Computer Systems Science & Engineering
DOI: 10.32604/csse.2022.028224
Article

Tech Science Press

Real-time Safety Helmet-wearing Detection Based on Improved YOLOv5

Yanman Li[1], Jun Zhang[1], Yang Hu[1], Yingnan Zhao[2,*] and Yi Cao[3]

[1]NARI Nanjing Control System Ltd., Nanjing, 210032, China
[2]School of Computer, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[3]Department of Electrical and computer Engineering, University of Windsor, Windsor, N9B 3P4, Canada
*Corresponding Author: Yingnan Zhao. Email: zh_yingnan@126.com
Received: 05 February 2022; Accepted: 30 March 2022

Abstract: Safety helmet-wearing detection is an essential part of the intelligent monitoring system. To improve the speed and accuracy of detection, especially small targets and occluded objects, it presents a novel and efficient detector model. The underlying core algorithm of this model adopts the YOLOv5 (You Only Look Once version 5) network with the best comprehensive detection performance. It is improved by adding an attention mechanism, a CIoU (Complete Intersection Over Union) Loss function, and the Mish activation function. First, it applies the attention mechanism in the feature extraction. The network can learn the weight of each channel independently and enhance the information dissemination between features. Second, it adopts CIoU loss function to achieve accurate bounding box regression. Third, it utilizes Mish activation function to improve detection accuracy and generalization ability. It builds a safety helmet-wearing detection data set containing more than 10,000 images collected from the Internet for preprocessing. On the self-made helmet wearing test data set, the average accuracy of the helmet detection of the proposed algorithm is 96.7%, which is 1.9% higher than that of the YOLOv5 algorithm. It meets the accuracy requirements of the helmet-wearing detection under construction scenarios.

Keywords: Safety helmet wearing detection; object detection; deep learning; YOLOv5; Attention Mechanism

1 Introduction

It is well known that a surveillance system is significant for construction site safety. Intelligent surveillance in construction sites has recently adopted artificial intelligence technologies like computer vision and machine learning. It can avoid a time-consuming labor-intensive task, point out the equipment fault and illegal worker operation in time and effectively prevent accidents. The safety helmet wearing detection is a standard inspection item on the construction site, closely related to the workers' lives. Realizing the automatic detection of safety helmet wearing in the intelligent monitoring system has become a current research hot spot.

Traditional safety helmet-wearing detection methods use a sliding window-based region selection strategy. The hand-designed feature extractor is not very robust to the diversity of background. For

example, Waranusast et al. [1] proposed an automatic detection system for motorcycle helmets based on K-Nearest-Neighbor (KNN). Li et al. [2] applied the Hough transform to determine the helmet's shape and used the corresponding histogram to train the support vector machine (SVM) for detection. Traditional methods usually do not consider complex background environments and are sensitive to medium and small objects, leading to a high false alarm rate [3].

Recently, deep learning methods [4,5] have been effective in target detection [6]. Object detection methods based on deep learning can be classified into two stages and one stage. The two-stage method is a series of R-CNN (Region-CNN) algorithms, which the core is CNN (Convolutional Neural Network) [7,8], such as R-CNN [9], Fast R-CNN (Fast Regions with CNN) [10], and Faster R-CNN (Faster Regions with CNN) [11]. This method first uses a heuristic method to search the CNN network to generate candidate regions and then performs classification and regression on the candidate regions. The other is the single-stage algorithm of SSD (Single Shot Multi-Box Detector) [12] and YOLO series [13–15]. The CNN network is used to predict the categories and positions of different targets directly. The single-stage method is faster than the two-stage method, but it reduces the accuracy, making it suitable for video surveillance. The above method can also be applied to detecting wearing a helmet. Literature [16] proposed a safety helmet detection method based on Faster R-CNN and achieved a detection accuracy of 90%. In [17], the author made improvements based on Faster R-CNN, combining online mining with multi-part detection to identify whether workers are wearing safety helmets. The helmet detection algorithm proposed by Wu et al. is based on the improved YOLOv3 (YOLO version 3) [15] model. It seems that the accuracy of YOLOv3 is slightly better than that of SSD and slightly inferior to Faster R-CNN. However, the speed of YOLOv3 is at least twice that of SSD and Faster-RCNN [18]. In April 2020, Bochkovskiy et al. [19] proposed YOLOv4 (YOLO version 4). This model uses CSP (Cross Stage Partial) Darknet-53 as the backbone network and uses PANet (Path Aggregation Networks) to replace the FPN (Feature Pyramid Networks) algorithm in the YOLOv3 network. In June 2020, Jocher [20] proposed YOLOv5, which added the Focus structure to the backbone network to achieve a new benchmark for the best balance of speed and accuracy.

Based on YOLOv5, this paper adds the Attention Mechanism, Mish activation function, and CIoU Loss function to enhance the original YOLOv5's detection ability in small targets and occlusion situations. Experiments show that the algorithm proposed in this paper has superior performance.

The rest of the paper is organized as follows: Section 2 gives the basic principles of YOLOv5; Section 3 describes the improved YOLOv5 algorithm proposed in this article; Section 4 is the experimental part, which is compared and analyzed with related algorithms; Section 5 summarizes and points out further research directions.

2 YOLOv5

There are four versions in the official code of YOLOv5, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Among them, YOLOv5s has the smallest volume, and the other three are constantly deepening and widening on this basis. The work of this paper is based on YOLOv5s.

YOLOv5 includes input, Backbone, Neck, and prediction components. The input of YOLOv5 uses the same Mosaic data enhancement method as YOLOv4 and adaptive anchor box calculation and image scaling. Among them, adding the most miniature black padding to the original image adaptive scaling can effectively improve the interactive speed of YOLOv5.

In the Backbone part, YOLOv5 uses the Focus structure, which is unavailable in previous versions. It adopts a slicing operation, which can keep the information of the image intact when the image is down-sampled. For example, after the original 608 × 608 × 3 image is sliced, it first becomes a 304 × 304 × 12 feature map, and then after a convolution operation of 32 convolution kernels, it finally becomes a 304 × 304 × 32 feature map. Here, the Focus structure of Yolov5s uses 32 convolution kernels, while the number of the other three structures has increased. In the Backbone part, Yolov5 also uses a CSP

structure. Unlike Yolov4, which is only used in the backbone network, Yolov5 designs two CSP structures, CSP1_X structure in the Backbone part and CSP2_X in the Neck part.

The Neck part of Yolov5 is the same as that of Yolov4, adopting the structure of FPN+PAN. However, the Neck structure of YOLOv4 (YOLO version 4) uses ordinary convolution operations, while the CSP2 structure designed by CSPnet is used in YOLOv5 to strengthen the ability of network feature fusion.

3 Improved YOLOv5

YOLOv5 has a small volume, excellent speed, and detection rate, but its performance in recognition of small targets and occluded targets has decreased. Therefore, this article improves the original YOLOv5 from three aspects; the feature extraction part adopts the channel attention mechanism, the introduction of the Mish activation function, the use of the CIoU Loss function, and the corresponding DIoU (Distance IoU) NMS (Non-maximum Suppression) algorithm. The overall framework of the improved YOLOv5 algorithm is shown in Fig. 1, which consists of three parts: feature extraction, feature fusion, and prediction results. The images are input through multiresidual blocks to extract features. Here CBM represents the Convolution, Batch normalization, and Mish cativate function. In the stage of feature fusion, feature maps with different sizes are obtained in the residual blocks, and the feature maps obtained by upsampling are concatenated to obtain feature maps with different sizes of receptive fields. Finally, the prediction results on feature maps are carried out.

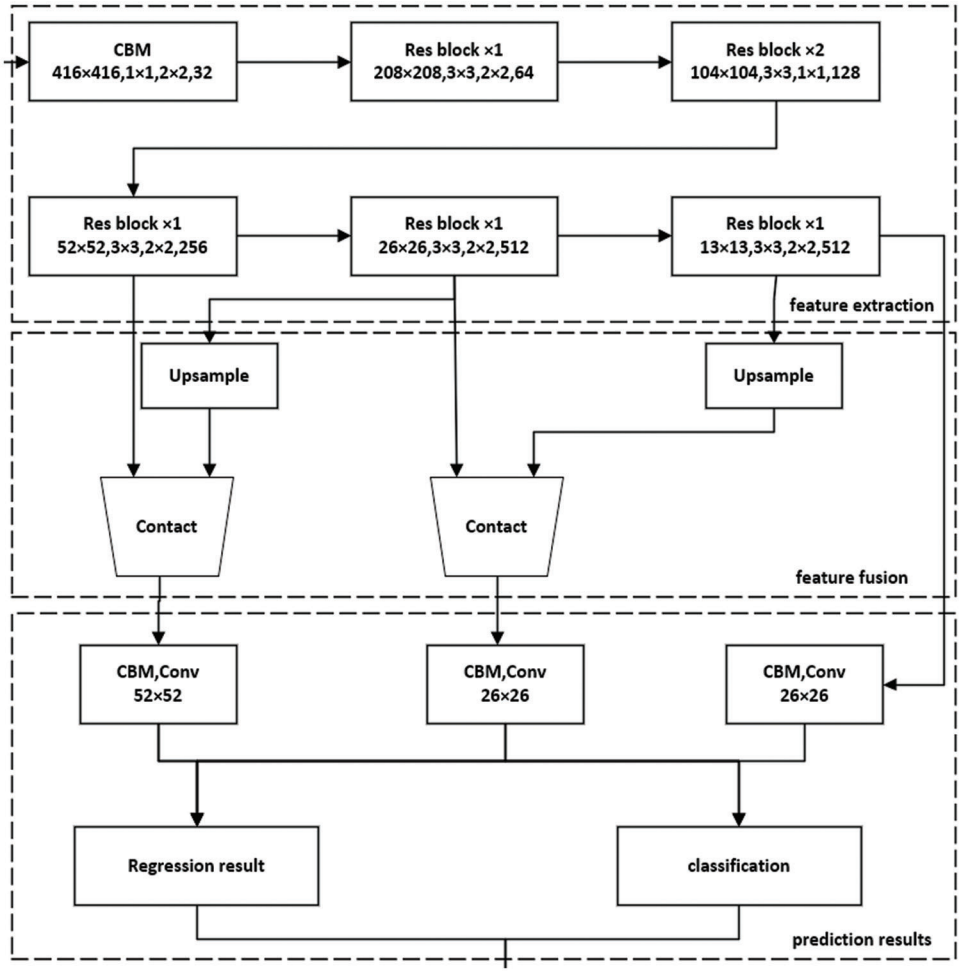

Figure 1: Framework of improved YOLOv5 algorithm

3.1 Attention Mechanism

The attention mechanism has shown great potential in object detection. Jaderberg et al. [21] proposed a spatial transformation to realize the spatial attention mechanism, and the spatial information in the image can be converted accordingly to extract critical information. The channel attention mechanism was proposed by Hu et al. [22], in which the importance between channels is calculated by two fully connected layers, and unimportant channel values are filtered out. Literature [23] introduced a residual attention network specifically designed for detection. The space and channel mechanisms are constructed by superimposing residual attention modules. S. Woo et al. developed the Convolutional Attention Module (CBAM) [24], multiplying feature maps and spatial attention mechanisms along the channel. CBAM can also be widely applied to other networks. Therefore, the attention mechanism has received extensive attention in target detection and has achieved excellent results.

The attention mechanism allows the neural network to focus on the shallow feature maps and allocate computing resources to more critical parts. Introducing the attention mechanism into the residual block can assign higher weights, thereby improving the ability of the entire network to express small targets.

In Fig. 2, the input feature map passes through a convolutional layer, the kernels of which are 1 × 1 and 3 × 3, respectively. F is the feature map of the channel attention mechanism. It uses the channel relationship between the features to generate the channel attention feature map, and the feature map F is weighted to obtain the channel feature map F1. Finally, it obtains the output feature map.

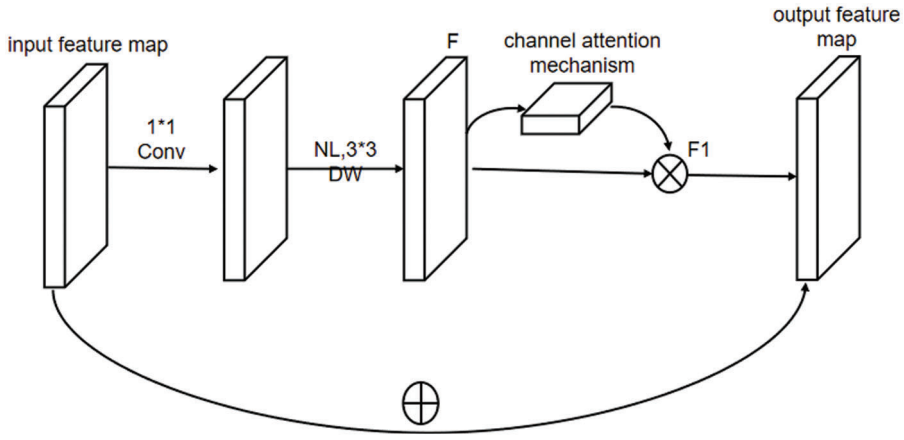

Figure 2: Channel attention mechanism fusion

The calculation formula of the channel attention module is as follows:

$$\begin{aligned} M_C(F) &= \delta(C1D_k(AvgPool(F) + MaxPool(F))) \\ &= \delta(C1D_k(F^C_{avg} + F^C_{\max})) \end{aligned} \tag{1}$$

where δ represents the activation function, $C1D_k$ represents the one-dimensional convolution, and k represents the adjacent channels of F, represents the fully connected layer. The second connection layer can receive k channel information from the first connection layer. F^C_{avg} and $F^C_{\max}$ respectively represent the feature maps of the channel attention module after average pooling and maximum pooling, where c is the dimension of the channel.

3.2 Activation Function

The activation function plays a vital role for the neural network model to understand complex and nonlinear functions, which can sufficiently express the nonlinear modeling ability of the network, making the neural network have a solid nonlinear learning ability. The YOLOv5 model uses the hidden layer activation function of Leaky ReLU [25]. Leaky ReLU assigns a small negative gradient value to all negative values of ReLU to solve the problem of gradient disappearance generated by the ReLU function when the input is negative. The formula is as follows,

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha x, & others \end{cases} \tag{2}$$

where the general value of α is 0.01, however, the effect of Leaky ReLU is not very stable, and it is not entirely better than ReLU in practical applications.

This paper uses the Mish activation function. The Mish activation function is relatively smoother, allowing better information to penetrate the neural network while allowing a slight negative gradient to flow into the negative value, achieving better accuracy and generalization ability. The mathematical formula can be expressed as

$$f(x) = x \cdot \tanh(\delta(x)) \tag{3}$$

where $\delta(x) = \ln(1 + e^x)$. It is a softmax activation function. The Mish activation function can significantly improve the accuracy of helmet detection. The Mish and Leaky ReLU activation functions are shown in Fig. 3. It depicts that Mish is smoother than Leaky ReLU in the negative part. A smooth activation function allows better information to penetrate the neural network, resulting in better accuracy and generalization.

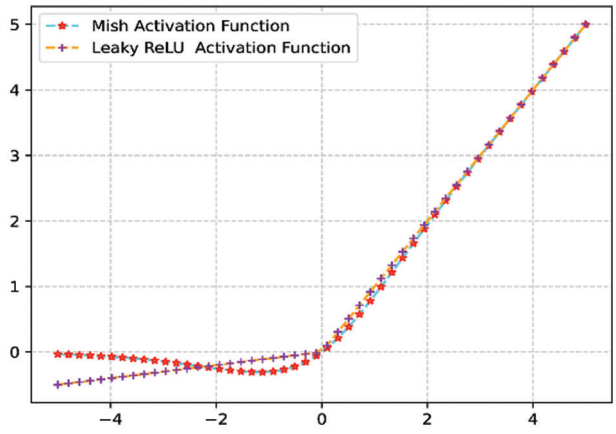

Figure 3: Mish and Leaky ReLU activation functions

3.3 Loss Function

The loss functions used in the YOLO series of algorithms are IoU(Intersection Over Union), GIoU (Generalized IoU), CIoU, and DIoU. Yolov5 applies GIoU_Loss as the loss function in the bounding box. However, it adopts CIoU_Loss in this scheme. Compared with GIoU, CIoU has faster converges of prediction frames and higher faster detection rates for small targets and occluded targets. The formula of the CIoU loss function can be as follows:

$$
\begin{aligned}
Loss_{CIoU} &= 1 - IoU + R_{DIoU} + \alpha v \\
R_{DIoU} &= \frac{\rho^2(b, b^{gt})}{c^2} \\
v &= \frac{4}{\pi^2}(\arctan\frac{W^{gt}}{h^{gt}} - \arctan\frac{W^P}{h^P})^2
\end{aligned} \tag{4}
$$

where IoU represents the intersection of the joint bounding boxes, R_{DIoU} is the distance between the two bounding boxes b and b^{gt} the center point, $\rho(\bullet)$ represents the Euclidean distance, and c represents the diagonal distance of the smallest rectangle formed by the two bounding boxes. α is the weight function, v is used to measure the similarity of the aspect ratio, as shown in the last formula in (5).

In the post-processing process of target detection, the screening of many target frames usually requires NMS, non-maximum suppression operation. YOLOv4 adopts the DIoU method on the basis of DIoU, and YOLO5 adopts the weighted NMS method. The DIoU_NMS method has a certain improvement effect on the recognition of occluded overlapping objects. Based on the CIoU loss function, we use DIoU-NMS for post-processing.

4 Experiment and Discuss

4.1 Data Construction

The data set is the prerequisite and primary condition for experiment development in deep learning detection. The only open-source safety helmet data set is Safety-Helmet-Wearing-Dataset [26]. The label data of the category of non-wearing helmets in this data set is monitoring images or photos taken by students in class in the classroom scene, not a standard. The construction site scenario data set does not meet the real-time monitoring requirements in the actual production environment. In order to solve this problem, this article self-made a helmet-wearing detection data set in the construction scene. The primary process of constructing the data set includes data collection, cleaning, and processing.

4.1.1 Data Collection

Public data set cleaning, construction site surveillance video, and network data collection. The collected data includes two types of pictures of workers wearing helmets and not wearing helmets in different environments, different resolutions, different colors of helmets, and different construction sites. In addition, to enhance the generalization ability of the model and increase the diversity of the data set, multiple sets of interference pictures are added to the data set. For example, construction workers wearing baseball caps, those who put their helmets on the table or the ground, those who hold their hands, those who wear sun hats, and those who wear police caps who wear non-safety helmets. Part of the data set samples collected this time are shown in Fig. 4.

4.1.2 Data Cleaning and Processing

Among the images collected from the surveillance video of the construction site or crawling on the Internet, many images do not contain the construction personnel, which is of no practical significance to the study. Therefore, it is necessary to clean this type of image data and select the images that meet the requirements to label them further.

Dimension reduction is an efficient preprocessing method, which can remove noise and unimportant features of high-dimensional data and subsequently improve data processing speed [27]. But here we just convert all the images that meet the requirements into .jpg format, and use the labeling tool labelImg to label each image, as shown in Fig. 5. Form the corresponding XML tag file. The file contains the four coordinates of the target in the frame and the given category. The format is PASCAL VOC [28].

(a) Normal samples

(b) Baseball cap samples

(c) Safety helmet samples on the ground

(d) Sun visor hat samples

(b) Police cap samples

(f) Samples on hand

Figure 4: Safety helmet samples

Figure 5: Labeling of helmet label

The data set has 10885 images, and the number of samples in the data set with and without helmets is shown in Tab. 1. The data set contains a variety of construction scenes, which can more fully reflect the actual construction scenes. The obtained data set is divided into a training set and a test set according to a specific ratio. In the last 10885 image data set, the number of samples in the training set is 9371, and the number of samples in the test set is 1514.

Table 1: Data set category assignment

Target category	Number of training set	Number of testing set
Wearing helmet category	5892	478
Not wearing helmet category	3479	1036

4.2 Experimental Environment and Network Training

4.2.1 Experimental Environment

The experiment in this article requires better hardware configuration and GPU (Graphics Processing Unit) to accelerate calculations. The model building, training, and result testing are all completed under the Pytorch framework, using the CUDA(Compuer Unified Device Architecture) parallel computing architecture, and at the same time integrating the cu-DNN acceleration library into the Pytorch framework to accelerate the computing power of the computer. The operating environment required for the experiment is shown in Tab. 2.

Table 2: Experimental environment

Type	Item	Edition
Hardware	Operating system	Ubuntu16.04
	Graphic card	RTX3090
	CPU	E5-2697
Software	Python	3.7
	Deep learning framework	Pytorch
	CUDA	11.1

4.2.2 Network Training

In the training process, the iteration number is 400, the weight attenuation coefficient is 0.0001, and the learning rate is 0.937 to prevent the model from overfitting. The maximum training batch is 16. The loss function value drops sharply from 0 to 240 times, and the loss number decreases slowly from 240 to 400 times. After 300 iterations, the loss value stabilizes around 0.015, and the model reaches the maximum excellent state. The training loss changes are shown in Fig. 6.

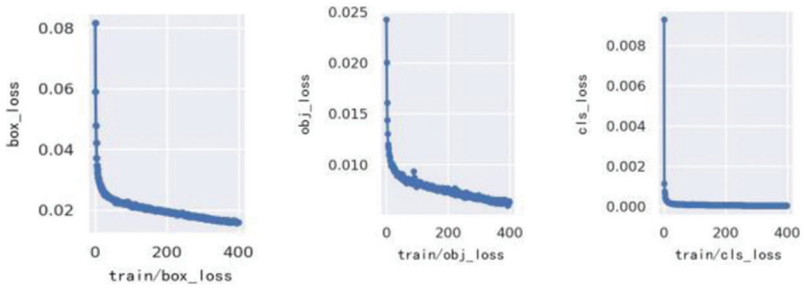

Figure 6: Convergence of improved YOLOv5

4.2.3 Result Analysis

In target detection, precision (P), recall (R), and mean average precision (mAP) are commonly used indicators to evaluate the performance and reliability of training models. This article also uses the above evaluation indicators to evaluate the performance of the helmet-wearing detection model. Therefore, it obtains two types of result images, including construction workers wearing safety helmets and construction workers not wearing safety helmets. It adopts True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) to describe accuracies. Specifically, TP_{hat} refers to the number of people not wearing safety helmets in the monitoring range of the construction site and the detection result is correct; FP_{hat} means the number of people who wear helmets but the result is wrong (that is, missed inspection); TN_{hat} represents the right detection result; FN_{hat} is the number of people who do not wear

safety helmet but are detected by mistake. The calculation of accuracy and recall is shown in the Eqs. (5)–(7).

$$recall_{hat} = \frac{TP_{hat}}{TP_{hat} + FN_{hat}} \tag{5}$$

$$precision_{hat} = \frac{TP_{hat}}{TP_{hat} + FP_{hat}} \tag{6}$$

$$AP_{hat} = \frac{TP_{hat} + TN_{hat}}{TP_{hat} + TN_{hat} + FP_{hat}} \tag{7}$$

where AP_{hat} is the average precision. Mean Average Accuracy (mAP) is the average of AP values under all categories, and the calculation formula is shown in (8), where Q is the total number of categories.

$$mAP = \frac{1}{Q}\sum\nolimits_{q \in Q} AP(q) \tag{8}$$

This paper applies the improved YOLOv5 algorithm to the detection of wearing safety helmets. In order to verify the superiority, we use the same number of test sets under the same configuration conditions, The baseline algorithms are Faster R-CNN, SSD, and YOLOv3, YOLOv4, YOLOv5, and improved YOLOv5. SSD and YOLO series are one-stage detection algorithms, and Faster R-CNN is a two-stage detection algorithm. The experimental results are evaluated by AP50 and mAP. Here, AP50 is the corresponding AP value when the IoU threshold is 0.5. Tab. 3 gives the experimental results. It implies that the improved YOLOv5 has significant superiority.

Table 3: The performance of different models

Models	AP50/%	mAP/%	Detection Time/ms	Model size/MB
Faster RCNN	81.7	62.1	260	182
SSD	79.2	73.6	119	188
YOLOv3	76.3	61.2	14	236
YOLOv4	79.5	65.3	9	246
YOLOv5	94.9	91.4	22	14
Improved YOLOv5	**96.7**	**93.1**	**23**	**15.3**

Some of the test results are shown in Fig. 7, where the word “hat” appears above the construction workers wearing safety helmets. The first column is the original samples, the second is the result of YOLOv5, and the third is the improved YOLOv5’s result. In Fig. 7a, shows the detection of targets of different sizes. YOLOv5 misses the detection of small targets in the distant view. Fig. 7b is the tiny target in the low-light construction scene. Due to insufficient illumination, the image is prone to blurring pixels of small objects. The YOLOv5 model misses more of this situation. However, the improved YOLOv5 performs better, illustrated in Fig. 7c. In the detection under intense light construction scene, it can be seen that YOLOv5 missed a construction worker wearing a helmet, and YOLOv5 improved the detection ultimately; Fig. 7d shows the detection of small targets in a long-distance construction scene. The comparison shows that the original YOLOv5 model has missed inspections for construction workers wearing safety helmets. The improved model has a better detection effect. Fig. 7e detects construction scenes obscured by steel bars. YOLOv5 has missed, while the algorithm proposed in the article has detected them all. From the above detection comparison in various construction scenarios, it shows that

the improved YOLOv5 algorithm is better for helmet detection in complex operating environments, but the reasoning time is relatively longer.

(a) Target detection of different sizes

(b) Detection of low-light construction scenes

(c) Detection of strong light construction scenes

(d) Detection far away from the construction scene

(e) Detection occluded by steel bars

Figure 7: Comparison of detection results of YOLOv5 and improved YOLOv5 algorithms in different construction scenarios

5 Conclusion

In this paper, an improved YOLOv5 algorithm is proposed to detect safety helmet wearing in construction scenes. The improved algorithm is based on the channel Attention Mechanism. The activation function and loss function is more suitable for detecting small targets and occlusion situations. Centralized verification on the self-made helmet wearing test data, the average accuracy of the helmet detection reached 96.7%. Compared with the YOLOv5 algorithm, the algorithm improves the average accuracy of wearing helmet detection by 1.9%, which meets the accuracy requirements for safety helmet-wearing detection in general construction environments.

In future work, we will attempt to improve the algorithm's accuracy and efficiency by employing some AM methods, such as spatial attention or channel and spatial hybrid attention.

Funding Statement: This work is supported by NARI Technology Development Co. LTD. (No. 524608190024).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Waranusast, N. Bundon, V. Timtong, C. Tangnoi and P. Pattanathaburt, "Machine vision techniques for motorcycle safety helmet detection," in *2013 28th Int. Conf. on Image and Vision Computing New Zealand (IVCNZ 2013)*, Wellington, New Zealand, pp. 35–40, 2013.

[2] J. Li, H. Liu, T. Wang, M. Jiang, S. Wang *et al.,* "Safety helmet wearing detection based on image processing and machine learning," in *2017 Ninth Int. Conf. on Advanced Computational Intelligence (ICACI)*, Doha, Qatar, pp. 201–205, 2017.

[3] B. Yogameena, K. Menaka and S. S. Perumaal, "Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system," *IET Intelligent Transport Systems*, vol. 13, no. 7, pp. 1190–1198, 2019.

[4] X. Zhang, X. Sun, W. Sun, T. Xu, P. Wang *et al.,* "Deformation expression of soft tissue based on BP neural network," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1041–1053, 2022.

[5] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2021. http://dx.doi.org/10.1109/TITS.2021.3130403.

[6] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time Small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 92, no. 6, pp. 1–16, 2021.

[7] S. Lee, "A study on classification and detection of small moths using CNN model," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1987–1998, 2022.

[8] R. Rajakumari and L. Kalaivani, "Breast cancer detection and classification using deep CNN techniques," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1089–1107, 2022.

[9] R. Girshick, J. Donahue, T. Darrell and J. Malik, "'Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, pp. 580–587, 2014.

[10] J. Zhu, Y. Guo, F. Yue, H. Yuan, A. Yang *et al.,* "A deep learning method to detect foreign objects for inspecting power transmission lines," *IEEE Access*, vol. 8, pp. 94065–94075, 2020.

[11] S. Ushasukhanya and M. Karthikeyan, "Automatic human detection using reinforced faster-rcnn for electricity conservation system," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1261–1275, 2022.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," in *Proc. of the European Conf. on Computer Vision*, Cham, Switzerland, Springer, pp. 21–37, 2016.

[13] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779–788, 2016.

[14] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6517–6525, 2017.

[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," CoRR, 1–6, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767.

[16] S. Chen, W. Tang, T. Ji, H. Zhu, Y. Ouyang *et al.,* "Detection of safety helmet wearing based on improved faster R-CNN," in *2020 Int. Joint Conf. on Neural Networks (IJCNN)*, Glasgow, UK, pp. 1–7, 2020.

[17] N. Li, X. Lv, S. Xu, Y. Wang, Y. Wang *et al.,* "Incorporate online hard example mining and multi-part combination into automatic safety helmet wearing detection," *IEEE Access*, vol. 9, pp. 139536–139543, 2020.

[18] F. Wu, G. Jin, M. Gao, HE. Z. and Y. Yang, "Helmet detection based on improved YOLO V3 deep model," in *2019 IEEE 16th Int. Conf. on Networking, Sensing and Control (ICNSC)*, Banff, AB, Canada, pp. 363–368, 2019.

[19] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2004. [Online]. Available: https://arxiv.org/abs/2004.10934.

[20] G. Jocher, "Yolov5," 2020. [Online]. Available: https://github.com/ultralyc-s/yolov5.

[21] M. Jaderberg, K. Simonyan and A. Zisserman, "Spatial transformer networks," in *Proc. of the Advances in Neural Information Processing Systems*, Montreal, Canada, pp. 2017–2025, 2015.

[22] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[23] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li *et al.,* "Residual attention network for image classification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6450–6458, 2017.

[24] S. Woo, J. Park, J.-Y. Lee and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 3–19, 2018.

[25] P. Ramachandran, B. Zoph and Q. V. Le, "Swish: A self-gated activation function," 2017. [Online]. Available: https://arxiv.org/abs/1710.05941v1.

[26] A. Howard, M. Sandler, B. Chen, W. Wang, L. C. Chen *et al.,* "Searching for MobileNetV3," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (south), pp. 1314–1324, 2019.

[27] X. R. Zhang, W. F. Zhang, W. Sun, X. M. Sun and S. K. Jha, "A robust 3-D medical watermarking based on wavelet transform for data protection," *Computer Systems Science & Engineering*, vol. 41, no. 3, pp. 1043–1056, 2022.

[28] M. Everingham and J. Winn, "The PASCAL Visual Object Classes challenge 2012 (VOC2012) development kit," 2012. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/devkit_doc.pdf.