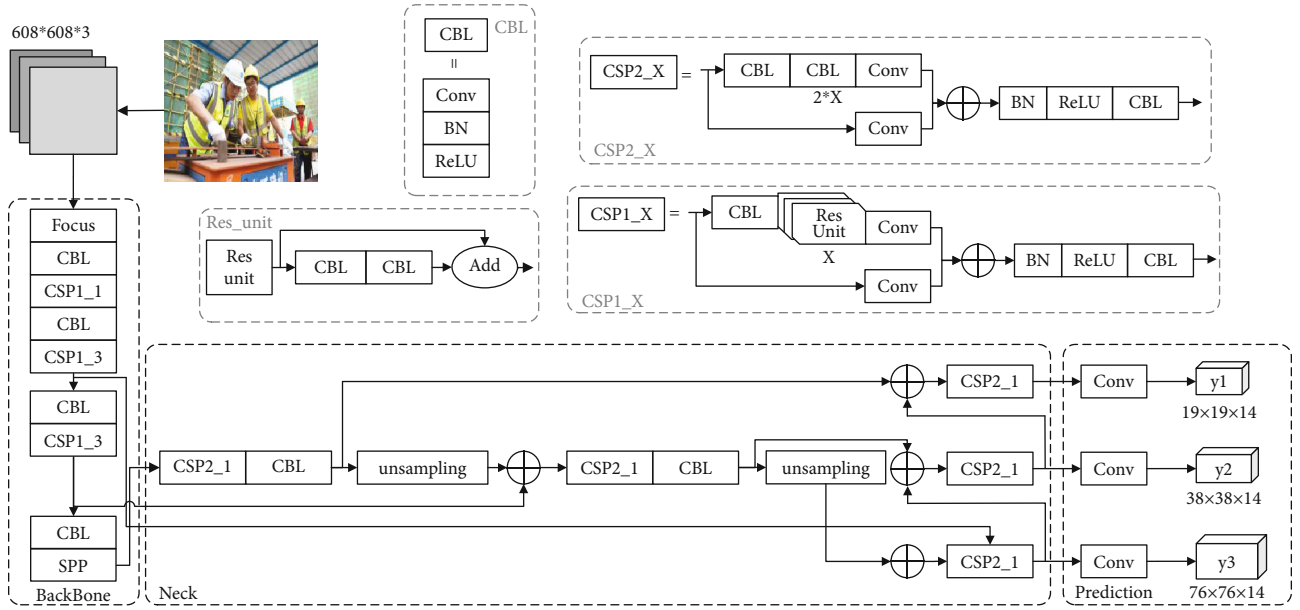


Figure 1: Network structure diagram of YOLO v5s.

better results in PASCAL VOC and COCO object detection tasks; so, this article uses the YOLOv5 detection network to detect the construction workers' helmet wearing.

The YOLOv5 object detection network official gave four network models: YOLOv5s, YOLOv5m, YOLOv51, and YOLOv5x. The three models of YOLOv5m, YOLOv51, and YOLOv5x are the products of continuous deepening and widening based on YOLOv5s. The YOLOv5 network structure is divided into four parts: input, backbone, neck, and prediction. YOLOv5 adds Mosaic data enhancement in the data input part; focus structure and CSP structure are used in the backbone; the FPN + PAN structure is added to the neck; the prediction part improves the bounding box loss function from CIOU_Loss to GIOU_Loss; YOLOv5 targets many in the postprocessing process of object detection. The screening of the target anchor frame adopts a weighted NMS operation.

Compared with YOLOv4, YOLOv5 has a new focus structure in the backbone network, which is mainly used for slicing operations. In the YOLOv5s network model, an ordinary image with a size of $3 \times 608 \times 608$ is input into the network, and after a focus slice operation, the feature map with a size of $12 \times 304 \times 304$ is converted, followed by the ordinary convolution operation of 32 convolution kernels. It is finally converted into a feature map with a size of $32 \times 304 \times 304$. Different from the YOLOv4 network model that only uses the CSP structure in the backbone network, the YOLOv5 network model has designed two new CSP structures. Taking the YOLOv5s network model as an example, the backbone network uses the CSP1_1 structure and the CSP1_3 structure, and the neck uses the CSP2_1 structure to enhance the feature fusion between networks. The network structure of YOLOv5s is shown in Figure 1.

2.2. DWCA Moduel. The traditional channel module is dedicated to constructing various channel importance weight functions. For example, SEnet [25] obtained a significant effect improvement by calculating channel attention with the aid of a 2D global pool and with a small computational overhead. However, SENet only considers the encoding of information between channels and ignores the importance of position information, which is essential for capturing the structure of objects in vision tasks. Coordinate attention [26] has achieved significant performance improvement by encoding the interchannel relationship and long-term dependence. ECANet [27] proposed a method that does not take dimensionality reduction measures to achieve cross-channel local interaction and a method that automatically adapts to select one-dimensional ordinary convolution, thereby achieving performance improvement. CBAM [28] and BAM [29] reduce the channel input dimension of the tensor and secondly use convolution to calculate spatial attention to use position information. However, convolution can only capture local relationships, but not what is needed for modeling long-term dependence on visual tasks.

To solve the above problems, we designed a new attention mechanism based on previous work, which integrates the position information in the feature space into the channel attention, so that the network can participate in a larger area and at the same time avoid a lot of model parameters overhead. The structure diagram of DWCA mechanism is shown in Figure 2.

To reduce the lack of relevant location information caused by two-dimensional global sharing, we use two one-dimensional global aggregation operations to decompose the channel attention into two aggregated features along with the vertical and horizontal directions and then aggregate the obtained features into two independent directional perception features map. To promote the module to capture the remote spatial interaction with precise location information, this paper decomposes the global pooling according to formula (1) and transforms it into a one-to-one dimensional feature encoding operation. The specific operation process is

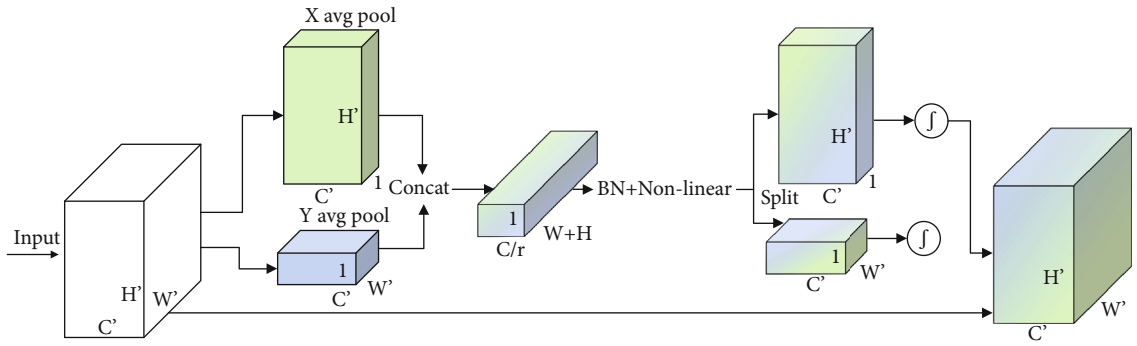


Figure 2: DWCA model network structure.

as follows: first, use a pooling kernel of size $(H, 1)$ or $(1, W)$ to encode the single dimension and horizontal and vertical coordinates of input X. Therefore, the cth channel of the output with a height of h can be seen below, and the details are shown in formula (2).

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

In the formula, Z_c is the output related to the cth channel, H is the height of the input X, and W is the width of the input X.

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i). \tag{2}$$

In the formula, $Z_c^h(h)$ is the specific output of the cth channel where the height is h, and W is the width of the input X.

By analogy, the specific output of the cth channel with width w can be seen below, see formula (3) for details.

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w). \tag{3}$$

In the formula, $Z_c^h(w)$ is the output of the cth channel at the width w, and H is the height of the input X.

By extending the above two features to the transformation of the aggregation of the two spatial dimensions, the direction-aware feature map is obtained, followed by CONCAT operation, and then use the shared 1×1 conventional convolution transformation function to transform it, such as the formula (4) shown.

$$f = \delta\left(F1\left(\left[z^h, z^h\right]\right)\right). \tag{4}$$

In the formula, $f \in R^{C/r \times (H+W)}$ is an intermediate feature map, which encodes the spatially related information in the vertical and horizontal directions, δ is a nonlinear activation function, and $[\cdot, \cdot]$ represents the splicing operation along the spatial dimension.

Then, follow the spatial dimension, and f is transformed into two independent tensors $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$, using two effective depthwise separable convolution transforms f^h and f^w and then transforms the tensors f^h and f^w with the same number of channels into input X, as shown in formulae (5) and (6).

$$g^h = \sigma\left(F_h\left(f^h\right)\right), \tag{5}$$

$$g^w = \sigma(F_w(f^w)). \tag{6}$$

In the formula, g^h and g^w are the attention weights to be expanded, σ is the Sigmoid function, and r is the reduction ratio of the number of channels.

Finally, the entire DWCA module can be expressed as follows, see formula (7) for details:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j), \tag{7}$$

2.3. Improve the YOLOv5 Algorithm

2.3.1. K-Means++ for Target Frame Optimization. Perform K-means dimensional clustering on the general target detection dataset COCO to obtain the initial a priori anchor frame parameters of YOLOv5. However, because the target types of the COCO dataset have 80 categories, the helmet detection types in this article only have two categories, which cannot be to meet the actual needs of helmet wearing detection, and the size of the a priori frame needs to be redesigned. Compared with the size of the anchor frame designed only relying on human prior knowledge, for the helmet wearing dataset, we select the K-means++ algorithm to perform multidimensional clustering on the marked target frame, resulting in different numbers and sizes. As far as possible, the accurate matching between the a priori anchor frame and the actual object is achieved, thereby further improving the accuracy of helmet detection. In the clustering process, the average intersection ratio (IoU) corresponding to the number of centers of different clusters is shown in Figure 3.

Observing Figure 3, we can get that when the number of prior anchor box clusters is 0 to 9, the average intersection ratio shows a rapid upward trend, but when the number of a priori anchor boxes is 9 to 12, the average intersection ratio

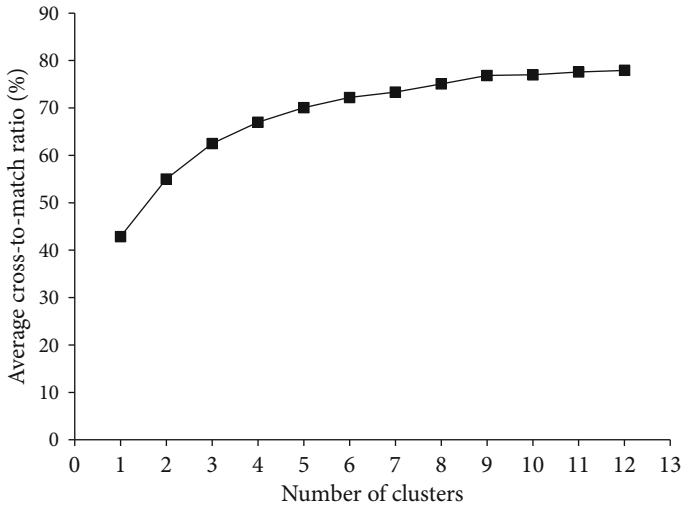


Figure 3: The number of centers of different clusters and the average intersection ratio.

increases gradually gentle. To balance the calculation accuracy and efficiency, 9 a priori anchor frames are finally selected and equally distributed to 3 prediction branches of different sizes. The determined a priori anchor frame sizes are normalized as shown in Table 1.

Table 2 illustrates the pros and cons of the model's performance. Among them, the clustering method of a priori anchor box is changed from K-means algorithm to K-means++ algorithm, mAP has a certain improvement, and the improved YOLOv5 algorithm changes due to the network structure and the detection accuracy. There is also a big improvement. At the same time, selecting the K-means++ clustering method and the improved YOLOv5 model is 3.2 percentage points higher than the original YOLOv5 algorithm. The average accuracy of the self-made helmet wearing detection dataset reaches 95.9%, which can accurately detect whether the construction personnel wears a hard hat.

Table 1: Prior anchor box scales.

Feature map scale	Anchor box size		
	Anchor 1	Anchor 2	Anchor 3
Small scale	(11.09,18)	(21.5,30.8)	(30.8,43)
Middle scale	(38.1, 60)	(52.3, 73.6)	(63,103.3)
Large scale	(89.2, 135)	(120, 207.5)	(209.4, 324)

Table 2: Effect evaluation of different models on the test set.

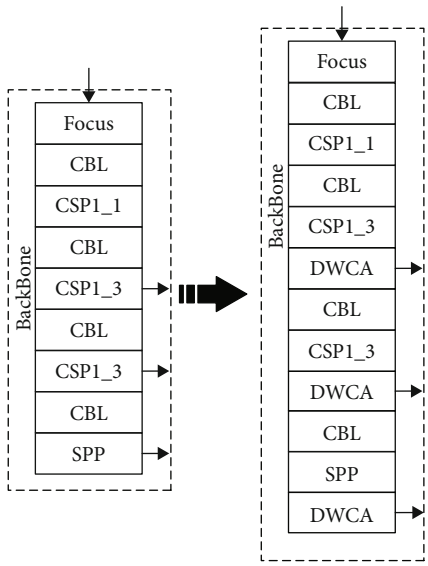
Detection model	Clustering method	AP50/%		mAP/%
		Hat	Person	
Original YOLOv5	K-means	93.3	91.7	92.7
Original YOLOv5	K-means++	94.4	92.8	93.6
Improved YOLOv5	K-means	95.5	94.6	95.1
Improved YOLOv5	K-means++	96.5	95.2	95.9

2.3.2. DWCA Module Fusion Design. For small target detection tasks, as the sum of the model network layers gradually increases, the feature information of small targets that can be collected gradually decreases. So, it is easy to cause the network model to false detection and miss detection of small targets. The DWCA module itself is to integrate the location information of the feature space with the channel features so that the network can grasp the "key points" of the target features during the training process. However, under specific circumstances, which position of the DWCA module to perform feature fusion in the network model is effective is still a question to be studied.

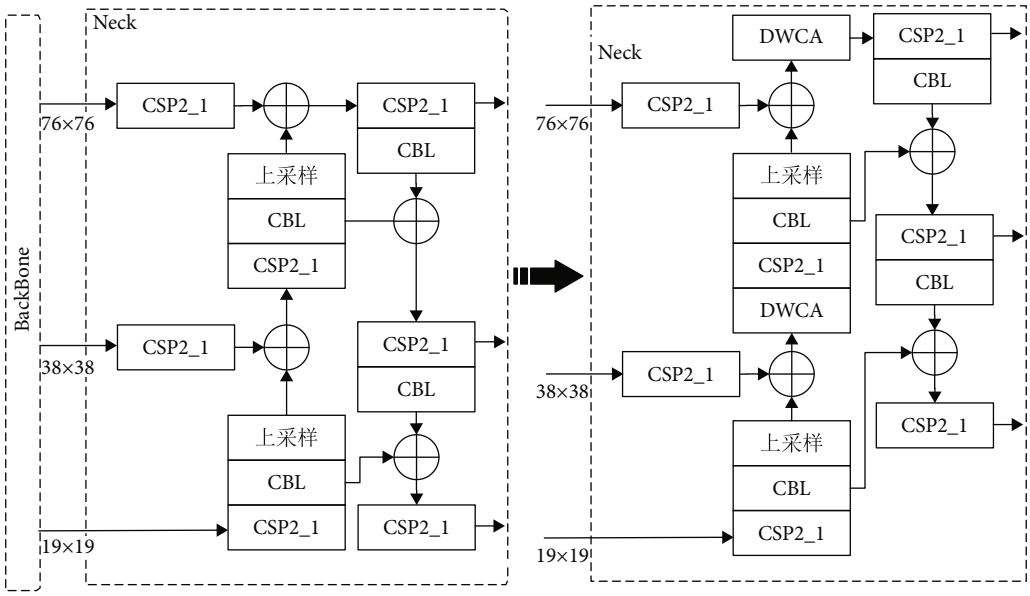
In this paper, the DWCA module is merged into different positions of the network model, and the detection results are studied. According to the structure of the YOLOv5s network model, this paper will integrate the DWCA module in the three areas of the backbone network, the neck, and the prediction module of YOLOv5s. Since the DWCA module is to enhance the relationship between channel information and channel information in the feature space, our embeds the DWCA module into each feature fusion area in the above three parts, thereby generating three new types based on the YOLOv5s algorithm. Network model is as follows: DWCA-YOLOv5s-backbone, DWCA-YOLOv5s-neck, and DWCA-YOLOv5s-prediction. Figure 4 shows the specific location where the DWCA module is integrated into the network.

In Figure 3(a), the DWCA module is integrated at CSP1_3 (i. e., the feature fusion) in the backbone network of YOLOV5s. In Figure 3(b), the DWCA module is integrated behind the CONCAT layer on the neck of YOLOV5s. In Figure 3(c), the DWCA module is integrated, respectively, before the convolution of each prediction in YOLOV5s. Table 3 shows the experimental results of whether the DWCA module is integrated with three different positions.

By visualizing the output of the same channel of the three fusion-designed networks, as shown in Figure 5 (only the channel output of the same feature map is visualized), the experimental results show that, compared to fusing the DWCA module into the network neck and network

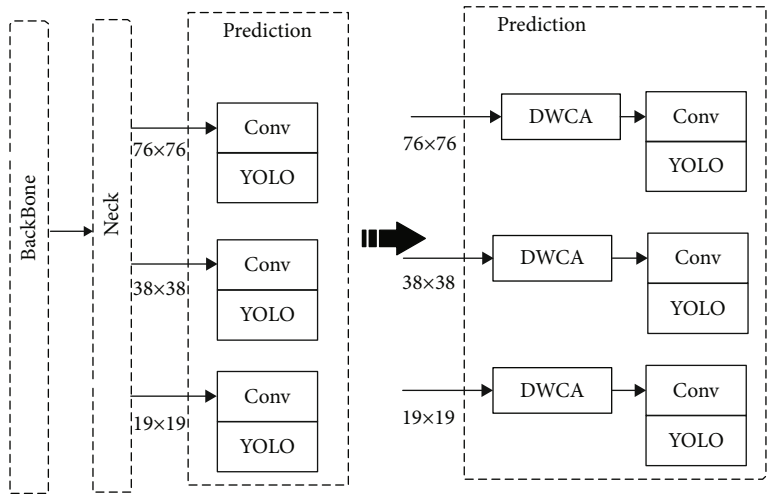


(a) DWCA-YOLOv5s-backbone



(b) DWCA-YOLOv5s-neck

Figure 4: Continued.



(c) DWCA-YOLOv5s-prediction

Figure 4: Three YOLOv5s modes embedded in the DWCA modules.

Table 3: Comparison of results of different detection models.

Network model	*P*/%	*R*/%	Model parameters/*M*	mAP/%
YOLOv5s	76.4	92.5	7.26	92.7
DWCA-YOLOv5-backbone	82.5	95.4	7.27	95.9
DWCA-YOLOv5-neck	70.9	93.7	7.26	91.6
DWCA-YOLOv5-prediction	72.5	92.8	7.27	92.4

prediction part, fusing DWCA into the backbone network can effectively strengthen the semantic information of the feature layer on the instance and pay more attention to the target hidden in the lower layer, which is easy to ignore. The texture information and contour information can effectively improve the network's attention to small targets.

3. Experimental Results and Analysis

3.1. Dataset Construction. In the detection direction, the dataset required by experiments has always been an essential basic condition. The safety helmet dataset that has been open sourced is only SHWD (SafetyHelmetWearing-Dataset). In this dataset, the category label data of not wearing a helmet is mainly derived from the SCUT-HEAD dataset. The SCUT-HEAD dataset is used by students in classroom scenarios monitoring diagrams or photos taken, so the dataset is not a standard construction site scene dataset, which does not meet the detection requirements of actual building construction scenarios. To solve this problem, this article self-made a helmet wearing detection dataset in construction scenarios. The main process of constructing this dataset includes data collection, screening, and processing.

3.2. Data Collection. The images required for the dataset in this article mainly come from the surveillance video framing of the construction site, self-collecting on the construction site, and Internet crawling. The collected data includes two types of pictures of workers wearing and not wearing helmets in different environments, different resolutions, and different construction sites. Multiple sets of interference pictures are added to the dataset, such as construction workers wearing baseball caps and safety helmets. Construction workers with hats placed on the table or in hand, construction workers wearing bamboo woven hats, etc., increase the diversity of the dataset, thereby enhancing the robustness of the network. The sample map of the dataset collected this time is shown in Figure 6.

3.3. Data Screening and Processing. The pictures collected from the surveillance video of the construction site are divided into frames or crawled on the Internet. Many of the pictures do not contain the construction personnel as the research object. They can be regarded as background pictures and have no practical significance for the study of this article. The picture data is confirmed as the background is deleted. This paper conducts a preliminary screening of the collected image and selects the images that meet the requirements as the annotation dataset.

Preprocess the data, convert the images that meet the requirements into.jpg format, and use the labeling tool labellmg to manually label each image, and the construction personnel in the image i under wearing a helmet (hat) and not wearing a helmet (person) These two categories are labeled, as shown in Figure 7; after processing, a corresponding XML tag file is formed, which contains the four

Figure 5: Model heat map comparison.

(a) Put on the table (b) Hand held (c) Hand held (d) Hand held

(e) Hand held (f) Baseball cap (g) baseball cap (h) Normal sample

(i) Bamboo braided hat (j) Police_cap (k) Normal sample (l) Sunhat

Figure 6: Safety helmet sample image.

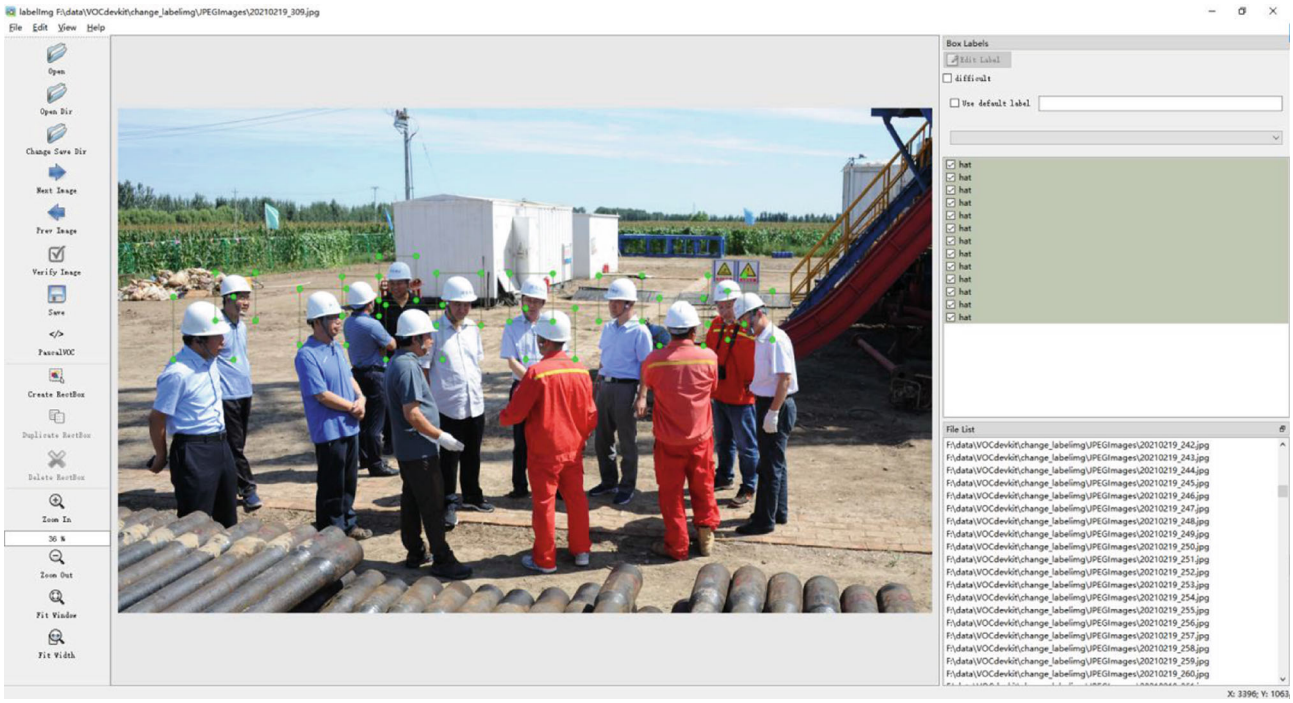


Figure 7: Safety helmet wearing status mark.

Table 4: Dataset category allocation.

Target category	Training set a target number	A test set the target number	Total number of labeled targets
Wearing helmets category	81836	11316	93152
Not wearing helmets category	98187	12021	110208

Table 5: Experiment operating environment.

Category	Entry	Version
Hardware configuration	System	Ubuntu 18.04
	GPU	GeForce RTX 2080 Ti
	CPU	AMD Ryzen 7 3800X 8-Core
Software configuration	Python version	3.8
	Deep learning framework	Pytorch
	CUDA	10.0

coordinates of the target in the frame and the given category (PASCAL VOC format).

The final dataset obtained in this paper has a total of 7076 images. Among them, the specific information of whether to wear a helmet in the dataset can be seen in Table 4. And the dataset contains a variety of construction scenes, which can more fully reflect the actual construction scenes. The final dataset is subdivided into training and validation in line with the 9 : 1 division ratio. The number of training set pictures in the final 7076 picture dataset is 6,370 pictures, and there are 706 pictures in the test set.

Table 6: Comparison of experimental results of multiple detection algorithms.

Detection model	AP50/%		mAP/%
	Hat	Person	
Faster RCNN	80.8	42.2	61.5
SSD	78.8	68.2	73.5
YOLOv3	89.12	80.7	84.9
YOLOv3 + SPP	90.5	86.3	88.4
YOLOv5m	94.8	93.1	93.9
YOLOv5l	95.1	93.5	94.3
YOLOv5x	95.6	94.3	95.0
YOLOv5s	93.3	91.7	92.7
Ours	96.5	95.2	95.9

3.4. Experimental Environment. During the experiment, this article has high requirements for the configuration of the operating environment, and GPU acceleration is required for the experiment. Table 5 shows the configuration instructions for the experiment operating environment of this article. The model building, training, and result testing are all completed under the PyTorch framework, using the CUDA parallel computing architecture and at the same time integrating the cuDNN acceleration library into the PyTorch framework to accelerate computer computing capabilities.

(a) Detection of strong light construction scene

(b) Detection of construction scenes occluded by steel bars

(c) Detection of targets of different sizes

(d) Detection of long-distance construction scenes

Figure 8: Comparison of test results of model parts under different construction scenarios.

4. Result Analysis

4.1. Evaluation Index. In object detection, detection accuracy, and recall, the average accuracy rate (mAP) is the basic index to test the training model's overall stability and performance. This article also uses the above evaluation indicators to detect helmet wearing model performance that is evaluated.

Apply the above evaluation indicators to the stability test of the helmet detection model and then the detection results of whether the construction worker wears a helmet are compared. Among them, TP_{hat} (true example), FP_{hat} (false positive example), TN_{hat} (true negative example), and FN_{hat} (false negative example) are key indicators used to describe accuracy. Specifically, TP refers to the sum of workers who did not wear helmets and whose test results were correct within the monitoring range of the construction site. FP indicates the sum of workers who wear helmets but are mistakenly detected, TN indicates that results are completely correct, and FN indicates the sum of workers who did not wear helmets but were mistakenly detected as wearing helmets. The calculation process of accuracy rate and recall rate is shown in formulae (8) and (9).

$$AP_{hat} = \frac{TN_{hat} + TP_{hat}}{TN_{hat} + TP_{hat} + FP_{hat}}, \tag{8}$$

$$Recall_{hat} = \frac{TP_{hat}}{FN_{hat} + TP_{hat}}, \tag{9}$$

$$Precision_{hat} = \frac{TP_{hat}}{TP_{hat} + FP_{hat}}. \tag{10}$$

$Precision_{hat}$ represents the ratio of real cases (TP_{hat}) to the sum of real cases and false real cases (TP_{hat}+FP_{hat}), and the sum of real cases and false real cases is the total number of helmets; $Recall_{hat}$represents the sum of real cases(TP_{hat})and real cases and false counterexamples (TP_{hat} + FN_{hat}). The ratio of true cases and false counterexamples is the actual number of helmets.

AP refers to the average value of all precisions obtained under all possible recall rates. The average precision of the mean is the average of the AP value in all categories, and the calculation formula is shown in (3).

$$m\mathrm{AP} = \frac{1}{C}\sum_{c\in C}\mathrm{AP}(c). \tag{11}$$

4.2. Result Analysis. This article uses the YOLOv5 algorithm for helmet wearing detection. To verify that the algorithm proposed has better results, the same number of test sets is used under the same configuration conditions, and several popular object detection networks at this stage are used for comparative experiments: faster RCNN, SSD, and YOLOv3. Among them, SSD and YOLOv3 are single-stage detection algorithms, and faster RCNN is a two-stage detection algorithm. The experimental results are evaluated using two evaluation indicators AP50 and mAP. The experimental results are shown in Table 6.

Observing Table 6, we can know that the DWCA-YOLOv5 algorithm can significantly improve the accuracy of detecting whether a worker is wearing a helmet. The average accuracy of the DWCA-YOLOv5 algorithm in this paper can reach 96.2% for the construction personnel who wear the helmet correctly and 95.1% for the construction personnel who do not wear the helmet. mAP (mean average precision) can reach 95.7%. Compared with faster RCNN and SSD, our model detection results are better. Compared with YOLOv3 and YOLOv5, the algorithm in this paper has a certain improvement in AP50 and mAP. This shows that the DWCA-YOLOv5 algorithm has an excellent performance in the accuracy of detection and detection of helmet wearing, and it can ensure the accuracy of helmet detection in a complex construction environment.

In addition, to more intuitively see the detection gap between different algorithms, this paper also collected 158 pictures of the construction work site as a test set. In this test set, we use YOLOv5 and our model to test separately. Some of the detection results are shown in Figure 8 below.

From Figure 8, we can observe that the operator who wears the helmet correctly is marked with a red frame, and the operator who does not wear the helmet is marked with a light green frame. Figure 8(a) shows the detection in a strong light construction scene. In comparison, the detection accuracy of the original YOLOv5 algorithm is much lower than our algorithm; Figure 8(b) shows the detection of small targets in the construction scene where steel bars are shielded. After observation, the original model missed a construction worker wearing a helmet who was behind the steel bars; Figure 8(c) shows the detection of targets with different sizes. The target size in close range is larger, and the target size in distant range is smaller. Our model has detected all the targets, while the original model missed the small targets in distant range and mistakenly detected steel pipes as two construction workers wearing safety helmets; Figure 8(d) shows the detection of small targets in a long-distance construction scene. The comparison shows that the original YOLOv5 model has missed detection of long-distance small helmets, and our model has a better detection effect. The original YOLOv5 model misses this situation. There are many inspections, but our model performs better. It can be seen from the detection comparison in the abovementioned various construction scenarios that the improved YOLOv5 model is better for the detection of safety helmets in a complex operating environment.

5. Conclusions

This paper proposes an improved YOLOv5 helmet wearing detection method. First, use the K-means++ method to perform dimensional clustering on the dataset of the self-made construction operation scene; secondly, so as to capture more detailed information, the DWCA attention mechanism is combined with the backbone network. According to the comparison of the final experimental results, our model can obtain high detection accuracy, which can meet the detection accuracy of helmets in the current complex operating environment. In the future, we will explore ways to keep the model detection accuracy as much as possible while reducing the weight of the model.

Data Availability

All data included in this study are available upon request by contact with the corresponding author.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work is partially supported by the Open Research Project of the State Key Laboratory of Industrial Control Technology (No. ICT2021B10), the Natural Science Foundation of Hunan Province (2021JJ30456), the Open Fund of Science and Technology on Parallel and Distributed Processing Laboratory (WDZC20205500119), the Hunan Provincial Science and Technology Department High-tech Industry Science and Technology Innovation Leading Project (2020GK2009), and the Scientific and Technological Progress and Innovation Program of the Transportation Department of Hunan Province (201927), etc.

References

[1] X. Chang and X. M. Liu, "Fault tree analysis of unreasonably wearing helmets for builders," *Journal of Jilin Jianzhu University*, vol. 35, no. 6, pp. 65–69, 2018.

[2] Z. Y. Wang, *Design and Implementation of Detection System of Wearing Helmets Based on Intelligent Video Surveillance*, Beijing University of Posts and Telecommunications, Beijing, China, 2018.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *in Proceedings of the IEEE conference on Computer*

Vision and Pattern Recognition, pp. 580–587, Columbus, OH, USA, 2014.

[4] R. Girshick, "Fast R-CNN. computer science," in *in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal network," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[7] W. Liu, D. Anguelov, and D. Erhan, "Single shot multibox detector," in *in Proceedings of the ECCV 2016: Computer vision ECCV 2016*, vol. 9905, pp. 21–37, Springer, Amsterdam, The Netherlands, 2016.

[8] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *in Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, Venice, Ital, 2017.

[9] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6517–6525, IEEE, Honolulu, HI, USA, 2017.

[10] J. Redmon and A. Farhadi, *YOLOv3: an incremental improvement*, 2018, http://arxiv.org/abs/1804.02767.

[11] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, *Yolov4: Optimal speed and accuracy of object detection*, 2020, https://arxiv.org/abs/2004.10934.

[12] G. JOCHER, *Yolov5*, Code repository, 2020, https://github.com/ultralytics/yolov5,2020.

[13] W. Chen, Y. Qiao, and Y. Li, "Inception-SSD: an improved single shot detector for vehicle detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 1–7, 2020.

[14] Y. Tian, G. Yang, Z. Wang, E. Li, and Z. Liang, "Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense," *Journal of Sensors*, vol. 2019, 13 pages, 2019.

[15] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?," in *in European conference on computer vision*, pp. 443–457, Cham, 2016.

[16] Z. Zhong, L. Sun, and Q. Huo, "Improved localization accuracy by LocNet for faster R-CNN based text detection in natural scene images," *Pattern Recognition*, vol. 96, p. 106986, 2019.

[17] J. Zhang, S. Chen, S. Tian, W. N. Gong, and G. S. Cai, "A crowd counting framework combining with crowd location," *Journal of Advanced Transportation*, vol. 2021, 14 pages, 2021.

[18] B. J. Cheng, J. Zhang, and Y. Wang, "Research on medical knowledge graph for stroke," *Journal of Healthcare Engineering*, vol. 2021, 10 pages, 2021.

[19] A. Kelm, L. Laußat, A. Meins-Becker 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.

[20] X. H. Liu and X. N. Ye, "Skin color detection and Hu moments in helmet recognition research," *Journal of East China University of Science and Technology (Nature Science Edition)*, vol. 40, no. 3, pp. 365–370, 2014.

[21] A. H. M. Rubaiyat, T. T. Toma, and M. Kalantari-Khandani, "Automatic detection of helmet uses for construction safety," in *in Proceedings of the 2016 IEEE ACM International Conference on Web Intelligence Workshops (WIW)*, ACM, Omaha, NE, USA, 2016.

[22] RRV E Silva, "Helmet detection on motorcyclists using image descriptors and classifiers," in *in 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 141–148, 2014.

[23] Q. R. Li, *A Research and Implementation of Safety-Helmet Video Detection System Based on Human Body Recognition*, University of Electronic Science and Technology of China, Chengdu, China, 2017.

[24] H. Wu and J. Zhao, "An intelligent vision-based approach for helmet identification for work safety," *Computers in Industry*, vol. 100, pp. 267–277, 2018.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *in Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[26] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2021.

[27] Q W, "ECA-Net: efficient channel attention for deep convolutional neural networks," in *in CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[28] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: convolutional block module," in *in Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[29] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, *Bam: Bottleneck Module*, 2018, https://arxiv.org/abs/1807.06514.