Journal of Network Intelligence ©2023 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **8**, Number **2**, May **2023**

One-Stage Safety Helmet Wearing Detection Based on Network-Structure-Improved YOLOv3

Zhi-Wei Gu[1], Wei Yang[1,2],Guang-Le Zhou[1], Xiao-Dan Jiang[3] and Zhe-Ming Lu*[4]

[1]Quzhou Guangming Power Investment Group Co., Ltd
Quzhou 324000, P. R. China
easteryang@163.com,858247@qq.com,quzhouguzhiwei@163.com

[2]State Grid Quzhou Power Supply Co.
Quzhou 324000, P. R. China
easteryang@163.com

[3]College of Electrical and Information Engineering
Quzhou University
Quzhou 324000, P. R. China
jxd@qzc.edu.cn

[4]School of Aeronautics and Astronautics
Zhejiang University
Hangzhou, 310027, China
zheminglu@zju.edu.cn

*Corresponding author: Zhe-Ming Lu

Received February 5, 2022, revised April 24, 2022, accepted June 10, 2022.

Abstract. *It is a common violation of regulations to not wear a helmet in the dangerous area of the substation, which will cause great hidden dangers to operators and power equipment. In order to automatically monitor whether workers are wearing safety helmets correctly and ensure safe operation in substations, this paper proposes a safety helmet wearing detection method based on the YOLOv3 algorithm with improved network structure. The two-stage detection method used in the current helmet detection research, which detects people first and then detects the helmet, has the problems of low detection efficiency and a large impact of cumulative error on accuracy. In response to this problem, this paper proposes a single-stage detection method, that is, the helmet and the worker's head are regarded as a whole, and the detection targets are divided into two categories, namely the head wearing a helmet and the head not wearing a helmet. At the same time, two types of targets are detected, which avoids redundant calculation steps and the influence of accumulated errors. On the other hand, according to the characteristics of helmet wearing status detection in construction scenes, this paper improves the network structure, the loss function and the priori anchor box sizes of YOLOv3. Our scheme builds a feature pyramid of four convolutional layers of different scales by adding a convolutional layer after the main network and the original three convolutional layers of YOLOv3, and performs upsampling with 2 times the step size at the same time. Deep residual network fusion improves the detection system effect. The experimental results show that the detection performance of the YOLOv3 model with the improved network structure is greatly improved, the mAP in the sample set constructed in this paper reaches 95.1%, and it also shows good robustness to the detection scene..*

Keywords: YOLO(You Only Look Once), Safety helmet wearing detection, Deep learning.

1. **Introduction.** In recent years, with the rapid growth of the world economy, the demand for electricity in various industries has continued to increase, and the scale of power grid construction has continued to expand. As an intermediate key link in the distribution of power resources, substations play an increasingly important role in the entire power production system. Therefore, it is necessary to ensure the safe and reliable operation of the substation for a long time. Although the traditional video surveillance system can display the working conditions of the staff in the substation in real time, it requires 24-hour monitoring and regular inspection by the power on-duty personnel. Due to the limited energy of the power on-duty personnel, they are prone to fatigue and distraction, which often makes it difficult for them to detect potential threats to safety in time. The failure of production and the prevention of safety production accidents directly cause the lag of accident handling. Therefore, it is more and more important to develop a set of unattended intelligent video surveillance system by using the historical video data of the storage equipment of the substation video surveillance system.

In the monitoring system, automatically detecting whether workers wear safety helmets and making corresponding feedback is essential to protect the country‘s labor force and improve safety performance. As a kind of target detection, helmet wearing detection has important application value. Therefore, the use of target detection methods to detect the wearing of helmets of substation operators has important engineering practical value. The automatic identification technology of helmet wearing status is an important link to realize the intellectualization of video surveillance system. At present, the mainstream helmet detection methods are basically divided into two types: 1) methods based on traditional target detection, including the HOG(Histogram Of Gradient) detector [1], the Deformable Part Model (DPM) [2] and other methods; 2) methods based on deep learning methods, including Faster R-CNN(Region-Convolutional Neural Network) [3], YOLO v3 [4], SSD (Single Shot multibox Detector) [5] and other methods.

For the traditional target detection method, the color and shape of the helmet, the resolution of the camera, and the change of the shooting angle and distance of the camera are the difficulties in the identification of the helmet of the video surveillance system. In response to this situation, many researchers have conducted in-depth research on the automatic identification technology of helmets. In 2013, Silva et al.[6] designed a helmet-wearing detection system for traffic scenes. The method is based on the detection of motorcycle drivers to detect whether a helmet is worn, the upper 1/5 of the motorcycle driver's object area is used as the potential area of the helmet, and local binary pattern features and HOG features are extracted. Then three classifiers, including Naive Bayes, Random Forest, and Support Vector Machine, were trained for comparative experiments. The results show that the trained random forest classifier has the best detection performance with an accuracy of 93.08%. In 2014, Silva et al. [7] first used Adaptive Mixing of Gaussians (AMG) to extract moving objects and then detected motorcycle drivers. Through the calculation of the sub-window, the human head is composed and converted into a grayscale image, and the mean filtering is performed to denoise, and the binarization conversion and Hough transform are performed to obtain a circular area. Then the LBP, HOG and W.T. features of the head region are extracted and any two feature pairs are combined. The experimental results show that the detection effect based on the combination of HOG and LBP features is the best, and the detection accuracy rate is as high as 94.04%. In 2018, Reference [8] proposed a feature fusion based helmet recognition method. First, the head image is extracted from the acquired video. Then, LBP (texture), H.U. extracts the moment invariant (geometry) and color histogram (color) feature vectors of the head image. Finally, head images are classified into four categories (red hard hat, yellow hard hat, blue hard hat, and no hard hat) using a hierarchical support vector machine (HSVM). The method in [8] can not only monitor whether a worker is wearing a helmet, but also further identify the color of the helmet. On the basis of traditional machine learning target detection algorithms, relevant researchers are required to conduct in-depth research on the detection field by designing specific and adaptable features for different detection tasks. These methods are optimized separately during the feature extraction and classifier training phases and do not affect each other. But it is vulnerable to environmental changes.

Above methods are based on traditional machine learning methods for object detection. These methods are mostly based on subjective feature selection, which required a solid professional foundation and rich experience. Moreover, feature selection is time-consuming, and its generalization ability is poor, hard to adapt to changes in conditions, such as lighting. Traditional target detection needs to be achieved by manually designing features. This method has low detection accuracy and is not robust. In recent years, deep learning has gradually gained the favor of scholars by relying on the advantages of convolutional neural network (CNN) in extracting image features without manually designing features. Correspondingly, many researchers have proposed a series of target detection algorithms based on deep learning. Chen et al. [3] proposed the improved Faster R-CNN algorithm to inspect the wearing of safety helmet. The Retinex image enhancement was introduced to improve image quality for the outdoor complex scenes in substations, and K-means++ algorithm is also adopted for better adaptation to the small size helmet. Wu et al. [4] adopts the advantage of Densenet in model parameters and technical cost to replace the backbone of the YOLO V3 network for feature extraction, thus forming the so-called YOLO-dense backbone convolutional neural network. Mohan et al. [9] demonstrated a novel implementation of the Faster R-CNN and SSD framework for accurate helmet detection in real-time low-quality surveillance videos. Vishnu et al. [10] developed a system to automatically detect helmet driver wearing in traffic scenarios. First, the dynamic objects of the video image are acquired using an adaptive image subtraction method. Two different convolutional neural networks are then used for motorcycle driver detection and helmet detection. The experiment used two datasets: one containing a single object per image, without small objects and blurred objects, and the other containing multiple objects per image, with occlusion and small objects. The average detection accuracy of the experiment is as high as 92.87%. The method is a helmet wearing test in a traffic scene, and the background, category and pose of the picture are relatively simple. Furthermore, it uses two deep convolutional neural networks, which are both tedious and computationally complex.

In general, the traditional safety helmet detection methods often do not consider the influence of the complex substation operating environment on the recognition results. The detection results are easily disturbed by light and small objects in the image background, and the false alarm rate is high. The existing deep learning-based helmet detection methods are highly adaptable to the environment, but ignore the relationship between the helmet and the person, that is, whether the worker is wearing the helmet correctly. Based on the improved YOLO v3 deep learning method, this paper designs a helmet wearing detection method to detect the relationship between helmets and workers in substation scenarios. The rest of this paper is organized as follows. Section 2 introduces the original YOLO v3 method. The detailed description of our proposed method is then presented in Section 3. In Section 4, the experimental results on datasets are reported. Finally, the conclusion is provided in Section 5.

2. **YOLO V3 Network Model.**

YOLOv3 is evolved from YOLOv1 and YOLOv2. It greatly improves the detection accuracy and the detection speed is also faster. Therefore, this paper selects the YOLOv3 network as the basic framework for safety helmet detection.

2.1. **Structure.**

As the latest algorithm of the YOLO series, YOLOv3 retains and improves the previous algorithms. Let's first analyze what is reserved on YOLOv3: 1) "Divide and conquer", starting from YOLOv1, the YOLO algorithm fulfils the detection task by dividing cells, but the number of cells is different. 2) Use "Leaky ReLU" as the activation function. 3) Train the network in an end-to-end manner. The training is completed with a single loss function, and the user only needs to pay attention to the input and output. Since YOLOv2, YOLO has used batch normalization as a method for regularization, accelerating convergence and avoiding over-fitting, connecting the BN layer and the Leaky ReLU layer to each convolutional layer. 4)Adopt a multi-scale training method. This method can obtain a compromise between the speed and accuracy. If you want to be faster, you can sacrifice the accuracy; if you want to be more accurate, you can sacrifice a little speed.

The improvement of each generation of YOLO is largely determined by the improvement of the backbone network, from darknet-19 of v2 to darknet-53 of v3. YOLOv3 also provides a replacement backbone called tiny darknet. For higher performance, Darknet-53 can be used as the backbone, while tiny-darknet can be used for light weight and high speed. In short, YOLO is inherently "flexible", so it is especially suitable as an engineering algorithm. The overall network architecture diagram of YOLOv3 is shown in Fig. 1. Among them, concat represents tensor concatenation to splice the darknet intermediate layer and the upsampling result of a later layer. The operation of splicing is different from the operation of adding in the residual layer. Splicing will expand the dimension of the tensor, while the adding operation is just a direct addition, which will not lead to a change in the dimension of the tensor. The interpretation of DBL and res_n in Fig. 1 is as follows:

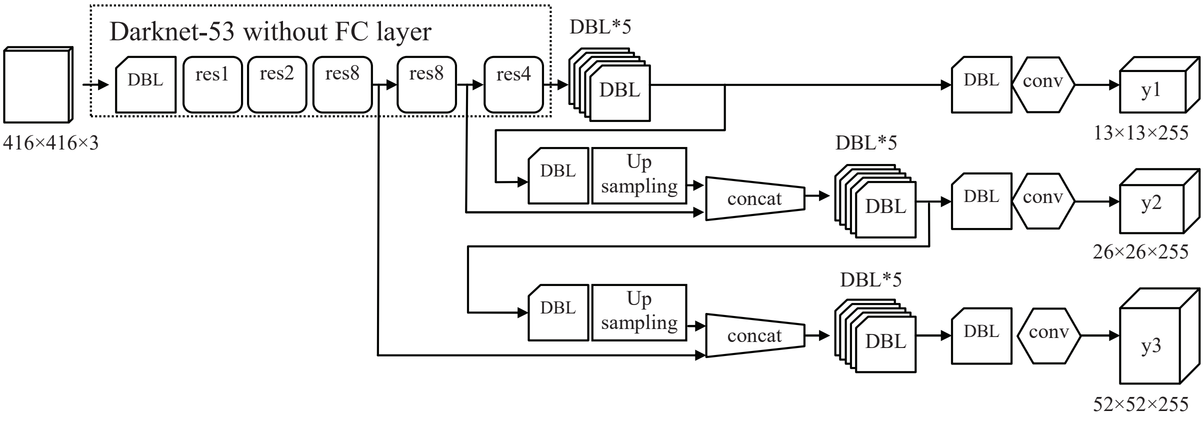


FIGURE 1. The overall structure of YOLOv3.

(1) DBL is shown in Fig. 2(a), which is Darknetconv2d_BN_Leaky in the code, it is the basic component of YOLOv3. It is convolution+BN+Leaky ReLU. For v3, BN and Leaky ReLU are already inseparable parts from the convolution layer (except for the last layer of convolution), which together constitute the smallest component.

(2) res_n is shown in Fig. 2(c), where n represents a number, res1, res2, …, res8, etc., indicating how many res_units are contained in this res_block. This is a large component

of YOLOv3. YOLOv3 began to introduce the residual structure of ResNet. Using this structure can make the network structure deeper (from v2's darknet-19 to v3's darknet-53, the former has no residual structure). For the explanation of res_unit, it can be seen intuitively from Fig. 2(b) that its basic components are also DBL. From the code level, the entire YOLOv3 body contains 252 layers, including 23 layers of "add" layers (mainly used for the composition of res_block, each res_unit needs an "add" layer, a total of 1+2+8+8+4=23 layers). In addition, the number of BN layers and LeakyReLU layers is exactly the same (72 layers), and the performance in the network structure is: each layer of BN will be followed by a layer of LeakyReLU. The convolutional layer has a total of 75 layers, of which 72 layers will be followed by a combination of BN+LeakyReLU to form the basic component DBL. From Fig. 1, it can be found that both upsampling and concat have 2 times. Each res_block is padded with the previous zero, for a total of 5 res_blocks.

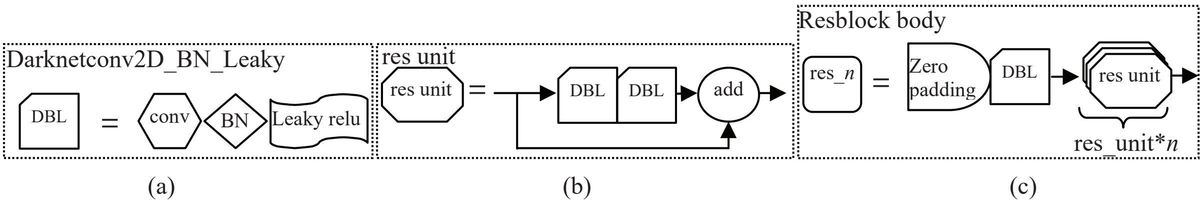
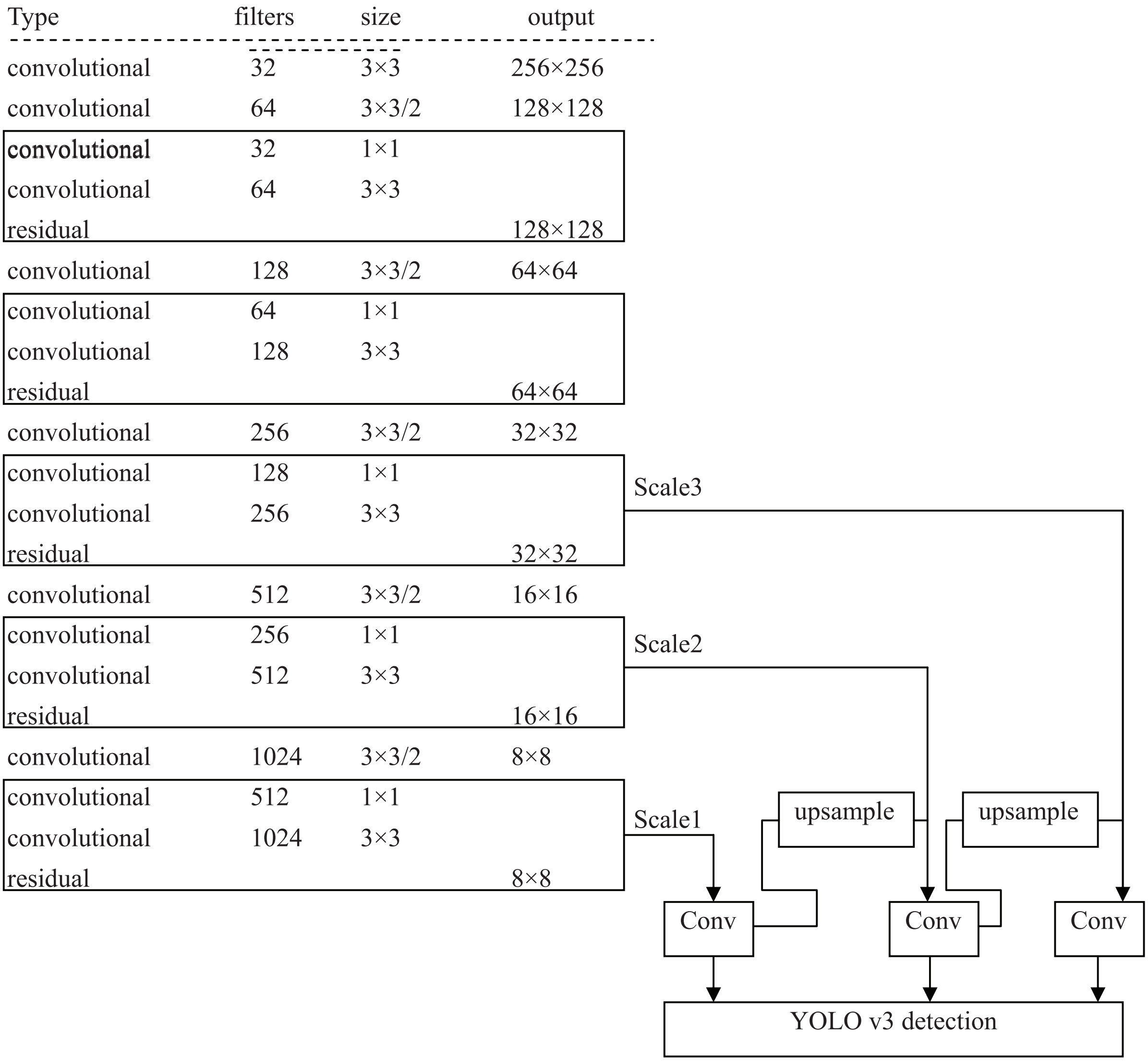


FIGURE 2. The components DBL and res_units of YOLOv3.

2.2. **Backbone.** There are no pooling layers and fully connected layers in the entire v3 structure. In the forward propagation process, the size transformation of the tensor is achieved by changing the step size of the convolution kernel, such as stride=(2, 2), which is equivalent to reducing the side length of the image by half (that is, the area is reduced to 1/4 of the original size). In YOLOv2, to go through 5 reductions, the feature map will be reduced to 2^5 of the original input size, i.e., 1/32. The input size is 416 × 416, the output size is 13 × 13 (416/32=13). YOLOv3 is also the same as v2 that the backbone will reduce the output feature map to 1/32 of the input. Therefore, the size of the input image is usually required to be a multiple of 32.

The overall architecture of YOLOv3 is shown in Fig. 3. YOLOv3 adds more convolutional layers to extract deep features of objects. YOLOv3 has a total of 75 convolutional layers, including jump connection and upsampling layers. In addition, it replaces the traditional maximum pooling layer with a 2-step convolutional layer. Compared with the pooling layer, convolution has more possibilities for change. The YOLOv3 network performs a total of 5 downsampling on the input image, and predicts the target in the last 3 downsampling.

In YOLOv2, the tensor size transformation in the forward process is performed by maximum pooling, a total of 5 times. And YOLOv3 is performed by increasing the step size of the convolution kernel, which is also 5 times. There is a global average pooling at the end of darknet-53, there is no such layer in YOLOv3, so the tensor dimension change only considers the first 5 times. This is also why a 416 × 416 input gets a 13 × 13 output. Darknet-19 has no residual structure (resblock, borrowed from resnet), and is the same type of backbone as VGG (belonging to the previous generation of CNN structure), while darknet-53 is a backbone comparable to resnet-152. Compared with other target detection algorithms, YOLOv3 has achieved better results in terms of detection accuracy and speed. The performance comparison between YOLO v3 network and other target



Type	filters	size	output
convolutional	32	3×3	256×256
convolutional	64	3×3/2	128×128
convolutional	32	1×1	
convolutional	64	3×3	
residual			128×128
convolutional	128	3×3/2	64×64
convolutional	64	1×1	
convolutional	128	3×3	
residual			64×64
convolutional	256	3×3/2	32×32
convolutional	128	1×1	
convolutional	256	3×3	
residual			32×32
convolutional	512	3×3/2	16×16
convolutional	256	1×1	
convolutional	512	3×3	
residual			16×16
convolutional	1024	3×3/2	8×8
convolutional	512	1×1	
convolutional	1024	3×3	
residual			8×8

Figure 3. The overall framework of YOLOv3.

Table 1. Comparison of feature extraction network performance.

Backbone	Darknet-19	ResNet-101	ResNet-152	Darknet-53
Top-1/%	74.1	77.1	77.6	77.2
Top-5/%	91.8	93.7	93.8	93.8
Bn Ops/109	7.29	19.70	29.40	18.70
BFLOP/s	1246	1039	1090	1457
Recognition frame rate/$\mathtt{f} \cdot \mathtt{s}^{-1}$	171	53	37	78

detection frameworks is shown in Table 1. It can also be seen from the above table that darknet-19 still has a great advantage in speed. In fact, it can be seen in other details (such as bounding box prior using $k = 9$), YOLOv3 does not pursue speed so much, but pursues performance on the basis of ensuring real-time performance (fps> 60). But as mentioned earlier, if you want to be faster, there is also a tiny-darknet as a backbone that can replace darknet-53, and you can switch the backbone with one line of code in the official code. Using tiny-darknet's YOLO, that is, tiny-YOLO is obviously state of the art level in terms of light weight and high speed, and tiny-darknet is a network comparable to squeezeNet. Therefore, with YOLOv3, you really don't need YOLOv2, let alone YOLOv1. This is also the reason why YOLO's official website(https://pjreddie.com/darknet/), after v3 came out, did not provide the v1 and v2 code download links.

2.3. **The Core Algorithm Flow of YOLOv3.** The detection process of the YOLOv3 model is as follows: first, the entire image is divided into $S \times S$ grids, each grid is only responsible for detecting objects in the grid; then, the classification model is used to classify each grid (output the predicted position and size information of the target box and the confidence of the target to be detected); finally output the target box larger than the confidence threshold, and use the non-maximum suppression algorithm (NMS) to remove the overlapping detection box to obtain the final prediction result.

The overall process of YOLOv3 bounding box prediction is as follows: First, YOLOv3 uses the K-means clustering method to initialize the bounding boxes, and obtains 9 bounding box sizes, according to the standard of large, medium and small receptive fields. The feature maps of different scales are matched; secondly, the regression algorithm is used to obtain the position information of the predicted bounding box, including the center coordinates and width and height of the bounding box; finally, YOLOv3 uses multiple independent Logistic classifiers to classify the categories of objects.

Each cell of image division in YOLOv3 needs to predict 3 bounding boxes, which is greatly increased compared to the number of bounding boxes generated by YOLOv2. Taking an image with a pixel of 416 × 416 as an example, YOLOv2 needs to predict 845 bounding boxes ($13 \times 13 \times 5 = 845$), and YOLOv3 needs to predict 10647(($13 \times 13 + 26 \times 26 + 52 \times 52) \times 3 = 10647$) bounding boxes, and based on this, the introduction of the multiscale prediction mechanism enables YOLOv3 to obtain better detection performance in small target detection scenarios than YOLOv2.

3. **Proposed Method.**

3.1. **General Idea of Helmet Detection in Power Construction Scene.** By analyzing the video data of the power construction site, it can be concluded that the characteristics of the helmet detection in this scene are as follows: First, the helmet is a very small target in the whole picture, and its height and width may only be a few pixels in the case of long-distance shooting by the camera; second, the background of the detection scene of outdoor construction is very complex, and the objects in the background, such as building materials and various components of construction machinery, are relatively close to the color and shape of the helmet, and false detections are prone to occur during detection.

In the design of detection ideas, by analyzing the disadvantages of the two-stage detection method, this paper considers designing a single detection idea, that is, through one detection, both workers can be detected and whether they are wearing helmets can be detected. Based on this, this paper regards the helmet and the worker's head as a target box, mainly based on the following two considerations: First, the proportion of the human head is in a fixed proportional relationship with the human body, and the head information can be regarded as a complete replacement of the human body information; Second, the background of the detection scene is more complicated. Compared with the head, the human body is more prone to occlusion. At the same time, the flexibility of the human body leads to the change of the posture of the human body. All kinds of interference will increase the difficulty of the detection model to recognize the human body, so that the model will not be able to recognize the human body. The detection accuracy of the human body is low, while the head area contains less background noise and is less affected by the flexibility of the human body. Therefore, this paper regards the helmet and the worker's head as a target box, which contains two types of targets: the head wearing a helmet and the head without a helmet. Data sets can be established for these two types of targets. In this way, a single detection idea can be realized, and workers and safety helmets can be detected at the same time, which not only avoids the correlation between the two detections, improves the detection accuracy of the model, but also saves the tedious detection process and greatly reduces the computational complexity.

3.2. **The Overall Improvement Idea of the Detection Model.** After experimental verification, YOLOv3 has poor performance in helmet detection in power construction scenarios. The reason is: on the one hand, the safety helmet is a very small object in the whole image, and on the other hand, the background of the power construction site is very complex, and the safety helmet is easily confused with other objects in the image, resulting in more false detections. Based on this, combined with the characteristics of power construction sites, this paper makes targeted improvements to the YOLOv3 model in terms of network structure, loss function and anchor size, and proposes a helmet detection model for power construction scenarios, which is called MYOLOv3 (Modified YOLOv3) model.

As we have known, the YOLOv3 algorithm achieves the best balance of detection accuracy and speed through residual network feature fusion and multi-scale prediction, but when the input image size is 416 × 416, the size of the minimum feature map is 13 × 13. Compared with the actual target to be detected, the receptive field is too large, which will easily cause the detection effect of medium or small size objects to be poor, resulting in problems such as false detection, missed detection or repeated detection. Therefore, this paper improves the YOLOv3 algorithm, proposes a new feature fusion algorithm, and changes the input image size to 448 × 448. The MYOLOv3 algorithm is a multi-target detection algorithm for detecting whether a helmet is worn correctly. The main network structure is a deep residual network. Since the detection targets of workers in the general video environment are in the small and medium range, there is no need for a large receptive field. At the same time, in order to strengthen the recognition of small and medium targets, it is necessary to integrate more shallow feature maps to make predictions together to improve the accuracy of small and medium-sized objects. The recognition rate of the target also takes into account the problems of time complexity and robustness, so this paper adds a convolutional layer after the main network of YOLOv3, and fuses the feature map with the shallow residual network, and combined with the original 3 convolutional layers of YOLOv3 to jointly construct a feature pyramid containing 4 convolutional layers of different scales, namely: 56 × 56, 28 × 28, 14 × 14 and 7 × 7 resolutions; at the same time, using the step size by 2 times, we perform an upsampling operation on the feature pyramid and fuse it with the previous deep residual network to form a deep fusion fast helmet detection model. MYOLOv3 takes the entire image as input, and outputs all detected locations of people and whether they are wearing helmets correctly. First, the algorithm uses a deep residual network to extract the characteristics of employees in production; then multiple convolutional layers are added after the residual network to divide it into 4 branches to form a multi-scale prediction network; in order to obtain more meaningful information, the algorithm fuses the feature map in the prediction network with the corresponding feature map in the deep residual network; finally, the non-maximum suppression method is used to remove the repeated bounding boxes, and the final helmet wearing detection result is obtained. Compared with the traditional YOLOv3, the MYOLOv3 algorithm is more effective, can prevent the false detection of employee helmets in complex scenes, and significantly improve the detection accuracy of employee helmets in complex work scenarios.

3.3. **Feature Extraction of Helmet Wearing Based on Deep Residual Network.** Inspired by residual learning, MYOLOv3 extracts employee's head features by building a deep residual network. The residual network consists of a series of residual blocks, each of which contains two branches: the identity map and the residual branch. As shown in Fig. 4, the two residual blocks are stacked in order, and the corresponding formula is defined as follows:

$$x_{t+1} = F_t(x_t) + x_t \tag{1}$$

where x_t and x_{t+1} are the input and output vectors of the t-th residual block, respectively, and $F_t(x_t)$ represents the transfer function, corresponding to the residual branch composed of stacked layers. The deep residual network composed in this way is easy to flow of information and easy to train.

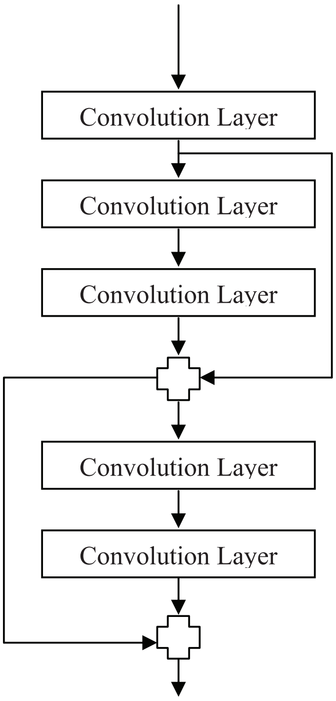
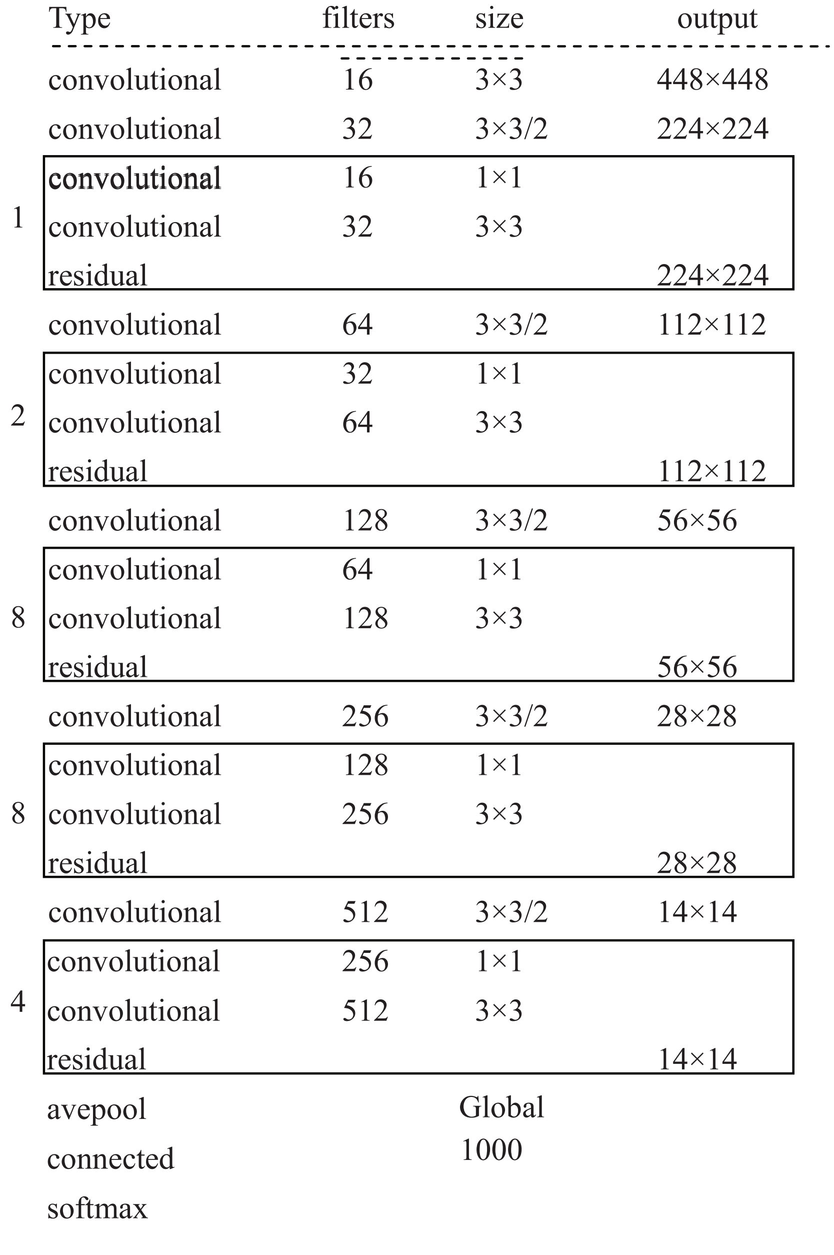


FIGURE 4. Schematic diagram of residual block structure.

The deep residual network structure in MYOLOv3 is shown in Fig. 5. The first convolutional layer uses 16 convolution kernels (filters) of size 3 × 3 to filter the input image with a resolution of 448 × 448; then uses 32 convolution kernels of size 3 × 3 with a stride of 2 pairs The input from the shallow layer is down-sampled and filtered, and a residual block similar to YOLOv3 is added to increase the depth of the network. The residual block consists of a 1 × 1 convolutional layer and a 3 × 3 convolutional layer. The size of feature map obtained at this time is 224 × 224; next, 4 groups of networks including 2×, 8×, 8×, 4× residual blocks are executed to obtain feature maps with 112 × 112, 56 × 56, 28 × 28, 14 × 14 resolutions respectively. In the network composed of these 4 groups of residual blocks, each residual block is similar except that the number of convolution kernels and the scale of feature maps are different. Specifically, the downsampling is specifically represented in each rectangular box as performing a downsampling operation with a step size of 2 in the convolutional layer above each residual block, and in order to standardize the network model and strengthen regularization, in all Batch normalization layers are added on top of the convolutional layers. In addition, MYOLOv3 selects feature maps with resolutions of 112 × 112, 56 × 56, 28 × 28, and 14 × 14 to fuse with the up-sampling feature maps proposed in the next section to form a feature pyramid for helmet wearing prediction.



	Type	filters	size	output
	convolutional	16	3×3	448×448
	convolutional	32	3×3/2	224×224
1	convolutional	16	1×1	
	convolutional	32	3×3	
	residual			224×224
	convolutional	64	3×3/2	112×112
2	convolutional	32	1×1	
	convolutional	64	3×3	
	residual			112×112
	convolutional	128	3×3/2	56×56
8	convolutional	64	1×1	
	convolutional	128	3×3	
	residual			56×56
	convolutional	256	3×3/2	28×28
8	convolutional	128	1×1	
	convolutional	256	3×3	
	residual			28×28
	convolutional	512	3×3/2	14×14
4	convolutional	256	1×1	
	convolutional	512	3×3	
	residual			14×14
	avepool		Global	
	connected		1000	
	softmax			

Figure 5. The backbone network structure of MYOLOv3.

3.4. **Anchor Box Reclustering Based on K-means Algorithm.** The YOLOv3 model clusters the COCO dataset to obtain 9 anchor boxes and outputs the dimensions of the 9 anchor boxes. Considering that the COCO dataset has a total of 80 types of data objects, including human body, vehicle, grass and other types of samples, there are many "prolate" boxes in the anchor box. The detection targets in this paper are helmets and workers' heads. Considering that the size of the anchor box of the target feature in this paper should be mainly “square-like” boxes, it is necessary to re-cluster the sample set data in this paper to obtain the suitable number and size of the most preferred boxes for the dataset in this paper. The above clustering method is used to perform cluster analysis on the sample data sets (helmet data set and worker head data set) in this paper. The relationship change curve between the number of clusters and the average IOU is shown in Fig. 6. It can be seen from Fig. 6 that as the number of clusters increases, the average IOU also changes. When the number of clusters increases gradually, the average IOU increases rapidly, and when the number of clusters increases further, the average IOU increases but changes slowly. Under the design of the 3-scale prediction mechanism of the YOLOv3 model, the clustering operation should be performed on its sample target box data set to obtain 9 different anchor sizes. Since our MYOLOv3 model adds a prediction scale, 12 anchor boxes with different sizes should be used. As shown in Fig. 6, from the relationship between the number of anchor boxes and the average IOU, it can be seen that the number of anchor boxes continues to increase on the basis of the 12 anchor boxes in this paper, which has a weak impact on the model accuracy, and it will reduce the detection efficiency of the model and affect the real-time detection of the model. According to the classification principle of the receptive field: the smaller the feature map, the larger the receptive field, and the more sensitive it is to large targets, so a large anchor box size is selected; the larger the feature map, the smaller the receptive field, and the more sensitive it is to small targets, so a small anchor box size is selected.

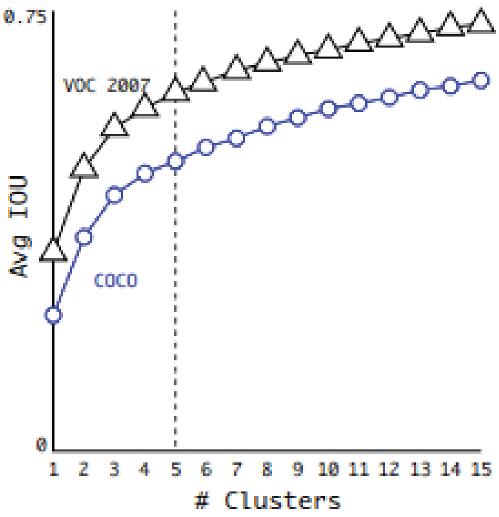


Figure 6. The relationship between the number of clusters and the average IOU.

3.5. **Multi-scale Network Prediction of Employee Location.** MYOLOv3 directly regresses and predicts the state of employees wearing helmets on the feature maps of multiple scales. The convolutional layers of different resolutions are divided into 4 branches (feature maps 14 × 14, 28 × 28, 56 × 56, 112 × 112 respectively), and each branch performs prediction independently. Specifically, each branch is equipped with multiple convolutional layers, performing upsampling at 2 times size for branches with resolutions of 14 × 14, 28 × 28, 56 × 56, 112 × 112. In order to enhance the representation ability of the feature pyramid, the upsampling feature and the feature map of the residual module in the corresponding backbone network are combined into a feature. In each branch, helmet detection is performed based on the pre- and post-semantic information, while the 4 detection branches share the features extracted from the residual network. When using MYOLOv3 for prediction, a three-dimensional tensor representing the box, head and neck objects, and helmet wearing category will be generated. The size of the feature map is $N \times N$ (the size of N is based on the size of the feature map), for each grid prediction 3 different bounding boxes, and the tensor can be represented as $N \times N \times [3 \times (4+1+2)]$, i.e. 4 bounding box offsets, 1 employee object and 2 detection types.

3.6. **Improvement of Loss Function Based on GIOU.** The full name of IOU is Intersection over Union. IOU calculates the ratio of the intersection to the union of the "predicted bounding box" and the "true bounding box". In the YOLOv3 algorithm, IOU is used to determine whether the model has detected a target. The model sets an IOU threshold. If the IOU of the detected target is greater than the threshold, the model judges that the target is detected, and if it is less than the threshold, the target is not detected. In the YOLOv3 model, the mean square error (MSE) is used as the loss function to perform the regression of the target box. The prediction results of different quality are sometimes indistinguishable by the MSE evaluation index. In view of the mismatch between the loss function and the model detection effect, this paper uses the loss function based on GIOU [11] to replace the bounding box regression loss function in the original YOLOv3 algorithm to better match the relationship between the loss function and the target detection evaluation. The calculation formula of the loss function based on GIOU is as follows:

$$\mathtt{I} = \mathtt{A_o} \tag{2}$$

$$\mathtt{U} = \mathtt{A_p} + \mathtt{A_g} - \mathtt{A_o} \tag{3}$$

$$\mathtt{IOU} = \frac{\mathtt{I}}{\mathtt{U}} \tag{4}$$

$$\mathtt{GIOU} = \mathtt{IOU} - \frac{\mathtt{A_c} - \mathtt{U}}{\mathtt{A_c}} \tag{5}$$

$$\mathtt{Loss_{GIOU}} = 1 - \mathtt{GIOU} = 2 - \frac{\mathtt{U}}{\mathtt{A_c}} - \frac{\mathtt{I}}{\mathtt{U}} \tag{6}$$

where $\mathtt{A_o}$ is the overlapping area, $\mathtt{A_p}$ is the area of the predicted box, $\mathtt{A_g}$ is the area of the real box, and $\mathtt{A_c}$ is the area of the smallest box that can cover the predicted box and the real box.

4. **Experimental Results.** In the above, we have completed the design of the helmet wearing detection algorithm, so the main work of this section is to train and test the performance of the helmet wearing detection model. The experimental environment is as follows: the operating system is Ubuntu 16.04, the CPU is Intel Xeon E5-2650 2.20GHz, the memory is 16G, and the GPU is NVIDIA GeForce GTX 1080Ti. Firstly, the process of data set establishment and model training is introduced in detail to complete the training of the helmet wearing detection model and the construction of the entire detection module. Then, this paper conducts a comparative experiment between the improved algorithm and the existing typical algorithms, and conducts an ablation experiment for the improved algorithm. The specific experimental data and visualization results are given, and the corresponding analysis and explanation are carried out.

4.1. **Data Set Establishment and Data Enhancement.** Deep learning is a typical data-driven method, and the detection performance of its model greatly depends on the quantity and quality of the training set. The dataset established in this paper is divided into two subsets, including a subset of 5000 images of workers wearing safety helmets in construction scenes, and a subset of 5000 images of workers' heads without safety helmets. This paper collects data through network pictures, construction site surveillance video and other channels.

The selection principles of sample data are as follows: 1) Cover images with large-scale targets, medium-scale targets and small-scale targets, especially to increase the proportion of samples with small-scale targets; 2) Collect images from different perspectives, including top-down, head-up and other angles ; 3) Contain multi-pose images of workers such as squatting and bending; 4) Add samples in bad weather such as cloudy and rainy days. In view of the unstable imaging quality of the camera due to light and weather, it is considered to add noise disturbance to some training samples to increase the robustness of the model in harsh environments.

4.2. **Selection of Detection Targets.** First, it is necessary to clarify the selection of the target region when training the model. At first glance, the purpose of this article is to detect whether the worker is wearing a helmet, so intuitively, it is necessary to locate the worker first, and then determine whether he is wearing a helmet. Most of the existing algorithms currently use this approach. However, in practical applications, we found that if the worker is located first and then judges whether the worker is wearing a helmet, two-step detection is required, which is not only cumbersome in steps and complex in algorithm, but also accumulates errors, and the detection effect is not ideal. Therefore, this paper designs a one-step strategy that can both detect the worker and determine whether he is wearing a helmet or not.

Based on the above considerations, this paper regards the helmet and the head as a target box. In a complex factory background, compared with the human body, the head position is less likely to be affected by interference such as occlusion and posture changes. In addition, considering that most workers in the factory wear safety helmets, under complex surveillance video, The head position-based target can better localize the worker, and there will be relatively less background noise in the target box, which improves the robustness of the system. Therefore, when annotating the dataset, this paper regards the helmet and the head and neck as a target box, and there are only two types of target boxes: the head wearing a helmet and the head not wearing a helmet. In this way, the two-step operation of the existing detection algorithm is simplified into one step, which not only detects the worker but also determines whether he is wearing a helmet, which reduces the complexity of the algorithm and improves the overall performance of the system.

4.3. **Implementation Details and Algorithm Comparison.** We train and test our proposed MYOLOv3 network on images of the same size. The training process is as follows: first, the sizes of the input images are all scaled to 448×448, the worker head features are extracted by a deep residual network stacked with multiple residual blocks, and 4 convolutional feature maps of different scales are used to predict the worker's position coordinates and types. On the self-made dataset in this paper, MYOLOv3 uses a total of 12 anchor boxes at 4 scales. In this paper, 12 anchor box sizes are automatically generated by running the K-means clustering algorithm, which are:(9,18), (13,17), (18,26), (30,36), (49,55), (57,70), (83,100), (96,117), (107,119), (130,167), (151,197) and (152,250). A total of 48,000 iterations were performed in the entire training process. The learning weight parameter was set to 0.0004, the momentum was set to 0.85, the batch size was set

Table 2. Performance comparison of different algorithms

Method	Network	Number of iterations	mAP
Faster R-CNN	VGG16	48000	85.7%
SSD	VGG16	48000	84.6%
YOLOv3	Darknet53	48000	90.4%
MYOLOv3	Modified Darknet53	48000	95.1%

Table 3. Ablation experiment results

Method	mAP
YOLOv3	90.4%
A_YOLOv3	91.6%
AN_YOLOv3	93.6%
MYOLOv3	95.1%

to 64, and the learning rate was initially set to 0.001, and then decreased to 0.0001 and 0.00001 in turn. Corresponding to the learning rate of each stage, the network iterates 20000, 16000 and 12000 times respectively.

In order to test the performance of the helmet wearing detection model obtained by using MYOLOv3, this paper conducts comparison experiments, and prepares a total of 2400 test images. The performance is compared in the same experimental environment, and the results are shown in Table 2.

It can be seen from the table that on the self-made dataset, the MYOLOv3 algorithm is the best, because it adds 3 convolutional layers to the original 3 convolutional layers of YOLOv3 to build the feature pyramid containing 4 convolutional layers of different scales, namely 112 × 112, 56 × 56, 28 × 28, 14 × 14; at the same time, it performs upsampling operations on the feature pyramid with 2 times step size, and fuse with the deep residual network to form a deep fused fast helmet detection model. At the same time, the fps is not significantly reduced, which fully meets the real-time requirements of the detection system.

4.4. **Ablation Experiment.** It should be noted that the detection scene in this paper is an outdoor construction scene with multiple lighting conditions, multiple human poses, and multiple scale targets. The results of the ablation experiment and the subsequent visualization results are the test results in this scene. In order to clarify the impact of each part of the improvement on performance, we compared the YOLOv3 model before improvement, the A_YOLOv3 model with anchor box reclustering, the AN_YOLOv3 model with anchor box reclustering and improved network structure, and the MYOLOv3 model with anchor box reclustering and improved network structure and loss function. The results are shown in Table 3.

It can be seen from the experimental results that the mAP of A_YOLOv3 is 1.2% higher than that of YOLOv3. The reason is that the A_YOLOv3 model uses the K-means algorithm to re-cluster the priori anchor box sizes of the data set in this paper to obtain priori anchor box sizes that are more suitable for the helmet samples, thereby improving the model's localization accuracy of the target in the image. The mAP of AN_YOLOv3 is 2.0% higher than that of A_YOLOv3. Among all the improved methods, the most important influence on the improvement of model performance is the improvement of the network structure. The reason is that considering that the heads with and without helmets are small objects in the whole image, and the environmental background is very complex, the AN_YOLOv3 algorithm adds a scale to the output layer of the feature map on the basis of the original network structure to obtain more sufficient shallow image information, thereby improving the detection accuracy of the model in small target detection. The mAP of MYOLOv3 is 1.5% higher than that of AN_YOLOv3, indicating that the GIOU-based loss function in this paper enhances the training effect of the model and improves the detection performance of the model compared to the MSE-based loss function. In summary, several improvement directions of the MYOLOv3 model proposed in this paper have a positive impact on the performance improvement of the model, which verifies that the improvement directions in this paper are reasonable and effective. At the same time, it can be seen from the experimental results that the mAP of the MYOLOv3 model reaches 95.1%, which has excellent detection performance and engineering application value.

4.5. **Visual Analysis.** Visual analysis of the detected image can more intuitively see the improvement effect of the MYOLOv3 model for the detection scene in this paper. Fig. 7 shows the comparison results of the detection effect of the MYOLOv3 model and the YOLOv3 model for the same detection scene, where the left side is the visualization result of the YOLOv3 model, and the right side is the visualization result of the MYOLOv3 model.

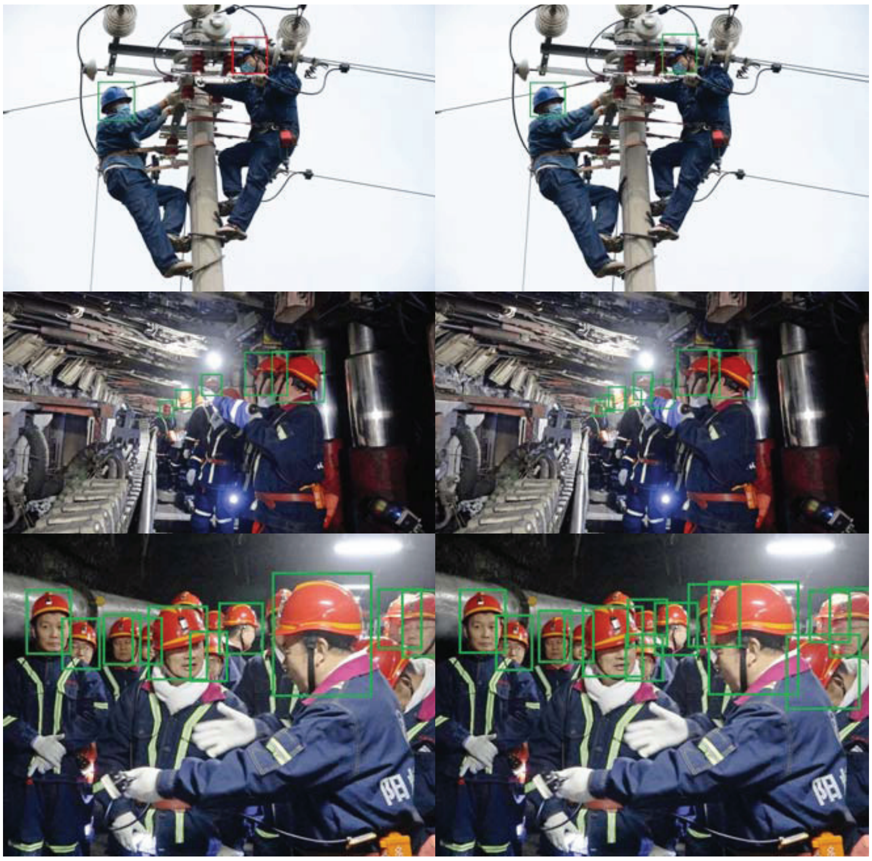


FIGURE 7. Visualization results of model detection.

As shown in Fig. 7, in complex scenes, the MYOLOv3 model shows stronger superiority in performance. The performance of the YOLOv3 model in complex detection scenarios is poor, there are a large number of false detections and missed detections, the target box positioning accuracy is low, and there are many stacking and accumulation situations. The MYOLOv3 model has better performance and more accurate positioning accuracy due to the targeted improvement of complex construction scenes in terms of network structure and loss function.

5. **Conclusions.** Aiming at the disadvantages of the two-stage detection method of helmets in terms of detection efficiency and error accumulation, this paper proposes a single-stage detection method, which treats the helmet and the head as a whole, and only needs to detect two types of targets in the model: the head of the worker wearing a safety helmet and the head of the worker without the helmet. On the one hand, the detection efficiency of the model is improved; on the other hand, the influence of the cumulative error brought by the two-stage detection method on the performance of the helmet detection model is avoided. Aiming at the problem that the YOLOv3 model has poor detection performance in small targets and complex backgrounds, the characteristics of the detection scene in this paper are firstly analyzed, and then the YOLOv3 model is improved in terms of network structure, loss function, and priori anchor box size according to these characteristics, and the MYOLOv3 model is proposed. The MYOLOv3 algorithm in this paper builds a feature pyramid of four convolutional layers of different scales by adding a convolutional layer after the main network and the original three convolutional layers of YOLOv3, and performs upsampling with 2 times the step size at the same time. Deep residual network fusion improves the detection system effect. The experimental results show that the average accuracy of the MYOLOv3 model for helmet detection reaches 95.1%, which is a significant improvement over the detection performance of the YOLOv3 model. At the same time, through the visual analysis of the experimental results, the results show that the MYOLOv3 model is more robust to the detection scene in this paper. However, the helmet wearing state detection model proposed in this paper does not consider the influence of weather on the image quality. Rain, snow, and haze weather will quickly degrade the image quality of the camera, and the detection target will also be blurred, which will greatly affect the detection performance of the model. Therefore, future work will consider applying image enhancement techniques for preprocessing in the front-end of the detection model to cope with detection under severe weather conditions. In the future, we will adopt the ideas from other literatures [12-16] to further improve our schemes.

Acknowledgment. This research work is partially supported by Ningbo Science and Technology innovation 2025 major project under Grants No. 2021Z010 and No. 2021Z063, the Public Good Research Project of Science and Technology Program of Zhejiang Province under Grants No.LGG21F020005, NO. LGF21F010002 and NO. LGN21C130001, and the Project for Science and Technology Project of Quzhou, China (NO. 2019K01)

REFERENCES

[1] M. Jin, J. Zhang, X. Chen, Q. Wang, B. Lu, W. Zhou, G. Nie, and X. Wang, Safety helmet detection algorithm based on color and HOG features, *IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing*, pp. 215–219, 2020.

[2] J. Jia, Q. Bao, and H. Tang, Method for detecting safety helmet based on deformable part model, *Application Research of Computers*, vol.33, no. 3, pp.953–956, 2016.

[3] S. Chen, W. Tang, T. Ji, H. Zhu, Y. Ouyang, and W. Wang, Detection of safety helmet wearing based on improved faster R-CNN, *International Joint Conference on Neural Networks*, pp. 1–7, 2020.

[4] F. Wu, G. Jin, M. Gao, Z. HE, and Y. Yang, Helmet detection based on improved YOLO V3 deep model,*IEEE 16th International Conference on Networking, Sensing and Control*, pp. 363–368, 2019.

[5] R. A. G. Vignesh, N. Manohar, and G. Dhyanjith, Helmet detection using single shot detector (SSD), *Second International Conference on Electronics and Sustainable Communication Systems*, pp. 1241–1244, 2021.

[6] R. Silva, K. Aires, T. Santos, K. Abdala, R. Veras, and A. Soares, Automatic detection of motorcyclists without helmet, *XXXIX Latin American Computing Conference*, pp. 1–7, 2013.

[7] R. R. V. e. Silva, K. R. T. Aires, and R. d. M. S. Veras, Helmet detection on motorcyclists using image descriptors and classifiers, *27th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 141–148, 2014.

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

[9] P. Mohan, P. Narayan, L. Sharma, and M. Anand, Helmet detection using faster region-based convolutional neural networks and single-shot multiBox detector, *The 8th International Conference on Smart Computing and Communications*, pp. 209–214, 2021.

[10] C. Vishnu, D. Singh, C. K. Mohan, and S. Babu, Detection of motorcyclists without helmet in videos using convolutional neural network,*International Joint Conference on Neural Networks*, pp. 3036–3041, 2017.

[11] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

[12] S. Kumar, A. Damaraju, A. Kumar, S. Kumari, and C.-M. Chen, LSTM network for transportation Mode Detection, *Journal of Internet Technology*, vol. 22, no.4, pp.891-902, 2021.

[13] L. Kang, R.-S. Chen, N. Xiong, Y.-C. Chen, Y.-X. Hu, and C.-M. Chen, Selecting hyper-parameters of gaussian process regression based on non-inertial particle swarm optimization in Internet of Things, *IEEE ACCESS*, vol. 7, pp. 59504-59513, 2019.

[14] E. K. Wang, S. P. Xu, C.-M. Chen, and N. Kumar, Neural-architecture-search-based multiobjective cognitive automation system, *IEEE Systems Journal*, vol. 15, no. 2, pp. 2918–2925, 2021.

[15] F. Zhang, T.-Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, Application of quantum genetic optimization of LVQ neural network in smart city traffic network prediction, *IEEE Access*, vol. 8, pp.104555–104564, 2020.

[16] Z.-F. Xu, H.-H. Bai, J.-M. Xiao, F.-R. Jie and Y. Zhao, Occluded and tiny face detection network for dense crowd, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 12, no. 4, pp. 162–174, December 2021.