Safety helmet wearing recognition based on improved YOLOv4 algorithm

Bin Wang
School of automation
Wuhan University of Technology
Wuhan, China
Email: 530725978@qq.com

Haojie Xiong
School of automation
Wuhan University of Technology
Wuhan, China
Email: xionghaojiephil@163.com

Lishou Liu
School of automation
Wuhan University of Technology
Wuhan, China
Email: 307343@whut.edu.cn

2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC) | 978-1-6654-3185-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITOEC53115.2022.9734707

***Abstract*—In safety helmet wearing detection, traditional methods have problems such as low detection accuracy due to small object size. In this paper, we propose an improved helmet wearing detection algorithm YOLOv4-P. Firstly, we use k-means clustering algorithm to readjust the prior bounding box parameters to improve the matching between the prior bounding box and the object, and secondly, we introduce the Pyramid Split Attention (PSA) model to further process the multi-scale feature information. The network structure of YOLOv4 is improved by adding a layer of network features to improve the detection accuracy through feature fusion. Experimental results show that the improved model YOLOv4-P has an average accuracy improvement of 2.15% over the YOLOv4 algorithm on the SHWD helmet target detection dataset, satisfying the accuracy of the helmet wearing detection task.**

***Keywords*—safety helmet wearing detection, YOLOv4 model, Pyramid Split Attention model, feature fusion**

I. Introduction

The safety helmet has a variety of protective capabilities, which can disperse the impact force when an accident occurs, thereby effectively reducing head and neck injuries. According to relevant statistics, among people with head injuries at construction sites, 47.3% of them were injured by falling objects or hits from objects because they did not wear helmets [1]. Therefore, whether workers wear safety during work Supervision by caps is an important part of ensuring safe production.

At present, object detection methods based on deep learning models can be divided into two main categories: One class is the two-stage detection algorithm, and the other class is the one-stage detection algorithm. The representative algorithms of two stages are R-CNN [2], Fast R-CNN [3], Faster R-CNN [4], etc. The detection process of these algorithms is divided into two steps, firstly, candidate region selection, and then classification and localization of the selected region. These detection methods have high detection accuracy, but the model structure is complex, with more computational parameters and poor real-time detection performance. One-stage algorithms, such as SSD [5], YOLO series [6–9], etc. this type of algorithm directly through the network structure, complete the selection of candidate frames and classification and positioning, saving the computational cost and improving the speed of detection, while this real-time detection characteristics are more advantageous in practical engineering.

Among the many algorithms for helmet wearing detection, the YOLO series models have good performance in terms of detection speed and accuracy. However, there are still deficiencies in the detection of small objects and multi-scale feature fusion, and there is still much room for improvement in detection accuracy and speed. Kai Xu et al [10] based on the YOLOv3 detection algorithm and compensated for the poor detection of small objects by the original YOLOv3 by adding feature maps. Li Wigang et al [11] based on the YOLOv4 algorithm for object detection in indoor scenes, introduced a depth-separable convolution module to replace the original 3x3 convolutional layers in the model and reduced the model parameters. Based on YOLOv4-tiny, Bing Wang et al [12] proposed a bottom-up multiscale fusion, which combines low-level information to enrich the feature hierarchy of the network and improve the feature utilization. These algorithms optimized for specific scenarios have good detection results for specific detection objects.

In this paper, the YOLOv4 algorithm is improved based on YOLOv4 for the safety consideration of whether the staff wears helmets in the operation, and for the problems of low detection accuracy, such as small object miss detection and occlusion, which exist in the detection task. Firstly, the prior bounding box is re-clustered, and secondly, introduce the Pyramid Split Attention (PSA) module ,add a layer of the output layer of the feature map to make the multi-scale information expression capability of the feature map richer and enhance the feature fusion capability. The algorithm proposed in this paper can significantly improve the effect of helmet wearing detection in complex backgrounds.

II. YOLOv4 object detection algorithm

The structure of the YOLOv4 network model is shown in Figure 1. Taking the image input size of 416×416 as an example, the Darknet-53 network performs feature extraction at different scales for the input image through convolution and pooling processes of each network layer, and then fuses the different scale feature maps outputted by the pair of feature extraction networks through the YOLO detector, and finally

performs result prediction in three different output dimensions to finally obtain the target location and category information.

Among them, the YOLO detection layer generates prediction results at three different scales, 13×13, 26×26 and 52×52, which are applicable to the detection of large scale targets, medium scale targets and small scale targets, respectively.

The overall YOLOv4 algorithm mostly extends the network structure of YOLOv3, and the performance of the detection is also a big breakthrough, but for the detection of targets in specific scenes, the YOLOv4 algorithm can obtain greater improvement in detection speed and accuracy.

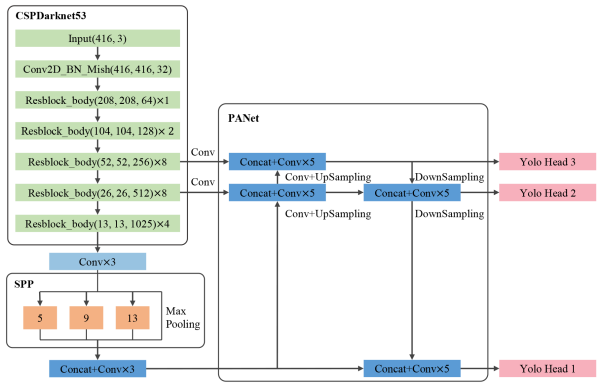


Fig. 1. YOLOv4 network structure

III. Improved YOLOv4 detection algorithm

A. prior bounding box re-clustering

The prior bounding box used by the original network of YOLOv4 is generated using a clustering algorithm on the COCO public data set. The COCO data set is mostly based on natural scenes. There are 80 object categories. There are only two categories in the helmet detection category in this article. Thus, the nine prior bounding box parameters in the original network corresponding to small, medium and large object detection cannot meet the practical needs of helmet detection.

By normalizing the size of the sample of the helmet data set, observe the distribution of the sample . If the position and size of the actual effective receptive field do not match the position and size of the prior bounding box, it will directly affect the accuracy of the detection algorithm, so the prior bounding box needs to be redesigned.

In this paper, k-means clustering algorithm is used to recluster the helmet data set to increase the matching degree between the prior bounding box and the actual object box as much as possible.

The figure shows the size of 9 prior bounding boxes obtained after clustering, where the prior bounding boxes of (5,10), (6,13), (7,15) are used for 52×52 features Map output detection, (9,19), (13,24), (19,32) prior bounding boxes are used for 26×26 feature map output detection and (27,47), (43,74), (81,140) The prior bounding box is used for 13×13 feature map output detection. These three different feature map outputs correspond to small, medium, and large helmet object detection.

B. Pyramid Split Attention

Existing studies have shown that attention mechanisms such as SENet [13], BAM [14], and CBAM [15] have good performance improvement on the network by embedding channel attention and spatial attention in the network. However, these attention mechanisms still need to be enhanced in the following two aspects: on the one hand, how to efficiently extract feature graph spatial information and enrich the feature space; on the other hand, how to improve remote channel dependencies instead of just acquiring local information. Accordingly, methods based on multi-scale feature representation and cross-channel information interaction have been proposed, such as PyConv [16], Res2Net [17] and HS-ResNet [18]. However, the above methods also make the complexity of the model high and the computational cost of the network large.

For the shortage of YOLOv4 network in the processing of extracting helmet wearing features, the Pyramid Split Attention (PSA) model [19] is used in this paper. The PSA module can process the spatial information of the input feature map at multiple scales and can effectively establish the remote dependencies among the multi-scale channel attention. As shown in Figure 2, the PSA module is implemented in four main steps.

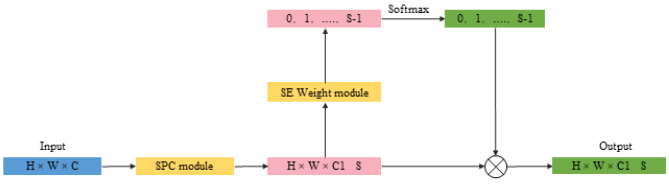


Fig. 2. Pyramid Split Attention (PSA) structure

First, the channels are sliced using the Squeeze and Concat (SPC) module, and the structure of the SPC module is shown in Figure 3. Second, multi-scale features are extracted for the spatial information on each channel feature map; second, the channel attentions of different scale feature maps are extracted using the SEWeight module to obtain the channel attention vector at each different scale; then the multi-scale channel attention vectors are feature recalibrated using Softmax to obtain the new multi-scale channel attention weights after interaction. Finally, the recalibrated weights and the corresponding feature maps are subjected to a dot product operation by element, and the output is obtained as a multi-scale feature map after attention weighting of feature information. The finally obtained feature map has a richer multi-scale information representation capability.

The SPC structure is the key to achieve multi-scale feature extraction in the PSA module. This structure extracts the spatial information of the feature map by using multiple

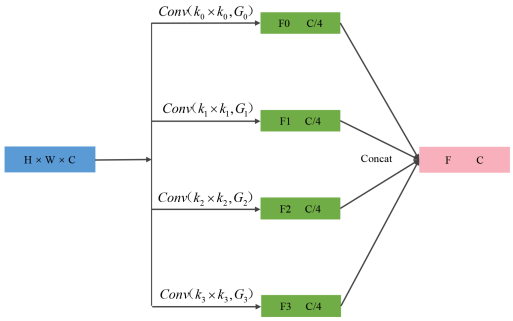


Fig. 3. Pyramid segmentation attention (PSA) structure

branches. Set the input channel dimension of each branch to C, so that a richer input tensor can be obtained Position information, and parallel processing on multiple scales. Thus, a feature map containing a single type of kernel is obtained. At the same time, the use of multi-scale convolution kernels in the pyramid structure can produce different spatial resolutions and depths. By compressing the channel dimensions of the input tensor, the spatial information of different scales on each channel feature map can be effectively extracted. S feature maps of different scales have the same channel dimension C', and the calculation formula is as follows:

$$C' = \frac{C}{S} \tag{1}$$

Each branch can learn multi-scale spatial information independently. In this way, the PSA module can fuse different proportions of contextual information and generate better pixel-level attention for feature maps. And, without destroying the original channel attention vector, the interaction of attention information is realized, and the dimensional vector is merged.

C. Increase feature layer processing

The background environment where the helmet wearers are located is often complex, containing construction materials, mechanical equipment, etc. and for such complex environments, the size of the helmet in which is relatively small, so enhancing the extraction of low-level features of the image in the network structure can improve the performance of the network model [20–22].

The network of YOLOv4 has three feature layers with different scales, 13×13, 26×26, and 52×52, and the feature layers are finally output through multiple convolution, upsampling, and downsampling feature fusion. There are limitations for helmet detection with complex scenes and small targets. Therefore, a shallow 104×104 feature output layer is added to the original network structure of YOLOv4 to improve the detection capability of the network for small targets through feature fusion.

The improved network structure is shown in Figure 4. The shallow layer features are more sensitive to the location information of construction personnel objects than the deep layer features, which makes the localization of personnel in construction images more accurate and makes the global information more focused, and strengthening the sensitivity to the shallow layer features can improve the detection performance of small objects.

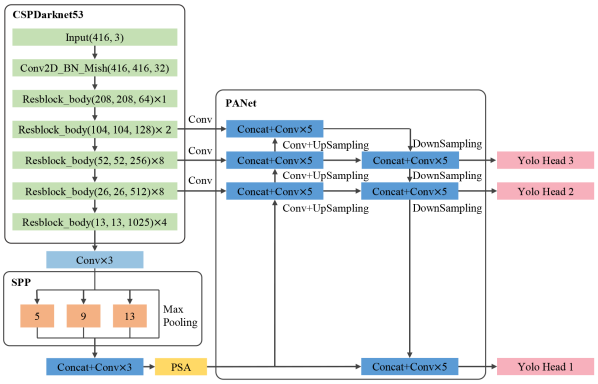


Fig. 4. The structure of the improved model YOLOv4-P

IV. Experimental Results and Analysis

A. data set

The data set used in this article comes from the Safety Helmet Wearing Dataset (SHWD), which is open source on the Internet. After data cleaning, 7581 images are finally obtained, including 9044 bounding boxes (positive types) wearing helmets, and 111514 Bounding box (negative type) without a helmet. In addition, part of the negative data in this data set comes from the SCUT-HEAD data set, which is used to determine whether it is a person who does not wear a helmet. Before training, the data set is divided into training set, validation set and test set at a ratio of 7:1:2.

B. Experimental platform and configuration

The operating system version of the experimental machine in this article is windows 10, the CPU model is AMD Ryzen 7 5800H, the GPU model is GeForce RTX 3060, the video memory size is 6GB, and the memory size is 16GB. All models are based on Pytorch 1.6 and use cuda 11.0 to accelerate the GPU.

In the analysis of the experimental results, the Mean Average Precision (mAP) is used as the model detection accuracy evaluation index [23–25], and the calculation formula is as follows.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$AP = \int_0^1 PR\mathrm{d}r \tag{4}$$

$$mAP = \frac{1}{n}\sum_1^n AP \tag{5}$$

In the above formula, TP refers to real cases, FP refers to false positive cases, FN refers to false negative cases, P refers to Precision (precision rate), R refers to Recall (recall rate), n is the number of categories, the helmet used in this article The number of data set categories is n=2.

C. Experimental process and result analysis

Since the features of the backbone feature extraction network are universal, in order to speed up the model training and prevent the weights from being destroyed at the beginning of the training, the experiment in this paper adopts the method of freezing training, first freezing the training of this part of the weights, and putting more resources on training The latter part of the network parameters, so that time and resource utilization can be greatly improved. The training model adopts the Adam optimizer, the initial learning rate of freezing training is set to 0.001, Epoch is set to 100, and Batch_size is set to 4.

In this paper, an ablation experiment is designed to measure the impact of the proposed k-means, PSA, and increased feature layers on the final detection results. During the test, the IOU threshold is selected as 0.3 and the confidence is set as 0.5. The experimental results are shown in Table I, where hat is The helmet is detected, and the person means that the helmet is not detected. Only improve prior bounding box re-clustering (YOLOv4+k-means); only improve pyramidal split attention module (YOLOv4+PSA); only improve feature layer fusion (YOLOv4+layer).

TABLE I
COMPARISON OF ABLATION RESULTS

Method	AP/%		mAP/%	FPS/($f \cdot s^{-1}$)
	hat	Person		
YOLOv3	92.49%	83.69%	88.09%	45.1
YOLOv4	92.09%	90.62%	90.62%	60.7
YOLOv4+k-means	92.66%	90.73%	91.69%	61.4
YOLOv4+PSA	92.23%	89.60%	90.92%	58.6
YOLOv4+layer	92.71%	91.36%	92.03%	57.5
YOLOv4-P	93.28%	92.26%	92.77%	55.6

By changing the prior bounding box parameters through clustering on the basis of the original YOLOv4 model and retraining the network model, the AP values of both hat and person were improved to some extent, and the mAP value was increased by 1.07%. The experimental results show that after clustering on the dataset by k-means clustering algorithm and using the new prior bounding box parameters, the match between the prior bounding box and the target frame is improved, which improves the detection accuracy and reduces the missed detection rate to some extent.

By conducting experiments on the safety cap dataset SHWD, the results show that the YOLOv4-P algorithm improves the mAP value by 2.15% and the AP values on hat and person categories by 1.19% and 1.64%, respectively, compared with the YOLOv4 algorithm. yOLOv4-P improves the recall rate of aircraft project in remote sensing images while ensuring the accuracy rate. Although the detection speed is 5.1 slower

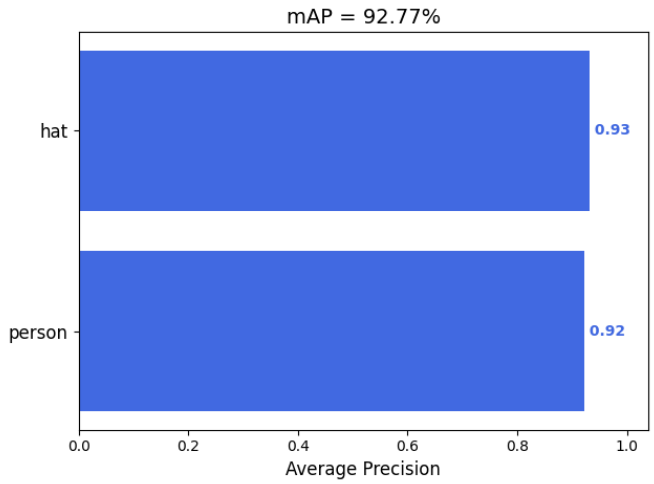
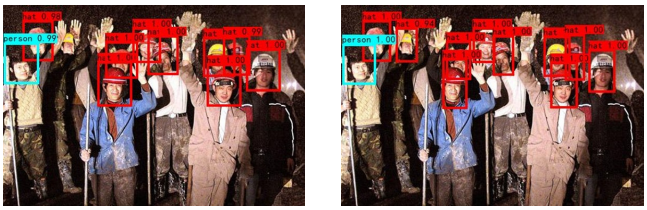


Fig. 5. mAP experimental results of YOLOv4-P model training

than the original YOLOv4, it still meets the requirement of real-time detection. That is, the fusion of the three by re-clustering the prior bounding box, adding the PSA module and improving the feature layer makes the model more sensitive to the shallow features of the helmet and has better detection performance for small project.

Figure 6 shows the comparison of the original YOLOv4 algorithm and the YOLOv4-P algorithm in the SHWD data set. It can be seen that the method in this paper has a lot of advantages in the detection of helmet wearing. In summary, it can be seen that the algorithm YOLOv4-P in this paper has a better detection effect than YOLOv4.



(a) (b)

Fig. 6. Comparison of detection results of YOLOv4 (a) and YOLOv4-P (b)

In the same environment, SSD, Faster R-CNN and RetinaNe algorithms were used for comparison, and mAP and recognition frames per second FPS were used as the evaluation index of detection effect, and the results are shown in Table II.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS OF VOC 2007 DATASET

Method	mAP/%	FPS/($f \cdot s^{-1}$)
SSD	83.68%	80.0
Faster R-CNN	87.35%	10
RetinaNet	85.31%	82.4
YOLOv4	90.62%	60.7
YOLOv4-P	92.77%	55.6

V. Conclusion

To solve the problems of small object detection and low detection accuracy of staff wearing helmets, this paper increases the match between the prior bounding box and the actual object box by re-clustering the prior bounding box using k-means clustering algorithm; secondly, a PSA module is added behind the SPP module to enhance the ability to process the spatial information of multi-scale feature maps; finally, the backbone feature extraction of YOLOv4 network by adding a 104×104 feature output layer to enhance feature fusion through multiple convolution and sampling, thus improving the detection capability of the model for small objects. The experimental results show that the YOLOv4-P algorithm improves the mAP value by 2.15% compared with the YOLOv4 algorithm, which improves the accuracy of helmet wearing detection.

References

[1] Hao Zhong, Wei Yanxiao. Analysis of standard statistical characteristics of 448 cases of construction accidents[J]. China Standardization, 2017,000(002): 245-247.

[2] GIRSHICK R. DONAHUE J. DARRELL T. Rich feature hierarchies for accurate object detection and semantic segmentation[C] //IEEE Conference on Computer Vision and Pattern Recognition(CVPR).2014.

[3] Girshick R. Fast R-CNN [C] //IEEE International Conference on Computer Vision, 2015:1440 -1448.

[4] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real time object detection with region proposal net-works [C] //28th International Conference on Neural Information Processing Systems, 2015:91 -99.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C] //European Conference on Computer Vision, 2016:21-37.

[6] BOCHKOVSKIY A.WANG C Y.LIAO H Y M.Yolov4: optimal speed and accuracy of object detection[C]//IEEEConference on Computer Vision and Pattern Recognition CVPR).2020.

[7] REDMON J. FARHADI A. Yolov3: an incremental improvement[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR).2018.

[8] REDMON J. FARHADI A. Yolo9000: better, faster, stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR).2016.

[9] REDMON J. DIVVALA S. GIRSHICK R. You only look once: unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016.

[10] Xu Kai, Deng Chao. Safety helmet wearing recognition algorithm based on improved YOLOv3 [J]. Progress in Laser and Optoelectronics, 2021, 58(06): 300-307.

[11] Li Weigang, Yang Chao, Jiang Lin, Zhao Yuntao. Indoor scene object detection based on improved YOLOv4 algorithm[J]. Progress in Laser and Optoelectronics, 2021.

[12] Wang Bing, Le Hongxia, Li Wenjing, Zhang Menghan. Improved mask detection algorithm for YOLO lightweight network [J]. Computer Engineering and Applications, 2021, 57(08): 62-69.

[13] Hu Jie et al. Squeeze-and-Excitation Networks.[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(8) : 2011-2023.

[14] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. In British Machine Vision Conference(BMVC), 2018.

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In European Conference on Computer Vision (ECCV), 2018.

[16] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Pyramidal convolution: rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:2006. 11538, 2020.

[17] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2):652-662, 2021.

[18] Pengcheng Yuan, Shufei Lin, Cheng Cui, Yuning Du, Ruoyu Guo, Dongliang He, Errui Ding, and Shumin Han. Hs-resnet: Hierarchical-split block on convolutional neural network. arXi preprint arXiv:2010.07621, 2020.

[19] Hu Zhang, Keke Zu, Jian Lu, et al. EPSANet: An efficient pyramid split attention block on convolutional neural network[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway. arXiv 2021.

[20] Yang Xiaogang, Gao Fan, Lu Ruitao, Li Weipeng, Zhang Tao, Zeng Jun. Lightweight aviation object detection method based on improved YOLOv5 [J/OL]. Information and control: 1-7 [2021-10-28].

[21] Zhou Weina, Ding Haowen, Zhou Ying. A real-time detection method for weak and small moving ships at sea[J]. Journal of Hefei University of Technology (Natural Science Edition), 2021, 44(09): 1187-1192.

[22] Li Fujin, Meng Luda, Ren Hongge. Object detection algorithm based on two-way fusion SSD[J]. Modern Electronic Technology, 2021, 44(19): 81-84.

[23] Zhang Chenchen, Jin Hong. Research on object detection technology based on improved YOLOv3-Tiny [J]. Ordnance Equipment Engineering Journal, 2021, 42(09): 215-218+312.

[24] Zhong Zhifeng, Xia Yifan, Zhou Dongping, Yan Yangtian. Lightweight object detection algorithm based on improved YOLOv4 [J/OL]. Computer application: 1-8 [2021-10-28].

[25] Li Wenjing, Xu Guowei, Kong Weigang, Guo Fengxiang, Song Qingzeng. Research on plant leaf and stem intersection object detection based on improved YOLOv4 [J/OL]. Computer Engineering and Applications: 1-8 [2021-10-28].