Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1319029, 14 pages
https://doi.org/10.1155/2022/1319029

Research Article

Lightweight Security Wear Detection Method Based on YOLOv5

Sitong Liu,[1] Nannan Zhang,[2] and Guo Yu[3]

[1]*Department of Software Engineering, Northeastern University, Shenyang 110000, China*
[2]*Department of Physical Education, Northeastern University, Shenyang 110000, China*
[3]*Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200000, China*

Correspondence should be addressed to Guo Yu; guoyu@ecust.edu.cn

Received 14 March 2022; Accepted 28 April 2022; Published 13 May 2022

Academic Editor: Xingsi Xue

Given a large number of network parameters of the existing security wear detection, it is difficult to run on the embedded platform. Based on the idea of deep separable convolution, a lightweight security wearable target detection network based on improved YOLOv5 is proposed. Specifically, the feature extraction network structure of YOLOv5 is lightweight improved to reduce computation of the proposed model, including the increase of the number of network layers and decrease of the number of parameters. In addition, an attention mechanism is introduced to weigh different channels of the feature map to improve detection accuracy. The model has been tested on PASCAL VOC dataset and security wear dataset. The experimental results show that the size of the proposed model is 8.0 MB, the number of parameters is 7.5*105, and the number of FLOPs is 7.5*105. Compared with the YOLOv5 model, the required memory is reduced by 44.8%, and the number of parameters decreased by 45.58%, FLOPs decreased by 54.54%. Accordingly, the results have demonstrated that the proposed method can significantly improve the detection speed while maintain the accuracy. Especially, we have successfully deployed the proposed model in the high-speed detection of security wear.

1. Introduction

Target detection is a classical task in the field of computer vision, such as scene content analysis and understanding [1]. In recent years, more and more artificial intelligence research has been applied to various intelligent systems to obtain greater benefits [2]. For example, radio frequency identification networks are widely used in the applications of the Internet of things (IOT) [3], biomedical domains [4], including retail production monitoring, supply chain management, localization, and navigation. However, for engineering industries such as steel mills, the production site is relatively fixed. Most of the work and business of construction enterprises take place on the construction site, but there are always many problems in the management and supervision of the construction site, such as difficult safety management, difficult monitoring of accidents that hide dangers, and difficult monitoring of work progress on the construction site. Generally, the larger the scale and the more complex the process of the project, the greater the requirements for the supervision and management of the construction site, and the greater the difficulty of management. With the development of science and technology, intelligent factories have become an important part of intelligent city construction and a typical application of AI, IOT, and other technologies in traditional engineering industries [5]. Many companies promote this technology and products on social networks, which can help users realize more intelligent and standardized management of personnel, mechanical equipment, materials, and materials on the construction site [6].

An intelligent factory changes passive monitoring into active monitoring in the form of visualization and data. As a result, the multilevel and all-around real-time monitoring

of project progress and standardized civilized construction will be carried out. This way can provide prewarning of the abnormal accidents and quickly analyze the accident, which will keep the factory safe and improve the efficiency of the management of the factory. As a part of smart construction sites, security wear detection is becoming the standard configuration of AI+ smart construction sites. In areas prone to falling objects, such as steel plants, power plants, construction sites, coal mines, and smelting, operators must wear safety helmets and protective clothes. Due to the lack of standard wearing of protective clothing and safety helmets, accidents such as falling object impacts, collisions, scalds, and smashes will cause great harm to the lives and safety of workers, while safety accidents on the construction site due to the lack of safety helmets and protective clothing occur frequently. At the beginning of the 20th century, safety helmets were gradually used on construction sites, reducing the annual accident death toll to less than 2.5%, which proved that wearing safety helmets correctly can effectively reduce risks [7]. However, for a long time, people in construction areas in China generally have had a weak awareness of self-safety protection [8] and often fail to realize the importance of wearing safety helmets. According to the data survey, among the 42 major accidents in the construction industry in 2016, the casualties caused by collapse accounted for 84% of the total casualties, a large part of which was due to the failure to wear protective measures such as safety helmets and protective clothing. Helmets and protective clothing are important protective tools for factory workers at the construction site, but many workers choose not to wear them because of the lack of comfort in helmets and cumbersome wearing of protective clothing, which will endanger the lives and safety of workers [9]. Therefore, it is important to check whether factory workers are wearing safety protection equipment correctly in real time. However, the working environment in some factories is very dangerous, so it is not suitable to use the manual area for real-time inspection and management. Therefore, consider using machines instead of manpower for security wear detection. These protective tools can prevent safety accidents to a certain extent and ensure the physical safety of factory workers. With the development of powerful computing power ES (Edge Storage) (ES means the edge storage, which is a tool to directly store data during the data collection.) [10] and computer vision, unmanned intelligent security detection methods have attracted people's attention because of their advantages of low detection cost and high efficiency [11]. Security wear detection has been studied by many scholars at home and abroad because of its complex working environment, shooting angle, distance from the target, and so on. Most researchers use the helmet as the primary research object. There are many types of research on helmet-wearing detection at home and abroad, mainly including the following two ideas.

The first method is through machine learning or deep learning algorithms. During the traditional helmet-wearing detection, the position of workers, pedestrians, faces in the image, or the position of the image foreground information is firstly extracted. Then, the extracted information is used to infer whether the target exists in the approximate area of the image. Finally, the circular Hough transform or SVM is applied to judge whether there is a helmet in this area. The traditional helmet-wearing detection algorithm mainly recognizes color and shape features. As discussed by Li [12], it was proposed to study how to locate the head area and calculate the color characteristics of the helmet to detect it. Liu and Ye[13] proposed using skin color to locate the face and intercept the area above the face, then taking the extracted Hu moments (Hu moment is a two-dimensional moment invariant theory [13] with a good invariance and anti-interference on the rotation and scaling changes of targets in an image, which is commonly used to effectively reflect the essential characteristics of the image [14].) [14] as the feature, and finally using the Hu moment feature extracted by SVM training to obtain the classifier that can detect the helmet. As discussed by Park et al. [15], they used the hog feature in the pedestrian detection stage, the color histogram in the helmet detection stage, and the spatial matching relationship between the human body and the helmet to judge whether the personnel were wearing the helmet. Feng et al. [8] proposed detecting the foreground with a Gaussian mixture model, then dealing with the connected domain, detecting the human body with a model-based method, and detecting the helmet with a SIFT feature and color statistical feature. Rubaiyat et al. [16] proposed combining the hog feature and the frequency domain information of the image to detect the workers, and then using the circular Hough transform and color information to judge whether a person is wearing a hat. Hence, the learning mechanisms of the individuals played a significant role in the algorithm's performance [17].

In a second way, the object detection algorithm based on a deep convolution network is used to train on the dataset to establish the model and to detect and identify the security wear of construction personnel. Huang and Pan [18] proposed using a parallel network to detect the human body on LeNet and then detect helmets through color features. Fang et al. [19] realized the detection of personnel's helmet-wearing in surveillance video by improving Faster R-CNN. As discussed by Zhang et al. [20], who proposed a helmet-wearing detection algorithm combining OpenPose and Faster-RCNN. First, OpenPose is used to detect the head and neck of people, and then, Faster-RCNN is used to detect the helmet in the image. Finally, the spatial relationship between the head, neck, and the helmet is analyzed to judge whether people wear the helmet. Zhang and Xu[21] improved SSD, used VGG as the backbone, chose the Adam optimizer to accelerate convolutional neural network convergence, and improved helmet-wearing detection accuracy by using characteristic diagrams of different scales. Fang et al. [22] improved its network structure based on the YOLO v2 target detection algorithm. By adding dense blocks to the original YOLO v2, the sensitivity of the network to small target detection is improved. Then, the deep separable convolution is used to compress the network, which increases the availability of the model. As discussed by Zhang et al. [23], they realized the detection of workers' helmet-wearing in the monitoring video through Faster-

RCNN. Wang [24] obtained the construction worker identification network and helmet identification network by improving the YOLO network structure, trained the network to obtain the generalization model, and then conducted semisupervised online learning on the obtained model, to obtain the helmet-wearing detection algorithm with high accuracy.

Compared with traditional algorithms, deep learning image recognition can extract more accurate image features and has a stronger recognition ability. However, the image recognition algorithm of deep learning needs to train on a large number of training datasets to learn the model weight, so an important premise of deep learning is to label the dataset. To solve the problems of large consumption of human and material resources and easy inspection gaps in manual supervision in the special working environment of some factories, after selecting the data collected by a factory for labeling and adding some filtered data from the open-source dataset HWD, the labeling data is added again to form a new security wear detection dataset. The target detection algorithm YOLOv5 is used to train and learn on the security wear detection dataset, and the security wear detection baseline model is established. In recent years, deep learning has become one of the most popular research methods for target detection because of its high accuracy and strong robustness. As discussed by Li et al. [11], they use the SSD [25] model to detect whether the helmet is worn. At present, the target detection algorithm based on deep learning mostly lays anchor frames of different sizes on the image and realizes target detection through regression and classification anchor frames. According to the generation method of the regression box, it is mainly divided into two stages and a single stage. The two-stage detector, such as Faster-RCNN [26], has high accuracy, but the detection speed is slow. Single-stage detectors, such as the YOLO deep learning algorithm, are particularly attractive because of their good recognition performance. Since the YOLO v1 [27] model in the field of target recognition was proposed by Redmon in 2016, the YOLO series has been constantly innovating. The YOLO v2 [28] model, YOLO v3 [29] model, YOLO v4 [30] model, and YOLO v5 are the new versions of the YOLO series. The innovative products are continuously integrated based on the YOLO series. In many current application scenarios, a problem usually involves multiple conflicting targets and is usually subject to a given set of constraints [31]. This paper selects the detection speed as the evaluation. Among them, the YOLO v5 model has the best performance and is suitable for practical engineering applications. The official YOLO v5 target detection network has given four network models: YOLO v5s, YOLO v5m, YOLO v5l, and YOLO v5x. The YOLO v5s network model is a YOLO network with the smallest depth and feature map width among the four sizes. Therefore, this paper uses the YOLO v5s detection network as the benchmark for security wear detection.

2. Related Work

2.1. The YOLO v5method. In 2017, the YOLO algorithm was optimized by Redmon and Farhadi [29]. Specifically, the convolution layer is used to replace the full connection layer on the basis of the idea of an anchor box in Faster-RCNN [26]. In addition, a BN (Batch Normalization) [32] layer is added to the convolution layer to further improve the detection effect, such as accuracy and speed. The architecture of YOLO v5 is shown in Figure 1. The YOLO v5 network consists of three parts: the backbone network, the neck, and the output. The main part focuses on extracting the feature information image from the input, fusing the extracted feature information to generate three scale feature maps, and the output part detects the object from these generated feature maps.

In the process of target detection and processing, the YOLO v5 algorithm adds a mosaic data enhancement function in this part of the data input. The backbone mainly adopts the Focus structure, SPP structure, and BottleneckCSP structure. Add the FPN + PAN (path aggregation network) structure to the neck. In the new version of the YOLOv5 network, the author transforms the bottleneck CSP (bottleneck layer) module into the C3 module. Its structure and function are the same as that of CSP, which includes three standard convolution layers and multiple bottleneck modules.

The difference between the C3 and CSP modules is that the Conv module after residual output is removed, and the activation function in the standard convolution module after the Concat module is also changed from LeakyRelu to SiLU. This module is the main module for learning the residual characteristics. Its structure is divided into two branches. One uses multiple bottleneck stacks and three standard convolution layers, while the other passes through only one basic convolution module, finally concatenating the two branches. In the YOLOv5s network model, the size of the image input is $3 \times 640 \times 640$, and the characteristic image with a size of $12 \times 304 \times 304$ is converted through one focus slice operation, and then, it is converted into the characteristic image with a size of $32 \times 304 \times 304$ through the ordinary convolution operation of 32 convolution cores.

2.2. Mosaic Data Enhancement. Mosaic [33] data augmentation is an advanced data augmentation method. The basic strategy of Mosaic data augmentation is to stitch together four security wear images and then perform data augmentation operations such as random scaling, random cropping, and random placement.

The advantages of mosaic data enhancement are as follows: there is no need to increase the size of the minibatch because Mosaic data augmentation can enrich the background and target of the security wear inspection object when calculating batch normalization. The GPU can calculate the data of four security wear images at a time, so that it can obtain a better detection effect. The security wear dataset used for detection can be significantly increased, making the network more robust. The function of the GPU can be simplified, and the performance can be greatly improved. An example of a security wear image enhanced by Mosaic data is shown in Figure 2. The images enhanced by Mosaic data are beneficial for better fitting the images in the training set during the training process. Mosaic data enhancement

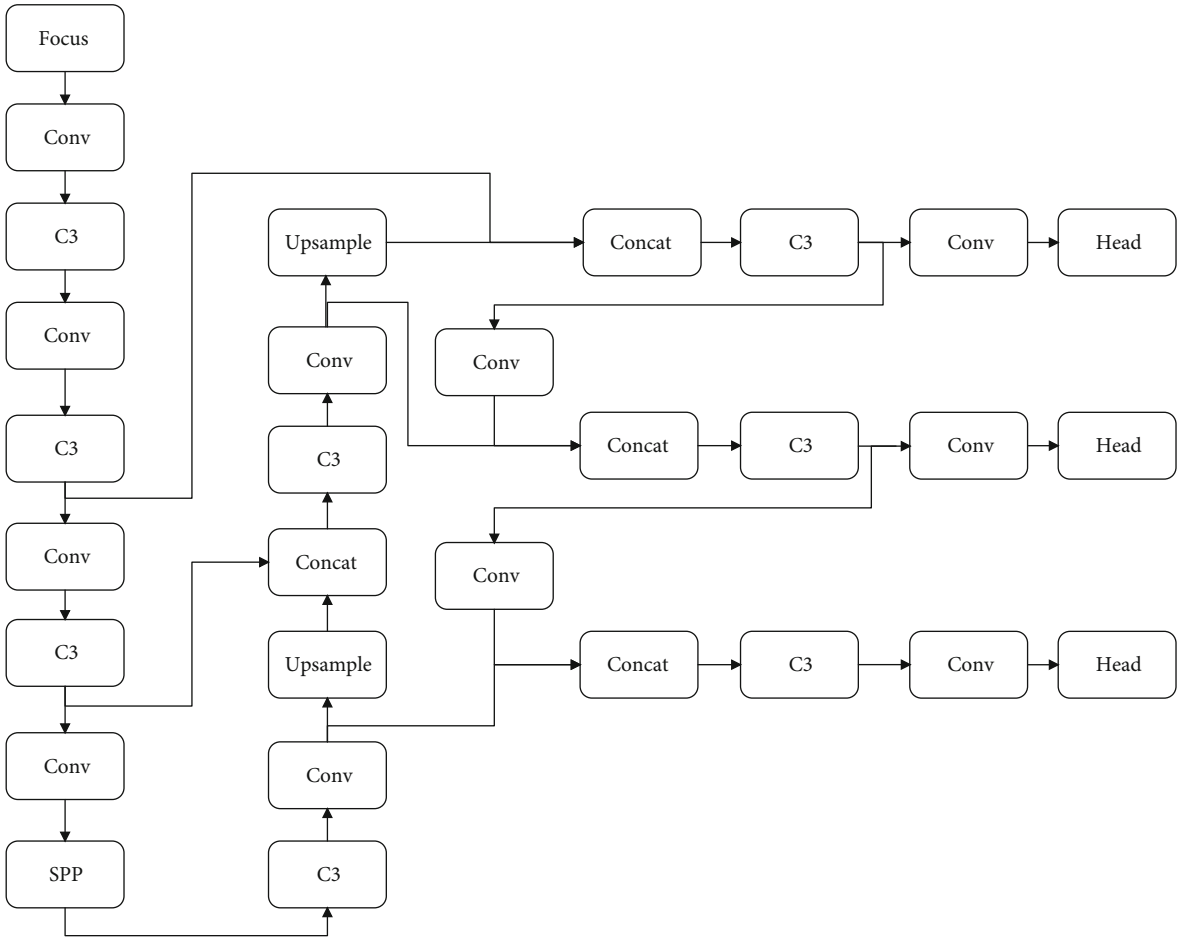

Figure 1: The architecture of the YOLOv5 method.

strategy is a training strategy that can improve the performance of the model, but it only needs to consume a little cost.

2.3. Loss Function. The calculation of loss in the YOLO series is based on objective score, class probability score, and bounding box region score. In this prediction, the loss function of the boundary anchor box is improved from CIOU (complete IOU) loss to a generalized IOU loss. The weighted NMS (Nonmaximum Suppression) operation is used to filter multiple target anchor frames. YOLO V5 uses GIOU loss as the loss of bounding box. The code uses nn. BCEWithLogitsLoss or FocalLoss evaluates the class loss and confidence loss of the target frame and prediction frame. The change of a parameter will have a great impact on the loss function [34]. The formula for BCELoss is as follows.

$$\text{BCELoss} = -\frac{1}{n}\sum(y_n \times \ln x_n + (1 - y_n) \times \ln(1 - x_n)). \quad (1)$$

In Equation (1), y is the target and x is the output value of the model.

The YOLO v5 code uses the IOU index to evaluate the position loss of the target frame and prediction frame. The YOLO v5 code selects the prediction box corresponding to the real box with the aspect ratio, and each real box corresponds to three prediction boxes. The YOLO v5 code uses the IOU value to evaluate the position loss between the prediction frame and the real frame. This paper introduces the CIOU index. Equation (2) is as follows.

$$\text{GIOU}_{\text{Loss}} = 1 - \text{CIOU} = 1 - \left(\text{IOU} - \frac{\text{Distance}_2{}^2}{\text{Distance}_c{}^2} - \frac{v^2}{(1 - IOU) + v}\right). \quad (2)$$

In Equation (2), IOU is the call union ratio of the prediction frame and the real frame. v is a parameter to measure the consistency of the aspect ratio.

3. Lightweight YOLOv5 Model

3.1. Depth-Wise Separable Convolution. Deep separable convolution is one of the methods that can miniaturize the network model at present. Depth-wise separable convolution is to decompose the standard convolution into two steps. In the first step, a convolution check should be applied to a channel, and a channel is extracted by only one convolution kernel. In the next step, n 1 by 1 convolution kernels are

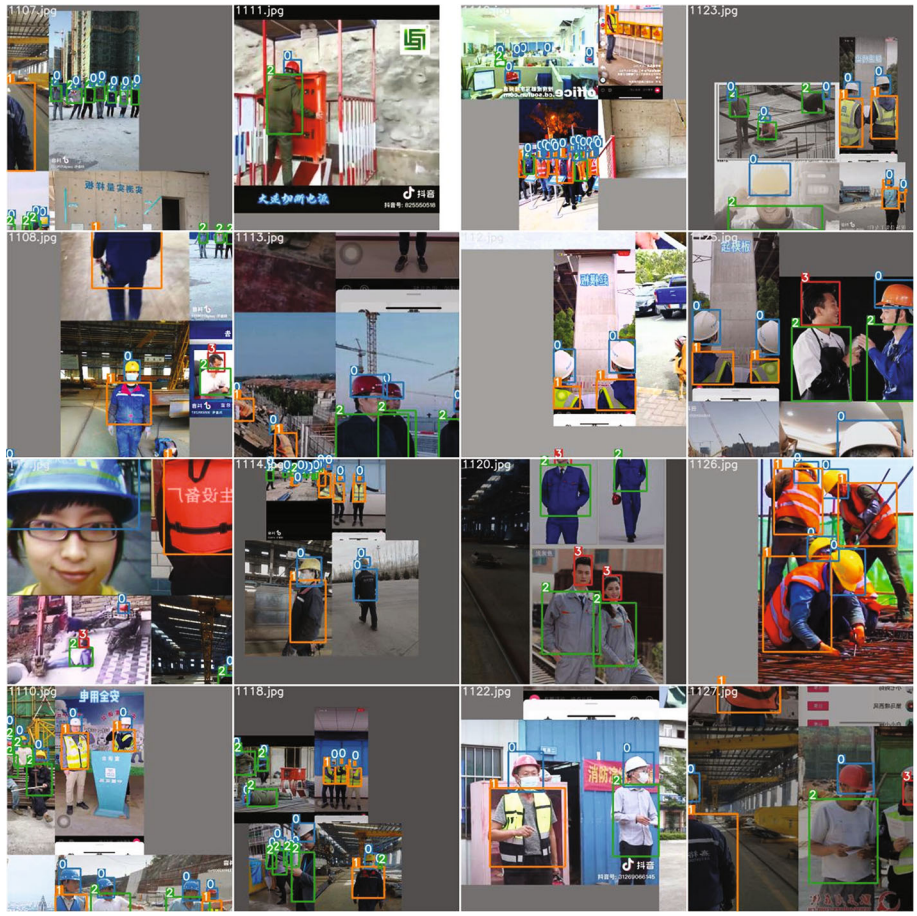

Figure 2: Mosaic data-enhanced image example.

used to connect the feature mapping obtained from the previous step to maintain the integrity of features [35]. The structure of deep separable convolution can reduce output channels and realize cross-channel information integration at the same time, keeping the algorithm performance unchanged while reducing the number of parameters [36]. Figure 3 shows the standard convolution and depth-wise separable convolution.

The ratio of parameters of depth-wise separable convolution to traditional standard convolution is shown in Equation (3).

$$\frac{D_k \times D_k \times M \times 1 + 1 \times 1 \times M \times N}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{1}{D_k^2}. \quad (3)$$

In the formula, D_k is the size of the convolution kernel. M is the input channel. N is the output channel. The deep separation convolution layer is used to replace the standard convolution layer in the convolution network model, which shows the amount of calculation required to convolute the same image to obtain the same dimensional image features is greatly reduced (see Figure 3). The advantage of deep separable convolution over conventional convolution is that it can significantly reduce the number of parameters.

In the new version of the YOLOv5 backbone, the author uses four-slice operations in the upper structure of feature extraction to form the Focus layer. The structure diagram of the Focus Layer is shown in Figure 4. For the Focus layer, every four adjacent pixels in a square generate a feature map with four times the number of channels, which is similar to the downsampling of the upper layer four times and concatenating the results. The main function is to reduce the parameters and accelerate the model without reducing the feature extraction ability of the model. When the parameters are reduced, the model is accelerated. However, there is a prerequisite for this acceleration, which can only be reflected by the use of the GPU. For this processing method of cloud deployment, the GPU does not need to consider the occupation of cache. That is, the method of fetching and processing makes the Focus layer very friendly on GPU devices. However, for chips, especially those without GPU and NPU acceleration, frequent slice operations will only seriously occupy the cache and increase the burden of computing processing. At the same time, during chip deployment, the transformation of the Focus layer is extremely unfriendly to novices. Therefore, the Focus layer is removed in this algorithm to avoid multiple slice operations.

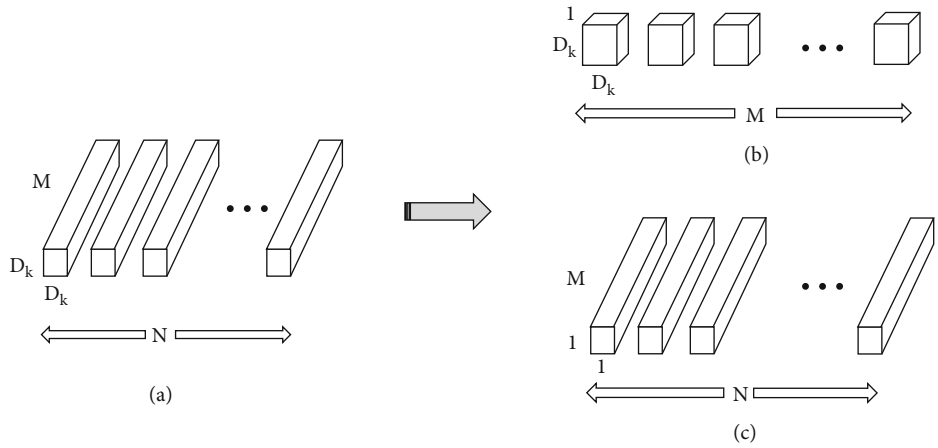

Figure 3: Standard convolution and depth-wise separable convolution.

C3 Layer is an improved version of CSPBottleneck proposed by YOLOv5 authors, which is simpler, faster, and lighter and achieves better results with nearly similar losses. However, the C3 Layer uses multiplexed convolution. Tests have shown that the frequent use of the C3 Layer and the C3 layer with a higher number of channels will occupy more cache space and reduce the running speed. Because of the G1 criterion of shufflenetv2 [37], the same channel size can minimize the amount of memory access. The higher the number of channels, the greater the step gap between hidden channels and c1 and c2. Imagine jumping down one step and ten steps. Although ten steps can be reached at one time, you need to run, adjust, and accumulate energy to jump up, which may take longer.

Therefore, lightweight YOLOv5 replaces the backbone of YOLOv5s with the DW network, modifies the model structure, increases the number of network layers, and improves the detection accuracy. The multiscale prediction structure of YOLOv5 is used to better detect different types of objects and improve detection accuracy. The improved network structure is shown in Figure 5. The reason for feature stitching is that the network can learn deep and shallow features at the same time, and the expression effect is better. Plots that show "CBR_BLOCK" are used to represent Conv2d + BN + LeakyRelu6. "DW_BLOCK" indicates deep separable convolution. The crosslink operation will be performed in "DW_BLOCK". If it is not the first layer, the cross-layer connection will be made. The output port of the network continues to use the output of YOLOv5s, and the $20 \times 20, 40 \times 40, 80 \times 80$ images of three different scales are predicted.

Using SPP [38] instead of SPP [39] can reduce flops, run faster, and realize local features as well as all other features. The SPPF network structure is shown in Figure 6. The fusion of local features and full moment features is conducive to the large difference in target size in the image to be detected in the security wear recognition image and can enrich the expression ability of the feature map. Especially for the complex multitarget detection in this paper, the detection accuracy can be greatly improved.

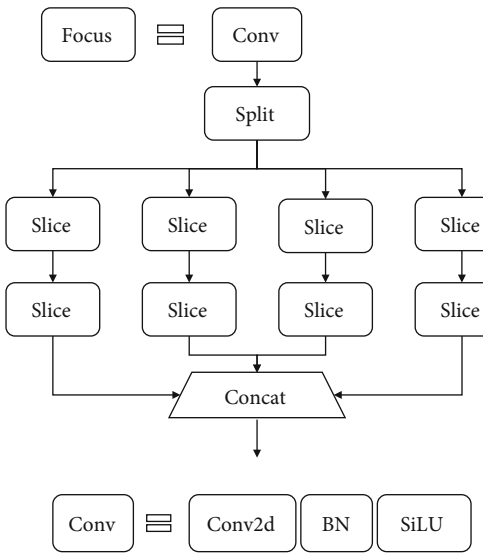

Figure 4: Focus structure diagram.

3.2. Attention Mechanism Module. The attention mechanism module [40] (convolution block attention module, CBAM) is a lightweight general module at present, focusing on channel dimension and spatial dimension and integrating two independent dimensions. It can be divided into two parts according to the spatial dimension and channel dimension. The first part is the channel attention module (CAM), and the second part is the spatial attention module (SAM). The attention mechanism module is a very simple module that can carry out end-to-end training with potential convolution at the same time to achieve good results. In two independent dimensions, the attention mechanism module can infer the attention map along different dimensions and carry out adaptive optimization based on the extraction of the feature map. In the process of attention inference, the lightweight 1×1 convolution is used to mark the position information ignored in the feature extraction process on the convolution

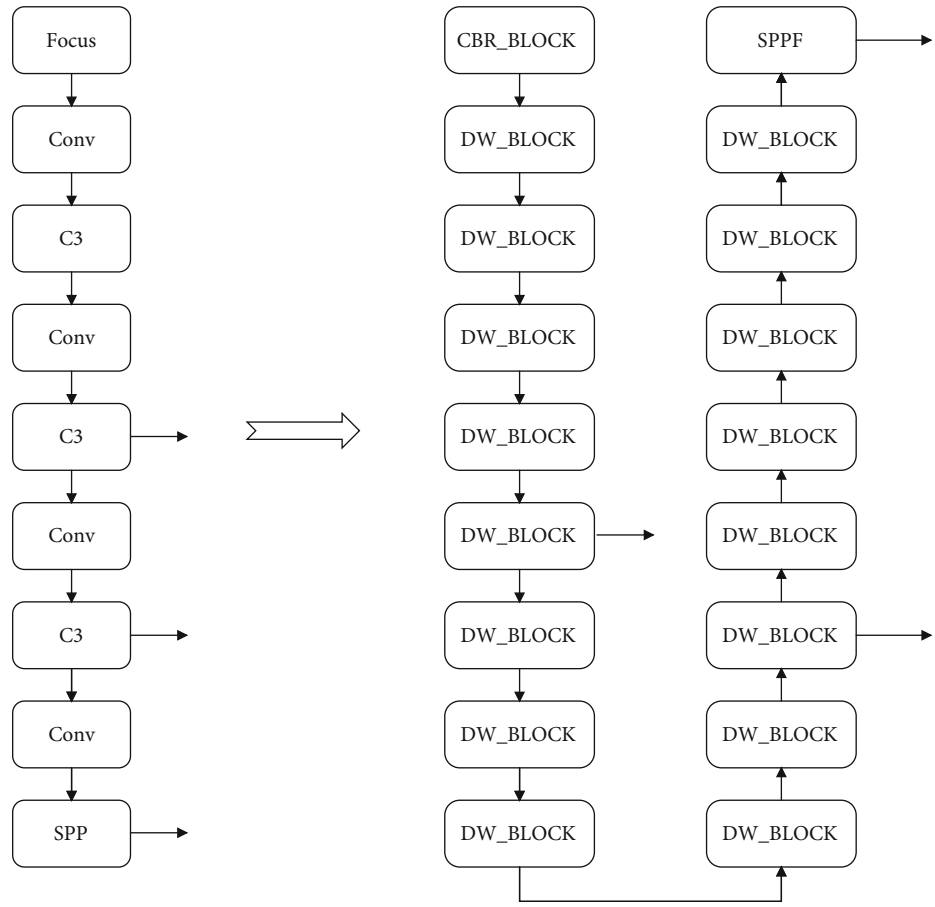

Figure 5: DW-YOLOv5s-BackBone.

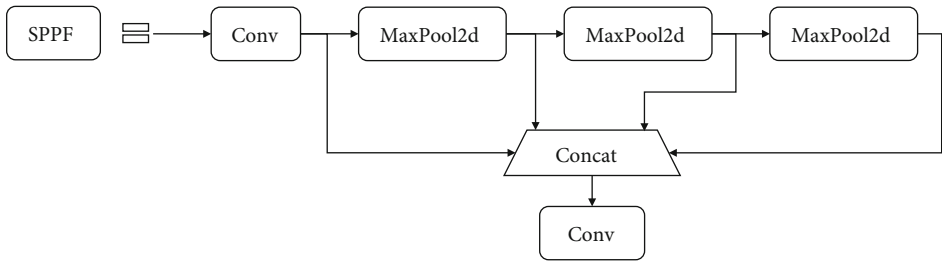

Figure 6: The SPPF network structure.

channel and strengthen the positioning ability of the model in the spatial dimension through the direction and feature information.

Firstly, this model uses the improved lightweight backbone network to obtain the coarse-grained target feature map, extracts its fine features in the feature fusion module, and embeds the attention mechanism CBAM to generate the attention map. Through the combination of attention map and coarse-grained features, it can enhance the feature information of security wear and realize the attention to the area of interest, that is, safety helmet and protective clothing, reducing the interference of irrelevant information on feature extraction [41].

Plots that show using DW-YOLOv5 as the basic network and adding a CBAM attention module in the neck of the network (between the backbone network and the detection layer) can better integrate the spatial features and channel features of small targets in the feature map, so as to enhance the feature information (see Figure 7). After model training, the test results are shown (see Table 1). It can be seen that after adding CBAM, the mAP of the model is increased from 80.5% to 82.1%.

After adding the attention mechanism module, there are some small gaps in training time, memory size, and detection speed of the model, but the mAP of the DW-YOLOv5-attention model is improved by nearly 1.6%.

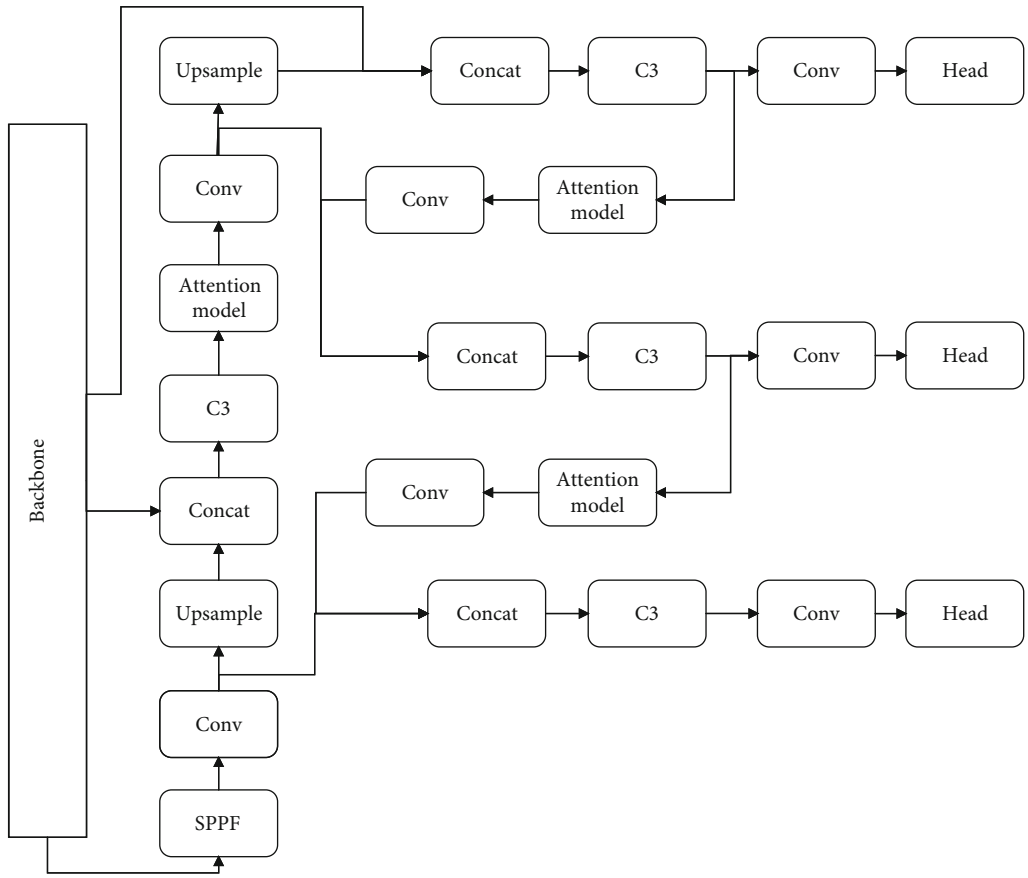

Figure 7: YOLOv5s- Attention-Neck.

Table 1: Comparison of results of multiple detection models on SWD dataset.

Model	Precision	Recall	mAP	FPS
Faster-RCNN	0.559	0.700	0.681	18.03
MobileNet-SSD	0.857	0.258	0.464	60.11
YOLO v5	0.877	0.809	0.849	83.33
DW-YOLO	0.869	0.760	0.805	90.90
DW-YOLO-attention	0.892	0.752	0.821	100.00

Experiments show that attention is applied to the model and improves the classification and detection performance of the model.

The security wear detection steps are shown in Figure 8. Using YOLO Auto Learning Bounding Box Anchors, the prior frames of these 9 scales are set in the DW-YOLO-attention module for training. Input the mosaic-enhanced image and adjust the image size, then enter the designed DW-YOLO attention module, and finally output the detected security wearing an image.

4. The Datasets and Evaluation Index

In the previous research on target detection, most of the public dataset PASCAL VOC dataset was used for experiments, but the object size in this dataset is mostly 300 to 500, which is quite different from the detection object size in this paper. If only the PASCAL VOC public dataset is used for experiments, it may have a certain impact on the practical application of this algorithm. Therefore, in this experiment, the PASCAL VOC 2007 public dataset and the self-made security wear dataset (SWD) are used for training and testing. The specific dataset is described as follows.

4.1. Pascal VOC. Select the training set of PASCAL VOC 2007 and VOC 2012 datasets as the training set and test set. This dataset is the benchmark dataset for evaluating image classification and target detection, including 20 types of label objects. The Train and Val datasets of VOC 2007 and 2012 are used for the training pictures, including 40025 objects of 16551 pictures, and the test dataset of the PASCAL VOC 2007 is used for the test pictures, a total of 12032 objects containing 4952 pictures.

4.2. Security Wear Dataset (SWD). Firstly, the monitoring video data of a factory's monitoring room for a week is collected. Considering the influence of the environment, the image information including different conditions, such as day, night, and inside and outside the factory, was selected in the construction process of the dataset. The selected scenes can ensure that the data has rich background information and that the model trained on this basis has good environmental generalization ability. By using openCV2 to

Figure 8: Security wear detection steps.

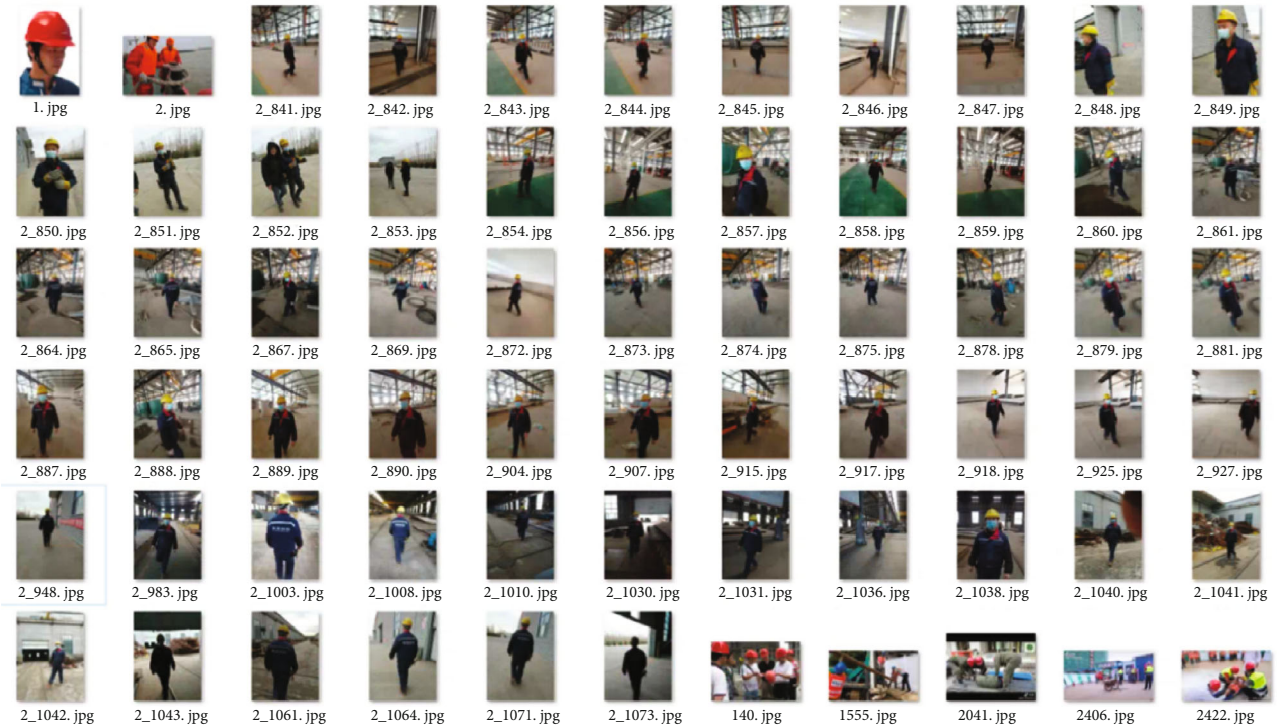

Figure 9: The sample of security wear datasets (SWD).

convert the video into image frames, the captured video is divided into several pictures. Considering the target size, illumination, and other factors, 3600 pictures, including helmets or heads, wearing protective clothing, and ordinary clothes, are selected. For the identification and detection of security wear, the main detection is whether the workers are wearing safety helmets and protective clothing. When the helmet is not worn, the detection category is the head. If you do not wear protective clothing, identify the category in which you are wearing ordinary clothes. According to the theory of literature [42], the sample data containing small targets can be added by replication. 110 images containing smaller targets were cut and copied from the training set. Select the open-source helmet dataset SHWD [43] (site scene helmet dataset), comprehensively consider screening 1250 pictures and adding them to the training set. After retagging with the open-source tagging tool, a dataset of 4950 security wear datasets is formed, which is made into the required YOLO dataset format. Some sample examples of security wear datasets are shown in Figure 9.

Table 2: Details of security wear datasets (SWD).

Label	Total
Safety_hat	8561
Reflective_cloth	4831
Other_cloth	2869
Head	926
Small($\mathbf{area} \le 32 \times 32$)	3970
Medium($32 \times 32 < \mathbf{area} \le 96 \times 96$)	7866
Large($\mathbf{area} > 96 \times 96$)	5414

The SWD datasets for personnel safety protection wear on special occasions are formed, with a total of 4950 pieces, which are divided into four categories: normal wearing a safety helmet, wearing protective clothing, wearing ordinary clothes, and head without a safety helmet. The dataset contains 3907 small targets (area $\le 32 \times 32$),

Table 3: Details of the three datasets.

Dataset	Number of pictures	Object type	Function
PASCAL VOC 2007 +12	16551	20	Training/ testing
Security wear dataset (SWD)	4950	4	Training/ testing

7866 medium targets ($32 \times 32 < \text{area} < 96 \times 96$), and 5414 large targets ($\text{area} > 96 \times 96$). The training set and verification sets are randomly allocated in the proportion of 7: 3. The number of instances include in each category is shown in Table 2.

The detailed information of VOC dataset and SWD data involved in this experiment (see Table 3).

4.3. Evaluation Indexes. In this paper, Precision and Recall [44] were selected as the evaluation indexes of this experiment.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

In Equations (4) and (5), TP is the probability that positive examples can be divided into pairs. FP is the probability of misclassifying negative cases into positive cases.

mAP is used to measure the accuracy of the detection model of personnel safety protection wear on special occasions. The mAP refers to the average value of the accuracy rate of all categories, the average value of AP of each category. For the test results of each specific category, sort according to the confidence level, and select each recall rate r_i and its corresponding maximum precision P_i in the recall rate set $\{r_0, r_1, \cdots, r_n\}$. The definition of AP is shown in Equation (6):

$$\text{AP} = \sum_{i=1}^{n} P_i(r_i - r_{i-1}), \tag{6}$$

$$m\text{AP} = \frac{\sum_{i=1}^{m} AP_i}{m}. \tag{7}$$

In Equation (7), the numerator is the average accuracy of each category, and m is the total number of categories of image detection.

FPS (frames per second) is used to measure the running efficiency of the model, that is, the number of pictures processed per second. Size measures the storage space occupied by the model, and FLOPs (floating-point operations) measures the complexity of the model.

5. Experimental Results and Discussion

5.1. Experiment Operating Environment. The experimental environment of this experiment is mainly carried out under the computer configuration of the Windows 10 operating system, Intel Core i7-8700k, 3.70 GHz, and 16 G RAM. The GPU adopts NVDIA RTX 2080 and 16G video memory. The experimental conclusions of this model and its comparative model are drawn in this experimental configuration environment.

Table 4: Settings of hyperparameters, where Input size means image input size, Lr initial learning rate, Lr_f cosine annealing hyperparameter, Weight_decay Weight attenuation coefficient, Warmup_epochs Warm up learning epoch, Warmup_momentum Warm up learning rate momentum, and Warmup_bias_lr preheating learning rate.

Hyperparameters	Default
Input size	640
Lr	0.01
Lr_f	0.2
Momentum	0.937
Weight_decay	0.0005
Warmup_epochs	3.0
Warmup_momentum	0.8
Warmup_bias_lr	0.1

Table 5: Object detection model performance comparison.

Model	FLOPs	Parameters	MB
YOLOv5s	16.5	7114785	14.5
DW-YOLO	7.4	3820257	7.9
DW-YOLO-Attention	7.5	3871751	8.0

Table 6: Comparison of different algorithms on PASCAL VOC dataset.

Algorithms	Input size	mAP/%
MobileNet-SSD	300×300	68.0
faster-RCNN	1000×600	70.0
YOLOv5s	640×640	81.6
DW-YOLO	640×640	77.8
DW-YOLO-attention	640×640	77.5

5.2. Hyperparameter Setting. The super parameter setting of the lightweight YOLO v5model is given. The training algebra of this experiment is 300 generations and the batch size is 18, the input size is 640. The initial momentum and initial learning rate (lr) are set to 0.937 and 0.01, respectively. Before the formal training in this paper, three generations of preheating learning are carried out, in which the preheating learning momentum is 0.8 and the Warmup lr is 0.1. The purpose is to make the model gradually stabilize after preheating learning and then carry out formal training. The effect of security wear recognition is better. The other superparameter settings are shown in Table 4.

5.3. Experimental Results and Analysis. Firstly, it analyzes the size of the model and compares it with the original YOLOv5 model. There are three measurement indexes

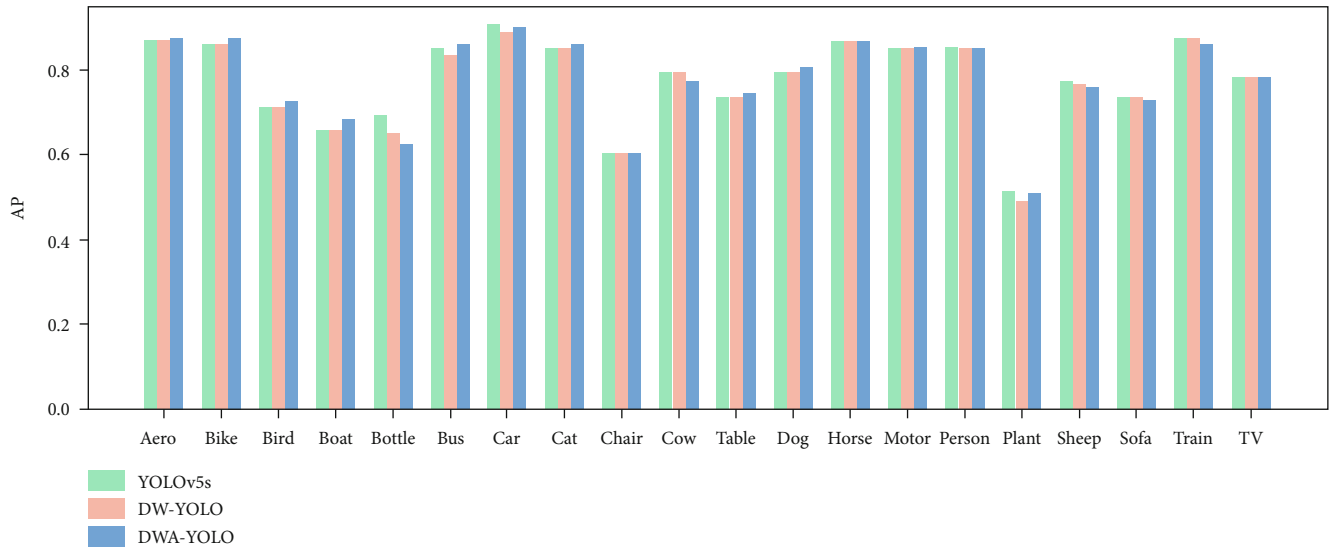

Figure 10: Comparison of different types of AP indexes in VOC datasets.

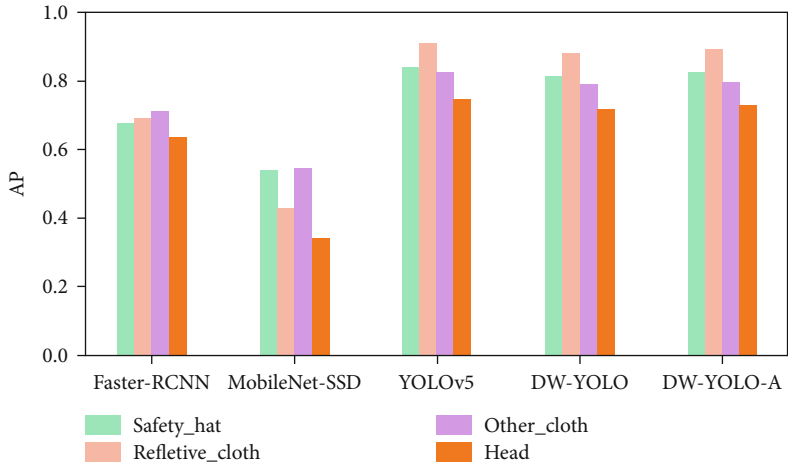

Figure 11: Comparison of different types of AP indexes in SWD datasets.

for the improved models (see Table 5). The memory required for lightweight DW-YOLO-Attention is 8.0 MB, the parameters are 7.5×10^5, and the FLOPs are 7.5. Compared with the YOLOv5 model, the memory required is reduced by 44.8%, respectively. The parameters were reduced by 45.58%. FLOPs are reduced by 54.54%, which reduces the complexity of the model and shows the superior performance of the lightweight DW-YOLO-Attention model in the field of security wear.

To fully evaluate the lightweight DW-YOLO-Attention detection algorithm in this paper, three commonly used detection algorithms are used for comparative experiments, and the mAP is selected as the evaluation index of different detection algorithms. The open datasets for model training are PASCAL VOC 2007+2012, and the test is PASCAL VOC 2007. The results are shown in Table 6.

Table 6 shows that compared with the advanced model algorithm, the mAP of the lightweight DW-YOLO model algorithm is good. The detection mAP can reach 77.8%, only 3.8% points lower than the baseline, 7.8% higher than the Faster-RCNN, and 9.8% points higher than the MobileNet-SSD. By improving the feature extraction backbone of DW-YOLOv5, the number of layers of the improved lightweight backbone network is deepened, and the extracted security wear image features become more detailed. The deep separation convolution ensures that the detection accuracy of the model is improved, but the number of model parameters is not increased, so a better detection effect is achieved.

Then, it compares the AP of DW-YOLOv5s and DW-YOLO-Attention in each category of the VOC dataset (see Figure 10). Plots that show DW-YOLO have high detection accuracy for cat, dog, table, bus, bird, etc., which shows that this model has a good recognition effect for objects of different sizes, especially small and medium-sized objects. It applies to the security wear detection in this paper.

In this paper, the improved lightweight YOLOv5 algorithm is applied to the identification of safety wear protection. To verify the better effect of the method proposed in this paper, under the same equipment and network parameter configuration, the same number of test sets is used, and several popular one-stage and two-stage target detection networks are used for the experiments: Faster-RCNN and MobileNet-SSD. The experimental results were evaluated by four evaluation indexes: Precision, Recall, mAP, and FPS. The experiment is shown in Table 1.

The table shows the results of different models in security wear recognition. It can be seen that on the SWD dataset, the mAP of the lightweight DW-YOLO-Attention proposed can reach 82.1%. Compared with Faster-RCNN and MobileNet-SSD increases by 20.55% and 76.93%, and decreases by about 3.52% compared with YOLO v5. The FPS of the lightweight DW-YOLO-Attention model is 100. Compared with Faster-RCNN and MobileNet-SSD, the FPS of DW-YOLO-attention is increased by 81.97 and 39.89. Therefore, the lightweight DW-YOLO improves the real-time performance of security wear recognition while maintaining a high mAP of 82.1%, which has a certain practical application value for security wear recognition. Compared with other algorithms, the lightweight DW-YOLO-Attention model in this paper has certain generalization and robustness.

Then, we compared the AP of Faster-RCNN, MobileNet-SSD, YOLOv5, DW-YOLO, and DW-YOLO-Attention in the security wear dataset (see Figure 11). The recognition performance of these five target models is not satisfied in the case of occlusion in the factory scene. It may even be misjudged due to factors such as ambient light. The accuracy of the above methods decreases when they detect small targets. By contrast, both the lightweight DW-YOLO and DW-YOLO-Attention are able to detect small objects in the images on the same test set. Both of them have higher detection accuracy on objects with larger targets and more obvious features like the protective clothing (Reflective_cloth) but have relatively lower accuracy on smaller targets with complex features like the head. The reason lies in the fact that DW-YOLO-Attention which updates YOLO with model lightweight is able to improve running speed and reduce the model training time. The above experimental results have demonstrated that the proposed DW-YOLO-Attention algorithm has competitive performance in reducing the missed detection rate of small and medium-sized targets. Notably, the accuracy and detection speed of the method are not degraded when DW-YOLO-Attention is used to detect small and medium-sized targets. Consequently, the proposed method can be applied in practical scenarios, and we have successfully utilized it in the factory for object detection.

6. Conclusions

In existing security wear detection, the model transplantation is challenging to perform on the embedded platform since the number of network parameters is huge, and the low computation power of the platform also triggers the degradation of the detection accuracy. Accordingly, we have modified the YOLOv5 network structure to handle the safety wear detection with three scales. In other words, a lightweight target detection network based on deep separable convolution is proposed. In our model, we replaced ordinary convolution with depth-separable convolution in order to reduce the number and scale of parameters. The attention part is used to weigh the different channels of feature mapping to improve the detection accuracy of the model. The experimental results have shown that the proposed model has higher accuracy in target detection and lower model calculations than the existing models. Especially, the proposed model has the fastest reasoning speed among all models. The line of our future research is to improve the ability of the proposed model to detect objects under employee occlusion and extreme shooting angles.

Data Availability

The code of the proposed algorithm and corresponding experimental data are provided. For interested readers, please visit https://github.com/lssttt/projectnewone.

Conflicts of Interest

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

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities under the grant no. N2117005, the Joint Funds of the Natural Science Foundation of Liaoning Province under grant no. 2021-KF-11-01, the Fundamental Research Funds for the Central Universities, the National Natural Science Foundation of China under the grant no. 62103150, and the project funded by China Postdoctoral Science Foundation under the grant no. 2021 M691012.

References

[1] T. Li, S. Xu, and Z. Yao, "Adaptive dim and weak target detection method based on DSP," *Computer Applications and Software*, vol. 35, no. 1, pp. 243–245, 317, 2018.

[2] X. Xue and J. Zhang, "Matching large-scale biomedical ontologies with central concept based partitioning algorithm and adaptive compact evolutionary algorithm," *Applied Soft Computing*, vol. 106, article 107343, 2021.

[3] L. Ma, X. Wang, M. Huang, Z. Lin, L. Tian, and H. Chen, "Two-level master-slave RFID networks planning via hybrid multiobjective artificial bee colony optimizer," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 5, pp. 861–880, 2019.

[4] X. Xue, J. Lu, and J. Chen, "Using NSGA-III for optimising biomedical ontology alignment," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 3, pp. 135–141, 2019.

[5] Q. He, X. Wang, Z. Lei, M. Huang, Y. Cai, and L. Ma, "TIFIM: a two-stage iterative framework for influence maximization in

social networks," *Applied Mathematics and Computation*, vol. 354, pp. 338–352, 2019.

[6] X. Xue, J. Chen, and X. Yao, "Efficient user involvement in semiautomatic ontology matching," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 2, pp. 214–224, 2018.

[7] R. Usukhbayar and J. Choi, "Critical safety factors influencing on the safety performance of construction projects in Mongolia," *Journal of Asian Architecture and Building Engineering*, vol. 19, no. 6, pp. 600–612, 2020.

[8] G. Feng, Y. Chen, and N. Chen, "Research on automatic recognition technology of safety helmet based on machine vision," *Mechanical Design and Manufacturing Engineering*, vol. 44, no. 10, pp. 39–42, 2015.

[9] J. Liu, S. Hou, and K. Zhang, "Real time vehicle detection and tracking based on enhanced tiny-YOLOv3 algorithm," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 35, no. 8, pp. 118–125, 2019.

[10] L. Ma, X. Wang, X. Wang, L. Wang, Y. Shi, and M. Huang, "TCDA: truthful combinatorial double auctions for mobile edge computing in industrial Internet of Things," *IEEE Transactions on Mobile Computing*, vol. 3064314, p. 1, 2021.

[11] M. Li, Q. Han, and T. Zhang, "Safety helmet detection method of improved SSD," *Computer Engineering and Applications*, vol. 57, no. 8, pp. 192–197, 2021.

[12] Q. Li, *Research and implementation of helmet video detection system based on human body recognition*, Chengdu University of Electronic Science and Technology of China, 2017.

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

[14] G. Li, X. Zhang, and F. Qin, "Paper cut pattern recognition based on moment invariants and BP neural network," *Computer Engineering and Application*, vol. 46, no. 29, pp. 158–160, 2010.

[15] M. W. Park, N. Elsafty, and Z. Zhu, "Hardhat-wearing detection for enhancing on-site safety of construction workers," *Journal of Construction Engineering and Management*, vol. 141, no. 9, p. 4015024, 2015.

[16] A. Rubaiyat, T. Toma, M. Kalantari-Khandani et al., "Automatic detection of helmet uses for construction safety," in *IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*, pp. 135–142, Omaha, NE, USA, 2016.

[17] L. Ma, S. Cheng, and Y. Shi, "Enhancing learning efficiency of brain storm optimization via orthogonal learning design," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 11, pp. 6723–6742, 2021.

[18] Y. Huang and D. Pan, "Helmet recognition based on parallel two-way convolutional neural network," *Enterprise Technology Development.*, vol. 37, no. 3, pp. 24–27, 2018.

[19] Q. Fang, H. Li, X. 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.

[20] B. Zhang, Y. Song, and R. Xiong, "Helmet wearing detection integrating human joint points," *Chinese Journal of Safety Science*, vol. 30, no. 2, pp. 181–186, 2020.

[21] Y. Zhang and X. Xu, "Helmet wearing detection method based on improved SSD," *Electronic Measurement Technology*, vol. 43, no. 19, pp. 80–94, 2020.

[22] M. Fang, T. Sun, and Z. Shao, "Fast helmet-wearing-condition detection based on improved YOLOv2," *Optical Precision Engineering*, vol. 27, no. 5, pp. 1196–1205, 2019.

[23] M. Zhang, Z. Cao, and X. Zhao, "Research on helmet wearing recognition of construction workers based on deep learning," *Journal of Safety and Environment*, vol. 2, pp. 535–541, 2019.

[24] Q. Wang, *Research on safety helmet wearing recognition of workers in construction site based on video stream*, Huazhong University of Science and Technology, 2018.

[25] W. Liu, D. Anguelov, D. Erhan et al., "SSD: Single Shot Multibox Detector," *Proceedings of the European Conference on Computer Vision*, pp. , 201621–37, 2016.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, Nevada, USA, 2016.

[28] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, Honolulu, Hawaii, USA, 2017.

[29] J. Redmon and A. Farhadi, "YOLOV3: an incremental improvement," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, Salt Lake City, Utah, USA, 2018.

[30] A. Bochkovskiy, C. Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of Object Detection," 2020, http://arxiv.org/abs/2004.10934.

[31] L. Ma, N. Li, Y. Guo et al., "Learning to optimize: reference vector reinforcement learning adaption to constrained many-objective optimization of industrial copper burdening system," *Cybernetics*, pp. 1–14, 2021.

[32] S. Wu, G. Li, L. Deng et al., "L1-norm batch normalization for efficient training of deep neural networks," *The IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 2043–2051, 2019.

[33] T. Lee and L. Luo, "Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis," *Neuron*, vol. 22, no. 3, pp. 451–461, 1999.

[34] L. Ma, M. Huang, S. Yang, R. Wang, and X. Wang, "An adaptive localized decision variable analysis approach to large-scale multiobjective and many-objective optimization," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–13, 2021.

[35] R. Qi, R. Jia, Q. Mao, H. M. Sun, and L. Q. Zuo, "Face detection method based on cascaded convolutional networks," *IEEE Access*, vol. 7, pp. 110740–110748, 2019.

[36] J. Bai, P. Hao, and S. Chen, "Traffic scene understanding using lightweight convolution neural network image semantic segmentation," *Journal of Automobile Safety and Energy Saving*, vol. 9, no. 4, pp. 433–440, 2018.

[37] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet V2: practical guidelines for efficient CNN architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, Munich, Germany, 2018.

[38] GitHub, "YOLOV5-Master," 2021, https://github.com/ultralytics/YOLOv5.git//.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[40] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, 2018.

[41] X. Xue and C. Jiang, "Matching sensor ontologies with multi-context similarity measure and parallel compact differential evolution algorithm," *IEEE Sensors Journal*, vol. 21, no. 21, pp. 24570–24578, 2021.

[42] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, http://arxiv.org/abs/1902.07296.

[43] GitHub, "Safety-Helmet-Wearing-Dataset," 2021, https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset/.

[44] I. Melamed, R. Green, and J. Turian, "Precision and Recall of Machine Translation," *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pp. 61–63, 2003.