

A safety helmet-wearing detection method based on cross-layer connection

Gang Dong

Xihua University

Weicheng Xie

scxweicheng@mail.xhu.edu.cn

Xihua University

Yefei Zhang

Xihua University

Research Article

Keywords: Safety helmet detection, YOLOv5, Cross-layer connection, Coordinate attention, NWD

Posted Date: June 16th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3046357/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Real-Time Image Processing on April 9th, 2024. See the published version at <https://doi.org/10.1007/s11554-024-01437-5>.

A safety helmet-wearing detection method based on cross-layer connection

Gang Dong¹ · Weicheng Xie^{1*} · Yefei Zhang¹

Received: date / Accepted: date

Abstract Given the current safety helmet detection methods, the feature information of the small-scale safety helmet will be lost after the network model is convolved many times, resulting in the problem of missing detection of the safety helmet. To this end, an improved target detection algorithm of YOLOv5 is used to detect the wearing of safety helmets. Firstly, a new small-scale detection layer is added to the head of the network for multi-scale feature fusion, thereby increasing the receptive field area of the feature map to improve the model's recognition of small targets. Secondly, a cross-layer connection is designed between the feature extraction network and the feature fusion network to enhance the fine-grained features of the target in the shallow layer of the network. Thirdly, a coordinate attention (CA) module is added to the cross-layer connection to capture the global information of the image and improve the localization ability of the target. Finally, The Normalized Wasserstein Distance(NWD) is used to measure the similarity between bounding boxes, replacing the Intersection over Union(IoU) method. The experimental results show that the improved model achieves 95.09% of the mAP value for safety helmet-wearing detection, which has a good effect on the recognition of small-sized safety helmets of different degrees in the construction work scene.

Keywords Safety helmet detection · YOLOv5 · Cross-layer connection · Coordinate attention · NWD

1 Introduction

In the early stage, the wearing of safety helmets was usually detected by manual inspection. Due to each construction link being scattered in different spaces, the problem of high mobility of personnel led to low efficiency of manual supervision. Therefore, the study of an end-to-end, high-performance, and high-accuracy safety helmet-wearing detection method plays a key role in maintaining the safety of the construction process, thereby reducing the occurrence of potential safety hazards, which is of great significance to the personal safety of workers.

Safety helmet-wearing detection belongs to the field of object detection. Before the development of deep learning, the machine vision-based safety helmet-wearing detection method needs to obtain the feature information of the worker's head first, mostly using traditional computational methods, and then judge whether the worker is wearing a safety helmet through the output features. For example, Liu et al [1] used a method of detecting skin color to locate the face, and determined the specific position of the safety helmet by comparing the color difference, so as to determine the overlapping relationship between the head and the safety helmet. Shrestha et al [2] adopted the Haar-like method to extract the face features, and then found the contour information of the safety helmet through the edge detection method; Rubaiyat [3] first used the histogram of gradient orientation (HOG) to detect the human body area, and obtained the worker's head after fusing the image frequency domain information, then used the method of circular Hough transform and color feature to discriminate the safety helmet. However, in the practical detection application, all of the above are used to identify the wearing condition of the safety helmet by

✉Weicheng Xie
E-mail: scxweicheng@mail.xhu.edu.cn

¹ School of Electrical Engineering and Electronic Information, Xihua University, Chengdu 610039, China

the traditional manual feature design method, and it is very difficult to obtain a high accuracy ratio.

Currently, many scholars have conducted in-depth research on safety helmet detection through deep learning methods. Fang Ming et al [4] used the structure of a densely connected network to add dense blocks to the network to fuse multi-layer features, and then used a lightweight network to compress the model. However, there would be missed detection in the scenario where the object and complex background feature information was similar. Chu Yuezhong et al [5] used SSD [6] algorithm to fuse the self-attention layer to calculate the interaction between pixels in feature maps, but their ability to capture object location information was poor. Zhang Bo et al [7] located the position of the human head and neck first, and then used the Faster R-CNN [8] to determine the spatial relationship between the head and the safety helmet. However, this method required judging the position relationship of two objects, resulting in increased errors in calculation. Han Kun et al [9] made a multi-scale prediction of YOLOv3 [10] model, and performed the object frame dimension clustering to improve the detection performance of the safety helmet, and finally used GIoU [11] as the loss function to improve the regression accuracy, but the increase of the prior frame hindered the screening of the object frame. Han Guang et al [12] used a cross-layer attention network in SSD to further refine the feature information of the object area, and introduced a feature pyramid module to improve the robustness of detecting object scale change. However, expansion convolution made information extraction discontinuous, resulting in the loss of small object information. Although the above methods have improved the accuracy of the model, they still have a high missed detection rate and false detection rate for small objects in complex scenes, and it is especially difficult to extract the feature information of extremely small objects.

In order to solve the above problems, the construction workers wearing safety helmets and those without safety helmets are divided into two types of objects to be detected. The YOLOv5s network with the smallest width and depth is used as the basic model for improvement. Firstly, the detection layer is added to improve the learning of small object features. Then using the added splicing layer fuses the shallow information and deep information of the network to prevent the loss of feature information of the object. At the same time, the CA module is introduced into the path of the connection of the shallow and deep network to locate the position of small objects effectively, which further improves the detection precision of the model and the accuracy of small-sized object detection in long-distance

images. The experimental results show that the improved YOLOv5 method can identify the small safety helmet object effectively in different states.

2 Improved YOLOv5 algorithm structure

2.1 Adding micro-scale detection layers

YOLOv5 performs detection at three scales and generates feature maps with scale sizes of 20×20 , 40×40 , and 80×80 in the detection layer respectively. These scales are accurately obtained by $32 \times$, $16 \times$, and $8 \times$ down-sampling of the input size of 640×640 images, respectively. During the experiment, detecting different sizes of safety helmets in different scale images has a better detection effect on a dataset with a larger object. However, the position of construction workers in the images taken at a distance is unevenly distributed, with different levels of distance and height, which are scattered in every corner of the camera shot, so there are also extremely small objects on the safety helmets mapped to the image. Especially in one image, there are multiple extremely small-scale safety helmet objects. The small-scale detection layer of YOLOv5 has poor generalization ability for the detection of these safety helmets. The feature information of small objects will be lost after multiple convolutions.

Therefore, to solve the above problems, a new small-scale detection layer is added to the network detection layer. The detection layer generates the feature map by integrating the shallow features obtained from the input image sampled twice under the network and the 160×160 feature layer sampled from the network neck. The new micro-scale detection layer enables the object detection network structure to have a better ability to learn the extremely small-scale object and makes full use of the feature information of the shallow network part. As shown in Fig. 1, which is suitable for detecting small safety helmet objects with scattered positions in remote images.

2.2 Design of cross-layer connections

In the process of continuous transmission of feature information from the line layer to the deeper layers in the network, the scale of the feature map will gradually shrink as it is transferred to the deeper layer, and the feature and gradient information will often become unclear, leading to the disappearance of features, resulting in negative effects such as misjudgment and object loss in the object prediction in the subsequent stage of the network.

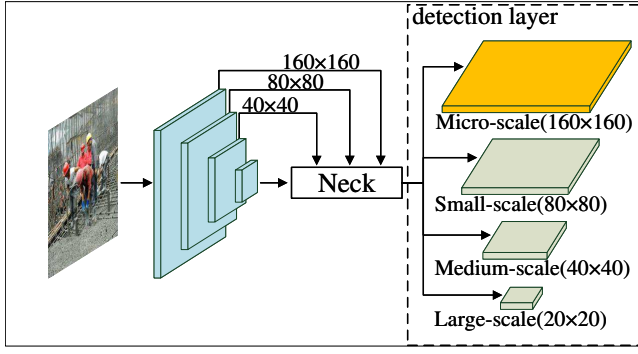


Fig. 1: Adding small-scale detection layers

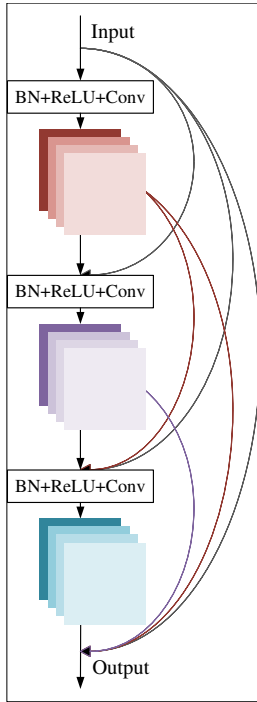


Fig. 2: DenseBlock module schematic

The DenseNet [16] network is the first to introduce the densely connected structure. The network structure is composed of a DenseBlock module and transmits feature information. In order to reduce the phenomenon of vanishing gradient, the output of the DenseBlock module will be passed to all subsequent modules. Fig. 2 shows the schematic diagram of the DenseBlock module.

Influenced by the transmission of feature information of the DenseBlock module, a cross-layer connection structure is introduced in the YOLOv5 algorithm. The image features of the feature extraction network are directly sent to the feature fusion network, and the feature maps are spliced and fused.

For the concatenation layer, the same height and width of the input must be required, and a consistent dimensionality is not required. After studying the structure of the YOLOv5 model, it is found that in the PAN structure, two feature maps from different layers are fused at the splicing layer and the fused feature information is output to the next layer structure. The two features information from different layers are $K_1 \times H_1 \times W_1$ and $K_2 \times H_1 \times W_1$ respectively. The output feature information of the splicing layer is as follows:

$$Z_{concat} = (K_1 + K_2) \times H_1 \times W_1 \quad (1)$$

K_1 and K_2 are the number of channels respectively input in different layers, while H_1 and W_1 are the height and width of the input feature map required by the splicing layer.

It can be found from equation (1) that since the feature information of different layers is received in the splicing layer for fusion, more comprehensive features can be obtained in the subsequent network layers. The cross-layer connection sends the new feature information $K_3 \times H_1 \times W_1$ to the corresponding splicing layer, and the feature information received by the splicing layer becomes:

$$Z_{concat} = (K_1 + K_2 + K_3) \times H_1 \times W_1 \quad (2)$$

Experiments show that an 80×80 scale splicing layer is added to the PAN structure after adding a new detection layer. The height and width of the output feature maps of the fifth and seventh layers in the feature extraction network of YOLOv5 correspond to the input scales of the two splicing layers in the PAN structure respectively. Cross-layer connections are introduced in these two layers, which fuse the feature information extracted from the shallow layer with the deep feature information. So as to enrich the feature information of small-scale and medium-scale objects.

2.3 Introduction of the Coordinate Attention Mechanism module

Although it is possible to introduce richer shallow feature information directly into the feature fusion network through cross-layer connections. Not all of this incoming feature information is important. Especially it is difficult to obtain the dependency of objects in space. Therefore, it is necessary to perform a weighting operation on the transmitted feature information to improve the accuracy of the YOLOv5 algorithm in obtaining object location information. So, the coordinate attention mechanism module is introduced into the network.

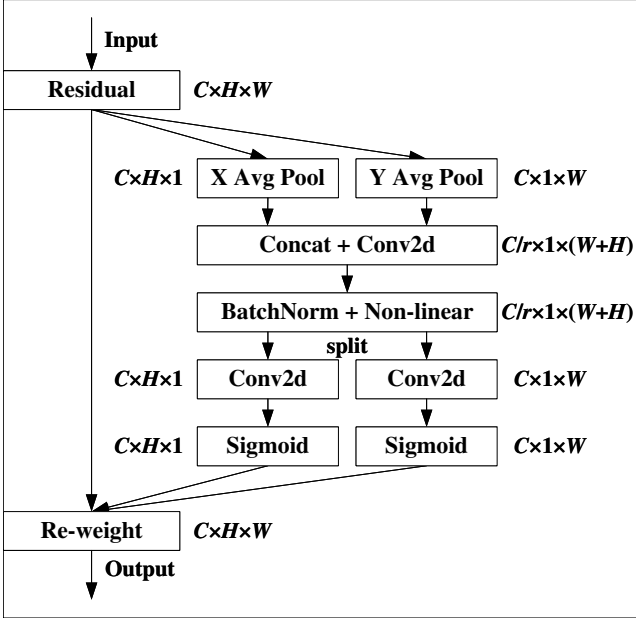


Fig. 3: CA module structure

Coordinate Attention (CA) [20] can obtain the object location information in channel attention to weight the object feature. In previous attention mechanisms such as SENet [21] and CBAM [22] can cause loss of location information, due to the use of two-dimensional global averaging pooling mainly. But the Coordinate Attention mechanism splits the channel attention into two parallel one-dimensional feature information encoding. Aggregating the feature in both horizontal and vertical directions to obtain the accurate location information of the object and establish long-range dependencies. The direction-aware and position-sensitive information can be obtained by encoding the feature map. The author of the CA module has used it in EfficientNet and MobileNet [23] with good results.

The structure of the CA module is shown in Fig. 3, which includes coordinate information embedding and coordinate attention generation respectively. The CA module, which is used to enhance the ability of feature representation, outputs a tensor $Y = [y_1, y_2, \dots, y_c]$ with the same dimensional length of the enhanced representation by taking an arbitrary intermediate tensor $X = [x_1, x_2, \dots, x_c] \in R^{C \times H \times W}$ as input, where C is the number of channels. The height and width of the input image are H and W respectively.

The first is the coordinate information embedding operation, to improve the ability of the attention module to capture the spatial long-distance dependency of the precise position information of the safety helmet. The network can locate the safety helmet object to be identified more accurately. The CA module performs a

pair of one-dimensional feature encoding operations instead of global pooling. The feature map with input size $C \times H \times W$ is set as X . The coding operation is taken along different directions of horizontal and vertical coordinates with pooling kernels of size $(H, 1)$ and $(1, W)$ respectively. The original method uses average pooling to encode each channel, but the cross-layer connection sends the feature map from the shallow network. In order to reduce the influence of the background and obtain more texture features of the object information, the coding of both directions is performed with maximum pooling. The output of the c th channel with height h and the c th channel with width w is obtained. The formula is shown in (3)(4).

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \leq i \leq W} X_c(h, i) \quad (3)$$

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \leq j \leq H} X_c(j, w) \quad (4)$$

The above operation aggregates feature along two different directions and compute a pair of direction-aware attention features Z^h and Z^w .

In coordinate attention generation, the module is divided into two feature layers before cascading, and then a shared 1×1 convolution is used for F1 transformation. The formula (5) is shown below.

$$f = \delta(F_1([Z^h, Z^w])) \quad (5)$$

In the formula, $f \in R^{C/r \times (H+W)}$ is the intermediate feature map in horizontal and vertical directions, and in such spatial information, r denotes the ratio of compression, which is generally set to 32 to control the number of channels in order to balance the global computational effort and performance. In the middle bracket is a computational operation connecting the spatial dimensions, and δ is a nonlinear activation function. Next, f is split into two independent tensors $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$ in the spatial dimension, and the two 1×1 convolutions are used to perform the upscaling operation to control the number of channels consistently, transforming the feature maps f^h and f^w to be the same as the input X , while normalizing the obtained parameters, and the results are shown in Equation (6)(7) below.

$$g^h = \delta(F_h(f^h)) \quad (6)$$

$$g^w = \delta(F_w(f^w)) \quad (7)$$

Finally, with g^h and g^w as the attention weight parameters, the weighted output features are shown in Equation (8).

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (8)$$

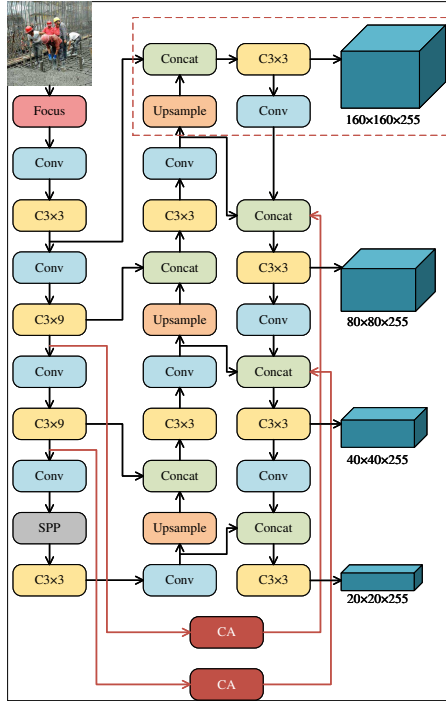


Fig. 4: Schematic diagram of the improved YOLOv5 algorithm

The coordinate attention mechanism module is used to perform the attention reinforcement operation on the feature maps sent through the cross-layer connection, which further strengthens the network for meaningful feature maps in the process of sending the shallow layer to the deep layer through accurate location information encoding channel relations and long-distance dependencies, especially for small object features, which can make the network better learn the feature information of the object in the same test image. The recognition features of the object can be captured more accurately and better recognition results can be achieved without increasing the training cost. The resulting embedding of the coordinate attention model improves the attention of small-scale objects and medium-scale objects and improves the feature information representation of objects under the influence of the background environment and other factors. The improved algorithm and its network structure are shown in Fig. 4.

3 Normalized Wasserstein Distance

When the predicted bounding box of large and small objects have the same pixel displacement on their ground truth, there is a significant difference in the sensitivity of the change in their IoU. The sensitivity of IoU to large objects is relatively small, while a small posi-

tional change in the predicted bounding box generated by a small object can cause a significant change in IoU, thereby affecting loss calculation and anchor matching. Therefore, it is necessary to improve the calculation of the prediction box loss.

Wang J et al[24] studied a method for calculating the similarity between the predicted bounding box and ground truth using Wasserstein distance. First, model the bounding boxes as 2-D Gaussian distributions, and then use NWD to measure the similarity of derived Gaussian distributions. For small targets with a head wearing a safety helmet, as neither the head nor the helmet is rectangular, there are often some background pixels in the predicted bounding box. The pixels in the predicted bounding box can be divided into two types: object pixels and background pixels. Object pixels are mainly distributed in the central region of the prediction box, while background pixels are mainly distributed in the edge region of the prediction box. Each pixel in the prediction box has different importance. To express this, a 2-D Gaussian distribution is used to represent the prediction box, and the weight of each pixel decreases successively from the center position to the boundary. To compare the similarity between two boxes, their shape and position are described using the calculated 2-D Gaussian distribution, and then NWD is used to calculate their differences. For the 2-D Gaussian distribution parameters $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$ and $\mu_2 = \mathcal{N}(m_2, \Sigma_2)$, the second-order Wasserstein distance between them is defined as follows:

$$W_2^2(\mu_1, \mu_1) = \|m_1 - m_2\|_2^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2} \Sigma_1 \Sigma_2^{1/2})^{1/2}) \quad (9)$$

The Gaussian distributions $\mathcal{N}_a$ and $\mathcal{N}_b$ used to model the prediction boxes $A = (cx_a, cy_a, w_a, h_a)$ and $B = (cx_b, cy_b, w_b, h_b)$ can be simplified as:

$$W_2^2(\mathcal{N}_a, \mathcal{N}_b) = \left\| \left(\begin{bmatrix} cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \end{bmatrix}^T, \begin{bmatrix} cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \end{bmatrix}^T \right) \right\|_2^2 \quad (10)$$

Since the $W_2^2(\mathcal{N}_a, \mathcal{N}_b)$ calculated distance is a length, it cannot be directly used as a measure of similarity. Instead, the exponential form of the distance can be used for normalization to obtain a new measure of similarity. This measure of similarity can then be combined with a constant C that is closely related to the dataset to obtain the Normalized Wasserstein Distance shown in equation (11).

$$NWD(\mathcal{N}_a, \mathcal{N}_b) = \exp \left(-\frac{\sqrt{W_2^2(\mathcal{N}_a, \mathcal{N}_b)}}{C} \right) \quad (11)$$

The Normalized Wasserstein Distance has scale invariance and is insensitive to position transformations. This makes the loss function threshold changes smooth during the generation of small-scale prediction boxes, allowing the network to converge better. Therefore, NWD is used to calculate the prediction box loss in the YOLOv5 network model, and the resulting loss function is shown in equation (12).

$$\mathcal{L}_{NWD} = 1 - NWD(\mathcal{N}_p, \mathcal{N}_g) \quad (12)$$

In the equation, $\mathcal{N}_p$ is the Gaussian distribution model of the prediction box, and $\mathcal{N}_g$ is the Gaussian distribution model of the ground-truth box. Even if the prediction box and the ground-truth box do not overlap at all, or if they overlap very little, their similarity can still be measured using NWD. The Normalized Wasserstein Distance is insensitive to the scale of the object, making it more stable for calculating distance thresholds between small-scale prediction boxes and ground-truth boxes.

4 Analysis of experimental results

4.1 Construction of the dataset

The dataset uses the public safety helmet-wearing detection dataset from the Kaggle machine learning competition platform, which has an image size of 416×416 and a total of 5000 images. The dataset is divided into two categories: wearing a safety helmet and not wearing a safety helmet, as shown in Fig. 5, and the images are resized to a uniform size by mirror stitching, which also expands the number of positive and negative samples in the images. The safety helmet dataset can reflect real construction scenarios, including many different construction scenarios, and there are construction workers wearing safety helmets in different postures in various construction environments. At the same time, since the negative sample data of non-wearing safety helmets are small, the network learning is improved and the generalization ability of the network model is enhanced by introducing negative sample data of different resolution sizes for expansion.

4.2 Experimental setup

The experimental OS is 64-bit Windows 11, CPU is Intel(R) Core(TM) i9-10900X 3.70GHz, GPU is GeForce RTX 3090 24GB, build Pytorch deep learning framework for model building, network training, and result testing, use Python to write Program code and call CUDA, Cudnn, OpenCV and other required libraries.



(a) Wearing safety helmet



(b) Failure to wear

Fig. 5: Safety helmet wearing category

Table 1: Experimental parameter settings

Parameters	Numerical values
Input image size	416×416
Initial learning rate	0.01
Number of images passed in at once	32
Number of training rounds	200

The safety helmet dataset is divided into training, validation, and test sets and divided according to the ratio 8:1:1.

To optimize the learning rate, stochastic gradient descent (SGD) is used as the optimization function during the network training. To prevent losses during training from generating oscillations or gradient explosions, the initial learning rate is reset to 0.01 after inputting images. The batch size is set to a suitable 32 without exceeding the GPU video memory size. Combined with the above parameter settings, the number of training rounds is set to 200 considering the final convergence of the model. Finally, the trained network model weight file is saved and the performance of the network model is evaluated using a test set. The specific settings of the hyperparameters for the network training are shown in Table 1.

4.3 Network Performance Evaluation

Experiments were conducted to evaluate the network performance using detection accuracy as an evaluation metric as a way to verify the feasibility of the improved network model. The two states of wearing a safety helmet and not wearing a safety helmet can generate four prediction methods: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP and FP are the numbers of safety helmet wearing detected correctly and incorrectly, respectively, and TN and FN are the numbers of not wearing safety helmets detected correctly and incorrectly, respectively. This prediction is used to determine the evaluation-related metrics for the performance of the two models, which are the precision (P) and recall (R), respectively, and the two metrics are defined specifically in equations (13) (14).

$$P = \frac{TP}{FP + TP} \quad (13)$$

$$R = \frac{TP}{FN + TP} \quad (14)$$

Accuracy and recall affect each other, but these two metrics cannot be used directly to evaluate the detection accuracy of the network model. For this reason, average precision (AP) is introduced to represent the detection accuracy of a single category, and in order to obtain a higher network model accuracy, a higher average precision (AP) is needed to verify it, as in equation (15), and then the mean average precision (mAP) is obtained by calculating the mean value of the average precision (AP) for all categories, as in equation (16).

$$AP = \int_0^1 P(r)dr \quad (15)$$

$$mAP = \frac{1}{c} \sum_{j=1}^c AP_j \quad (16)$$

4.4 Experimental comparison analysis

4.4.1 Analysis of experimental results

The specific effect after detection is shown in Fig. 6, with the detection data set on the left, the original YOLOv5s detection graph in the middle, and the improved YOLOv5s detection effect graph on the right.

Figure 6(a) shows the dense safety helmet small object image, Figure 6(b) shows the small object safety helmet image under a complex pose, Figure 6(c) shows the small object at a distance, and Figure 6(d) shows the small object with occlusion. It can be seen that

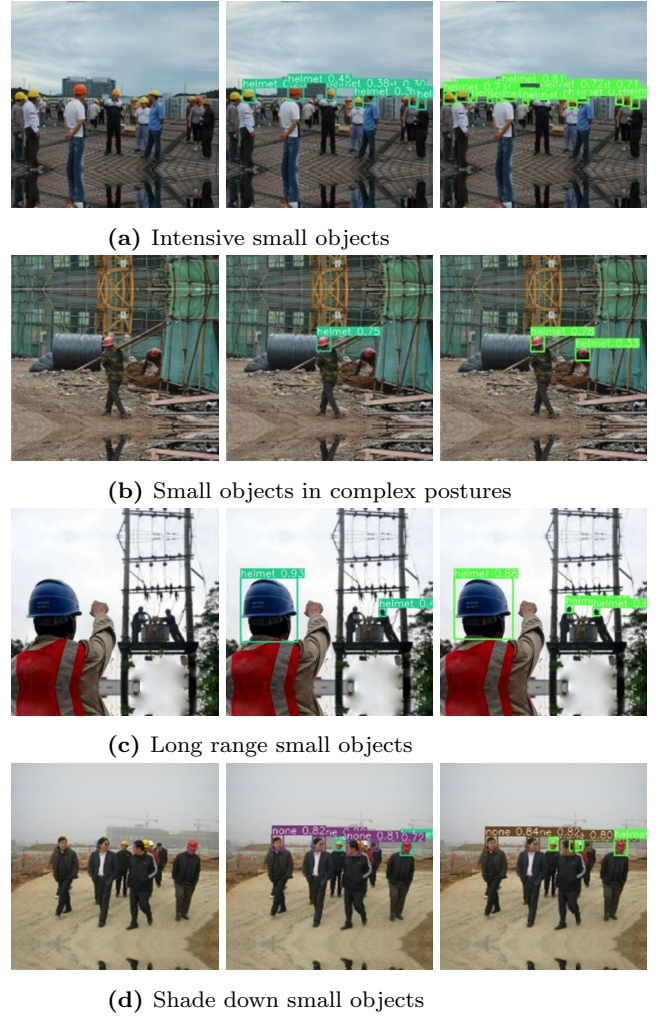


Fig. 6: Comparison of testing results

the improved model in Fig. 6(a) detects multiple objects wearing safety helmets at a distance missed by the original model while improving the category confidence score for each object; the improved model in Fig. 6(b) detects the object wearing a safety helmet in a bent posture on the right side missed by the original model; and Fig. 6(c) contains the large-scale object and the small object safety helmet images at a distance. By observation, it can be seen that the original YOLOv5s model did not detect the small distant object on the left side; finally, it can be seen in Figure 6(d) that the improved model detects the small object wearing a safety helmet in the occlusion case. For small object recognition in different states, the original model is even more serious in missing detection, while the improved model can still accurately identify the unrecognized small objects and most of the confidence scores have been improved. It shows that the improved model is able to detect small

Table 2: Experimental parameter settings

Models	P/%	R/%	mAP@0.5
YOLOv5s	90.56	88.85	92.27
Adding detection layer	89.33	89.23	93.78
Adding detection layer + Cross-layer connection	91.02	90.76	94.71
Adding detection layer + Cross-layer connection + CA	91.35	93.02	95.09
Adding detection layer + Cross-layer connection + CA + NWD	92.29	92.24	95.92

Table 3: Comparison with different safety helmet detection algorithms

Algorithms	AP/(IOU=0.5)		mAP/%	FPS/(frame/s)	Model size/MB
	Wearing safety helmet	Failure to wear safety helmet			
Han Kun et al (2021)	94.79	90.91	92.85	52.6	236
Han Guang et al (2021)	91.62	88.60	90.11	41.5	98.4
YOLOv5s	96.22	88.32	92.27	69.3	14.1
Ours	98.20	93.64	95.92	59.2	15.3

objects in many different scenarios and has a strong generalization ability.

4.4.2 Comparative analysis of ablation experiments

The ablation experiments were conducted to compare and analyze the improved methods. The details are shown in Table 2. The YOLOv5s model without any improvement was used as the standard, and it was trained by three different improvement methods in turn, and the final obtained metrics were evaluated with the average precision mean as the final evaluation metric. After comparison, it was found that the accuracy, recall, and average accuracy of the original YOLOv5s were 90.56%, 88.85%, and 92.27%, respectively. Using this metric as the standard, the mAP values derived from each improved and adjusted method increased, and the addition of detection layers slightly reduced the accuracy rate due to the stitching of the feature map and the deep network after 4-fold down-sampling of the input image, which was too small to recognize large objects but kept the fine-grained features of small objects. By improving the network architecture, the accuracy, recall, and mAP metrics have been improved by 0.79, 4.17, and 2.82 respectively compared to the original YOLOv5s. On this basis, after using NWD instead of IoU, mAP further increased by 0.83, reaching 95.92%. This ablation experiment further demonstrates the feasibility of the improved method.

4.4.3 Comparative analysis of object detection algorithms

Finally, the improved method was validated, and the comparison experiments were conducted with other cur-

rent algorithms with better performance of safety helmet detection algorithms, which are Han Kun et al [9] and Han Guang et al [12]. The experimental results of each algorithm are shown in Table 3. As can be seen from Table 3, the improved algorithm outperforms the other comparison algorithms in terms of overall mAP value, the accuracy of both object categories is improved, and the overall mAP value is improved, and it can be seen that the average accuracy improvement of detecting the object of not wearing a safety helmet is larger, and it is improved by 5.32 percentage points on the base network, which is because the feature information of the head of not wearing a safety helmet in a complex environment is not obvious. The improved algorithm can capture more head feature information, which again verifies the feasibility of the improved model; compared to the improved algorithm in Han Kun et al [9], the mAP value is improved by 3.07%, although the method in this literature is faster than the two-stage network in terms of detection speed, the complexity of the network structure causes its model file to be too large, which is time-consuming and laborious in the deployment of the model; Han Guang et al [12], although the average accuracy of the mean in comparison with the basic YOLOv5s network, although the speed decreases slightly, transmitting about two images per second less than the original YOLOv5s, the general processing speed reaches 30 frames, which is the industrial standard. The more images per second processed on this benchmark, the more efficient the model is in detecting the object, which can meet the demand of real-time engineering.

5 Conclusion

A YOLOv5-based safety helmet-wearing detection algorithm is improved to solve the problem of identifying missed detections on small objects by current safety helmet-wearing detection methods. By adding a new detection layer to the prediction head of the network to improve the feature information of small objects under multiple scales of detection, the problem of small objects being unpredictable due to too large sensory fields is solved. By adding two new paths that lead to new branches after the two C3 layers in the backbone network and then sending them to the splicing layer of PAN structure in the neck network for fusion respectively, and adding CA modules to the paths, the problem of localization loss of small and medium-sized objects is solved. Finally, using NWD solves the problem of IoU being sensitive to small objects. The improved YOLOv5 network model reduces the missed detection of small-sized safety helmet objects and improves the classification confidence score.

Data availability

The datasets generated and analysed during the current study are available in the <https://www.kaggle.com/datasets/andrewmvd/hard-hat-detection>

Funding

This work was supported in part by the Chunhui Project of Ministry of Education in China under Grant z2018087, and in part by the Science and Technology Achievements Transfer and Transformation Demonstration project of Sichuan province in China under Grant 2020ZHCG0099.

Declarations

Conflict of interest Authors in this paper declare that they have no conflict of interest.

References

1. Liu X, Ye X.: Application of skin colour detection and Hu moments in helmet recognition. *Journal of East China University of Science and Technology: Natural Science Edition* 40(3),365-370(2014)
2. Shrestha K, Shrestha P P, Bajracharya D, Yfantis E A.: Hard-hat detection for construction safety visualization. *Journal of Construction Engineering* 2015(1),1-8(2015).<https://doi.org/10.1155/2015/721380>
3. Rubaiyat A H, Toma T T, Kalantari-Khandani M, Rahman S A, Chen L, Ye Y, Pan C S.: Automatic detection of helmet uses for construction safety. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops, pp 135-142(2016).<https://doi.org/10.1109/wiw.2016.045>
4. Fang M, Sun T, Shao Z.: Fast helmet wear detection based on improved YOLOv2. *Optical precision engineering* 27(05), 1196-1205(2019)
5. Chu Y, Huang Y, Zhang X, Liu H.: SSD image target detection algorithm based on self-attention. *Journal of Huazhong University of Science and Technology (Natural Science Edition)* 48(09),70-75(2020)
6. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg A C.: Ssd: Single shot multibox detector. In: *European Conference on Computer Vision*, pp 21-37(2016)
7. Zhang B, Song Y, Xiong R, Zhang S.: Safety helmet wear detection incorporating human articulation points. *Chinese Journal of Safety Science* 30(02),177-182(2020)
8. Ren S, He K, Girshick R, Sun J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39(6),1137-1149(2017).<https://doi.org/10.1109/tpami.2016.2577031>
9. Han K, Li S, Xiao Y.: YOLOv3-based helmet wearing status detection in construction scenarios. *Journal of Railway Science and Engineering* 18(01),268-276(2021)
10. Redmon J, Farhadi A.: Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*(2018)
11. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 658-666(2019).<https://doi.org/10.1109/cvpr.2019.00075>
12. Han G, Zhu M, Zhao X, Gao H.: Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection. *Computers & Electrical Engineering* 95,107458(2021)
13. Redmon J, Divvala S, Girshick R, Farhadi A.: You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 779-788(2016).<https://doi.org/10.1109/cvpr.2016.91>
14. Xue J, Zheng Y, Dong-Ye C, Wang P, Yasir M.: Improved YOLOv5 network method for remote sensing image-based ground objects recognition. *Soft Computing* 26(20),10879-10889(2022). <https://doi.org/10.1007/s00500-022-07106-8>
15. He K, Zhang X, Ren S, Sun J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 770-778(2016). <https://doi.org/10.1109/cvpr.2016.90>
16. Huang G, Liu Z, Van Der Maaten L, Weinberger K Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 4700-4708(2017)
17. Tan M, Le Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*, pp 6105-6114(2019)
18. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S.: Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 2117-2125(2017)
19. Liu S, Qi L, Qin H, Shi J, Jia J.: Path aggregation network for instance segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 8759-8768(2018).<https://doi.org/10.1109/cvpr.2018.00913>

20. Hou Q, Zhou D, Feng J.:Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13713-13722(2021).<https://doi.org/10.1109/cvpr46437.2021.01350>
21. Hu J, Shen L, Sun G.:Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132-7141(2018).<https://doi.org/10.1109/cvpr.2018.00745>
22. Woo S, Park J, Lee J-Y, Kweon I S.:Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision, pp 3-19(2018). https://doi.org/10.1007/978-3-030-01234-2_1
23. Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H.:Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv:170404861](https://arxiv.org/abs/1704.04861)(2017)
24. Wang J, Xu C, Yang W, Yu L.: A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. ArXiv, [abs/2110.13389](https://arxiv.org/abs/2110.13389)(2021).