Smaller Target Detection Algorithms Based on YOLOv5 in Safety Helmet Wearing Detection

2022 4th International Conference on Robotics and Computer Vision (ICRCV) | 978-1-6654-8170-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRCV55858.2022.9953233

Kai-Yang Cao
Department of Computer Science and Technology
Yanbian University
Yanji, Jilin, China
396343802@qq.com

Xu Cui
Department of Computer Science and Technology
Yanbian University
Yanji, Jilin, China
xcui@ybu.edu.cn

Jin-Chun Piao*
Department of Computer Science and Technology
Yanbian University
Yanji, Jilin, China
*Corresponding Author: jcpiao@ybu.edu.cn

***Abstract*—In many construction scenarios, the frequency of workers' safety accidents has gradually increased. The main reason is that workers do not wear safety helmets. Safety helmets are a kind of safety guarantee for workers, which can effectively prevent workers' head injuries. Therefore, wearing a helmet correctly can effectively reduce the frequency of safety accidents. With the continuous development of the field of computer vision, deep learning has made remarkable achievements in safety helmet wearing detection. For safety helmet wearing detection, we can detect whether workers in construction sites wearing safety helmet by establishing a detection model. Aiming at the shortcomings of the existing helmet detection algorithms, it is difficult to detect small targets, occluded targets, and dense targets. This paper proposes a helmet wearing detection algorithm based on YOLOv5. Add a smaller target detection layer P2 to improve the efficiency of small target detection, and we use the FReLU activation function to replace the original SiLU activation function. Finally, the coordinate attention is added to the backbone network to enhance the network's learning of the details of the safety helmet target. The experimental results show that under the SHWD dataset, compared with the original model, the average accuracy of the model reaches 95.2%, an improvement of about 1.9%. The improved algorithm has good accuracy and practicability for the helmet detection task, and can better meet the actual needs of the construction site.**

Keywords—Deep learning, Safety helmet wearing detection, YOLOv5, The smaller target detection layer, Coordinate attention, Activation function

I. INTRODUCTION

In recent years, more and more people's research on artificial intelligence has been applied to real life. And the safety helmet wearing detection is one of them. Safety helmets are a kind of safety guarantee for workers. Safety helmet can protect workers' heads from external injuries. If the head is injured, the result is often fatal. Although safety helmets can effectively protect workers, there are still many workers who do not wear safety helmets in the construction environment. So there are many potential safety hazards, resulting in safety accidents. Although it is mandatory for workers to wear safety helmets, in the actual construction site. It is inevitable that many workers do not wear safety helmets for various reasons. This is mainly because workers lack self-safety awareness, so they ignore the importance of safety helmets.

Today, we usually use manual methods to determine whether workers are wearing safety helmets, which inefficient, costly, time-consuming, and prone to errors. We can move artificial intelligence technology into helmet detection. So that the detection efficiency is much higher than that of manual detection. Finally the task of helmet detection can be efficiently completed.

In previous studies, we have conducted research on target detection with the YOLO series [1]. In this paper, we choose YOLOv5 [2] as the baseline of the model. As a more advanced single-stage detector, it has achieved good results by itself. But there are many deficiencies in the direct application of the helmet detection task. In order to make the model more suitable for the safety helmet wearing detection, we have made many improvements to the original model.

II. RELATED WORK

In recent years, some improved models for safety helmet wearing detection have not only brought certain contributions to the task of safety helmet wearing detection, but also exposed some shortcomings. More and more scholars at home and abroad take advantage of the accuracy and real-time nature of target detection technology to integrate it into the task of helmet detection. Target detection algorithms based on deep learning are mainly divided into two-stage detection algorithms based on candidate boxes and one-stage detection algorithms that can complete detection and classification tasks with only one network. The two-stage target detection algorithm has the advantage of high accuracy, but its detection speed is slow. The advantage of the one-stage object detection algorithm is that it is faster, but the accuracy is lost.

In recent years, some helmet detection methods based on improved models have not only brought certain contributions to the helmet detection task, but also exposed some shortcomings. Benyang Deng et al. [3] proposed an improved model based on YOLOv4. They used a multi-scale training strategy to improve the adaptability of the model under different detection scales. The detection strength of the target is not enough. Bin Wang et al. [4] proposed an improved model based on YOLOv4, named YOLOv4-P, they used the K-mean

This work was supported in part by the Jilin Provincial Natural Science Foundation (No. YDZJ202201ZYTS566); in part by the Education Department of Jilin Province of China (No. JJKH20210570KJ); and in part by the National Natural Science Foundation of China (No. 62062064).

clustering algorithm to readjust the parameters of the prior bounding box to improve the matching degree between the prior bounding box and the object, and introduced Pyramid Split Attention (PSA) model to further process multi-scale feature information, but the performance is not outstanding in accuracy. Huang Li et al. [5] proposed an improved model based on YOLOv3. The algorithm performs pixel feature statistics on the predicted anchor box, and then multiplies it by the weight coefficient to output the confidence level of the helmet standard wearing in each predicted anchor box area. Finally, It can judge whether the worker wears a helmet. It improves the training speed, and realizes feature enhancement. Its performance is not good in a complex environment, so it is difficult to apply to the more demanding helmet detection task. ZP Xu et al. [6] proposed a YOLOv5 model with the squeeze-and-excitation module added. This method can not only obtain the weight of the image channel, but also accurately separate the foreground and background of the image. The disadvantage is that the amount of parameters is large, and it is difficult to deploy on the mobile terminal.

In view of the shortcomings of previous scholars' research, this paper improves the helmet detection model. Through a series of comparative experiments, the YOLOv5 algorithm with superior performance is finally selected as the model baseline. And corresponding improvements are made in combination with the difficult problems in the helmet detection task. So that it can effectively perform the task of safety helmet wearing detection, and has higher detection accuracy and stronger generalization ability.

III. Methodologies

YOLOv5 shows good performance in target detection, soit satisfies the task of safety helmet wearing detection. But there are also problems such as too high model complexity for single target detection task and poor detection effect on smaller targets. Based on this problem, we improve the YOLOv5 model to make it better used in the safety helmet wearing detection.

The improved YOLOv5 algorithm uses the CSPDarkNet53 network as the backbone network and PANet as the feature fusion network. After inputting the image into the backbone network CSPDarkNet53, four feature layers are obtained. The four feature layers have information of different dimensions. As a more advanced single-stage detector, YOLOv5 itself has achieved good results, making it more suitable for helmets. Make it a better model for the safety helmet wearing detection, thus adding a coordinate attention to the convolutional layers. And use the FReLU activation function to replace the original SiLU activation function. And through the top-down and bottom-up bidirectional fusion backbone network PANet, the features extracted by the backbone network multi-scale fusion. Finally sent them to the detection layer, thus enhancing the representation ability of the backbone network. The improved YOLOv5 algorithm proposed in this paper improves the accuracy of helmet detection.

A. The Smaller Target Detection Layer P2

One reason why YOLOv5 small target detection effect is not good is because the size of small target samples is small, and the down-sampling multiple of YOLOv5 is relatively large. And it is difficult for deeper feature maps to learn the feature information of small targets. Therefore, we proposed adding a smaller target detection layer P2 to detect after splicing the shallow feature map and the deep feature map.

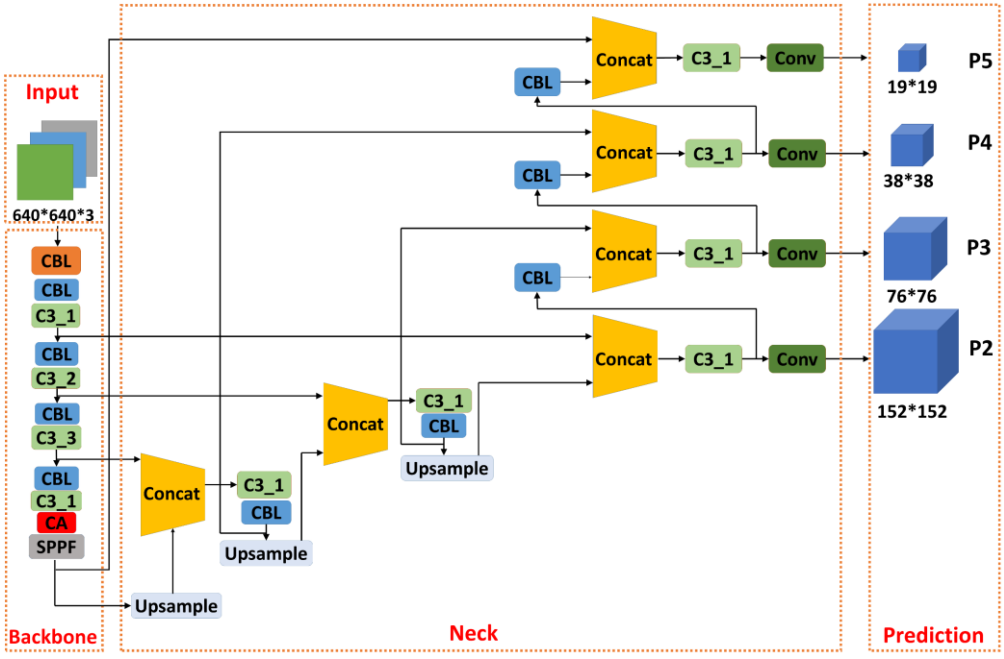

Fig 1. Smaller target detection algorithms based on YOLOv5 in safety helmet wearing detection

The original model of YOLOv5 has only three detection layers, namely P3, P4, and P5. So they correspond to three sets of initialized anchor values. Continue to do the up-sampling

operation in the feature pyramid, and perform further feature fusion with the backbone. A regression and classification convolution block is added to deal with small layers, namely P3, P4, and P5. So they correspond to three sets of initialized anchor values. Continue to do the up-sampling operation in the feature pyramid, and perform further feature fusion with the backbone. A regression and classification convolution block is added to deal with small target detection. In the improved YOLOv5 model, the input image size is 640×640. And the feature maps corresponding to the three detection layers P3, P4, and P5 of the original model are 80×80, 40×40, and 20×20, which used for detection size 8×8, 16×16, 32×32. The overall structure of the improved YOLOv5 network proposed in this paper is shown in Fig. 1.

In this paper, a new smaller target detection layer P2 is added. And the corresponding detection feature map size is 160×160, which is used to detect targets with a size of 4×4 or more. After the improvement, although the calculation amount and detection speed have increased. The detection accuracy of small targets has been significantly improved.

B. *Coordinate Attention*

The coordinate attention [7] mechanism decomposes channel attention into two 1D feature encoding processes that aggregate features along two spatial directions, respectively. Its structure diagram is shown in Fig. 2.

In order to obtain the attention on the width and height of the image and encode the precise position information. The input feature map is firstly divided into two directions of width and height for global average pooling respectively. And the feature maps in the two directions of width and height are obtained respectively. The output of the t channel of height g is expressed as:

$$z_t^g(g) = \frac{1}{K} \sum_{0 \le i \le K} x_t(g, i) \tag{1}$$

The output of the t channel at width k can be expressed as:

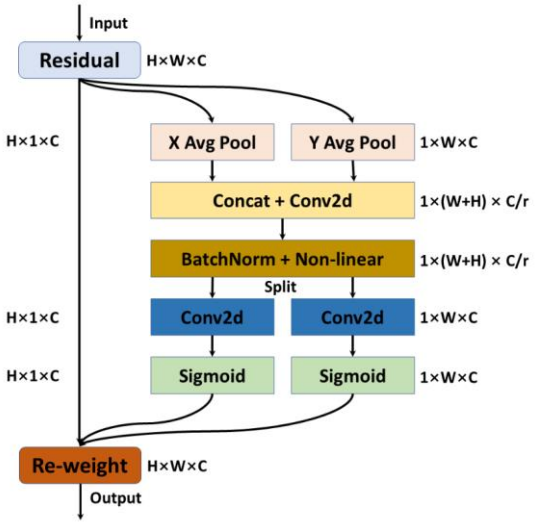

Fig. 2. Structure of coordinate attention

$$z_t^k(k) = \frac{1}{G} \sum_{0 \le j \le G} x_t(j, w) \tag{2}$$

The above two transformations aggregate features from different orientations to generate a pairs of orientation-aware feature maps z^g and z^k, which then fed into a shared 1×1 convolutional transformation function F_1 to output an intermediate feature map f.

$$f = \delta\left(F_1([z^g, z^k])\right) \tag{3}$$

Among them,$[\cdot,\cdot]$represents the connection operation, and is the nonlinear activation function. $f \in R^{\frac{T}{r} \times (G+K)}$encodes the horizontal and vertical direction information, and r is a scaling factor used to control the size of the SE (Squeeze-and-Excitation) block.

Next, f is decomposed into two independent tensors $f^g \in R^{\frac{T}{r} \times G}$ and $f^k \in R^{\frac{T}{r} \times K}$ by spatial direction, and then utilize two 1×1 The convolutional transformations F_g and F_k respectively transform these two tensors into tensors with the same number of channels to the input, resulting in:

$$\varphi^g = \sigma\left(F_g(f^g)\right) \tag{4}$$

$$\varphi^k = \sigma\left(F_k(f^k)\right) \tag{5}$$

σ is the sigmoid activation function.

After the above calculation, the attention weight is φ^g in the height direction and in the width direction φ^k. In the direction of the input feature map will be obtained.

Finally, the feature map with the attention weight is obtained by multiplying the weighted calculation on the original feature map:

$$y_t(i, j) = x_t(i, j) \times \varphi_t^g(i) \times \varphi_t^k(j) \tag{6}$$

C. *FReLU*

In view of the complexity of the network, we use the FReLU [8] activation function specially used for computer vision tasks is used to replace the original SiLU activation function. As shown in Fig. 3.

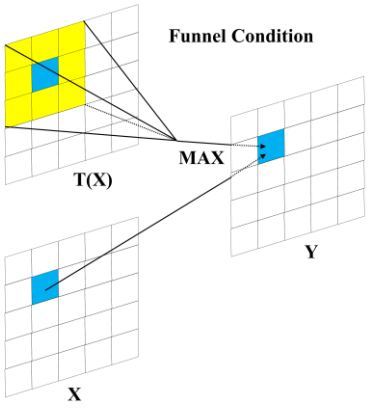

Fig. 3. FReLU activation function

The ReLU functions are extended by adding a spatial condition to enhance the sensitivity of the activation space and significantly improve image vision. FReLU is a two-dimensional funnel-like condition $T(X)$, which depends on the spatial context, and the visual condition helps to extract the fine spatial layout of objects. It improves vision tasks by adding a negligible computational overhead. The activation function has the following form:

$$F = max\big(x, T(x)\big) \tag{7}$$

where $T(x)$ represents the spatial context feature extractor. By using spatial condition, FReLU can simply extend ReLU to a visually parametric ReLU with pixelated modeling capabilities.

IV. Experiments

A. Experimental Environment and Datasets

In order to make the algorithm run effectively, we built a deep learning environment on Windows 10 system, PyTorch1.8.2 and Python3.7. All experimental procedures are evaluated on a GTX1060 with 6GB memory.

In terms of datasets, we use the SHWD dataset with a total of 7581 images (train set:validation set:test set=7:2:1).

B. Different Attention Mechanisms

TABLE I. Comparison of Different Attention Mechanism

CA	CBAM	SE	mAP
			93.3%
√			**93.6%**
	√		**93.4%**
		√	**93.0%**

In Table I, Compared with SE [10] and CBAM [11] attention, the CA attention with better performance is finally selected.

C. Different Activation Function

TABLE II. Comparison of Different Activation Function

SiLU (Original)	ReLU	FReLU	ACON	mAP
√				**93.3%**
	√			**93.4%**
		√		**93.6%**
			√	**93.1%**

In Table II, We compared with ReLU [12], ACON [13] activation function. Finally, we chose the FReLU activation function, which achieved best results.

D. Ablation Experiment

TABLE III. Ablation Experiments on SHWD

P2-Layer	CA	FReLU	mAP
			93.3%
√			**94.8%**
	√		**93.6%**
		√	**93.7%**
√	√		**95.0%**
√	√	√	**95.2%**

P2-Layer module: By adding the P2-Layer, the performance of small target detection is improved.

CA module: It can make the network pay more attention to the target to be detected and improve the detection effect.

FReLU module: FReLU has better context capture capability. And it has a better understanding of small targets. Therefore, it is beneficial to detect smaller targets.

In Table III, we performed ablation experiments. It can be seen from the table that for different improvements. The model detection accuracy has been improved to varying degrees. Finally, the three improved models are used at the same time, which is about 1.9% higher than the original model accuracy.

E. Comparative Experiment

TABLE IV. Experimental Results of Different Methods

Methods	SHWD			
	Precision	*Recall*	*F1-Score*	*mAP*
Improved YOLOv3[4]	90.5%	88.6%	89.6%	89.4%
YOLOv4-P[3]	92.1%	88.1%	90.1%	92.8%
Improved YOLOv4[2]	91.8%	90.8%	91.3%	92.7%
Improved YOLOv5[5]	92.7%	90.9%	91.8%	94.5%
ConvNeXt[9]	91.3%	88.1%	89.7%	92.2%
YOLOX[14]	93.8%	84.2%	88.4%	92.8%
YOLOv5[1]	92.6%	88.8%	90.7%	93.3%
Ours	**93.1%**	**91.6%**	**92.3%**	**95.2%**

To more intuitively evaluate the performance of our proposed model, above models are used for comparative experiments. In Table IV, we reproduced these models and used the SHWD dataset to compare the precision, Recall, F1 and mAP of the above models. On the whole, our model outperformed other comparative models, and achieved 95.2% mAP. This shows that our improvement has the effect of improving the feature extraction ability of the model, making it more capable of processing the details of the image, especially the detection effect of small targets is better.

Through comparative experiments, the effectiveness of the algorithm is further proved.

F. Sample of Results

Fig. 4 shows the detection results of part of the test set (IoU=0.5). (a) shows the detection results of the original YOLOv5 model, and (b) the prediction results of the improved YOLOv5 model. It can be seen that in the original model, some small targets are prone to false detection and missed detection. But after the improvement, these problems are effectively alleviated. Therefore, the improved YOLOv5 model proposed

in this paper can be more adaptable to changing environments and small target detection.

(a)YOLOv5

(b)Ours

Fig 4. Comparison example of YOLOv5 and ours test results

V. Conclusion

In order to better protect the safety of workers on the construction site, this paper proposes an smaller target detection algorithms based on YOLOv5 in safety helmet wearing detection. Corresponding improvements are made to the defects of YOLOv5 performance in safety helmet wearing detection. Based on the original YOLOv5 with only three detection layers, the smaller target detection layer P2 is added. And a regression and classification convolution block is added to deal with small target detection. And good detection results have been achieved in small target detection. At the same time, the coordinate attention mechanism is added to the backbone network, which further improves the model's learning of sample details. Finally, we use the FReLU activation function specially used for computer vision tasks is used to replace the original SiLU activation function. A series of experimental results shows that this algorithm achieves high accuracy in small target detection. But it produces a large amount of computation. In the future, we need to do some lightweight processing appropriately to improve the detection speed.

Acknowledgment

This work was supported in part by the Jilin Provincial Natural Science Foundation (No. YDZJ202201ZYTS566); in part by the Education Department of Jilin Province of China (No. JJKH20210570KJ); and in part by the National Natural Science Foundation of China (No. 62062064).

References

[1] Zhou, Li-Qun and Sun, Peng and Piao, Jin-Chun and others, "A Novel Object Detection Method in City Aerial Image Based on Deformable Convolutional Networks," IEEE Access, vol. 10, pp. 31455-31465, 2022.

[2] G. Jocher, Ultralytics/Yolov5: V6.1 Bug Fixes and Performance Improve-ments, Feb. 2022, doi: 10.5281/zenodo.4154370. [Online]. Available:https://github.com/ultralytics/yolov5

[3] Benyang, Deng and Xiaochun, Lei and Miao, Ye, "Safety helmet detection method based on YOLOv4," in 2020 16th International Conference on Computational Intelligence and Security (CIS), 2020, pp. 155-158.

[4] Wang, Bin and Xiong, Haojie and Liu, Lishou, "Safety helmet wearing recognition based on improved YOLOv4 algorithm," in 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC), 2022, pp. 1732-1736.

[5] Huang, Li and Fu, Qiaobo and He, Meiling and Jiang, Du and Hao, Zhiqiang, "Detection algorithm of safety helmet wearing based on deep learning," Concurrency and Computation: Practice and Experience, vol. 33, no. 13, pp. e6234, 2021.

[6] Xu, ZP and Zhang, Y and Cheng, J and Ge, G, "Safety Helmet Wearing Detection Based on YOLOv5 of Attention Mechanism," in Journal of Physics: Conference Series, 2022, pp. 012038.

[7] Hou, Qibin and Zhou, Daquan and Feng, Jiashi, "Coordinate attention for efficient mobile network design," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713-13722.

[8] Ma, Ningning and Zhang, Xiangyu and Sun, Jian, "Funnel activation for visual recognition," in European Conference on Computer Vision, 2020, pp. 351-368.

[9] Liu, Zhuang and Mao, Hanzi and Wu, Chao-Yuan and Feichtenhofer, Christoph and Darrell, Trevor and Xie, Saining, "A convnet for the 2020s," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976-11986.

[10] Hu, Jie and Shen, Li and Sun, Gang, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[11] Woo, Sanghyun and Park, Jongchan and Lee, Joon-Young and Kweon, In So, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.

[12] Glorot, Xavier and Bordes, Antoine and Bengio, Yoshua, "Deep sparse rectifier neural networks," in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 315-323.

[13] Ma, Ningning and Zhang, Xiangyu and Liu, Ming and Sun, Jian, "Activate or not: Learning customized activation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8032-8042

[14] Ge, Zheng and Liu, Songtao and Wang, Feng and Li, Zeming and Sun, Jian, "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.