A Method of Wisdom Site Safety Helmet Detection Based on Deep Learning

Jiaxiang Guo[1], Huiyi Zhang[1], Tao Tao[1] and Rencai Jin[2]

[1] College of Computer, Anhui University of Technology, Maanshan, China
[2] China Technical Quality Department, China MCC17 Group Co., LTD, Maanshan, China

Abstract

Aiming at the real-time detection of helmet wearing in construction sites with small targets, incomplete features and many interference factors, the multi-scale and multi-branch feature extraction and feature reconstruction module are applied to YOLOv5s for model reconstruction, so that each level contains convolution networks with different sizes and depths, which can capture the details of different sizes of receptive fields in the scene. The feature reconstruction is used to extract finer grained features improve the robustness of the model. Experiments on self-made construction site data sets show that compared with the YOLOv5s model, the improved algorithm improves the recognition effect of helmet wearing in the detection of long-distance small pixels and the presence of a large number of occlusion, and the real-time performance is not affected, which can meet the application needs of smart construction sites.

Keywords

YOLOv5s; feature reconstruction; target detection; wisdom site

1. Introduction

With the help of image recognition technology, automatic detection and management of safety helmet are carried out, which is one of the main means of wisdom site construction. The construction site environment is complex and the detection targets will be blocked by various objects or mutual occlusion between people, so in the actual scene, the helmet detection will be a lot of interference. Due to the fixed position of the camera, the different recognition distance of the target will also increase the interference. The real-time automatic detection of the helmet is a small target when the remote detection is performed.

The helmet detection method has experienced the stages of radio frequency identification technology and image processing technology. Early mobile radio frequency identification technology[1] could not confirm whether the helmet was worn because the reader had a limited working range and could only detect whether the helmet was close to the worker. RibGaiya and Silva algorithm[2] combined the frequency domain information of the image with the histogram of orientation gradient to detect the human body, and then used the ring Hough transform algorithm to detect the helmet wearing, which solved the problem that it is difficult to distinguish the skin color of the human body and the helmet. However, it is easy to be interfered by many occlusions and light in the actual scene, which affects the detection accuracy. In reference[3], a hybrid descriptor consisting of local binary pattern, color histograms, and Hu moment invariants is proposed to extract the features of hard hats, and then hierarchical support vector machine is constructed to classify hard hats, which reduces the influence of environmental changes. But this method is not accurate enough to detect small objects such as safety helmets. Based on the improved YOLOv3 model method[4], the latitude clustering of the target box is used to optimize the selection of the target box, which improves the accuracy of the detection helmet, but the model is complex and the response speed is slow.

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China
770984322@qq.com (Jiaxiang Guo); hyzhang@ahut.edu.cn (Jiaxiang Guo); taotao@ahut.edu.cn (Tao Tao); 1195154491@qq.com (Rencai Jin)

YOLOv5 model has fast response speed and high detection accuracy, which is considered as one of the effective algorithms for real-time image recognition, and is widely used in unmanned driving, wisdom sites and other fields. However, YOLOv5 network model is also susceptible to noise in some multi-objective scenarios.

To solve this problem, based on YOLOv5s of YOLOv5 series, this paper applies a multi-scale feature extraction module in its model head to improve the extraction ability of features of different sizes. Furthermore, the feature reconstruction module is proposed to improve the ability of the model to extract fine-grained features and improve the robustness of the algorithm. Make it more suitable for wisdom site safety helmet inspection application.

2. Algorithm design

Figure 1 (a) shows the model block diagram after the feature extraction module and feature reconstruction module with additational scales are added to the predicted position of the head of YOLOv5s model. Figure 1 (b) shows the structural annotation of some module names in the model block diagram.

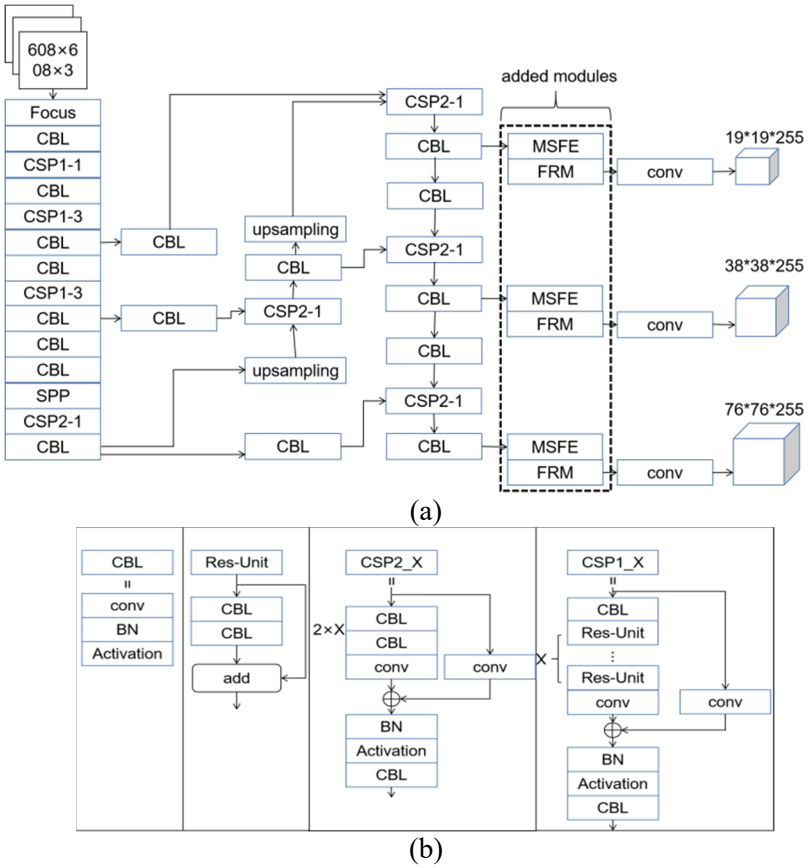


Figure 1. An improved YOLOv5s model block diagram with some module annotations

2.1.Multi-scale feature extraction method

Since people wearing safety helmets have different positions relative to the camera and different shooting angles, if the network model can have different receptive fields when extracting features, the recognition accuracy of the model for the target object will be effectively improved. The YOLOv5s model adopts ordinary convolution with a single type of kernel, and the calculation process is shown in figure 2. The input feature is x, and if the size of the convolution kernel in a single space is K_1^2, the depth is equal to the number of input feature maps FM_i. Applying a large number of FM_o cores with the same spatial resolution and depth to the input feature map FM_i can get a large number of output feature maps FM_o.

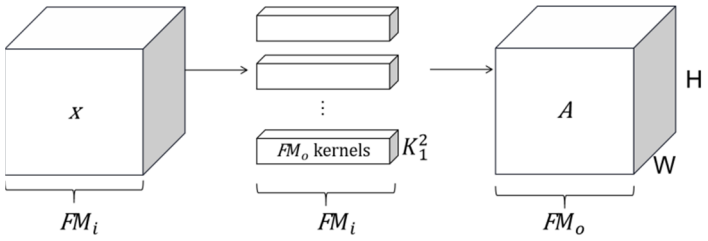


Figure 2. YOLOv5s model ordinary convolution calculation flow diagram

Figure 3 shows the schematic diagram of the feature multi-scale convolution calculation module(Hereinafter referred to as MSFE) adopted in this paper, which contains multiple layers of different types of convolution kernel to deal with the input feature[5]. The input features pass through multiple layers of convolution kernels $FM_{o1}, FM_{o2}, \cdots, FM_{on}$ with different sizes and types to obtain a large number of output features, and finally concatenated together. The convolution kernel in each layer of the multi-scale feature extraction module has different sizes, and the depth of the convolution kernel gradually decreases with the increase of the number of convolution kernel layers. Generally, the small convolution kernel has a smaller receptive field, so that local details can be obtained. The larger receptive field of convolution kernel can get the global semantic information of large target.

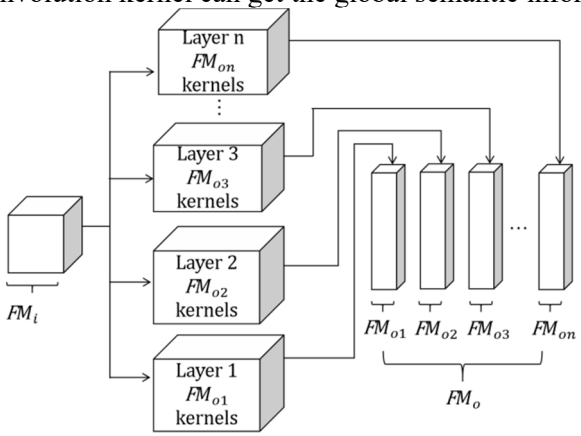


Figure 3. Flow diagram of multi-scale feature convolution calculation

In order to achieve multi-scale feature extraction without reducing the computational speed of the network model, the features from the neck part of YOLOv5s network were divided into different groups and independently calculated by convolution. As shown in figure 4, the number of channels of each group of feature maps in a module is related to the number of layers of the module. Although the number of channels in each group is different, the output dimension of convolution in different groups is the same, and then the output features are concatenated.

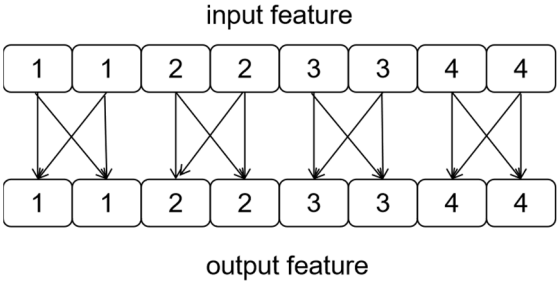


Figure 4. Schematic diagram of block convolution

Assuming that the input of the multi-scale extraction module contains C_i channels, the convolution kernel resolution of each layer is $K_1^2, K_2^2, \cdots, K_n^2$, and the depth is $C_i, \frac{C_i}{(\frac{K_2^2}{K_1^2})}, \cdots, \frac{C_i}{(\frac{K_n^2}{K_1^2})}$. The corresponding output feature dimension is $C_{o1}, C_{o2}, \cdots, C_{on}$. The dimensions of the final output feature is $C_o = C_{o1} + C_{o2} + \cdots + C_{on}$.

Then the parameter usage and floating point operations per second required by ordinary convolution are:

$$para = K_1^2 \times FM_i \times FM_o \tag{1}$$

$$flops = K_1^2 \times FM_i \times FM_o \times (W \times H) \tag{2}$$

The parameter usage and floating point operation times per second of the multi-scale feature extraction module are:

$$Para = K_1^2 \times C_{o1} \times C_i + K_2^2 \times C_{o2} \times \frac{C_i}{(\frac{K_2^2}{K_1^2})} + \cdots + K_n^2 \times C_{on} \times \frac{C_i}{(\frac{K_n^2}{K_1^2})} \tag{3}$$

$$flops = Para \times (W \times H) \tag{4}$$

If the number of output channels of each layer of multi-scale extraction method is the same, then the number of parameters and computational complexity of each layer will be distributed evenly. The measured results show (see Table 1) that after adding the multi-scale feature extraction and feature reconstruction module, the number of parameters increases from 16.3×106 to 29.6×106, and the number of iterations per second decreases by 0.31, which improves the detection ability of the model for small target objects at the minimum cost.

Table 1. Test results of parameter number and running speed in model experiment

The model name	The number of arguments	Iterations per second
YOLOv5s	16.3×10^6	2.44
YOLOv5s+ multi-scale feature extraction	21.1×10^6	2.26
YOLOv5s+ Multi-scale feature extraction + feature reconstruction	29.6×10^6	2.13

In order to solve the problem of gradient disappearance caused by increasing the depth of deep neural network, hopping residual connection structure is adopted in the model, and the original features are added after multi-layer convolution, as shown in figure 5.

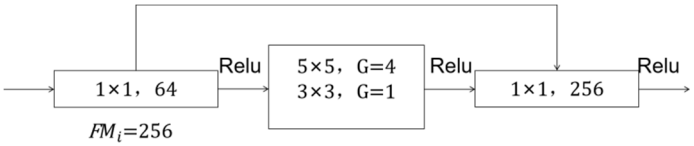


Figure 5. Flow chart of the characteristic jump connection network

The input features are outputted by two groups of different convolution kernels and then concatenated to get the output features. The final output features is A:

$$A = pyconv(x) + x \tag{5}$$

x represents the input feature of the multi-scale feature extraction module, $pyconv()$ represents the multi-scale convolution, the network model adopts residual link, and the output feature is $A \in \mathbb{R}^{H\times W\times C}$.

2.2.Reconstructed feature module

In order to improve the extraction effect of fine-grained features, a Feature Reconstruction Module (FRM) is further constructed. FRM takes the output of the multi-scale Feature extraction Module $A \in \mathbb{R}^{H\times W\times C}$ as the input and introduces the attention mechanism, as shown in figure 6. The feature reconstruction module includes three convolution layers: shift convolution, ordinary convolution, cyclic grouping convolution and an attention mechanism layer. The results of the three-part

convolution are summed with the output of a certain weight and attention mechanism, and the features are reconstituted into feature maps of the same size as before.

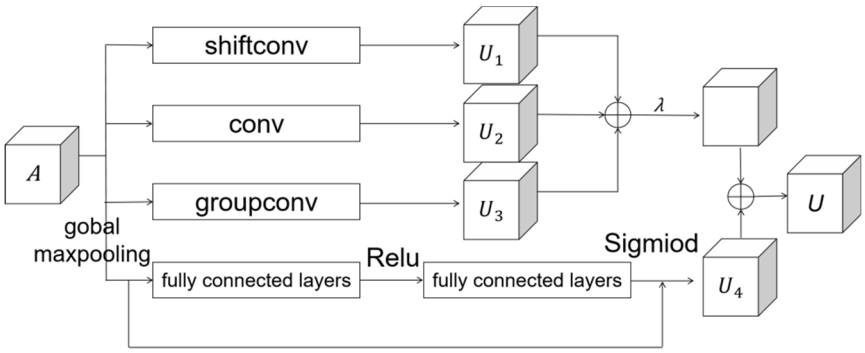


Figure 6. Feature reconstruction module structure diagram

Shift convolutional layer cuts the input feature map into two parts and rejoins them. The purpose is to force the network to learn the disconnected feature map, so that the network can pay attention to the small features that cannot be noticed under normal conditions.

$$U_1 = shiftconv(x_1, x_2) \tag{6}$$

x_1, x_2 represents the two parts of the original feature, $shiftconv()$ represents the reassembly of the feature map x_1, x_2 and then convolution.

Ordinary convolutional layer:

$$U_2 = conv(A) \tag{7}$$

$conv()$ represents ordinary convolution operation.

The convolution method used in the cyclic grouping convolutional layer is to extract information by using convolution check of different scales inside a convolutional layer[6], and different expansion rates are adopted for each input channel. At the same time, different convolution kernels and expansion rates are used repeatedly, and finally, block convolution is also used to improve the computational efficiency.

Let $A \in \mathbb{R}^{C_{in} \times H \times W}$ denote the input feature, $J \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}$ denote the convolution kernel.N、$M \in \mathbb{R}^{C_{out} \times H \times W}$ represents the output features of ordinary convolution and cyclic block convolution respectively. Then the conventional convolution is defined as:

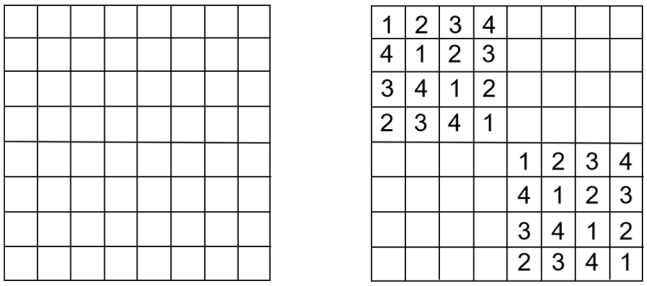
$$N_{c,x,y} = \sum_{k=1}^{C_{in}} \sum_{i=-\frac{K-1}{2}}^{\frac{K-1}{2}} \sum_{j=-\frac{K-1}{2}}^{\frac{K-1}{2}} (J_{c,k,i,j} A_{k,x+i,y+j}) \tag{8}$$

And the circular block convolution method is:

$$M_{c,x,y} = \sum_{k=1}^{C_{in}} \sum_{i=-\frac{K-1}{2}}^{\frac{K-1}{2}} \sum_{j=-\frac{K-1}{2}}^{\frac{K-1}{2}} (J_{c,k,i,j} A_{k,x+iD_{(c,k)},y+jD_{(c,k)}}) \tag{9}$$

In equation (9), $D_{(c,k)}$ represents $D \in \mathbb{R}^{C_{in} \times C_{out}}$, which is a matrix composed of the expansion rates of channel level and filter level of two orthogonal dimensions. $D_{(c,k)}$ associated with a particular channel in a filter, the entire matrix D can therefore be interpreted as a mathematical representation of a lattice of convolution kernels in its expansion rate subspace.

Figure 7 (a) shows ordinary convolution, where each square represents the connection relationship between input and output, and there are connections between each input channel and output channel. Figure 7 (b) shows grouped cyclic convolution, where grids labeled 1, 2, 3, and 4 represent receptive fields of different sizes. In the convolution operation, the packet convolution network recycled-uses convolution kernels of different sizes to deal with the features. The convolution kernels with four different receptive fields are arranged together to indicate that every four convolution kernels complete a cycle. It not only ensures the sensitivity of the network to fine-grained features, but also does not reduce the computational efficiency.



(a) (b)

Figure 7. Comparison of ordinary convolution and grouped cyclic convolution

That is:

$$U_3 = groupconv\ (A) \tag{10}$$

$groupconv()$ represents the grouped cyclic convolution, $A \in \mathbb{R}^{H\times W\times C}$ is the input feature, and $U_3 \in \mathbb{R}^{H\times W\times C}$ is the output feature.

In order to increase the sensitivity of the network model to key targets, an additional feature attention mechanism layer[7] is added to the feature reconstruction module. Figure.8 shows its calculation process.

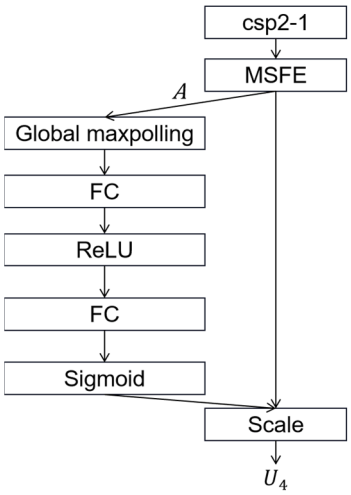


Figure 8. The attention mechanism outputs the feature calculation flow

In Figure 8 Csp2-1 is the existing module in YOLOv5s, and the output features after multi-scale feature extraction are taken as the input of the attention layer. FC represents the fully connected layer. A global maximum pooling operation is performed on the input feature A first, and the feature's dimension becomes 1×1×C. After dimensionality reduction by the first fully connected layer, the feature is activated at the site to the rectified linear unit (ReLU), and the original dimension is restored by the latter fully connected layer. Then the activation value of each channel is multiplied by the original feature to be used as the input feature of the next level. The principle is to enhance the important features and weaken the unimportant features by learning the weight coefficients of each channel, so as to make the extracted features more directional.

The final output feature $U \in \mathbb{R}^{H\times W\times C}$ of the feature reconstruction module is:

$$U = \lambda(U_1 + U_2 + U_3) + U_4 \tag{11}$$

λ is a hyper-parameter. The reconstructed features U is then input into the YOLOv5s model to realize the improvement of its application in real-time detection of smart construction site safety hats.

3. Analysis of experimental results

3.1. Data sources and preprocessing

The experimental data set in this paper is processed and made based on the pictures taken by the camera in the construction site, the relevant pictures climbed from Google and baidu, and the pictures in the Safety-Helmet-Wearing-Dataset that has been publicly released. The Safety-Helmet-Wearing-Dataset contains a large number of classroom self-learning images taken by cameras, which do not conform to the detection task in real scenes. Therefore, these images are deleted to clean the open Data set. And add more than 500 images with a lot of disturbing factors. Compared to the publicly available hardhat datasets, the newly created data set includes images with more occlusions. The collected data also includes workers wearing helmets and those not wearing helmets in different environments, at different resolutions, and at different construction sites. Include more pictures with helmets as small targets and pictures with occlusions.

The data set consists of 7,495 images. The data set was annotated in XML format (PASCAL VOC format[8]) with the labeling software LabelImg, where people wearing helmets were labeled as hat and people without helmets were labeled as people. And use a specific script to convert to YOLO format. The ratio of training set to test set is divided according to 8:2. 6874 images are used as training set, and 1621 images are used for testing. These include human helmet wearing objects (front) and normal head objects (not wearing or profile). Once annotated, each image corresponds to an XML file with the same name as the image, which is converted to a TXT file in YOLO format. Each line in the TXT file represents an instance of the tag. The TXT file has 5 columns, from left to right respectively represents the label category, the ratio of the tag box central abscissa to image width, the ratio of the tag box central ordinate to image height, the ratio of tag box width to image width, the ratio of tag box height to image height.

3.2. Analysis of experimental results

The performance of the proposed algorithm is evaluated by the common evaluation indexes in object detection algorithms, such as mean average precision(mAP), precision rate(P) and recall rate(R). Through experiments, the first added multi-scale feature extraction module can adopt the double-layer structure to achieve the best effect. The module contains two groups of convolution kernels with different sizes, and the size of convolution kernels for each channel is 3×3 and 5×5 respectively, which can make the model achieve better results. In addition, hyper-parameters in the multi-scale feature extraction module can be flexibly mobilized, such as the number of layers, the number of output channels in different layers, different depths, and different number of groups to adapt to different detection tasks. The hyper-parameterλin the reconstruction module is set to 0.1. The experimental results are shown in Table 2.

Table 2. Comparison of experimental results of model performance parameters

The model name	Class	Images	Labels	P	R	mAP	mAP upgrade
YOLOv5s	all	1621	24564	0.895	0.883	0.917	-
	hat	1621	2020	0.859	0.873	0.903	
	person	1621	22544	0.93	0.894	0.931	
YOLOv5s+multi-scale feature extraction	all	1621	24564	0.898	0.885	0.921	0.4%
	hat	1621	2020	0.867	0.873	0.908	
	person	1621	22544	0.93	0.897	0.935	
YOLOv5s+Multi-scale feature extraction+feature reconstruction	all	1621	24564	0.899	0.884	0.924	0.7%
	hat	1621	2020	0.862	0.874	0.91	
	person	1621	22544	0.936	0.894	0.938	

As can be seen from Table 2, the algorithm in this paper can effectively improve the detection accuracy of safety helmets and workers who are not wearing safety helmets. In the original YOLOv5s model, the Average mAP(Mean Average Precision) of people wearing and not wearing hard hats was 91.7%. After adding the multi-scale feature extraction module and feature reconstruction module(FRM), the mAP increased by 0.7% to 92.4%. Among them, the accuracy of detecting workers without helmets increased by 0.6 percent to 93.6 percent.

A total of 210 images with obstructions and long distance (small field of view) in the data set were used in the model anti-interference experiment (Table 3), and the mAP of the improved model was improved by 1.9%. It shows that the detection accuracy of the proposed algorithm is more excellent than that of the YOLOv5s model in the scenarios of different distances, pixel changes, obstacles and so on. It can meet the accuracy requirements of helmet inspection in complex working environment.

Table 3. Experimental results of disturbance rejection adaptability of the model

The model name	Class	Images	Labels	mAP	mAP upgrade
YOLOv5s	all	210	971	0.784	-
	hat	210	740	0.852	
	person	210	231	0.716	
YOLOv5s+ Multi-scale feature extraction + feature reconstruction	all	210	971	0.803	1.9%
	hat	210	740	0.854	
	person	210	231	0.752	

4. Conclusion

YOLOv5s is currently recognized as one of the effective real-time image detection algorithms, which is widely used in application fields with high real-time requirements. In order to apply it to the real-time detection of helmet wearing in construction sites with small targets, incomplete features and many interference factors, this paper applies the multi-scale and multi-branch feature extraction and feature reconstruction method to YOLOv5s for model reconstruction. Experiments show that compared with the YOLOv5s model, the improved algorithm has better scene recognition effect and real-time performance in the detection of remote small pixels and the detection in the presence of large number of occlusions. And can meet the application requirements of wisdom construction sites.

In order to further improve the detection accuracy, multiple cameras can be arranged to detect the same scene from different angles, and then the composite processing can be performed.

5. References

[1] KELM A, LAUSSAT L, MEINS-BECKER A, 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, 2013, 36: 38- 52.

[2] LI Q R. A Research and Implementation of Safety-helmet Video Detection System Based on Human Body Recognition. Chengdu：University of Electronic Science and Technology of China, 2017, 1-6, 34-59.

[3] WU H, ZHAO J S. An intelligent vision-based approach for helmet identification for work safety. Computers in Industry, 2018, 100: 267- 277.

[4] SHI H, CHEN X Q, YANG Y, et al. Safety helmet wearing detection method of improved YOLOv3. Computer Engineering and Applications, 2019, 55: 213- 220.

[5] LIU L, ZHU F, SHAO L. Pyramidal convolution: rethinking convolutional neural networks for visual recognition. (2020-06-20)[2022-8-31]. URL: https://arxiv.org/pdf/2006.11538.pdf.

[6] LI D, YAO A B, CHEN Q F. PSConv: squeezing feature pyramid into one compact poly-scale convolutional layer. In: European Conference on Computer Vision. 2020:615-632.

[7] HU J, SHEN L, SUN G. Squeeze-and-excitation networks//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2018:7132-7141.

[8] EVERINGHAM M, WINN J. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. (2019-03-20) [2022-8-31]. URL: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/VOCtrainval_11-May-2012.tar.