Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 4, No. 2, 2022

Detection Method of Safety Protective Equipment Based on Deep Learning

Zhu Shi[1, *], Hao Wu[1, 2], Zhongyang Jin[1, 2], Hong Song[3]

[1]Sichuan University of Science & Engineering, School of Automation and Information Engineering, Sichuan Yibin, 644000, China
[2]Key Laboratory of Artificial Intelligence in Sichuan Province, Sichuan Yibin, 644000, China
[3]Aba Teachers College, Sichuan Aba 624000, China

Abstract: The construction site belongs to the high-risk operation area, and whether the safety protection articles are correctly worn will directly affect the personal safety of operators. Aiming at the problems of weak anti-interference and low detection accuracy of existing detection methods, a new detection method for safety appliances based on improved YOLOv5 is proposed. This method embeds the BoT module into YOLOv5 backbone network, and uses its self attention mechanism to further mine image feature clues, so as to enhance the network's ability to extract features from target regions. The experimental results show that the average precision mAP of the method based on the improved YOLOv5 is 87.1%, and the recall rate is 80.8%. Compared with the original YOLOv5, the precision is improved by two percent.

Keywords: Safety appliances, YOLOv5, BoT, Self attention mechanism.

1. Introduction

At the construction site, the correct wearing of safety protective equipment by the staff can effectively avoid the occurrence of safety accidents. In order to improve the level of project safety supervision, intelligent supervision has been gradually used to replace the previous manual inspection method. At present, many domestic colleges and research institutions have carried out in-depth research on this, and proposed real-time detection technology for the wearing of safety protective equipment on the construction site. The traditional safety appliance wear detection algorithm mainly selects and designs features of the target to be detected in the input image manually, and then carries out comparative detection through an appropriate classifier [1]. For example, Liu Yunbo et al; Rubaiyat et al. [3] combined the frequency domain information of the image and the gradient direction histogram (HOG) feature to locate the human body, and then used the color feature and the circular hough transform (CHT) to detect the helmet; Wang Zhen et al. [4] proposed a dual weight detection model. By introducing the weight of the pixel center and the weight of the background weakening, and calculating the distance of the centroid change of different target pixels, they can detect the wearing of work clothes. Although the traditional detection algorithm is fast, it relies heavily on manual selection and design features, and its generalization ability is relatively poor. It is unable to accurately detect the wearing of safety appliances in the complex and changeable construction site.

With the rapid development of deep learning technology, target detection technology based on convolutional neural network has been widely used in the field of safety appliance wear detection with its strong feature extraction ability and algorithm stability. Compared with traditional safety appliance detection methods, the safety appliance wear detection method based on deep learning mainly realizes the detection of safety appliances by automatically extracting target features through convolutional neural network [5]. For example, Li Mingshan et al. [6] introduced a feature fusion branch network on the basis of the SSD network to strengthen shallow semantic information, and introduced variable parameters to adjust the priori frame to effectively improve the target detection accuracy; Zhang Wukang et al. [7] introduced a multi-scale feature extraction network into RetinaNet, which enhanced the network's ability to extract feature maps and also improved the detection accuracy; Zheng Haiyang et al. [8], on the basis of YOLOv3 network, fused multi-level feature information in the feature pyramid structure, and optimized the candidate box with K-means algorithm, effectively improving the detection accuracy of the wearing of insulating gloves.

In view of the special environmental characteristics of the construction site and the requirements for wearing safety appliances, this paper selects the YOLOv5 algorithm with strong flexibility and speed, and adds the BoT module to the YOLOv5s backbone network, which brings better performance and reduces the number of parameters. Finally, the effectiveness of this algorithm is verified through comparative experiments.

2. YOLOv5 Algorithm Principle

YOLOv5 is a single-stage detection algorithm. Compared with YOLOv4 algorithm, it has greatly improved in flexibility, speed and accuracy. Compared with YOLOv4 algorithm, it has faster speed and simpler structure. The network structure of YOLOv5 algorithm is shown in Figure 1.

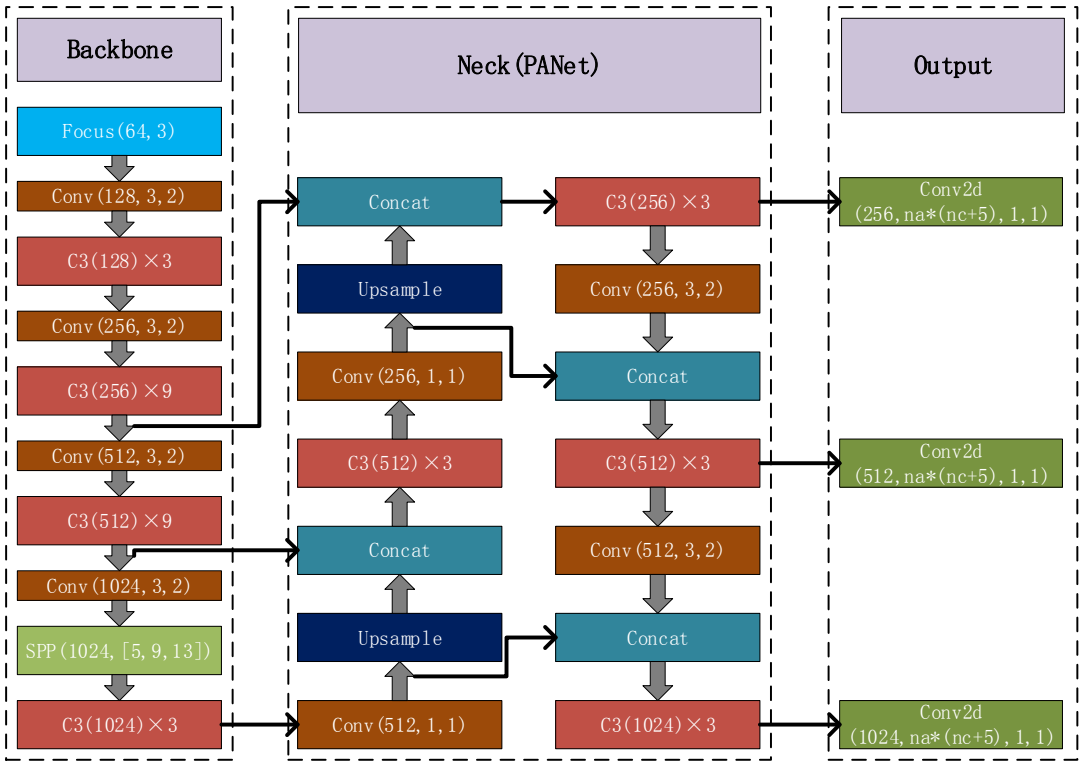

Figure 1. YOLOv5 Network Structure

YOLOv5 network mainly includes four models: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. In this paper, YOLOv5m network is used to detect the wearing of safety appliances on the power site. YOLOv5m model consists of four parts: input, backbone, neck and output. The input terminal adopts Mosaic data enhancement to scale the image adaptively. Backbone uses multiple convolution and pooling operations to obtain feature maps of different sizes in the image, mainly composed of Conv, C3, SPP and other modules. Neck is a feature fusion part, which is composed of a feature pyramid FPN and a path aggregation network PAN structure. The output terminal mainly outputs the prediction results.

(1) Input terminal

YOLOv5 network uses Mosaic mode to enhance the input data and adaptively scale the image to improve the generalization of the network. Mosaic data enhancement method randomly selects four images, scales the images randomly, combines them into one image by cutting, overlapping and other methods, and simultaneously obtains the real frame information corresponding to the target in the image, which can effectively enrich the semantic information of the input image, save more computing time for the network, and improve the operation speed of the network.

(2) Backbone

YOLOv5 uses the CSPParknet53 structure as the backbone network. The main function of the backbone network is to extract a series of feature maps of different scales of the input image. It uses multiple convolution and pooling operations to obtain feature maps of different sizes in the image, which is mainly composed of Focus, Conv, C3, SPP and other modules. Focus structure, Conv is the basic convolution unit of YOLOv5 network, which performs two-dimensional convolution, regularization, activation and other operations on the input image. C3 is composed of several classical residual Bottleneck modules. The SPP module is a spatial pyramid pooling layer, which performs 5 steps for the input image × 5, 9 × 9, 13 × 13 Pool operations with three different sizes, and connect the output results with Concat to keep the depth of the network unchanged.

(3) Neck

The neck (Neck), as the intermediate connection part of the network, mainly processes the feature information of the extracted images of different sizes, processes the images to the same size, and then performs feature fusion to generate feature maps of three scales, and then transfers them to the output. The Neck network is composed of the Feature Pyramid Network (FPN) and the Pyramid Attention Network (PAN) structures, which together form the PANet structure. The FPN structure connects the feature maps from top to bottom, transferring the category features of high-level big targets to the lower level, and the PAN structure transfers the location features of low-level big targets and the category and location features of small targets to the upper level from bottom to bottom, They complement each other and overcome their respective limitations, realizing the fusion and complementarity of high-level features and low-level features, thus strengthening the feature extraction capability of the model.

(4) Output terminal

The output terminal is mainly used for the final stage of the detection process. It associates the anchor frame mechanism with the processed feature map. For the feature map output by Backbone, it is predicted and classified after features are fused by the Neck network. Finally, the prediction information of the detection target is output. The most important part of the output layer is the calculation of the loss function and the analysis of the prediction results. YOLOv5 uses GIoU Loss as the loss function of the network, and uses Non Maximum Suppression (NMS) to analyze the prediction results

3. The Design of Inspection Model for Safety Protective Equipment

BoTNet [5] is a simple but powerful backbone network. It combines CNN and transformer, and integrates self attention into a variety of computer vision tasks, including image classification, target detection, instance segmentation, etc. The overall idea of BoT is relatively simple. For tasks requiring high input image size, such as target detection, ordinary transformer (ViT) [6] is difficult to achieve large resolution input in terms of computation load and video memory occupation. For networks such as DeTR, the characteristics of CNN output are directly used as the input of transformer, which cannot achieve end-to-end training. Therefore, the author of BoTNet proposes to directly replace 3x3 convolution with MHSA [7] in the bottomleneck of ResNet, which is the so-called BoT block. Information is summarized on the premise of effectively reducing model parameters. In the BoT block, the channel compression ratio is still 4 times. The BoT module structure is shown in Figure 2.

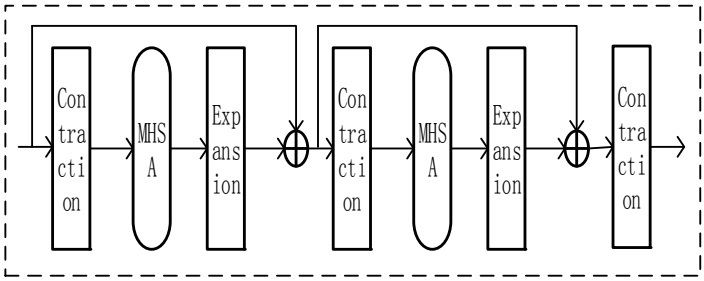

Figure 2. BoT Module Structure

In this paper, the C3 module of layer 8 in the backbone network is deleted, and the BoT module is added behind the SPPF module. The essence of the BoT module is to focus on the problem that CNN only pays attention to the local part of the image, combine CNN with transformer, replace the 3 * 3 convolution layer in ResNet Bottleneck with self attention, use its dynamic self attention mechanism to focus on the global information of the image, and establish connections in multiple spaces. Using transformer as a bottomleneck not only reduces the number of parameters, but also skilfully avoids the problem of image patch processing, and brings better performance. The improved backbone network is shown in Table 1.

Table 1. Improved Backbone Network Structure

Layer No	Repetitions	modular	parameter
0	×1	Conv	[64,6,2,2]
1	×1	Conv	[128,3,2]
2	×3	C3	[128]
3	×1	Conv	[256,3,2]
4	×6	C3	[256]
5	×1	Conv	[512,3,2]
6	×9	C3	[512]
7	×1	Conv	[1024,3,2]
8	×1	SPPF	[1024,5]
9	×3	BoT	[1024]

4. Model Training and Experimental Results

(1) data set

In the process of deep learning object detection, the data set used in the experiment has always been an essential part. Most of the electric power operation sites are located in the suburbs. Due to the uncertain operation time and large changes in light, the collected image effects vary greatly. Because the current open source data set does not have a standard power field data set, it does not meet the detection requirements in the actual production environment. In order to solve this problem, according to the images collected at a power operation site and obtained by the web crawler, this paper makes a data set of 5015 images for illegal wear detection at the power operation site. According to the data set, we have defined eight label categories (see Table 2). We use the YOLO format annotation tool Make Sense to annotate the image, add a detection box to the position to be detected, and select the corresponding label category. Finally, the data set is divided into training set and verification set according to 8:1 ratio for model training and performance verification.

Table 2. Label Category

Tagname	describe
Gloves	Wear gloves correctly
No gloves	Not wearing gloves
Seat belt	Wear the safety belt correctly
No seat belt	Not wearing safety belt
Work clothes	Wear work clothes correctly
No work clothes	Not wearing work clothes
Helmet	Wear safety helmet correctly
No helmet	Not wearing a helmet

(2) Network training

In order to achieve the best performance of the model, in the training process, the number of iterations is set to 500, the initial learning rate is 0.01, the learning rate attenuation weight is set to 0.0005, and the learning rate momentum is set to 0.937 to prevent the model from over fitting. The training batch size is set to 32, so as to fully call the GPU. The loss decline curve is shown in Figure 3.

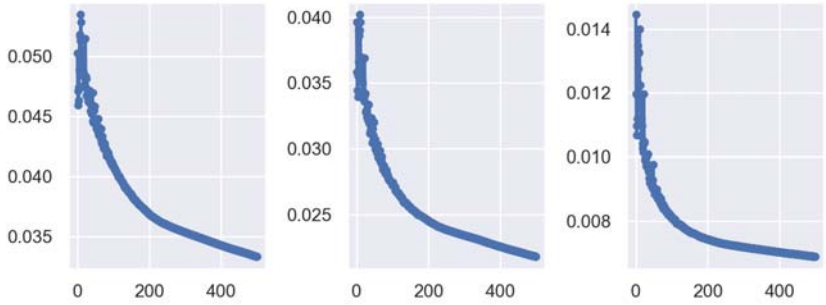

Figure 3. Loss decline curve

(3) experimental result

In this paper, YOLOv5 algorithm is applied to the detection of safety appliances at the construction site. In order to verify the superiority of the improved algorithm in this paper, the same data set is used under the same configuration, compared with YOLOv5 network, and each detection algorithm is evaluated using the mAP index. The comparison experiment results are shown in Table 3. The comparison of the detection results of the model for each category is shown in Figure 6.

Table 5. Comparison of mAP Values of Various Types of Targets (/%)

	Gloves	No gloves	Seat belt	No seat belt	Work clothes	No work clothes	Helmet	No helmet	All
YOLOv5	62.9	69.3	95	94.8	98.7	96.7	95	68.1	85.1
OURS	69.6	72.1	95.2	95.5	98.6	97.2	95.8	73.5	87.1

According to the experimental results in Table 5, the algorithm in this paper can effectively improve the detection accuracy of the wearing of safety appliances of workers at the power operation site. The mAP value of this algorithm for staff wearing violations detection is 87.1%. Compared with the original YOLOv5, the mAP value of each category detection has been improved to a certain extent. This shows that the algorithm in this paper performs well in the detection accuracy of safety appliances at the power operation site, and can meet the accuracy requirements of illegal wear detection under complex power operation environment.

In addition, in order to see the detection gap between different algorithms more intuitively, some detection results are selected as shown in Figure 4. It can be seen from the figure that YOLOv5's detection performance is only inferior to the algorithm in this paper. Because of its deep FPN network and PAN network, it has good performance in detecting small and weak targets, and can detect most long-distance and short-range wearing targets, but there are still missing glove and helmet targets. This method can accurately detect all small and weak targets and occluded targets in Figure 4, This shows that the security appliance detection method proposed in this paper can fuse the feature information of high and low levels well, so as to improve the detection accuracy of the algorithm for security appliance targets. It can be seen from the comparison of the above multiple network detection results that the improved YOLOv5 network model has a good detection effect on illegal wear in complex electric power operation environment.

Figure 4. Comparison Diagram of Test Results

5. Conclusion

In this paper, aiming at the problem that traditional target detection technology is difficult to detect illegal wear in complex electric work scenes and has a high rate of missed detection, a method based on improved YOLOv5 for detecting illegal wear in electric work sites is proposed. This method is based on YOLOv5 network framework and embedded with BoT module, which effectively improves the feature extraction capability of the network. Experimental results show that the algorithm can achieve better detection accuracy.

Acknowledgment

The authors gratefully acknowledge the financial support from Postgraduate Innovation Fund of Sichuan University of Science & Engineering (y2021060).

References

[1] Ge Qingqing, Zhang Zhijie, Yuan Long, Li Xiumei, Sun Junmei. Safety helmet wearing detection by integrating environmental features and improving YOLOv4 [J]. Chinese Journal of Image Graphics, 2021,26 (12): 2904-2917.

[2] Liu Yunbo, Huang Hua. Research on Monitoring Technology of Helmet Wearing on Construction Site [J]. Electronic Science and Technology, 2015, 28 (04): 69-72.

[3] Rubaiyat A H M, Toma T T, Kalantari-Khandani M, et al. Automatic detection of helmet uses for construction safety[C]//2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW). IEEE, 2016: 135-142.

[4] Wang Zhen, Jin Yong, Wang Zhaoba, Chen Youxing. Workwear detection and tracking based on dual weight color histogram [J]. TV Technology, 2016, 40 (01): 131-134.

[5] Zhong Xinhao Research on Real time Monitoring Method of Helmet Wearing in Metro Construction Scenarios [D]. Hunan University of Technology, 2021.

[6] Li Mingshan, Han Qingpeng, Zhang Tianyu, Wang Daolei. Improving the safety helmet detection method of SSD [J]. Computer Engineering and Application, 2021,57 (08): 192-197.

[7] Zhang Wukang, Pan Lizhi, Guo Zhibin, Lin Xuan, Tu Xiaotong. RetinaNet based visual detection method for abnormal state of insulating gloves in power scenarios [J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2022,37 (01): 85-91.

[8] Zheng Haiyang, Song Chunhe, Wu Tingting, Liu Shuo, Zhou Zhongran. Small target detection and matching algorithm for insulating gloves wear detection [J/OL]. Small microcomputer system: 1-10 [2022-10-11].

[9] Srinivas A, Lin T Y, Parmar N, et al. Bottleneck transformers for visual recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16519-16529.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems Long Beach, CA, Dec 4-9, 2017.Cambridge: The MIT Press, 2017: 6000-6010. [5]SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model [J]. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.