Research on Safety Helmet Detection Algorithm of Power Workers Based on Improved YOLOv5

Desu Fu1, Lin Gao1*, Tao Hu1, Shukun Wang1, Wei Liu1

1School of Information Engineering, Hubei MinZu University, Enshi, Hubei, China

*email: 2003001@hbmzu.edu.cn

Abstract. The traditional helmet detection algorithm in power industry has low precision and poor robustness. In response to this problem, the helmet detection algorithm based on improved YOLOv5 (You only look once) is put forward in this paper. Firstly, the YOLOv5 network structure is improved. By increasing the size of the feature map, one scale is added to the original three scales, and the added 160*160 feature map can be used for the detection of small targets; Secondly, the K-means is used for re-clustering the helmet data set to get more suitable priori anchor boxes. The experimental results illustrate that the average accuracy of the improved YOLOv5 algorithm is increased by 2.9% and reaching 95% compared with the initial model, and the accuracy of helmet recognition is increased by 2.4% and reaching 94.6%. This algorithm reduces the rates of missing detection and misdetection of small target detection in original network, and has strong practicability and advanced nature. It can satisfy the requirements of real-time detection and has a certain role in promoting the safety of power industry.

1. Introduction

In the working process of power workers, our first impression of them is that they are wearing safety helmets, whether it is in sunny, rainy or snowy. If the power workers do not wear safety helmets during operation, they may be hit by objects falling from above, hurt their head due to falling from a height, or their heads may suffer from electric shock. Therefore, safety helmet is the safety guarantee for workers in power industry.

Power workers must wear safety helmets to enter the operation area, but manual supervision is time consuming and laborious, and there are risks in close range supervision in some work scenarios. So the intelligent real-time safety helmet detection system of power workers is particularly important. It can not only realize the automation and digitization of safety supervision and monitoring, but also improve the safety of power workers, which has practical development significance.

The development of target detection technology is divided into two periods, which can be called the traditional detection period and the deep learning-based detection period [1]. In the traditional target detection period, VJ (Viola-Jones) face detector [2], HOG + SVM (Histogram of oriented gradient + Support Vector Machine) algorithm [3] and DPM (Deformable Part Model) algorithm [4] are the representations. For example, in 2014, Liu Xiaohui [5] combined SVM and skin color detection to Identify helmets. The detection period based on deep learning is represented by R-CNN (Region-Convolutional Neural Networks) [6], Fast R-CNN [7], Faster R-CNN [8], SPP-Net (Spatial Pyramid Pooling-Net)[9], YOLO [10], SSD (Single Shot MultiBox Detector)[11] algorithm, etc. These algorithms are divided into two-stage and one-stage. The former mainly includes R-CNN, SPP-Net, Fast R-CNN and Faster R-CNN, and the latter mainly includes YOLO, SSD, etc. In the two-stage algorithm, the first-level network extracts features from the candidate area, and the
second-level network classifies and accurately regresses the selected area. The detection accuracy is high, but the running speed is slow. In the single-stage algorithm, the tasks of classification and regression can be completed only by the first-level network, and the step of candidate regions is not required. The running speed is fast and the detection accuracy is slightly lower. The YOLO is a representative of single-stage algorithm. Because of its fast running speed, it is suitable for real-time detection. Very popular in practice. The YOLOv5 is the latest and best performance version, and its application and research are very important.

2. Principle of YOLOv5 algorithm

2.1. Brief description of the development of YOLOv5

The YOLO series have become the current hot target detection algorithms. Compared with other algorithms, they have the characteristics of fastness and real-time, and their structure are relatively simple. The method is to first extract features, the input image is then divided into s * s grids, and finally detect the target whose center point falls on the grid.

After YOLOv1[10] appeared in 2015, in order to continue to improve its performance, YOLOv2[12], YOLOv3[13], YOLOv4[14] appeared, and they have been updated to the YOLOv5[15]. Take the comparison between YOLOv5 and YOLOv4, the former is more flexible and quicker than the latter without reducing its accuracy. It includes four models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Their parameter size and accuracy increase in order, From Bottleneck to distinguish, there are some mechanisms like EfficientNet [16] to select a model of the appropriate size.

The version in YOLOv5 is updated very quickly, mainly including the 3.0, 4.0 and 5.0 versions. The comparison of the models trained by YOLOv5s in the 3.0, 4.0, and 5.0 versions is shown in Table 1. The 5.0 version is not much updated compared to the 4.0 version, and the main update is the version 5.0 can directly test online videos. Compared with version 3.0, version 4.0 has updated the new activation function SiLU to replace the original LeakyReLU and Hardswish, as shown in Figure 1 below. This SiLU was also introduced in Pytorch 1.7, the model is more streamlined, and a convolutional layer is removed in each bottleneck, as well as the reconstruction of the utils module and so on. This article uses version 5.0 of the YOLOv5s model.

| YOLOv5s | size (pixels) | mAPval 0.5:0.95 | mAPtest 0.5:0.95 | mAPval 0.5 | Spend v100(ms) | params (M) | FLOPS 640(B) |
|---------|---------------|----------------|-----------------|-----------|----------------|------------|-------------|
| 3.0     | 640           | 37.0           | 37.0            | 55.4      | 2.4            | 7.5        | 13.2        |
| 4.0     | 640           | 36.8           | 36.8            | 63.1      | 2.2            | 7.3        | 17          |
| 5.0     | 640           | 36.7           | 36.7            | 66.9      | 2.0            | 7.3        | 17          |

2.2. Model network architecture of YOLOv5

The network structure of YOLOv5 includes four parts: Input, Backbone, Neck, and Prediction. The structure of YOLOv5 is shown in Figure 1.
Fig. 1 Network structure of YOLOv5

Firstly, mosaic data enhancement method is used for its capture in the Input terminal, and it also has an automatic anchor frame mechanism system, which is different from the separate anchor frame mechanism of YOLOv3 and YOLOv4.

Secondly, the Backbone is divided into two structural domains. One is the Focus structure, which is unique in YOLOv5. Its slicing operation is a very critical step. For example, the image input of 640 * 640 * 3, after the slice operation, it will turn into a characteristic map of 320*320*12, and then pass through 32 convolution kernels, eventually become 320*320*32 Feature map. The other is the CSP (Cross Stage Partial network) structure. YOLOv5 has designed two CSP structures, the CSP structure is applied to the backbone and the neck network. The SPP network module is also used in the Backbone part. The SPP network was proposed by He Kaiming [9] in 2015. The purpose of using SPP is that regardless of the input size, it can generate a fixed size output, and it can also use multiple pooling windows.

Thirdly, the Neck part of the current version of YOLOv5 adopts the structure of FPN+PAN (Feature Pyramid Networks + Path Aggregation Networks)[17]. The FPN transmits semantic information from high dimension to low dimension (the big goal is more clear), and the PAN transmits semantic information again from low dimension to high dimension (the small goal is also more clear). In the neck, the CSP structure designed by CSPNet[18] is also adopted and used for reference.
enables the ability of network feature fusion. In addition, other parts of the network have also been adjusted.

Finally, the Prediction of YOLOv5 is innovative to a certain extent. It increased PANNet [19] to better complement the underlying and high-rise feature advantages, effectively solve the multi-scale problem.

3. Improved YOLOv5 helmet detection algorithm

The YOLOv5 is a real-time, fast and accurate algorithm. However, for the detection of some small targets, YOLOv5 may have false detection or missed detection. In addition, due to the operation of convolution and down sampling, the number of feature maps will be reduced, and it is easy to lose feature information in the transmission process, so it is easy to produce gradient disappearance. For the sake of improving the detection accuracy and effect, a safety helmet detection algorithm for power workers according to improved YOLOv5 is put forward. The main improvements are as follows:

- Improve and optimize the feature fusion layer and multi-scale detection layer, and add a fusion scale layer for small target detection, which greatly improves the detection ability of small targets or even dense small targets.
- Because there are only two categories for identification and detection of helmets, and the priori anchor frame obtained in the original YOLOv5 algorithm clustering by K-means for COCO data set is not suitable for the actual recognition and detection of helmets, so K-means is accustomed to re-cluster the priori anchor frame that is more suitable for detecting and identifying helmets.
- Based on the experimental data, this paper selects between GIOU Loss and CIOU Loss, and finally selects the CIOU Loss with better effect as the loss function of bounding box regression. Because CIOU Loss also considers the scale information of the width height ratio of the bounding box.

3.1. Improvement of feature extraction network

In the YOLOv5 algorithm, its feature fusion layer and detection layer use the FPN and PAN network structure to enhance the ability and increase the accuracy of image recognition. The minimum feature scale layer size output is 20*20, and the smaller feature map contains a lot of semantic information, but the error will be large for the information judgment of the predicted position. The detection of the helmets may be affected by different positions, distances, weather conditions, and location occlusion. Small targets such as safety helmets are prone to misdetection or missed detection. The purpose is to better improve the detection ability of small targets, we have added a scale layer for small targets on the original three output scales. After fusion with other feature maps, four output scale layers are used to identify and detect safety helmets.

The original three scale output detection layers are 20*20, 40*40 and 80*80, which are used to detect large, medium and small targets respectively. If the detection of safety helmets belongs to small targets, this paper makes the following changes on the original basis: first, because the detection is small targets, in the case of the original three priori frames, a priori box for small target detection is added, and the network of the initial YOLOv5 model is changed. The difference between the improved YOLOv5 network model and the original algorithm is that if the size of the input picture is 640*640*3, after the second up sampling splicing in the neck part, another up sampling will be carried out through C3 module and CBS module (The two modules are shown in Figure 1). It is spliced with the 160*160*64 features output in the backbone part, which becomes the size of 160*160*128 after the operation of C3 module. Through the up sampling operation in the feature fusion layer stage, the image is finally output to the size of 160*160*255.

Therefore, compared with the output of the original YOLOv5 network model, the improved network model increases a 160*160 output layer. That is, the improved output layer size is 160*160, 80*80, 40*40, 20*20. The feature map size of 160*160 is 4 times downsampling of the input image 640*640, the receptive field is relatively small, suitable for detection and identification of relatively small targets such as helmets. The helmet detection model trained on the above 4 scale output layers
can be accurately detected on multiple scales even if the size of the helmets in the screen even have some changes. The improved YOLOv5 model is shown in Figure 2.

![Network Structure](image)

**Fig. 2** The network structure of our improved YOLOv5 (The red box is the improved part)

### 3.2. K-means dimension re-clustering

K-means is a widely used clustering algorithm. The central idea is to divide each point into clusters, which are then represented by the nearest cluster center with a given K value and K first-category center points. Finally, it specifies a point and iteratively update the cluster center point until the cluster center does not change much or reaches the specified number of iterations. By the time YOLOv2, K-means has been used, which can recognize more types of target and has better performance than YOLOv1. So it has also continued to the current YOLOv5.

The improved YOLOv5 model has one more output scale than the original one. The original 3*3 a priori anchor frame is no longer applicable. Therefore, it is necessary to re-cluster the improved YOLOv5 by K-means to make the accuracy recognition and detection better. After many experiments and research, according to the Avg IOU, it is concluded that when K=12, the final detection accuracy is better.

The results of the Avg IOU ratio under different K values are shown in Figure 3.
Fig. 3 Re-clustering results using K-means

The distribution of the candidate boxes after clustering is illustrated in Table 2 below. The increased feature map scale of 160*160 can be used for the detection of smaller targets, while the 80*80, 40*40, 20*20 in the original YOLOv5 can be used for detecting small, medium or large targets, by assigning different sizes of a priori boxes, the ability of the network models to detect helmets of different sizes can be further enhanced.

| Feature map size       | Anchor frame size | Anchor frame size | Anchor frame size |
|------------------------|-------------------|-------------------|-------------------|
| 160*160 (Smaller scale)| (9,11)            | (12,14)           | (16,19)           |
| 80*80 (small scale)   | (22,26)           | (29,34)           | (38,44)           |
| 40*40 (Medium scale)  | (49,56)           | (64,76)           | (86,101)          |
| 20*20 (Large scale)   | (123,142)         | (188,206)         | (306,349)         |

3.3. Loss function

The YOLOv5 uses the binary cross-entropy loss function to compute the loss of category probability and target confidence score. This paper chooses between GIOU Loss and CIOU Loss through experimental results, and finally chooses the CIOU Loss as Location loss function, The loss of the bounding box is equal to 1-CIOU, where the formula of CIOU Loss is:

\[ l_{CIOU} = 1 - IOU + \frac{\rho^2(b,h)}{c^2} + \alpha \gamma \]  

\[ \gamma = \frac{4}{\pi^2} \left( \arctan \frac{a}{b} - \frac{w}{h} \right)^2 \]  

\[ \alpha = \frac{\gamma}{(1-IOU)+\gamma} \]

Among them, \( \gamma \) is a parameter that measures the consistency of the aspect ratio, \( \alpha \) is a parameter used to make trade-offs, \( \frac{a}{b} \) is the aspect ratio of the bounding box, and \( \frac{w}{h} \) is the aspect ratio of the predicted frame. CIOU considers the overlap of the frame on the basis of the IOU, and the center The scale information of distance and aspect ratio makes the final prediction effect better.

4. Experimental results and comparative analysis

The operating system of the experimental computer in this article is Windows10, the CPU model is Intel(R) Core(TM)i7-7700k CPU@4.2GHz, the GPU model is GeForce RTX 1070, the video memory size is 8GB, and the memory size is 16GB. All network models are based on Pytorch 1.9, and use Cuda 11.0 and Cudnn 8.0.4 to accelerate the GPU.

4.1. Data set production and processing

In deep learning, the quality of the data set will greatly affect the quality of the final experimental results. We have obtained a data set of more than 18,500 operation images of power workers by means
of web crawlers, of which the training set is 15,900 images, the verification set is more than 2,000, and the test set is more than 500. The labels are labeled with the Labelme tool. There are two types of labels, namely "helmet" and "no helmet". The distribution of data set label operation, training set, test set and data samples are illustrated in Figure 4.

For the sake of reduce the problem of over-fitting during network training, data enhancement was carried out before training, and methods such as rotation, contrast change, flipping, cropping, and scaling were used to enhance the self-made data set.

4.2. Network training

Before network training, first configure the hyperparameters of the model, Such as the original learning rate is $lr_0=0.01$, the final learning rate is $lr_0*lr_f=0.002$, the weight decay coefficient is $\text{weight}\_\text{decay}=0.0005$, the learning rate momentum is $\text{momentum}=0.937$, the warmup initial momentum is 0.8, and its bias learning rate is 0.1 and so on. The SGD is used as the optimizer, and the training period is set to 150 epochs. After configuring these initial hyperparameters for training, we will get the initial pre-model, that is, the model effect of the original YOLOv5.

After several fine-tunings, the improved network model of us is added, the environment for training our network model is configured. The K-means dimension is used to re-cluster, and the rectangle is used to fill the training to accelerate the model inference process. We use the training method of adding some weights to the images that were not well trained in the previous round, and adjusting the cosine annealing function value in the hyperparameters and changing the image clipping ratio, flip direction, rotation angle, and zoom size, learning rate momentum value, mixup coefficient, etc. After many experiments, we finally got the improved network model.

4.3. Comparison of YOLOv5 initial model and improved model

The performance of the model requires a good evaluation method. For the sake of reduce the problem of over-fitting during network training, data enhancement was carried out before training, and methods
such as rotation, contrast change, flipping, cropping, and scaling were used to enhance the self-made data set. Prevent the uneven distribution of the sample targets, we use the precision and the recall rate to measure. The precision is mainly for the level of prediction results, and the recall rate is mainly for its own samples. The relevant formula is as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (4)
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \quad (5)
\]

In the formulas, TP (True positives) can be understood as judging the positive class as the positive class, that is, the amount that the prediction in the model is correct, and FP (False positive) can be understood as the negative class being judged as the positive class, that is, the amount that the prediction in the model is wrong, FN (False Negative) can be understood as a positive class judged as a negative class, that is, the amount of the model that was originally a positive sample but was missed.

All the experimental results in this article are warmup training. This is to maintain the stability of the model structure and will not cause oscillation effects due to the high initial learning rate of the model. After passing the warmup stage, the cosine annealing algorithm is accustomed to update the learning rate, so as to achieve a better network model. The curve of the relative change of the YOLOv5 original model and improved model is Precision (accuracy), mAp_0.5, mAp_0.5:0.95, as shown in Figure 5, 6, and 7, where the higher curve is the consequence of the improved model, and the lower one is the consequence of original model.

![Fig. 5 Precision of YOLOv5 initial model and improved model](image-url)
It’s observed that the precision and mAP value of the improved YOLOv5 network model are higher than those of the original, which can effectively reduce the misdetection rate and missed detection rate. The comparison of the helmet detection effect of the improved model and the original model is illustrated in Table 3.

**Tab. 3** Comparison of detection performance of YOLOv5 initial model and improved model

| Algorithm       | Parameter/piece | Model size/Mb | Precision/% | Recall/% | Speed/fps |
|-----------------|-----------------|---------------|-------------|----------|-----------|
| the initial YOLOv5 | 7066239         | 13.7          | 92.2        | 94       | 32.6      |
| the improved YOLOv5 | 7851796         | 15.3          | 94.6        | 97       | 29.2      |
From the table, we can read that the parameters of the improved model are more than those of the initial one, but the model size is not much different. The improved detection speed is 28.6fps, which is only 3.4fps less than the original, and it will not have a great impact on the detection. However, the improved detection precision rate has increased by 2.4% compared to the original, reaching 94.6%, and the recall rate has increased by 3%, reaching 97%. By the comparison and analysis of the results, the improved YOLOv5 fully meets the real-time detection requirements, and the detection precision and recall rate are better than those of the original. Therefore, the improved YOLOv5 network model in this article is effective in the detection task of whether power workers wear helmets or not.

4.4. Algorithm detection capability comparison

For the sake of fully evaluate and compare the improved YOLOv5 recognition and detection algorithm in this article, we conducted the following types of experiments:

- The mAp values of the improved YOLOv5 and the original YOLOv5 by using SGD and Adam are compared, as illustrated in Table 4. The consequence indicate that the SGD optimizer is superior to the Adam.
- Although the Adam optimizer converges faster, it will make the learning rate converge quickly and reduce the final mAp value. Therefore, the SGD optimizer is used by default in the subsequent experiments.

| Algorithm      | SGD   | ADAM       |
|----------------|-------|------------|
|                 | mAp@0.5 | mAp@0.5:0.95 | mAp@0.5 | mAp@0.5:0.95 |
| YOLOv5          | 0.921  | 0.557      | 0.905   | 0.497         |
| Improved YOLOv5 | 0.95   | 0.568      | 0.921   | 0.534         |

- The values of average precision and detection speed between YOLOv3, YOLOv5, and improved YOLOv5 were compared, and the consequence are illustrated in Table 5. The results show that the average precision rate of the improved YOLOv5 is 2.2% higher than YOLOv3 and 2.9% higher than YOLOv5, which significantly improves the detection precision. Although the speed is 3.4fps slower than YOLOv5, it is 10.5fps faster than YOLOv3. The comprehensive comparison shows that the improved network model has been able to meet real-time safety helmet detection, and has reached a high precision rate.

| Algorithm | Average precision/% | Detection speed/fps |
|-----------|---------------------|---------------------|
| YOLOv3    | 92.8                | 18.7                |
| YOLOv5    | 92.1                | 32.6                |
| Improved YOLOv5 | 95.0   | 29.2                |

- The differences between the improved YOLOv5 and the original one for detection under special conditions are compared. Some examples are shown in Figure 8-13. From left to right, the input image, the original YOLOv5 detection result and the improved YOLOv5 detection result are shown. Judging from the comparison of image detection effects in six different situations, the improved YOLOv5 model has obvious advantages in reducing the missed detection rate and misdetection rate of object detection in various situations, which improves the precision of detection. It is advanced in safety helmet detection algorithm of power workers.
**Fig. 8** The effect of the initial YOLOv5 and the improved model in low light background

**Fig. 9** The effect of the initial YOLOv5 and the improved model in the presence of obstruction

**Fig. 10** The effect of the initial YOLOv5 and the improved model in the case of missed detection

**Fig. 11** The effect of the initial YOLOv5 and the improved model in the case of false detection

**Fig. 12** The effect of the initial YOLOv5 and the improved model with remote small targets
5. Conclusion
Aiming at the poor detection effect and low accuracy of power workers helmets, this paper put forward an improved YOLOv5 algorithm, which optimizes the feature fusion layer and multi-scale detection layer, and adds a fusion scale for small target recognition Layer. greatly improving the detection ability of small targets or even dense small targets. By increasing the size of the feature map of 160*160, the detection accuracy of the network model for small targets is improved obviously, and the rate of misdetection and missed detection is reduced. The K-means clustering method is also accustomed to find the suitable detect candidate anchor frames for small targets such as helmets. Through comparative experimental analysis, the improved YOLOv5 network model still meets real-time detection, and the final precision rate is higher, the recall rate is higher, and the stability of the algorithm is improved obviously.

Acknowledgments
This work was supported by the National Natural Science Foundation of China under Grant 61562025, and Grant 61962019.

References
[1] Zou Z X, Shi Z W, Guo Y H and Ye J P 2019 Object detection in 20 years: A survey J. Computer Vision and Pattern Recognition
[2] Viola P and Jones M 2001 Robust real-time face detection Proceedings Eighth IEEE International Conference on Computer Vision (ICCV) pp 747-747
[3] Llorca D F, Arroyo R and Sotelo M A 2013 Vehicle logo recognition in traffic images using HOG features and SVM Proceedings of International IEEE Conference on Intelligent Transportation Systems (ITSC) pp 2229-2234
[4] Felzenszwalb P F, Girshick R B, McAllester D and Ramanan D 2010 Object Detection with Discriminatively Trained Part-Based Models J. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9) pp 1627-1645
[5] Liu X H and Ye X N 2014 The application of skin color detection and Hu moment in helmet recognition J. East China University of Science and Technology (Natural Science Edition) 40(03) pp 365-370
[6] Kido S, Hirano Y and Hashimoto N 2018 Detection and classification of lung abnormalities by use of convolutional neural network(CNN) and regions with CNN features (R-CNN) International Workshop on Advanced Image Technology (IWAIT) pp 1-4
[7] Girshick R 2015 Fast R-CNN IEEE International Conference on Computer Vision (ICCV) pp 1440-1448
[8] Ren S, He K, Girshick R and Sun J 2017 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks IEEE Transactions on Pattern Analysis and Machine Intelligence vol 39 pp 1137-1149
[9] He L, Zhang X, Ren S and Sun J 2015 Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence vol 37 pp 1904-1916
[10] Redmon J, Divvala S, Girshick R and Farhadi A 2016 You Only Look Once: Unified, Real-Time Object Detection IEEE Conference on Computer Vision and Pattern Recognition
[11] Poirson P, Ammirato P, Fu C, Liu W, Košcká J and Berg A C 2016 Fast Single Shot Detection and Pose Estimation Fourth International Conference on 3D Vision (3DV) pp 676-684
[12] Redmon J and Farhadi A 2017 YOLO9000: Better, Faster, Stronger IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 6517-6525
[13] Redmon J and Farhadi A 2021 YOLO v3: an incremental improvement
[14] Bochkovshiy A, Wang C and Liao H 2020 YOLOv4: optimal speed and accuracy of object detection J. Computer Vision and Pattern Recognition
[15] Jocher G 2020 Yolov5 https://github.com/ultralytics/yolov5
[16] Tan M X and Le Q V 2019 EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks International Conference on Machine Learning vol 97 pp 6105-6114
[17] Huang Z, Zhong Z, Sun L and Huo Q 2019 Mask R-CNN With Pyramid Attention Network for Scene Text Detection IEEE Winter Conference on Applications of Computer Vision (WACV) pp 764-772
[18] Wang C, Liao H M, Wu Y, Chen P, Hsieh J and Yeh I 2020 CSPNet: A New Backbone that can Enhance Learning Capability of CNN IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp 1571-1580
[19] Yang J, Fu X, Hu Y, Huang Y, Ding X and Paisley J 2017 PanNet: A Deep Network Architecture for Pan-Sharpening IEEE International Conference on Computer Vision (ICCV) pp 1753-1761