International Journal of Innovative Computing, Information and Control
Volume **18**, Number **3**, June **2022**

ICIC International ©2022 ISSN 1349-4198
pp. **973–988**

SAFETY HELMET WEARING DETECTION BASED ON AN IMPROVED YOLOV3 SCHEME

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

[1]Quzhou Guangming Power Investment Group Co., Ltd.
No. 328, Tonghe Road, Quzhou 324000, P. R. China
{ easteryang; quzhouguzhiwei }@163.com; 858247@qq.com

[2]State Grid Quzhou Power Supply Co., Ltd.
No. 6, Xinhe Road, Quzhou 324000, P. R. China

[3]College of Electrical and Information Engineering
Quzhou University
No. 78, North Jiuhua Road, Quzhou 324000, P. R. China
jxd@qzc.edu.cn

[4]School of Aeronautics and Astronautics
Zhejiang University
No. 38, Zheda Road, Hangzhou 310027, P. R. China
*Corresponding author: zheminglu@zju.edu.cn

Received December 2021; revised March 2022

Abstract. *During the construction process, ensuring construction safety is an important link to improve production efficiency, enhance corporate efficiency, and ensure employee safety. Real-time checking whether workers wear safety helmets is naturally a key task in ensuring safe production. In order to reduce the incidence of safety accidents caused by not wearing a helmet, a helmet wearing detection method based on the improved YOLO (You Only Look Once) v3 algorithm is proposed. The improvement of this paper is reflected in two aspects: one is the improvement of the YOLOv3 algorithm itself, and the other is the improvement of the overall process. Aiming at the improvement of YOLOv3, this paper improves the feature fusion steps of YOLOv3, where we use upsampling to fuse high-level features with low-level features. From the perspective of the overall process, this paper first calibrates the relevant data set, and divides the data set into four categories. For the calibrated data, transfer learning is used to train the improved YOLOv3 network. Then, these parameters and model are used to detect the categories and positions of human figures and helmets on surveillance video data. Finally, we calculate several intersection-over-unions among the four detected types, and use this to judge whether the worker is wearing a helmet correctly. Experimental results show that the proposed algorithm satisfies the real-time performance of the detection task in the helmet wearing detection, and has a higher detection accuracy rate. The mAP (mean average precision) reaches 95.13%, and this detection accuracy rate is higher than that of the traditional SSD (single shot multibox detector), Faster R-CNN (regional-convolutional neural network) and YOLOv3.*

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

DOI: 10.24507/ijicic.18.03.973

1. **Introduction.** With the accelerating process of urbanization and the continuous development of infrastructure construction, construction accidents occur frequently, and the concept of construction safety is becoming stronger. Construction scenarios such as substations, chemical plants, and mine work areas are more complicated and present certain risks. Unsafe behavior of workers can easily lead to accidents, causing casualties and economic losses. In the construction site, safety helmets are the guarantee of life, and the wearing of safety helmets can reduce workers' operational risks to a certain extent. In order to protect the personal lives of workers' safety and reduce the incidence of accidents caused by not wearing a helmet, the research on detecting whether workers wear safety helmets or not has important significance and application value.

Traditional target detection needs to be achieved by manually designing features. This kind of methods has low detection accuracy and is not robust. In recent years, deep learning has gradually gained the favor of scholars by relying on the advantage of convolutional neural network (CNN) in extracting image features without manually designing features [1]. Accordingly, many researchers have proposed a series of target detection algorithms based on deep learning. Girshick et al. [2] used a "region proposal + CNN" in 2014 to replace the "sliding window + manually-designed" features used in traditional target detection, and designed a regional-convolutional neural network (R-CNN). Based on the VOC 2012 data set, the average accuracy of target detection (mAP) was increased by 30% to 53.3%. Girshick [3] and Ren et al. [4] respectively proposed fast regional-convolutional neural network (Fast R-CNN) and faster regional-convolutional neural network (Faster R-CNN), which not only improved the accuracy rate, but also increased the detection speed (the frame rate can reach 5 f/s). In 2016, Redmon et al. [5] proposed the YOLO (You Only Look Once) detection algorithm, which reached the speed (45 f/s) such that it can be used to detect videos. In 2016, Liu et al. [6] proposed the SSD (single shot multibox detector) detection algorithm, which achieved good results in detection accuracy and detection time. At the same time, based on YOLO, Redmon and Farhadi successively proposed YOLOv2 [7] and YOLOv3 [8] detection algorithms. Among them, YOLOv3 has a better detection effect. On the COCO data set, it achieves mAP = 57.9% in 51 ms, which is comparable to RetinaNet in 198 ms. That is to say, the performance is similar but the speed is 3.8 times faster. It can be seen that YOLOv3 can guarantee the accuracy and detection speed at the same time in the field of target detection, and achieve better detection results.

Due to the rapid development of deep learning in the field of target detection, many scholars began to use deep learning technology in practical application scenarios, and achieved good results. In [9], deep convolutional neural networks were used to conduct related research on target detection. In [10], convolutional neural networks were used in the field of pedestrian detection, which improved the accuracy of pedestrian detection. In [11], convolutional neural networks were used to extract features in the field of vehicle recognition in order to achieve rapid vehicle recognition. It can be seen that deep learning methods are currently receiving more and more attention, and combining them with actual application scenarios is a hot research direction at present.

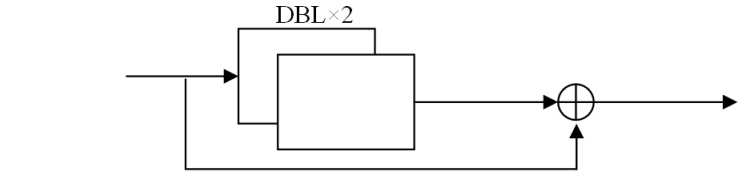
Safety helmet wearing detection is also a type of target detection problems. At present, some scholars have conducted related research on the automatic identification technology of safety helmets. In the traditional detection methods of safety helmets, Waranusast et al. [12] used machine vision related methods to conduct in-depth research on automatic identification of safety helmets. Jin et al. [13] used the SVM trained by HOG features to detect the safety helmet and ultimately to realize the judgment of workers wearing a safety helmet. Li et al. [14] proposed an innovative and practical safety helmet wearing detection method based on image processing and machine learning. In [15], a hybrid descriptor for

features extraction was proposed based on local binary pattern, histograms of oriented gradients and the Hough transform descriptors for automatic detection of motorcyclists without helmet. Hu and Wang [16] put forward the helmet recognition neural network model based on the analysis of wavelet transform and BP neural network in the helmet recognition. In terms of helmet detection based on deep learning methods, Chen et al. [17] 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. [18] adopted 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. [19] demonstrated a novel implementation of the Faster R-CNN and SSD framework for accurate helmet detection in real-time low-quality surveillance videos. Patrik et al. [20] concentrated on the modeling and implementation of object detection and navigation system for quadcopter drone.

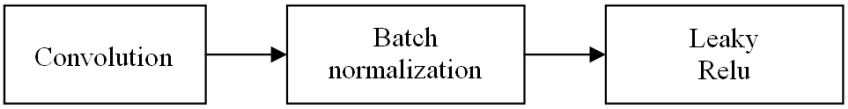
Most of above traditional helmet detection methods have problems such as low accuracy and high environmental requirements. At the same time, the detection speed is slow and cannot meet the real-time detection requirements in the production environment. Two-stage detection methods are commonly used in existing deep learning helmet detection research. This type of detection methods has two obvious disadvantages: first, the cumbersome detection steps lead to a large amount of model calculation and low detection efficiency; second, the accumulation of errors caused by this method will greatly reduce the detection accuracy of the model, and the detection of helmets. The accuracy is almost completely limited by the detection accuracy of the human body or face. In summary, for the task of helmet recognition, there is currently a lack of a highly robust classification algorithm. Aiming at the video recognition of safety monitoring measures for electric power construction sites, which requires high accuracy while meeting high real-time performance, an improved method using the YOLOv3 algorithm is proposed to identify construction site surveillance videos and realize the detection of construction personnel in the surveillance videos whether to wear safety equipment or not. In the YOLOv3 algorithm, the feature maps of three scales output by the feature extraction network Darknet-53 are used for feature fusion. The feature maps of these three scales correspond to different levels of feature information, and the span between these feature maps is very large. After going through multiple convolutional layers, the feature information of the previous scale is seriously lost, which in turn will lead to the problem of poor feature fusion quality. In response to this problem, this paper improves the feature fusion steps of YOLOv3, where we use upsampling to fuse high-level features with low-level features. This paper uses the construction site surveillance video data, carries on corresponding data processing to it, produces the helmet wearing detection data set, and divides the data set into four categories overall. For the calibrated data, the improved YOLOv3 network is trained using transfer learning. Then, we use the parameters and model to detect the types and positions of human figures and helmets on surveillance video data. Finally, we calculate the relevant IoU (intersection over union, which is defined as the ratio of the area of overlap to the area of union) among the four types of detection, and use this to judge whether the worker is wearing a helmet correctly. The algorithm innovatively combines the improved YOLOv3 network model and the IoU algorithm, and then refines the category and location information output by the network, which improves the accuracy of detection and reduces the false recognition rate. The test results show that the algorithm can meet the real-time requirements in the detection of wearing helmets, and at the same time can accurately detect persons who are not wearing helmets and notify relevant personnel.

2. **YOLOv3 Network Structure Model.** YOLO network is designed for the purpose of target detection. Due to the end-to-end characteristics of YOLO itself, its calculation speed is faster than most deep networks and has good real-time performance. In addition, the network can maintain a high accuracy rate under the premise of good robustness. Based on the above two advantages, it is considered to meet the accuracy and real-time requirements of most industrial inspection scenes, and has good industrial application prospects, so it has been widely studied in recent years. The YOLOv3 network is the version with better performance after several improvements based on the original YOLO network.

2.1. **Feature extraction network Darknet-53.** YOLOv3 uses a new network to implement feature extraction. The structural unit of Darknet-53 is shown in Figure 1. The YOLO network only uses convolutional layers and is a fully convolutional network. This speeds up the speed of the network while reducing the parameter variables. Compared with other neural networks, the YOLO series neural network has successfully transformed the target detection problem into a regression problem through a reasonable design, and thus directly generates the location and category information of the object through the network. For other mainstream networks, most of them need to reprocess images that have been processed and output by neural networks. The YOLOv3 main convolutional network is based on the structure of Darknet-53 [8]. It can be seen from Figure 1 that the Darknet-53 structure is built by 53 convolutional layers. In order to prevent the disappearance of gradients and the explosion of gradients, a residual unit is added in the Darknet-53 network, which allows the network to train with deeper layers. Because too many residual units will lead to other undesirable results, Darknet-53 chose to add five residual units. The construction of each residual unit is shown in Figure 1(a). The original input of the upper layer is not only input to the lower layer through two DBL (Darknetconv2d_BN_Leaky, BN: batch normalization) units, but also skips the DBL unit directly to the lower layer. In other words, the lower layer will receive the original upper layer data and the processed upper layer data. In this way, a residual unit is constructed. The DBL unit structure includes a convolutional layer, a batch normalization layer, and a leaky Relu activation function layer, with a total of 3 layers.



(a) Residual unit



(b) DBL unit

FIGURE 1. Darknet-53 structure units

From the point of view of the network structure, compared to the Darknet-19 network in YOLOv2, the residual unit is added, using continuous 3×3 and 1×1 convolutional layers. YOLOv3 expands it to 53 layers and calls the network Darknet-53 [8]. The network structure contains 53 convolutional layers and 5 maximum pooling layers. At the same time, batch normalization and dropout operations are added after each convolutional layer

to prevent over-fitting. The network is more powerful than Darknet-19 and more effective than ResNet-101 or ResNet-152. The performance test results under the ImageNet data set are shown in Table 1. It can be seen from Table 1 that Darknet-53 has a better detection effect than ResNet-101, and the speed is increased by 1.5 times. Darknet-53 has similar performance than ResNet-152, and the speed is increased by 2 times. Compared with Darknet-19, the detection accuracy has been greatly improved. Although the speed is not as fast as Darknet-19, it still meets the real-time requirements.

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/10^9	7.29	19.70	29.40	18.70
BFLOP/s	1246	1039	1090	1457
Recognition frame rate/(f·s^{-1})	171	53	37	78

2.2. **Design ideas of YOLOv3.** YOLOv3 has made some improvements compared to the Darknet-53 architecture. The overall architecture of YOLOv3 is shown in Figure 2. 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. The output of the last 3 downsampling can be understood as the feature maps that contain 3 scale target detections. When the scale is larger, the extracted feature map is smaller. These feature maps of different sizes also have their own functions: small feature maps provide in-depth semantic information, and large feature maps provide target location information. In addition, a path is prepared in the YOLOv3 network in advance, so that the small feature map can be up-sampled and merged with the large feature map, so that the large feature map contains the feature information of the small feature map. Even if the detection target set by the model is larger, the loss of small features in the image is less. Therefore, the YOLOv3 network has a good positioning effect for targets of different sizes.

YOLOv3 uses multiple scale fusion methods to make predictions. Using a similar feature pyramid network fusion approach, position and category predictions are performed on feature maps of multiple scales, which improve the accuracy of target detection. In YOLOv3, dimensional clustering is also used as a priori box to predict the bounding box. Use the k-means method to cluster the target blocks in the data set to obtain 9 prior blocks of different sizes, and divide them into multiple scale feature maps. The larger scale feature maps use smaller ones. This scheme can finally obtain more priori boxes than YOLOv2, and the feature extraction effect is better. In category prediction, YOLOv3 does not use softmax to classify each box, but uses multiple independent logical classifiers. During the training process, binary cross-entropy loss is used for category prediction.

Based on the above design ideas, compared with other target detection algorithms, YOLOv3 has achieved better results in detection accuracy and speed. The performance comparison between YOLOv3 network and other target detection frameworks is shown in Table 2.

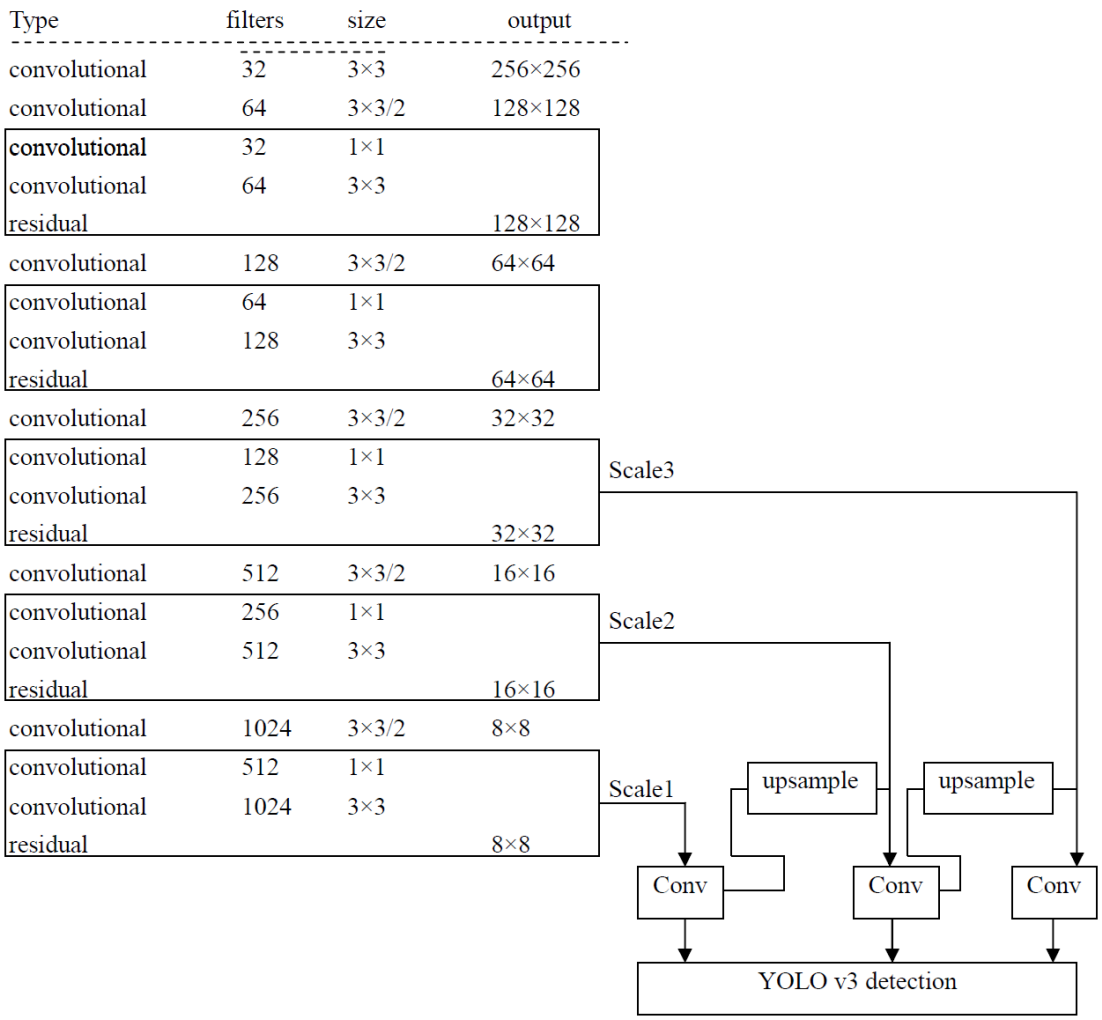


FIGURE 2. YOLOv3 overall structure

TABLE 2. Performance comparison of YOLOv3 and other networks

Method	mAP-50/%	Time/ms
SSD321	45.4	61
DSSD321	46.1	85
R-FCN	51.9	85
SSD513	50.4	125
DSSD513	53.3	156
FPN FRCN	59.1	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-50-800	57.5	198
YOLOv3-320	51.5	22
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

3. **Proposed Scheme.** In the YOLOv3 algorithm, the feature maps of the three scales output by the feature extraction network Darknet-53 are used for feature fusion. The feature maps of these three scales correspond to different levels of feature information, and the

span between these feature maps is very large. After going through multiple convolutional layers, the feature information of the previous scale is severely lost, which in turn will lead to the problem of poor feature fusion quality. On the other hand, YOLOv3 is a fully convolutional network. The convolutional layer used in the feature extraction part needs a lot of sample training to extract deep features that are easier to classify. It is foreseeable that if you train the YOLO network directly with this data set, it will inevitably be difficult to obtain generalized results. Based on these two problems, we propose the improved schemes in two aspects: one is the improvement of the feature fusion step in YOLOv3 algorithm, and the other is the improvement of the overall process including the training process with transfer learning. In detail, our contribution lies in four aspects. The first aspect is that the image pyramid structure is first used to fuse features at different levels to obtain feature maps of different scales for location and category prediction, and the number of anchor boxes is increased from five to nine, so that the model can obtain more object edge information. The second aspect is that in the training process, multi-size images are used for training, so that the model can adapt to images of different resolutions. The third aspect is that the relevant data sets are calibrated, and the data sets are divided into four categories. For the calibrated data, the YOLOv3 network is trained by the training method of transfer learning. The fourth aspect is that the YOLOv3 network model is innovatively combined with the IoU algorithm. The classification and location information are further refined, which improves the detection accuracy and reduces the misrecognition rate.

3.1. **Improved feature fusion method.** First, in order to better understand the unique scale transformation process of YOLOv3, the principle of downsampling of the YOLO network is explained with examples. When the input image size is (416×416) pixels, if you downsample 5 times, then the 3×3 convolution kernel is moved 5 times with a step size of 2, so the image size becomes $1/32$ that of the original, and a feature map with a size of 13×13 is finally obtained in the detection layer. Similarly, when down-sampling 4 times and 3 times, the image becomes smaller by 16 and 8 times respectively, and the detection layer obtains feature maps with sizes of 26×26 and 52×52, respectively. When these large-scale images are restored upwards, the corresponding units can be repeated several times. In this way, although it seems that the shallow pixel is lost, it does not actually lose the deep feature information obtained at a large scale. Through the above algorithm, YOLOv3 uses different feature information obtained at multiple scales to detect targets of different sizes, and then can detect large and small objects at the same time.

The feature maps of different scales pass through the bottom detection layer of YOLOv3, and the output is a final feature map containing multi-scale information. In the final feature map, the grid cell of each size contains the data of $[B \times (5 + C)]$. Among them, B is the number of prediction boxes in each unit, which is used to predict a certain type of specific object. Each prediction box has 5 data (i.e., x, y, w, h, $S + C$). Among them, (x, y) is the offset of the center of the prediction box relative to the cell, (w, h) is the width and height offset of the prediction box relative to the corresponding anchor point, and S is the confidence value, and $S = P_{\text{object}} \times U$, where P_{object} is the probability of the object in the predicted box, 1 is the existence, and 0 is the absence; U is the intersection ratio of the predicted box and the real box (intersection over union, IOU), C is the conditional probability of C categories $P_{\text{class|object}}$. Generally, the number of prediction frames in each YOLO network unit is $B = 3$. Finally, the regression algorithm is used for these prediction boxes, and the judgment results are rationalized through non-maximum suppression (NMS) to complete the target detection task.

In addition, the YOLOv3 algorithm uses an anchor box to improve the accuracy of detection. Therefore, this method is also used in our helmet identification task. This part of the content learning method is somewhat different from the image feature learning in the previous two paragraphs, so it will be explained at the end. In order to reasonably determine the initial size of the initial anchor box, this paper uses the k-means clustering method, and uses the marked bounding box size in the training sample as the sample to determine the anchor box size. Select k clustering centers ($k = 9$), and use equation $d_{\text{bbox,kmeans}} = 1 - U_{\text{bbox,kmeans}}$ as the distance function for clustering. Among them, $U_{\text{bbox,kmeans}}$ represents the intersection ratio of bounding box and cluster size.

YOLOv3 uses the Darknet-53 network to extract features. The low-level features have rich details and positioning information, and the high-level features have rich semantic features. From low-level to high-level, detailed information is continuously reduced, and semantic information is constantly increased. For position prediction, more low-level feature information is needed. For category prediction, more high-level local information is needed. Therefore, an image pyramid-based model can be considered, using upsampling to combine high-level features with low-level features. Fuse to obtain feature maps of different scales for location and category prediction. The feature pyramid on the right in Figure 3 is generated by the feature pyramid on the left. The whole process is first deep convolution of the input image, then convolution operation on the features on layer L2, and upsampling on the features in layer L4. Make it the same size, then perform convolution and operation on the processed layer L2 and layer L4, and input the obtained result into layer L5. In the same way, feature fusion is performed on multiple layers accordingly to obtain multiple sets of feature maps for prediction. Based on this scheme, the processed low-level features and high-level features are accumulated. The purpose of this scheme is to provide more accurate location information because the low-level features can cause errors in the positioning information of the deep network due to multiple down-sampling and up-sampling operations. Therefore, they are combined and used to construct a deeper feature pyramid, which integrates multi-layer feature information and makes predictions on different feature maps.

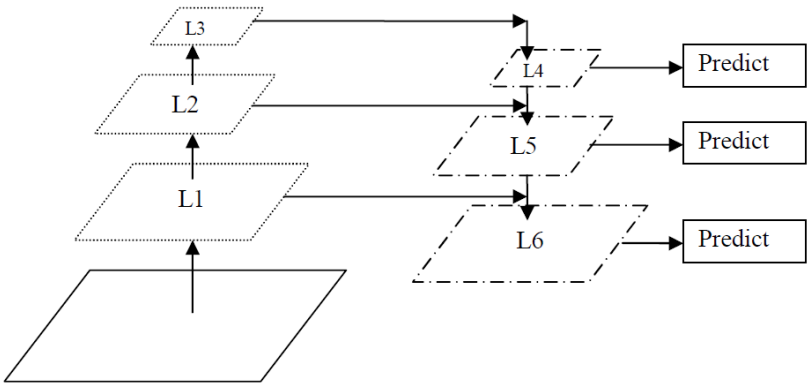


FIGURE 3. Feature fusion pyramid structure

Based on the above feature fusion ideas, we improved the YOLOv3 algorithm, using upsampling to fuse high-level features with low-level features, and finally 3 sets of feature maps were obtained, and these 3 sets of feature maps of different scales were used for prediction. The improved network structure is shown in Figure 4. The details of the specific network structure improvement are as follows: First, the feature pyramid is obtained through Darknet-53, and the conv53 layer is subjected to continuous 1×1 and 3×3 convolution operations to obtain a set of YOLO layers to be processed, and then perform

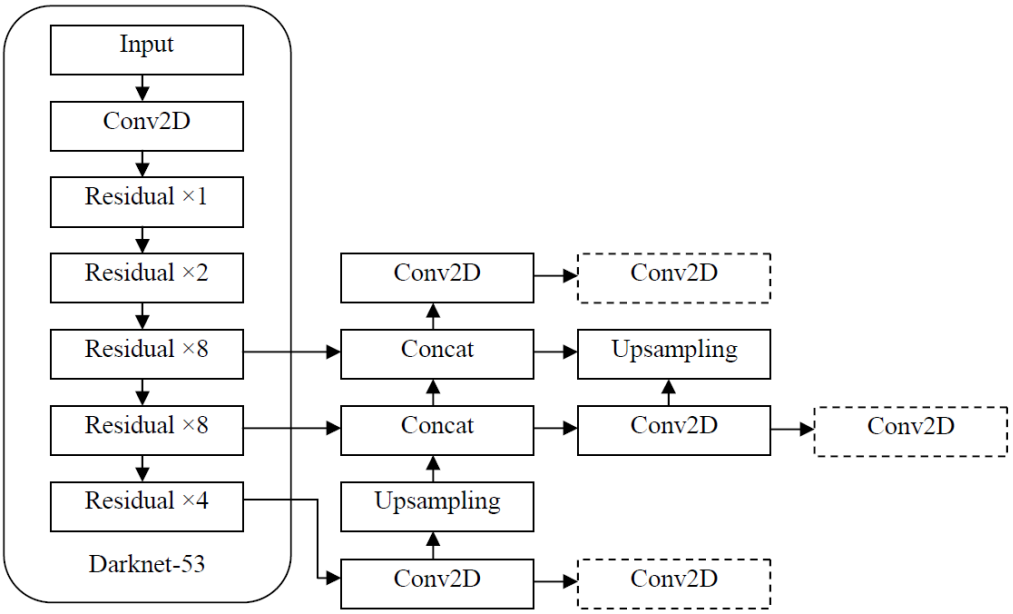


FIGURE 4. Modified network structure

a set of 1×1 and 3×3 convolution operations to obtain a small-scale YOLO layer; at the same time, this layer is up-sampled, convolutions are performed with the conv45 layer in Darknet-53, and continuous 1×1 and 3×3 convolution operations are also used. Get the second set of to-be-processed YOLO layers, perform a set of 1×1 and 3×3 convolution operations on this layer to get the mesoscale YOLO layer; at the same time, upsample this layer to the conv29 layer in Darknet-53 to perform convolution, also use successive 1×1 and 3×3 convolution operations to obtain the third set of YOLO layers to be processed, and perform a set of 1×1 and 3×3 convolution operations on this layer to obtain large-scale YOLO layer. After the above operations, 3 sets of YOLO feature layers of different scales are obtained, and these 3 sets of feature layers are used for positioning and category prediction.

3.2. **Use of transfer learning to train the network.** If each network is trained from beginning to end until it can be applied to practical engineering tasks, then two conditions are necessary. One is a large number of training data sets, and the other is a validation set with no entry and back propagation. The former is to train the network performance to obtain deeper and more effective image features; the latter is to prevent over-fitting and avoid the situation that the accuracy of the training set becomes higher and the accuracy of the actual task becomes lower.

In order to apply the YOLOv3 network to the task of helmet recognition, the first step is to obtain the above two types of data. In this study, the original images were obtained by calling the on-site monitoring video of the workers' operations in the factory; then, the data set of the identification and detection of helmet wearing was established by the way of manual annotation. However, due to the time-consuming manual labeling of information, coupled with fewer original valid images, there are a total of 1328 images in the labeled data set. Therefore, even if the data set is not divided into training set and test set but all applied to training, the required training data set is far from enough. Not only that YOLOv3 is a fully convolutional network, but also the convolutional layer used for feature extraction requires a lot of sample training to extract deep features that are easier to classify. It is foreseeable that if the YOLO network is trained directly with this data set, it will inevitably be difficult to obtain generalized results.

In response to the above problems, this paper uses the transfer learning method to train the convolutional neural network. Since YOLOv3 has tested thousands of classification tasks on Image-Net, each node of the YOLOv3 network has actually been parameter training. Therefore, the classification data set of ImageNet can be used as the source domain. First, load these parameters at each parameter point of YOLOv3. This is because after training the classification network using the ImageNet data set, the weights of the convolution kernel have been trained to have the ability to extract generalized features. Then, use the fine-tuning method of pattern recognition to freeze most of the network layer, and only enable back propagation for the last few layers (especially the last used to convert the feature vector into the probabilistic output softmax layer) to update the data of the node parameters. This is to allow the last few deep convolutional layers to extract deep features under a specific task, so that YOLOv3 can be applied to this task. For this research, the desired feature information is the computer vision features of the helmet and its related pixels. After using transfer learning, a part of the data set can be divided into the verification set to prevent the occurrence of over-fitting. After using transfer learning, a large number of training sets are no longer needed to train the network performance to meet the requirements. In this task, the training set has 1000 sample sets, which are used to train the convolution kernel parameters of the regression prediction part; the test set has 328 samples, which do not enter the back propagation but directly enter the network calculation results for objective evaluation of the actual performance.

3.3. **Design of safety helmet detection scheme.** This paper uses the labelImg tool to calibrate the data set into four categories, namely, category A, category B, category C and category D during the stage of preparing the data set before the experiment. Type A represents the upper part of the human body, type B is the avatar wearing a helmet correctly, type C is the avatar wearing other hats, and type D is the avatar without a hat. Put the calibrated data set on the server, use the above method to train the improved YOLOv3 network, and finally get the model parameters of the network. In the test phase, the frame image to be tested is taken from the video image every N seconds, the image to be tested is preprocessed, the data format of the image to be tested is converted, and the image size is adjusted; the preprocessed image is loaded into the improved YOLO model that has been trained, and the model outputs type A, type B, type C and type D. The algorithm flow is shown in Figure 5.

4. Experimental Results.

4.1. **Test data set.** For deep learning, the test data set is an important condition for the effectiveness of the algorithm. However, the data set for helmet detection does not currently exist in the public data set. Therefore, this paper uses factory surveillance video to create a data set of safety helmets, which mainly includes the following steps. 1) Data collection: The data mainly comes from on-site surveillance videos of workers in the factory, as well as head images without helmets. Among them, on-site surveillance video in the factory is the main focus. 2) Data preprocessing: The factory area uses a high-definition video camera, so the data preprocessing needs to use the OpenCV video development library to convert the video files into the PNG format of the image. The converted image is used as an image data set. The uniform size of the image is (640×480) pixels. 3) Data calibration: Use labelImg data calibration tool to mark the pictures of the helmet wearing detection database. The tool needs to manually mark out the custom target of the picture, and finally can generate the relevant configuration file according to the input mark information.

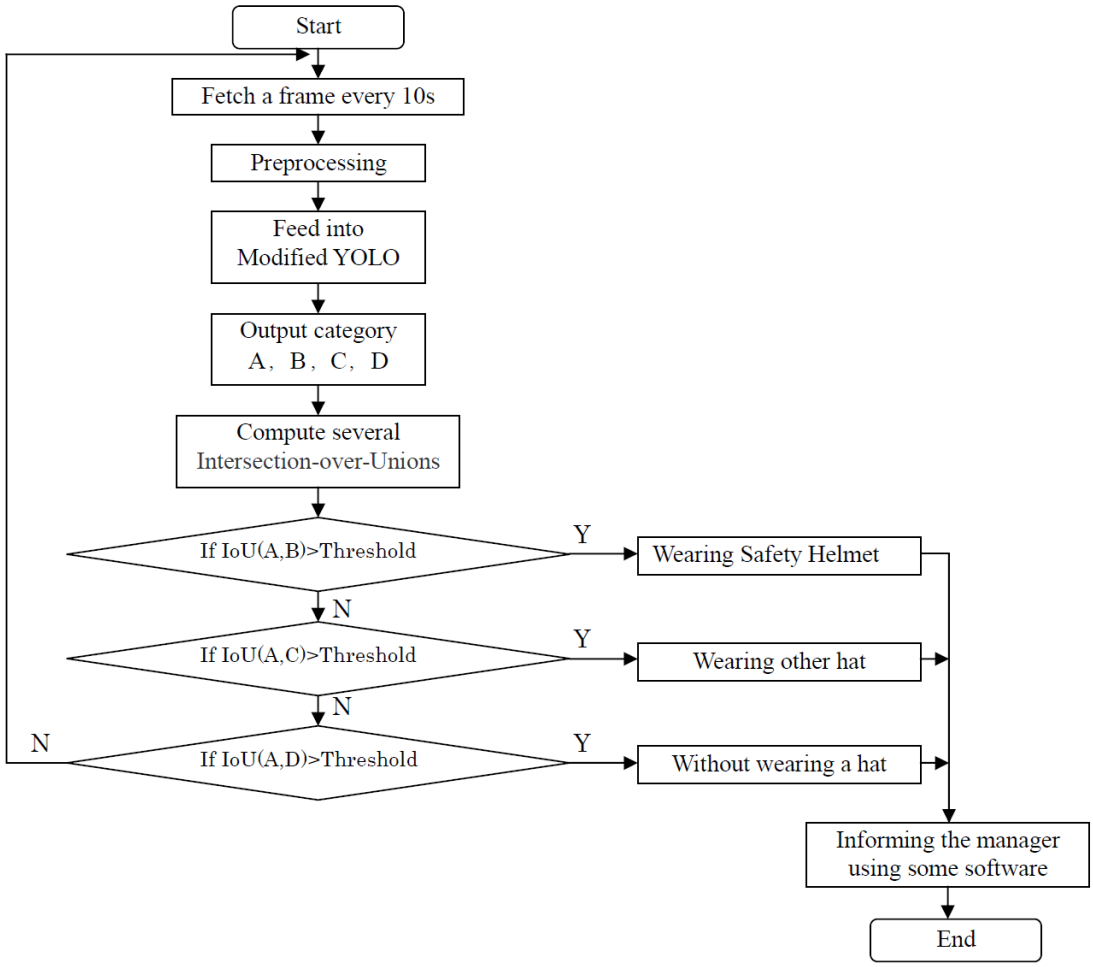


FIGURE 5. The algorithm flow

This test is divided into model training phase and detection phase. The model training stage has very high requirements on the hardware environment, and the GPU server of GTX3080Ti is used for calculation. The detection stage also has certain requirements for the hardware environment: GTX1060 GPU desktop is used, and an environment for testing is built, including common environments such as Ubuntu16, CUDA, python, OpenCV, and gcc. YOLOv3 uses the Darknet-53 framework. Use the weight parameters provided on the YOLOv3 official website as the initialization parameters of the network training, and fine-tune the network parameters of the images in the calibrated helmet wearing detection training data set, so that the entire network detection can achieve the best results.

4.2. **Generating anchor boxes.** In the YOLOv3 algorithm, the anchor boxes are used to roughly determine the target location and size so that the subsequent algorithm can perform regression refinement. Therefore, the quality of the anchor frame directly affects the effect of target detection. This paper uses the k-means clustering algorithm to perform 9 clustering on 1328 image data to determine the most suitable anchor box for the current data set, and then assign it to the feature map of the corresponding scale according to the size of the anchor box. In this paper, we adopt $k = 9$ based on detailed experimental results. Figure 6 shows the relationship between the number of clusters and IoU, and the two curves represent the test results of the VOC and COCO data sets, respectively. Finally, combined with the impact of different k values on the recall rate and the complexity

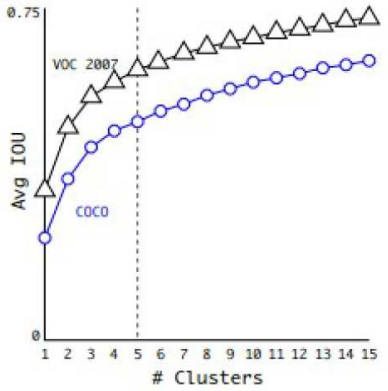


FIGURE 6. The relationship between the number of clusters and IoU

of the algorithm, we choose $k = 9$. The results are $(59, 22)$, $(68, 30)$, $(75, 35)$, $(88, 37)$, $(90, 40)$, $(99, 45)$, $(108, 54)$, $(119, 57)$, $(128, 63)$. We arrange their areas from small to large, and divide them into 3 feature maps of different scales. Among them, the feature maps with larger scales use smaller a priori boxes. Finally, this cluster center will be used for helmet wearing detection experiments.

4.3. **YOLO model training.** This paper divides the self-built data set into a training set and a test set with a total of 1328 images at a ratio of 8 : 2. Perform back propagation transfer learning on the training set. The resolution of the data is the same as that of the self-made data set, both of which are (640×480) pixels. The detection categories include Type A, Type B, Type C and Type D. The gradient descent method is the stochastic gradient descent method, the momentum is 0.9, the weight attenuation is 0.0005, and the 4×16 small batch training method is used. The learning rate starts from 0.001. The entire training process is a total of 10,000 iterations, and the learning rate is adjusted as needed. The learning rate is reduced by 10 times when the iteration reaches 80% (8000) and 90% (9000). The calibrated data set is trained on the server using the improved YOLOv3 network, and finally the model parameters of the network are obtained. In order to compare the effect with the original algorithm, the original YOLOv3 model was trained in the same environment.

4.4. **Analysis of experimental results.** In this paper, when the intersection ratio of the target bounding box predicted by the YOLO neural network and the manually calibrated bounding box is greater than or equal to 0.5, the task prediction is considered to be successful. This paper mainly uses the improved YOLOv3 algorithm for experiments. The improvements are multi-scale feature fusion and transfer learning training. At the same time, the SSD in [6], the Faster R-CNN in [4] and the YOLOv3 in [8] are used for comparison. This paper uses mean average precision (mAP), and frame per second (FPS) commonly used in the field of target detection as evaluation indicators to evaluate the algorithm's performance on helmet wearing detection.

To explain how to calculate the mAP, we give a simple example. Assume there are two categories, the first category has four images, and the second category has five images. Based on a certain threshold (In general, IoU = 0.5), for these nine images, a certain algorithm can classify four images from Category 1 into Category 1, whose ranks are 1, 2, 4, 7, and can classify three images from Category 2 into Category 2, whose ranks are 1, 3, 5. Thus, for Category 1, the average precision (AP) is

$$\mathrm{AP}_1 = \frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{4} + \frac{4}{7}}{4} = 0.83 \tag{1}$$

and for Category 2, the average precision (AP) is

$$\mathrm{AP}_2 = \frac{\frac{1}{1} + \frac{2}{3} + \frac{3}{5} + 0 + 0}{5} = 0.45 \tag{2}$$

Therefore, we can get the mAP as follows

$$\mathrm{mAP} = \frac{\mathrm{AP}_1 + \mathrm{AP}_2}{2} = \frac{0.83 + 0.45}{2} = 0.64 \tag{3}$$

In our paper, we test on the test set when IoU is set to 0.5, and the results are shown in Table 3.

Table 3. Comparison results of different algorithms

Algorithm	mAP/%	Frame per second
SSD	86.15	41
Faster R-CNN	94.62	0.4
YOLOv3	91.24	56
Our improved YOLOv3	95.13	60

It can be seen from the experimental results that due to the introduction of feature fusion, the algorithm in this paper increases the amount of model calculations, and its operating efficiency is slightly lower than that of YOLOv3, but in terms of detection accuracy, the detection accuracy is greatly improved compared to YOLOv3. The improved YOLOv3 algorithm has a highest mAP value of 95.13%, and Faster R-CNN has the second highest mAP value, reaching 94.62%. Not only the detection accuracy rate is better than Faster R-CNN, but the detection speed is 150 times faster than it. Therefore, for real-time detection tasks, the improved YOLOv3 performs better in terms of performance. The SSD algorithm and the original YOLOv3 algorithm are inferior to the improved YOLOv3 algorithm in mAP and recognition frame rate. It can be seen that the improved YOLOv3 algorithm takes account of the detection accuracy and detection rate at the same time, and can better implement the task of wearing a helmet.

In addition, in order to more intuitively experience the detection differences between our algorithm and the original YOLOv3, this paper selects some detection images for comparative analysis. Figure 7 shows the detection results of the original YOLOv3, and Figure 8 shows the detection results of our improved YOLOv3. It can be seen that the improved YOLOv3 algorithm can all correctly detect the target category, but YOLOv3 mistakenly detects some hats. It can be seen that the YOLOv3 algorithm has missed the detection of small distant targets, and the improved YOLOv3 algorithm performs better for multi-target detection. For scenes with multiple targets and small targets in the image, it can be seen that the YOLOv3 algorithm before the improvement has missed detection of small targets in the distance, while the improved YOLOv3 algorithm performs better for multi-target detection. Furthermore, for scenes with partial occlusion in the image, it can be seen that the pre-improved YOLOv3 algorithm misses detection of some occluded small targets, while the improved YOLOv3 algorithm does not miss detection. In general, the improved YOLOv3 is better than the original YOLOv3 algorithm in terms of target accuracy. Because Faster R-CNN needs to build an RPN network in the target detection process, it involves a lot of calculations, so the detection speed is not as good as the improved YOLOv3. Therefore, for helmet wearing detection tasks, the improved YOLOv3 algorithm proposed in this paper can meet the real-time detection requirements while maintaining a high detection rate. In the actual test, the algorithm can detect that the target is not wearing a helmet or wearing other hats, with an accuracy rate of more

FIGURE 7. Some detection results of the original YOLOv3 algorithm

FIGURE 8. Some detection results of the improved YOLOv3 algorithm

than 95%. At the same time, in the actual test, the improved YOLOv3 helmet detection algorithm is also effective for multi-target detection.

5. **Conclusions.** This paper presents a method for detecting helmet wearing based on the improved YOLOv3 algorithm. Use construction site entrance and exit surveillance video as a data set to carry out helmet wearing detection experiments, and improve the YOLOv3 network by adopting methods such as image pyramid-based multi-scale feature detection and transfer learning. At the same time, it can still have a faster detection speed, which basically meets the accuracy and real-time requirements of the safety helmet wearing detection in the monitoring video of the working environment. The main shortcoming is

that our algorithm cannot confirm if the worker correctly wears the helmet, although his helmet is on his head. Future work will be concentrated on further detecting the wearing status of the helmet on head.

Acknowledgement. This research work is partially supported by Ningbo Science and Technology innovation 2025 major project under Grants No. 2021Z010 and No. 2021Z063, and the Public Good Research Project of Science and Technology Program of Zhejiang Province under Grant No. LGG21F020005.

REFERENCES

[1] Y. Tian, Artificial intelligence image recognition method based on convolutional neural network algorithm, *IEEE Access*, vol.8, pp.125731-125744, 2020.

[2] R. Girshick, J. Donahue, T. Darrell and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.580-587, 2014.

[3] R. Girshick, Fast R-CNN, *Proc. of the IEEE International Conference on Computer Vision*, pp.1440-1448, 2015.

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

[5] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You Only Look Once: Unified, real-time object detection, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.779-788, 2016.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg, SSD: Single shot multibox detector, *European Conference on Computer Vision*, pp.21-37, 2016.

[7] J. Redmon and A. Farhadi, YOLO9000: Better, faster, stronger, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.6517-6525, 2017.

[8] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.89-95, 2018.

[9] J. Tian and J. Hu, Image target detection based on deep convolutional neural network, *International Conference on Communications, Information System and Computer Engineering*, pp.461-464, 2019.

[10] F. Ahmed, B. A. Topu and S. M. M. Islam, HOG and Gabor filter based pedestrian detection using convolutional neural networks, *International Conference on Electrical, Computer and Communication Engineering*, pp.1-6, 2019.

[11] W. J. San, M. G. Lim and J. H. Chuah, Efficient vehicle recognition and classification using convolutional neural network, *IEEE International Conference on Automatic Control and Intelligent Systems*, pp.117-122, 2018.

[12] R. Waranusast, N. Bundon, V. Timtong, C. Tangnoi and P. Pattanathaburt, Machine vision techniques for motorcycle safety helmet detection, *The 28th International Conference on Image and Vision Computing New Zealand*, pp.35-40, 2013.

[13] 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.

[14] J. Li, H. Liu, T. Wang, M. Jiang, S. Wang, K. Li and X. Zhao, Safety helmet wearing detection based on image processing and machine learning, *The 9th International Conference on Advanced Computational Intelligence*, pp.201-205, 2017.

[15] 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.

[16] T. Hu and X. Wang, Analysis and design of safety helmet recognition system based on wavelet transform and neural network, *Software Guide*, no.23, pp.37-39, 2006.

[17] 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.

[18] 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.

[19] 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.

[20] A. Patrik, G. Utama, A. A. S. Gunawan, A. Chowanda, J. S. Suroso and W. Budiharto, Modeling and implementation of object detection and navigation system for quadcopter drone, *ICIC Express Letters*, vol.13, no.6, pp.461-468, 2019.

Author Biography

Wei Yang received the Bachelor's degree of Computer Science and Technology from Northeast Electric Power College (now Northeast Electric Power University), China, 2003; the Master's degree of Electrical Engineering from Zhejiang University, China, 2011.

Mr. Yang is currently a full-time senior engineer at State Grid Quzhou Power Supply Company, China. His main research interests include Application of Technology Informatization in Power Systems, Smart Grid, etc. He is an Enterprise Level 1 Human Resources Teacher.

Guang-Le Zhou received the Bachelor's degree of Electrical Engineering and Automation from Chongqing University, China, 2006; the Master's degree of Electrical Engineering from Zhejiang University, China, 2016.

Mr. Zhou is currently engaged in science and technology management and innovation research and development in Quzhou power supply company of State Grid Zhejiang Electric Power Company (Futeng technology branch of Quzhou Guangming Power Investment Group Co., Ltd.).

Zhi-Wei Gu received the Bachelor of Engineering degree in Computer Science and Technology from Wuhan University, China, 2002; the Master's degree in Electrical Engineering from Zhejiang University, China, 2012.

Mr. Gu is currently a full-time senior engineer at State Grid Quzhou Power Supply Company, China. His main research interests include Application of Technology Informatization in Power Systems, Smart Grid.

Xiao-Dan Jiang received the Bachelor of Engineering degree in Computer Science and Technology from Chongqing Technology and Business University, China, 2003; the Master's degree in Computer Technology from Zhejiang University of Technology, China, 2010; the Ph.D. student in Zhejiang University of Technology, Now.

Ms. Jiang is currently a full-time associate professor at the College of Electrical and Information Engineering, Quzhou University, China. Her main research interests include digital watermarking and image processing. She has published over 20 papers in journals and conferences. She is hosting some research projects funded from the Project for Public Interest Research Projects of Science and Technology Program of Zhejiang Province, China, etc.

Zhe-Ming Lu received the B.S. degree in Electrical Engineering from Harbin Institute of Technology, China, 1995; the M.S. degree in Electrical Engineering from Harbin Institute of Technology, China, 1997; the Ph.D. degree in Instrument Science and Technology from Harbin Institute of Technology, China, 2001.

Prof. Lu is currently a full-time professor at the School of Aeronautics and Astronautics, Zhejiang University, China. His research interests include multimedia single analysis and processing, information hiding and astronautics signal processing, etc. He has published over 300 papers in journals and conferences.