



Lightweight safety helmet detection algorithm using improved YOLOv5

Hongge Ren¹ · Anni Fan¹ · Jian Zhao¹ · Hairui Song¹ · Xiuman Liang²

Received: 30 January 2024 / Accepted: 13 June 2024 / Published online: 5 July 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

In response to the challenges faced by existing safety helmet detection algorithms when applied to complex construction site scenarios, such as poor accuracy, large number of parameters, large amount of computation and large model size, this paper proposes a lightweight safety helmet detection algorithm based on YOLOv5, which achieves a balance between lightweight and accuracy. First, the algorithm integrates the Distribution Shifting Convolution (DSConv) layer and the Squeeze-and-Excitation (SE) attention mechanism, effectively replacing the original partial convolution and C3 modules, this integration significantly enhances the capabilities of feature extraction and representation learning. Second, multi-scale feature fusion is performed on the Ghost module using skip connections, replacing certain C3 module, to achieve lightweight and maintain accuracy. Finally, adjustments have been made to the Bottleneck Attention Mechanism (BAM) to suppress irrelevant information and enhance the extraction of features in rich regions. The experimental results show that improved model improves the mean average precision (mAP) by 1.0% compared to the original algorithm, reduces the number of parameters by 22.2%, decreases the computation by 20.9%, and the model size is reduced by 20.1%, which realizes the lightweight of the detection algorithm.

Keywords Safety helmet detection · Distribution Shifting Convolution · Bottleneck Attention Mechanism · Lightweight

1 Introduction

Construction sites represent intricate and perilous work environments, exposing workers to a multitude of high-risk activities [1]. Wearing a safety helmet can effectively reduce the occurrence rate of accidents and ensure safety. The safety of workers is of great significance to construction workers.

There are various target detection algorithms for safety helmets, and sensor-based detection methods require the use of sensors to collect data for detection, and sensors are placed inside the safety helmet. This technology is difficult and has a large number of parameters. The detection method based on traditional computer vision technology uses a sliding window to scan each pixel in the image, statistically analyze the features of the target to be detected, describe them, and use classification methods or establish models based on extracted features to determine whether the target is wearing a safety helmet. However, this type of method lacks detection accuracy and robustness, and the inference speed is very slow.

With the rapid development of high-performance computing, it has become possible for security monitoring to be automated, real time, and intelligent [2]. At present, target detection algorithms can be divided into two main directions. One type is the single-stage algorithm based on regression strategy, such as the YOLO series algorithms [3–6] and the SSD algorithm [7]. Their main characteristic is the direct localization and classification of objects in a single image, without the need for multiple stages of processing.

✉ Jian Zhao
zhaojiantcu@163.com

Hongge Ren
renhg_tcu@163.com

Anni Fan
fanannijiyau@163.com

Hairui Song
2669174618@qq.com

Xiuman Liang
13230857556@126.com

¹ School of Control and Mechanical Engineering, Tianjin Chengjian University, Tianjin 300384, China

² College of Electrical Engineering, North China University of Science and Technology, Tangshan 063210, Hebei, China

In contrast, another common type of target detection algorithm is the two-stage algorithm, mainly represented by the R-CNN series [8–10], which typically generates candidate regions first, and then classifies and precisely locates objects within these regions. Kerdvibulvech [11] proposed a motion analysis and hand tracking method based on an adaptive probability model, which integrates a deterministic clustering framework and particle filters to achieve efficient hand tracking. Singh et al. [12] proposed a method that combines Zernike moments (ZMs) and local binary patterns (LBP)/local ternary patterns (LTP) to address the issue of insufficient feature sets in facial recognition tasks. Mithun et al. [13] proposed a sterile and intuitive context integration system for improving continuous gesture recognition and the discovery and exploration of MRI through hand gesture navigation during nerve biopsy. A-masiri et al. [14] combined the Japanese anime industry with facial recognition and tested the ability to detect and recognize anime characters by comparing two images. The results showed that the program could recognize anime faces, but there were some limitations. Ge et al. [15] proposed a Convolutional Visual Self-Attention Network (CVSAN) to improve the performance of Masked Face Recognition (MFR), which utilizes self-attention mechanism to enhance the convolution operator. By connecting local features with self-attention feature maps modeled with long-range dependencies, the performance of the network is significantly improved compared to other algorithms.

In recent years, object detection algorithms based on deep learning technology have gradually been applied to the detection of helmet wearing. In recent years, object detection algorithms based on deep learning technology have gradually been applied to the detection of helmet wearing. Currently, many scholars at home and abroad have conducted relevant research on safety helmets. Silva [16] proposed a helmet-less motorcycle detection system that utilizes circular Hough transform and directional gradient histogram descriptors to extract image attributes, and enhances detection accuracy through a multi-layer perceptron classifier. Chen et al. [17] introduced a WHU-YOLO lightweight facial assistance model for welding caps, which modified YOLOv5s by incorporating a Ghost module and a Bidirectional Feature Pyramid Network (Bi FPN). The model remained lightweight while maintaining unchanged detection performance. Zhao et al. [18] proposed BDC-YOLOv5, which incorporates BiFPN, additional detection layers, and CBAM modules into YOLOv5 to reduce model false positives and false negatives, while enhancing the detection capability of small-scale objects. Xu et al. [19] introduced the MCX-YOLOv5 helmet detection algorithm, integrating a coordinate-space attention module in YOLOv5 to effectively filter spatiotemporal data in feature inputs. They also implemented a multi-scale asymmetric convolution down sampling module to improve

the algorithm's sensitivity to feature scale variations. Jin et al. [20] presented the YOLO-ESCA safety helmet detection algorithm, which utilizes efficient intersection loss functions, Soft-NMS non-maximum suppression, and Convolutional Block Attention Modules to enhance the speed and accuracy of safety helmet detection. Wang et al. [21] combined GIoU with the objective function of YOLOv3, achieving local optimization of the objective function but without improving speed. Zhao et al. [22] introduced MobileNetv2 into the YOLOv5s network, compressing the model and pruning redundant channels, improving recall rate and mAP, but increasing the model parameters and weights, which cannot meet the practical needs of production safety. Song et al. [23] introduced the CoordAtt coordinate attention mechanism module into the backbone network of the network, considering global information and improving the detection capability of small objects. They also replaced the residual blocks in the backbone network with residual blocks in the Res2NetBlock structure to enhance the fusion ability of YOLOv5s at a fine granularity, achieving more accurate, lightweight, efficient, and real-time detection of safety helmet wearing. Song et al. [24] combined the multi-object tracking algorithm DeepSort with YOLOv5 in the environment of small and dense targets, improving the detection speed and accuracy of safety helmets. Zhang et al. [25] introduced the DWCA attention mechanism into the YOLOv5s backbone network to enhance feature learning and improve the detection accuracy of safety helmets. Sun et al. [26] introduced the MCA attention mechanism into YOLOv5s to reduce the miss detection rate of small helmet objects and improve detection accuracy. Although the above detection methods have improved the detection accuracy of safety helmets to some extent, they have not changed the problems of complex detection algorithms, high number of parameters, slow inference speed, large computational complexity, and some lightweight models can effectively reduce model parameters, but cannot achieve a balance between accuracy and speed.

To address the existing problems in safety helmet detection algorithms, this paper aims to propose a lightweight detection algorithm based on improved YOLOv5s. The contributions of this paper are as follows:

- 1) Introduce and propose a new hybrid model, which combines DSConv (Distribution Shift Convolution) layers and SE (Squeeze-and-Excitation) attention mechanisms. This model replaces the original convolution modules and some C3 modules to improve feature extraction and representation learning capabilities.
- 2) Perform parallel multi branch on the Ghost module, and then perform serial skip layer connections. This model replaces some of the C3 modules. This model is proposed to reduce the computational cost and parameters

while enhancing the fusion of multi-scale features and enriching the semantic features.

- 3) Lastly, incorporate the attention mechanism BAM to enhance the weights of important information in feature maps and address the issues of missed detection.

The remaining sections of this paper are as follows: Sect. 2 describes the framework and implementation details of the proposed model. Section 3 introduces the experimental environment and datasets. Section 4 discusses the experimental results. Section 5 provides concluding remarks and future work.

2 Improving the YOLOv5 algorithm

YOLOv5 [27] has n, s, and m versions, etc. We chose YOLOv5s which has balanced detection speed and accuracy on the GDUT-HWD dataset.

The YOLOv5s network consists of three main components: Backbone, Neck, and Head. The Backbone is composed of CBS module, C3 module, and the Fast Spatial Pyramid Pooling (SPPF), which are primarily used for feature extraction. These functions are encapsulated in the CBS module: Convolution (Conv2d), Batch normalization and SiLU [28] is used as the activation function. The Neck

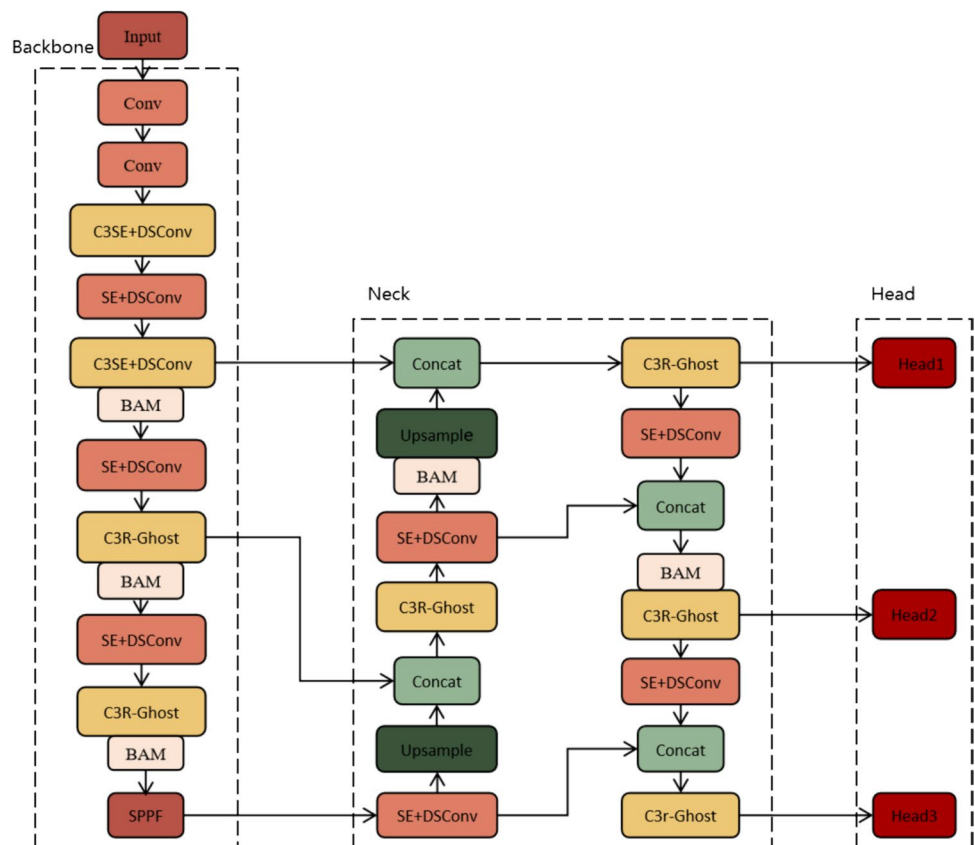
consists of the Feature Pyramid Network (FPN) [29] and the Perceptual Adversarial Network (PAN) [30], which are responsible for feature fusion. The Head performs the final predictions on the image.

The improved YOLOv5s network model structure is depicted in Fig. 1. The traditional convolution layers in the Backbone and Neck layers are replaced by the SE + DSConv module, which enhances the feature extraction and representation learning capabilities. In addition, the C3SE + DSConv module is used to substitute certain C3 modules in the main network. To reduce the number of model parameters and maintain accuracy while preserving the feature extraction, the R-Ghost module replaces some C3 modules in the main network and C3 modules in the Neck layer. Furthermore, five Bottleneck Attention Module (BAM) modules are incorporated into the network model. This addition strengthens the localization and recognition of regions of interest in the feature maps without significantly increasing the parameter count, thereby improving recognition accuracy.

2.1 SE + DSConv module

Distribution Shifting Convolution (DSConv) [31, 32] is a novel convolutional layer designed to enhance the memory efficiency and speed of standard convolutional layers. However, traditional convolution operations typically demand

Fig. 1 The improved YOLOv5s network structure



significant memory resources in terms of computation and storage. When deployed on resource-constrained devices, their extensive parameter counts and computational requirements of these models present significant challenges. To address these issues, DSConv introduces a novel convolutional operation that optimizes traditional convolution operations through quantization and distribution shifts. Specifically, DSConv decomposes the convolution kernel into two parts: Variable Quantized Kernel (VQK) and distribution shifts. VQK stores only integer values and quantizes them based on the distribution of the original weights, thus reducing storage space consumption. Meanwhile, distribution shifts are utilized to adjust the distribution of VQK to match that of the original weights, thereby preserving the model’s performance. This is achieved through distribution shifts in the kernel (KDS) and allocation shifts in the channel (CDS).

We propose a method for quantizing weights that share one floating-point value for each block of size B , along the depth dimension of each weight tensor filter. The memory saved per tensor weight is calculated as follows:

$$p = \frac{b}{32} + \frac{\left\lceil \frac{c_i}{B} \right\rceil}{C_i} \tag{1}$$

where C is the number of channels and b is the selected hyper parameter setting.

As shown in Fig. 2, to prevent potential information loss during the quantization process of converting floating-point weights to integer values, we integrate the SE (Squeeze-and-Excitation) attention mechanism [33], which recalibrates the feature responses of the convolution layers. By dynamically adjusting the importance of each channel, DSConv further enhances its capture of the importance of input features.

The SE attention mechanism consists of three key operations: Squeeze operation, Excitation operation, and Scale operation. From Fig. 2, it can be observed that the original image X undergoes a convolution operation F_{tr} to generate the feature map U . However, the entire F_{tr} operation is only performed within a local spatial region, which means that the feature map U cannot capture global feature information and thus struggles to establish relationships between channels [34]. Therefore, SE attention mechanism proposes a compression operation that utilizes global average pooling to compress the spatial dimensions of the feature map U , effectively reducing the spatial features of each channel into a single global feature. This can be expressed mathematically as shown in Eq. (2).

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i,j). \tag{2}$$

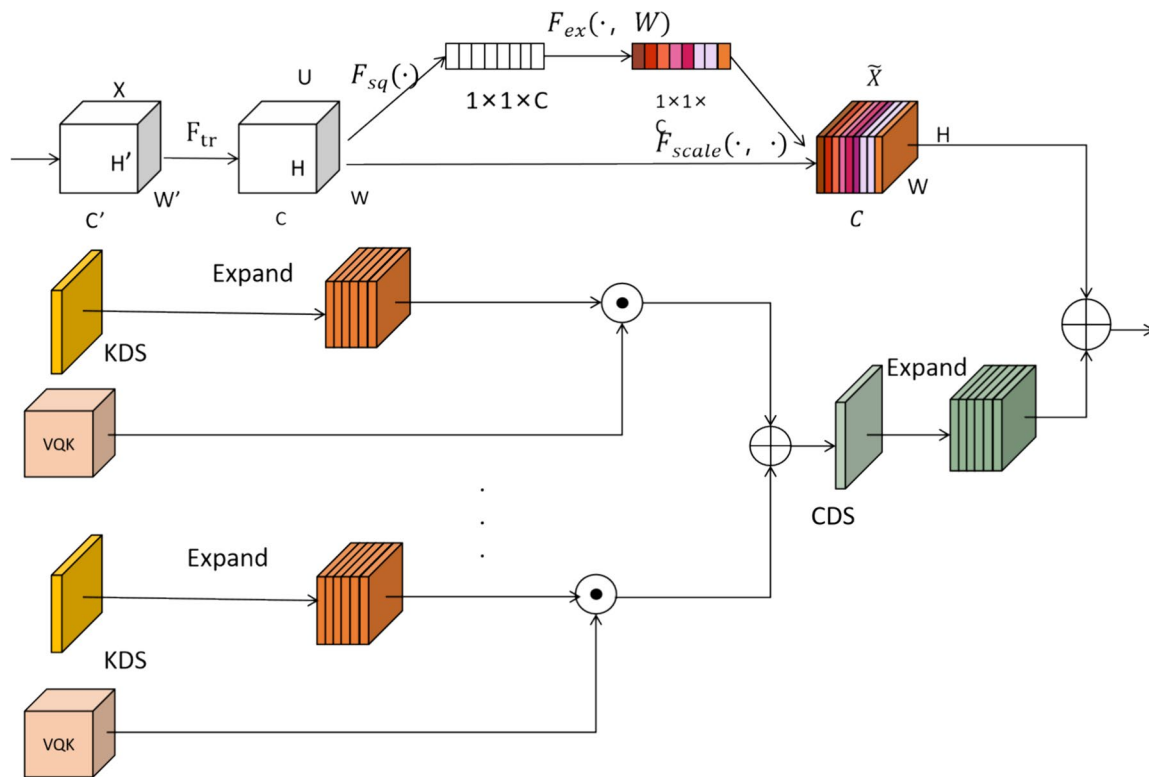


Fig. 2 SE+DSConv network structure

After performing the Squeeze operation, the global features Z of the obtained feature map are acquired. The Excitation operation establishes the inter-channel dependencies using the global features Z and computes the importance weights for each channel domain. The Excitation operation utilizes the Rectified Linear Unit (ReLU) function and the Sigmoid function for activation. The expression can be represented by Eq. (3).

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(w_2 \delta(w_1 z)). \tag{3}$$

In this context, z refers to the output result of the Squeeze operation, σ refers to the Sigmoid function, ReLU refers to the ReLU function, W_1 refers to the first fully connected layer, and W_2 refers to the second fully connected layer. The Scale operation is the process of weighting the weights of each channel domain onto the original feature channels, thus achieving the reevaluation of channel domains. This expression can be represented by Eq. (4).

$$\tilde{X}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{4}$$

where the term $F_{scale}(u_c, s_c)$ refers to the multiplication of feature map u_c with scalar weight s_c , performed channel wise. The DSConv can be expressed as Eqs. (5), (6):

$$\text{Output} = \xi \odot VQK \tag{5}$$

where ξ is the KDS value for that block.

$$\text{Output}' = \varphi * \text{Output} \tag{6}$$

where φ is the CDS value tensor. Complete the SE+DSConv formula first passes through ‘SELayer’, as shown in Eq. (7).

$$X' = \text{SELayer}(X). \tag{7}$$

Next, the convolution operation is performed using ‘Modified Weight’, as shown in Eq. (8).

$$O = \text{Conv}(X', \text{Output}'). \tag{8}$$

In Fig. 3, the first 1×1 SE+DSConv layer in the SE+DSConvBottleneck is utilized for dimension transformation and feature dimension reduction. By appropriately reducing the number of channels in the input feature map, it effectively maintains computational and parameter efficiency while aiding the model in retaining crucial feature information [35]. The second 3×3 SE+DSConv layer, following

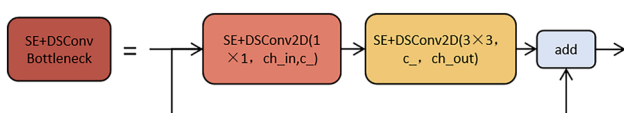


Fig. 3 SE+DSConvBottleneck module

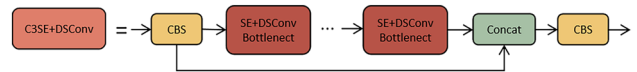


Fig. 4 C3SE+DSConv module

the dimension transformation, performs deeper feature extraction through the use of a 3×3 convolution kernel. This layer contributes to the model’s understanding of a broader spatial structure and patterns, capturing more complex features and improving the model’s perceptual capabilities. Combining the C3 module with the SE+DSConvBottleneck results in the C3SE+DSConv module (Fig. 4), enhancing the ability of feature learning and subsequently improving detection performance. This module enables efficient detection within resource-constrained environments.

2.2 R-Ghost module

In practical application environments, the performance of the YOLOv5s network model is susceptible to the constraints imposed by hardware memory and computational complexity. To meet the demands of mobile and embedded devices, certain convolutional (Conv) layers in the original network are replaced with Ghost modules [36].

The Ghost model initially employs a limited number of convolutional kernels to extract features from the input feature map. Subsequently, it executes more cost-effective linear transformation operations on this portion of the feature map, ultimately generating the final feature map through concatenation. This approach replaces conventional convolution methods by combining a small number of convolution kernels with more economical linear transformation operations, thereby lowering the learning cost of non-critical features and effectively reducing the demand for computing resources.

In the process of safety helmet detection, many helmets in the scene are of very small size, occupying a small proportion of the entire surveillance screen, or there may be significant size differences among helmet objects. To prevent the potential loss of crucial feature information during linear transformation operations, which could result in lower detection accuracy of safety helmets, improvements have been made to the network structure based on multi-scale feature fusion to enhance object detection capability.

First, the input features are divided into two parts after 1×1 convolution. The first part is passed through directly without any processing, while the second part undergoes Ghost convolution before being propagated forward. Finally, features from both parts are concatenated and sent to another 1×1 convolution for complete information fusion.

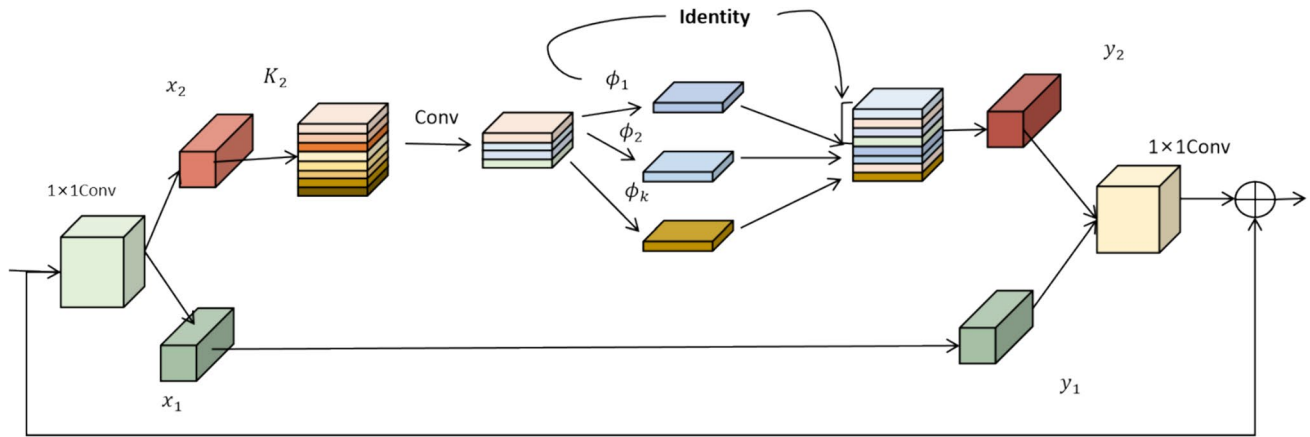


Fig. 5 R-Ghost network structure

Second, skip connections are introduced by incorporating short connections within the network to merge shallow features with deep ones. This approach not only helps the network better utilize feature information from different levels but also effectively handles scale variations, thus improving the model’s adaptability across different scales [37], thereby enhancing the performance of object detection.

As shown in Fig. 5, after a 1×1 convolution, the feature maps are evenly divided into two subsets, denoted as, where $i = 1, 2$. Each feature subset has the same spatial size as the input feature maps, but with half the number of channels.

The formula for the computation of floating-point operations (FLOPs) in a regular convolution is as Eq. (9):

$$F_1 = n \times h' \times w' \times c \times g \times g. \tag{9}$$

The formula for the floating-point operation (FLOPs) of the R-Ghost module is presented as Eq. (10):

$$F_2 = \frac{s}{2} \times h' \times w' \times c \times g \times g + \frac{n}{2s} \times h' \times w' \times c \times g \times g + (s - 1) \times h' \times w' \times \frac{n}{2s} \times d \times d. \tag{10}$$

In the given formula, n represents the number of output channels, h' represents the height of the output features, w' represents the width of the output features, c represents the number of input channels, g represents the size of the convolutional kernel, s represents the number of feature maps generated in the R-Ghost module, and d represents the size of the convolutional kernel used for linear operations.

$$\frac{F_1}{F_2} = \frac{n \times c \times g \times g}{\frac{s}{2} \times c \times g \times g + \frac{n}{2s} \times c \times g \times g + (s - 1) \times \frac{n}{2s} \times d \times d} = \frac{2s}{s + 1} \approx 2. \tag{11}$$

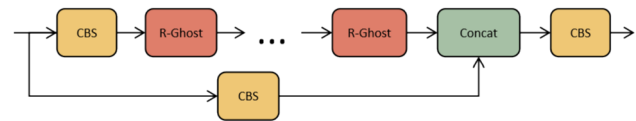


Fig. 6 C3R-Ghost module

Among them, $s \ll c$, as Eq. (11), the computational cost of regular convolution is approximately twice that of the R-Ghost module. The current study presents the integration of R-Ghost and the Bottleneck module from YOLOv5s, forming a novel architecture called R-Ghost-BoTteNeck. This architecture replaces the Bottleneck module in C3, resulting in C3R-Ghost. The integration achieves a reduction in parameter quantity while simultaneously enhancing the network’s ability to perceive different features, as shown in Fig. 6.

2.3 Bottleneck Attention Module

This mechanism can selectively emphasize information-rich features and suppress useless features by learning global information. In industrial production and transportation operations, where the surrounding environment is complex and variable, BAM (Bottleneck Attention Module) [38] is introduced to allow the network to focus more on safety helmets while ignoring background information.

Unlike traditional convolution operations, BAM simultaneously focuses on the channel dimension and the spatial dimension of feature maps to obtain richer feature information. The core idea of BAM is to combine channel attention and spatial attention to better understand image features. While enhancing performance, it incurs negligible overhead in terms of model computational complexity.

The module structure is shown in Fig. 7. The input feature F is subjected to global average pooling to obtain a feature

Fig. 7 Channel attention module

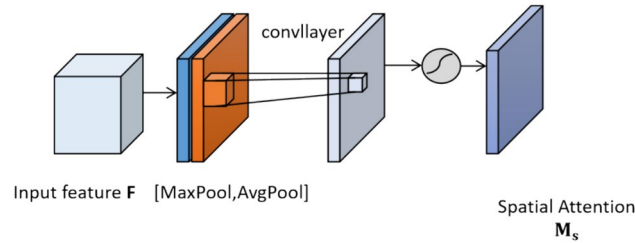
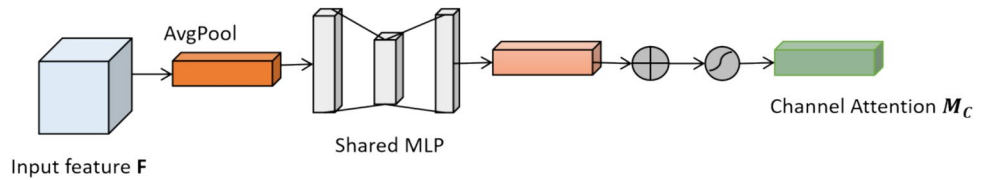


Fig. 8 Spatial attention module

descriptor, which is then fed into a multi-layer perceptron (MLP) consisting of three fully connected layers. The result is passed through the sigmoid activation function to obtain one-dimensional channel attention weights. The calculation formula is as Eq. (12):

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F))). \tag{12}$$

In the given statement, F represents the output features obtained after processing the image through the YOLOv5s model. AvgPool denotes the global average pooling function, while MLP refers to the multi-layer perceptron, represents the Sigmoid function.

The attention to spatial details focuses on various locations within the feature maps. The module structure, as depicted in Fig. 8, involves initially applying average pooling and global max pooling to the input feature F . The resulting two feature descriptions are then concatenated, followed by convolution and activation function processing [39]. The formula for calculating the two-dimensional spatial attention weights, is given as Eq. (13):

$$M_s(F) = \sigma(\text{Conv}(\text{AvgPool}(F) \oplus \text{MaxPool}(F))). \tag{13}$$

The function AvgPool represents spatial average pooling, MaxPool represents spatial maximum pooling, and Conv represents the convolution function in the given formula [40]. The symbol \oplus denotes the operation of feature merging.

Due to the inefficiency of using fully connected layers in channel attention to extract spatial features while increasing the computational burden of the network, this study proposes a channel attention mechanism that only performs global average pooling. Compared to the

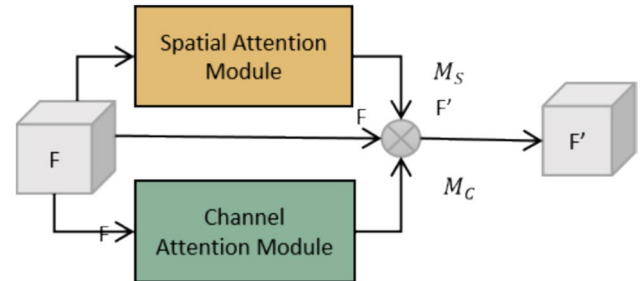


Fig. 9 Bottleneck Attention Module

commonly used channel attention mechanism that simultaneously performs global average pooling and max pooling, this approach has lower computational cost, faster inference speed, and improves the stability and robustness of the model. The BAM module combines channel attention and spatial attention. This fusion can be achieved through element-wise multiplication or addition. The overall calculation formula of the attention module for input features is given by Eq. (14).

$$F' = (M_c(F) \otimes M_s(F)). \tag{14}$$

The input features F in the given equation represent the spatial attention module and the improved channel attention module. The BAM structure, as depicted in Fig. 9, combines these modules. This combination allows the network to simultaneously focus on feature information from different channels and positions, resulting in a better representation of the image content. In summary, the BAM branch attention mechanism integrates both channel and spatial information, enabling the convolution neural network to focus more on the important aspects of the image features, thereby achieving better performance in various image processing tasks.

3 Experiment

3.1 Training environment and details

During this experiment, the execution of the code uses Google Colab, a web-based notebook that allows for writing and executing arbitrary Python code through a browser. The

Table 1 Hyperparameter setting

parameter Name	Parameter value
Learning	0.01
Momentum	0.937
Weight decay	0.0005
Batch size	16
Epoch	100

training of the model utilized the NVIDIA T4 Tensor Core GPU provided on Google Colab. PyTorch version 1.9.0 and Python version 3.10.12 were used for programming, with CUDA version 11.8. The training hyperparameters were set according to Table 1. The model was trained with 16 Batch size, initial learning rate of 0.01, and 100 training epochs.

3.2 Datasets

This study was tested on the publicly available GDUT Hardhat Wear Detection (GDUT-HWD) dataset [41]. The dataset consists of 3,174 images, with 2,177 images used for training and 997 images used for testing, divided in an 7:3 ratio. It contains a total of 18,893 instances. The labels include blue safety helmets (blue), white safety helmets (white), yellow safety helmets (yellow), red safety helmets (red), and no safety helmet (none), comprising a total of five detection categories. Furthermore, GDUT-HWD categorizes safety helmets into 3 sizes: small (safety helmet area less than 322 pixels) accounting for 47.4%, medium (safety helmet area greater than 322 pixels and less than 962 pixels) accounting for 41.9%, and large (safety helmet area greater than 962 pixels) accounting for 10.7%. Instances of small safety helmets constitute the majority, which increases the difficulty of safety helmet detection and raises the requirement for the model to detect small objects. Each instance is annotated with a class label and its corresponding bounding box. Its basic characteristics are depicted in Fig. 10.

3.3 Model evaluation

The evaluation metrics employed in this text predominantly encompass the mean average precision (mAP), parameter quantity, floating point operation count (FLOPs), and model size.

$$P = \frac{TP}{TP + FP}. \quad (15)$$

$$R = \frac{TP}{TP + FN}. \quad (16)$$

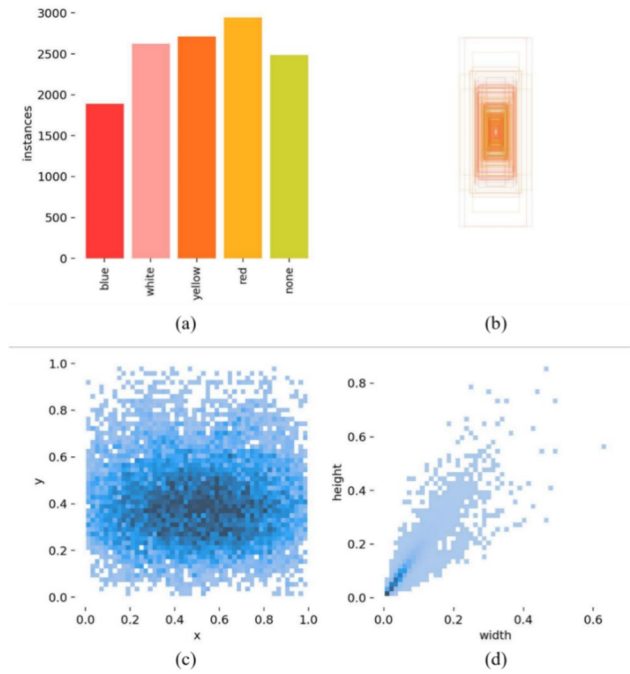


Fig. 10 The fundamental characteristics of the dataset are as follows. (a) Displays the number of instances wearing safety helmets versus those not wearing safety helmets; (b) showcases the dimensions and quantity of the boxed samples; (c) illustrates the basic characteristics of target sizes in the dataset, with darker areas indicating a higher concentration of small targets; 10(d) provides information on the distribution of class positions

$$AP = \int_0^1 P(R)dR. \quad (17)$$

$$mAP = \frac{1}{N} \sum_{j=1}^N AP_j. \quad (18)$$

$$F_1 - \text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

In the given equations, TP (true positive) represents the classification of positive samples as positive samples, FP (false positive) represents the classification of negative samples as positive samples, and FN (false negative) represents the classification of positive samples as negative samples. N represents the sample category. P (precision) represents the precision, R (recall) represents the recall, and AP (average precision) represents the average precision of a certain class comparison. The complexity of the model is measured by the number of parameters and the computational complexity FLOPs. F1 score is a measure of classification problems, which is the harmonic mean of accuracy and recall, with a maximum of 1 and a minimum of 0.

4 Results

4.1 Ablation experiments results

This study conducted a comparative analysis of ablative experiments to detect the optimization effects of various improvement points, as the results are shown in Table 2.

It is shown that after adding the SE + DSCConv module, the mAP (0.5) and FPS values increased by 0.7% and 2.7%, respectively, while maintaining almost the same number of parameters and computational loads.

When the R-Ghost module was added separately to YOLOv5s, the number of parameters decreased by 25.4%, the FLOPs decreased by 22.2%, and the model size decreased by 25%, resulting in a 0.6% improvement in model mAP (0.5) value. With the introduction of BAM only, mAP (0.5) compared to YOLOv5s, while the number of parameters and FLOPs increased only slightly. From the comparison in Table 2, it can be concluded that the R-Ghost module significantly reduces the complexity of the model without affecting its average precision. Although the addition of the BAM module can improve the mAP value, it slightly increases the complexity of the model. The SE + DSCConv module maintains the complexity of the model while also improving the mAP value to some extent. With the simultaneous improvement of these two modules and the introduction of the BAM attention mechanism, compared to the original YOLOv5s model, the number of parameters was reduced by 22.2%, FLOPs by 20.9%, the model size decreased by 20.1%, and mAP (0.5) value increased by 1.0%.

4.2 Comparison of experiments results

To further validate the effectiveness of the model, a series of comparative experiments is conducted in this section. As shown in Table 3, we compare the performance of different versions of YOLOv5. Due to the balanced detection speed and accuracy of the YOLOv5 model on the GDUT-HWD dataset, YOLOv5s is adopted as the detection model in this study, and improvements are made based on it.

Table 3 Performance comparison of different models of YOLOv5

Model	mAP (0.5)	Parameters	FPLOPs(G)	Weight	FPS
YOLOv5s	88.2%	7.02 M	16.0	14.4 MB	86.2
YOLOv5m	90.2%	20.87 M	47.9	42.2 MB	71.0
YOLOv5n	86.9%	1.77 M	4.1	3.8 MB	113.6

Furthermore, the improved attention mechanism BAM is compared with other mainstream attention mechanisms to validate the effectiveness of the attention proposed in this study. Comparative experiments are conducted on the GDUT-HWD dataset to compare the improved BAM attention with the SE (Squeeze-and-Excite) module, CBAM (Convolutional Block Attention Module) [42], and EMA (Expectation–Maximization Attention) [43] module.

According to Table 4, compared to the more advanced EMA attention model, the improved BAM used in this study outperforms in terms of parameter quantity, computational complexity, and detection speed. Although the SE attention has the optimal mAP value and detection speed, it falls behind the improved BAM attention module in terms of parameter quantity and model size. Furthermore, the improved BAM attention introduced channel attention and spatial attention, similar to the BAM and CBAM attention modules. However, the improved BAM attention outperforms BAM and CBAM attention in terms of model size and detection speed, making it more suitable for lightweight applications.

To further validate the effectiveness of the improved LOYOv5, we conducted comparative experiments using the latest method of replacing the backbone network with a lightweight network. Our findings indicate that EfficientViT (Lightweight Multi-Scale Attention) [44] has the fastest detection speed, but falls short in terms of accuracy. Although EfficientViT (Cascaded Group Attention) [45] has lower computational complexity compared to the improved YOLOv5s, it is inferior in terms of precision, GFLOPs, model size, and FPS. The results are presented in Table 5. The actual measurement of confidence, precision, recall, and F1 value for detection of the helmet of our

Table 2 Ablation experiments

Method	SE + DSCConv	R-Ghost	BAM	mAP (0.5)	Parameters	FLOPs (G)	Weight
YOLOv5s				88.2%	7.02 M	15.8	14.4 MB
A	√			88.8%	7.02 M	15.8	14.5 MB
B		√		88.7%	5.24 M	12.5	10.8 MB
C			√	88.9%	7.27 M	16.1	14.9 MB
D	√	√		88.5%	5.30 M	12.3	11.0 MB
E	√		√	89.1%	7.26 M	16.1	15.0 MB
F		√	√	89.6%	5.46 M	12.6	11.3 MB
G	√	√	√	89.1%	5.46 M	12.6	11.5 MB

Table 4 Comparison of attention models

Model	mAP (0.5)	Parameters	FLOPs (G)	Weight	FPS
YOLOv5s-improved BAM	88.9%	7.27 M	16.1	14.4 MB	82.0
YOLOv5s-BAM	89.1%	7.75 M	16.0	15.8 MB	73.5
YOLOv5s-SE	89.3%	7.33 M	16.0	15.0 MB	76.9
YOLOv5s-CBAM	88.9%	7.26 M	16.0	14.9 MB	84.75
YOLOv5s-EMA	88.9%	7.26 M	17.7	14.9 MB	66.23

Table 5 Comparison of state-of-the-art lightweight models

Model	mAP (0.5)	Parameters	FLOPs (G)	Weight	FPS
Reference [44]	71.4%	3.75 M	6.9	7.8 MB	86.2
Reference [45]	85.2%	5.32 M	10	15.2 MB	71.0
Reference [46]	84.9%	0.54 M	4.8	1.3 MB	156
Reference [18]	92.9%	8.1 M	18.4	17.1 MB	79
Improved YOLOv5s	89.1%	5.46 M	12.6	11.5 MB	72.5

model is shown in Fig. 11. Furthermore, the comparison of loss, precision, recall, mAP0.5, and mAP0.5–0.95 curves between the original YOLOv5 model and the improved YOLOv5 model can be seen in Fig. 12. It is evident that the improved model outperforms the original YOLOv5 model

in all performance indicators, demonstrating the high practicality of the improved model in detecting safety helmets in complex environments.

The comparison between YOLOv5s and the improved YOLOv5s detection algorithm for safety helmet detection on the GDUT-HWD dataset is shown in Fig. 13. Five sets of images show the detection results of safety helmets of different sizes. The improved detection confidence is similar to before, but the problem of missed detections has been solved, proving the practicality of the improved algorithm.

To further demonstrate the superiority of the proposed algorithm, a comparison is made with the improved detection algorithm SDD network model presented in this paper, as well as the lightweight models YOLOX-Tiny [35] and YOLOv7-tiny from the YOLO series, and the classical YOLO detection algorithms YOLOX-s and YOLOv7 [36]. Table 6 compares the model parameters, model size, FLOPs and detection speed. Table 7 compares the average precision (AP) of each class and the mAP of all classes. The mAP

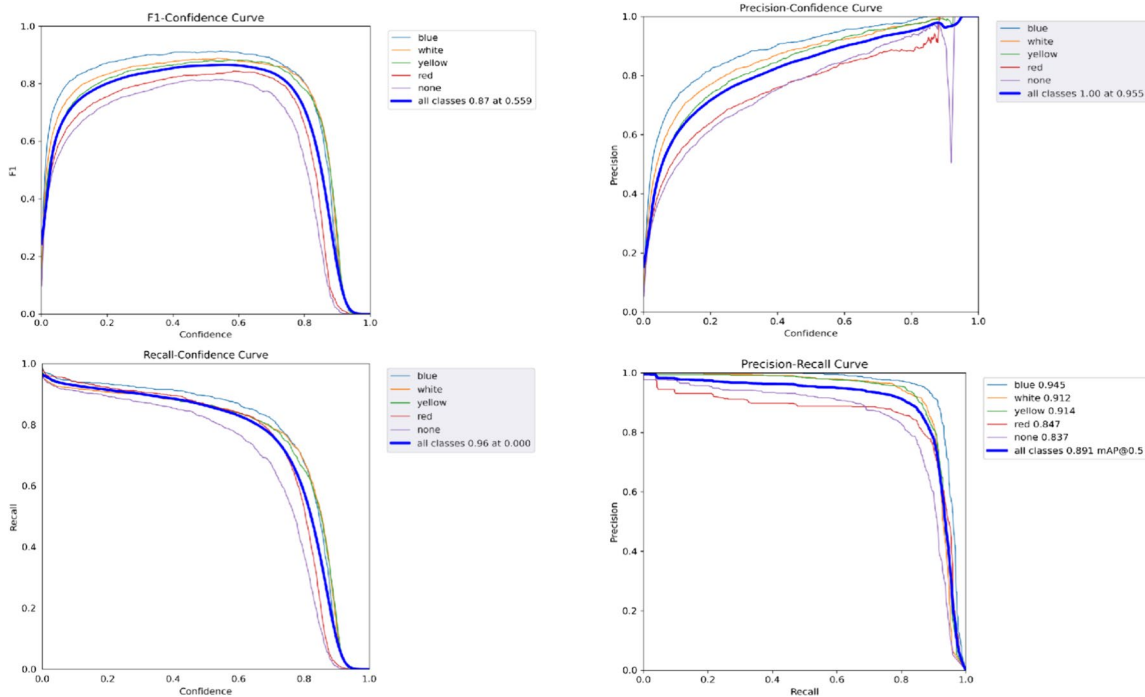
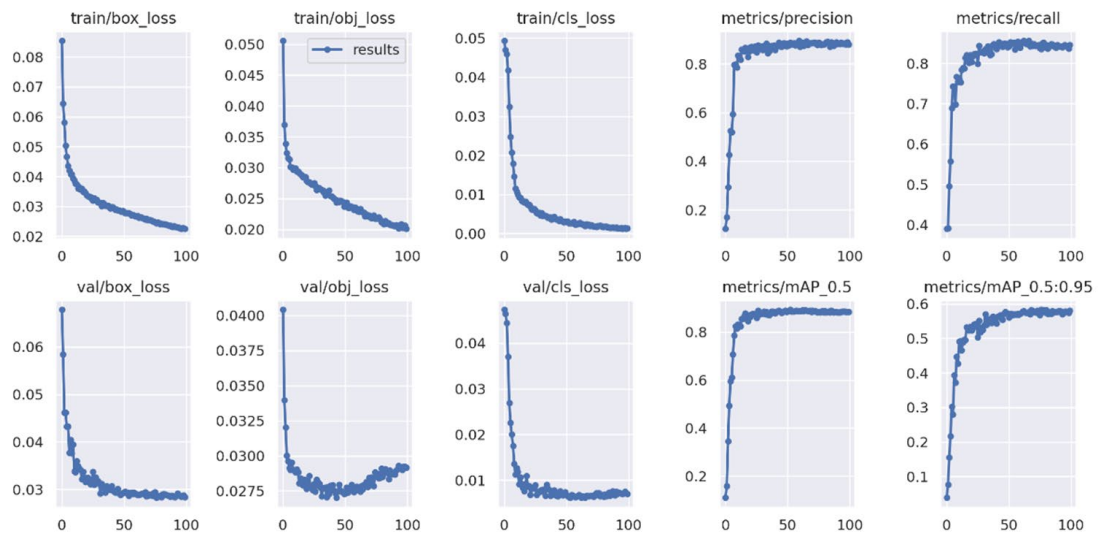
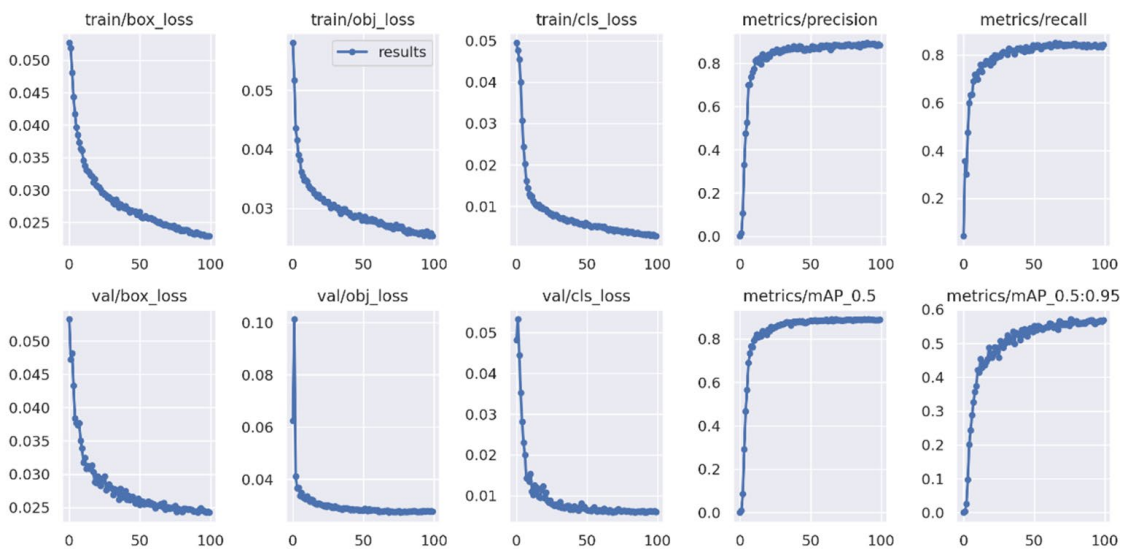


Fig. 11 P curve, R curve, PR curve, F1 curve. (a)YOLOv5s, (b) improved YOLOv5s



(a)YOLOv5s



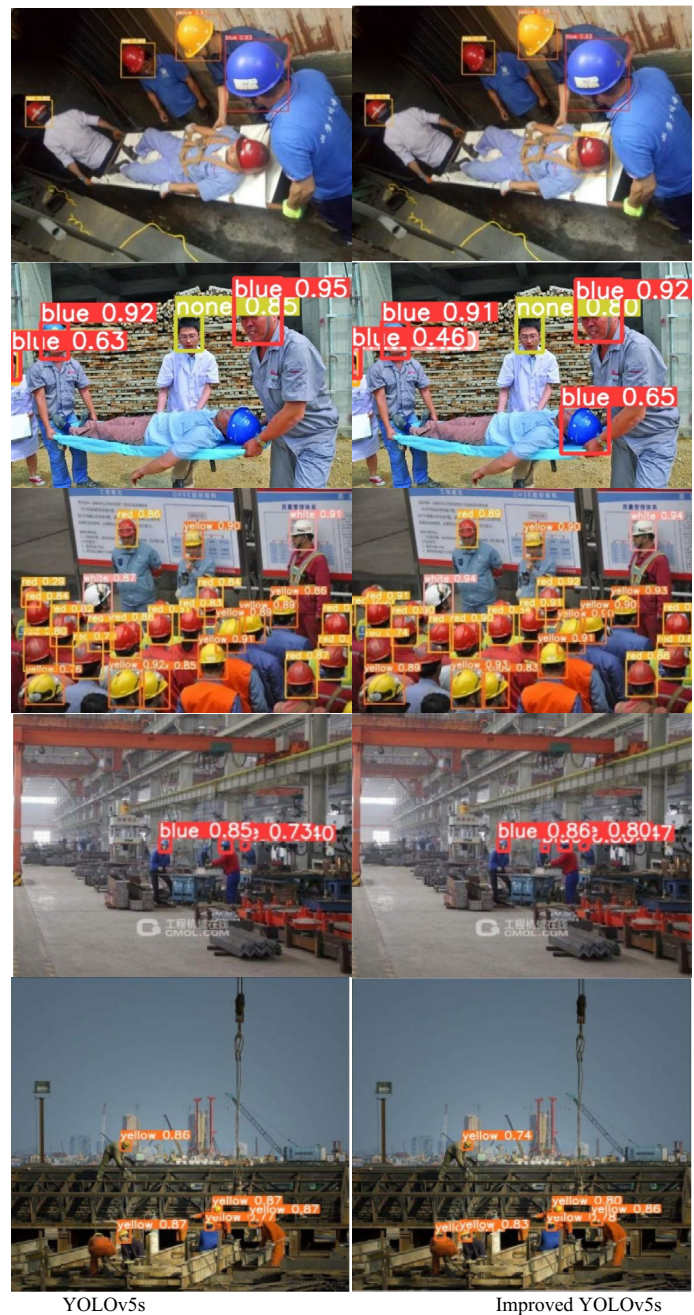
(b)Improved YOLOv5s

Fig. 12 Box_loss, obj_loss, cls_loss, precision, recall, mAP0.5 and mAP0.5–0.95 curves. The x-axis represents the number of experimental epochs, while the y-axis represents the probability

value of the SSD network model is 84.9%, but its network structure and model size are large, not suitable for lightweight applications. YOLOX-Tiny achieves a mAP value of 88.6% with a model size of 36.9 MB, showing good performance, but the model is complex. YOLOv7-tiny has a mAP value of 61.3% with a parameter size of 6.01 M, showing a significant difference from the improved YOLOv5s model. The improved YOLOv5s achieves a mAP of 89.1%, and the complexity of the model is smaller than other lightweight

models. Although the mAP value of the YOLOX-s detection algorithm is close to the improved algorithm at 80.6%, its computational complexity is 26.8 GFLOPs, slightly larger than the lightweight algorithms. The mAP value of YOLOv7 is 79.8%, but its model complexity does not meet the requirements for lightweight applications. YOLOv8s has the best accuracy but is not suitable for lightweight applications. Overall, the performance of the improved model is significantly better than other network models.

Fig. 13 Detection and comparison



YOLOv5s

Improved YOLOv5s

Table 6 Contrast experiment

Model	mAP (0.5)	parameters	FLOPs (G)	Weight	FPS
YOLOv5s	88.2%	7.02 M	16.0	14.4 MB	86.2
SSD	84.9%	26.3 M	273.7	91.1 MB	57.2
YOLOX-Tiny	88.6%	5.06 M	10.4	36.9 MB	74.5
YOLOv7-Tiny	61.3%	6.01 M	13.2	12.3 MB	85.9
YOLOX-s	80.6%	8.93 M	26.8	34.9 MB	66.4
YOLOv7	79.8%	37.20 M	105.1	74.8 MB	26.7
YOLOv8s	90.6%	11.12 M	28.4	22.5 MB	59.5
Improved YOLOv5s	89.1%	5.46 M	12.6	11.5 MB	72.5

Table 7 Detailed results for five categories of the dataset

Class	YOLOv5s	SDD	YOLOX-Tiny	YOLOX-S	YOLOv7	YOLOv8s	Improved-YOLOv5s
Blue	93.5%	91.3%	93.8%	66.9%	84.9%	95.3%	94.5%
White	91.5%	87.2%	91.4%	63.0%	80.7%	93.4%	91.2%
Yellow	92.4%	88.6%	91.8%	63.3%	81.9%	93.4%	91.4%
Red	84.2%	82.8%	84.4%	58.5%	80.3%	86.4%	84.7%
NONE	79.3%	74.6%	81.6%	54.9%	75.2%	84.2%	83.7%
mAP	88.2%	84.9%	88.6%	61.3%	80.6%	90.5%	89.1%

5 Conclusion

This article presents an improved lightweight YOLOv5s detection algorithm for safety helmets. The SE + DSCConv module is used to replace the original convolution module and some C3 modules, which not only maintains computational complexity but also improves accuracy. Parallel multi-branching of the Ghost module followed by serial skip layer connections reduces computational and parameter costs while improving detection accuracy. Finally, an improved attention mechanism called BAM is introduced to enhance the weight of important information in the feature map and address the issues of missed detection and false detection. Through comparative ablation experiments, the advantages of the improved modules are demonstrated. Finally, compared with other lightweight algorithms, the improved detection algorithm exhibits excellent performance in various indicators, meeting the requirements of lightweight model and is more suitable for use in mobile devices and embedded systems. In the future, the focus will be on the detection of small targets such as safety helmets, and improvements need to be made to the improved algorithm to improve accuracy without compromising speed, to better apply it to real-time detection and recognition of safety helmet wearing in production and construction sites.

Acknowledgements This research is sponsored by National Natural Science Foundation of China [Grant No.: 61203343].

Author contributions HR provided guidance throughout the research process and managed the funding. AF conceived the study and wrote the manuscript. JZ reviewed the manuscript. HS collected the data. XL analyzed the data. All authors reviewed and approved the final manuscript.

Funding This study is supported by the National Natural Science Foundation of China, 61203343.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request. The

publicly available dataset utilized in this research can be accessed via the following link: <https://github.com/wujixiu/helmet-detection>.

Declarations

Competing interests The authors declare no competing interests.

Conflicts of interest The authors declare no conflict of interest exists.

References

- Li, Y., Wei, H., Han, Z., Huang, J., Wang, W.: Deep learning-based safety helmet detection in engineering management based on convolutional neural networks. *Adv Civil Eng* (2020). <https://doi.org/10.1155/2020/9703560>
- Chen, Z., Zhang, F., Liu, H., et al.: Real-time detection algorithm of helmet and reflective vest based on improved YOLOv5. *J. Real-Time Image Proc.* **20**, 4 (2023). <https://doi.org/10.1007/s11554-023-01268-w>
- Wu, J., Cai, N., Chen, W., et al.: Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. *Autom. Constr.* **106**, 102894 (2019). <https://doi.org/10.1016/j.autcon.2019.102894>
- Adarsh, P., Rathi, P., Kumar, M. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020; pp. 687–694. <https://doi.org/10.1109/ICACCS48705.2020.9074315>.
- Bochkovskiy, A., Wang, C-Y., Liao, H-Y. M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020; <http://dx.doi.org/https://doi.org/10.48550/arXiv.2004.10934>, arXiv preprint arXiv:2004.10934.
- Redmon, J., Divvala, S., Grishick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016; pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>.
- Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; pp. 7263–7271. <https://doi.org/10.48550/arXiv.1612.08242>.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C-Y., Berg, A.C.: SSD: Single Shot Multi Box Detector. arXiv. Org. 2015; https://doi.org/10.1007/978-3-319-46448-0_2.

9. Girshick, R.: Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015; 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>
10. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2013; 580–587. [arXiv.Org. https://arxiv.org/abs/1311.2524v5](https://arxiv.org/abs/1311.2524v5).
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2017). <https://doi.org/10.1109/TPAMI.2016.2577031>
12. Kerdvibulvech, C.: A methodology for hand and finger motion analysis using adaptive probabilistic models. *J Embedded Systems* **2014**, 18 (2014). <https://doi.org/10.1186/s13639-014-0018-7>
13. Singh, C., Mittal, N., Walia, E.: Complementary feature sets for optimal face recognition. *J Image Video Proc* **2014**, 35 (2014). <https://doi.org/10.1186/1687-5281-2014-35>
14. Mithun, G.J., Juan, P.W.: Context-based hand gesture recognition for the operating room. *Patt Recogn Lett* **36**, 196–203 (2014)
15. A-masiri, P., Kerdvibulvech, C.: Anime face recognition to create awareness. *Int J Inf Technol* **15**, 3507–3512 (2023). <https://doi.org/10.1007/s41870-023-01391-8>
16. Ge, Y., Liu, H., Du, J., Li, Z., Wei, Y.: Masked face recognition with convolutional visual self-attention network. *Neurocomputing* **518**, 496–506 (2023)
17. Silva R. R. V. e., Aires K. R. T., Veras R. d. M. S.: Helmet Detection on Motorcyclists Using Image Descriptors and Classifiers. 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images, Rio de Janeiro, Brazil, 2014, pp. 141–148, <https://doi.org/10.1109/SIBGRAPI.2014.28>.
18. Chen, W., Li, C., Guo, H.: A lightweight face-assisted object detection model for welding helmet use. *Expert Syst. Appl.* **221**, 119764 (2023)
19. Zhao, L., Tohti, T., Hamdulla, A.: BDC-YOLOv5: a helmet detection model employs improved YOLOv5. *SIVIP* **17**, 4435–4445 (2023). <https://doi.org/10.1007/s11760-023-02677-x>
20. Xu, H., Wu, Z.: MCX-YOLOv5: efficient helmet detection in complex power warehouse scenarios. *J. Real-Time Image Proc.* **21**, 27 (2024). <https://doi.org/10.1007/s11554-023-01406-4>
21. Jin, P., Li, H., Yan, W., Xu, J.: YOLO-ESCA: a high-performance safety helmet standard wearing behavior detection model based on improved YOLOv5. *IEEE Access* **12**, 23854–23868 (2024). <https://doi.org/10.1109/ACCESS.2024.3365530>
22. Wang, B., Li, W., Tang, H.: Improved YOLO v3 algorithm and its application in helmet detection. *Comput. Eng. Appl.* **56**(9), 33 (2020). <https://doi.org/10.3778/j.issn.1002-8331.1912-0267>
23. Zhao, H., Tian, X., Yang, Z., Bai, W.: YOLO-S: a novel lightweight model for safety helmet wearing detection. *J East China Normal Univ* **5**, 12 (2021). <https://doi.org/10.3969/j.issn.1000-5641.2021.05.01>
24. Song, X., Wu, Y., Liu, B., Zhang, Q.: Safety helmet wearing detection using improved YOLOv5s algorithm. *Comput. Eng. Appl.* **59**(2), 194–201 (2023)
25. Song, H., Zhang, X., Song, J., et al.: Detection and tracking of safety helmet based on DeepSort and YOLOv5. *Multimed Tools Appl* **82**(7), 10781–10794 (2023). <https://doi.org/10.1007/s11042-022-13305-0>
26. Zhang, J., Qu, P., Sun, C., Luo, M., Yan, G., Zhang, J., Liu, H.: DWCA-YOLOv5: an improve single shot detector for safety helmet detection. *J Sensors* **2021**, 1–12 (2021)
27. Sun, C., Zhang, S., Qu, P., Wu, X., Feng, P., Tao, Z., Zhang, J., Wang, Y.: MCA-YOLOV5-light: a faster, stronger and lighter algorithm for helmet-wearing detection. *Appl. Sci.* **12**, 9697 (2022). <https://doi.org/10.3390/app12199697>
28. Ultralytics. YOLOv5. 2021; <https://github.com/ultralytics/yolov5>.
29. Ramachandran, P., Zoph, B., Le, QV.: Searching for Activation Functions. *arXiv.Org.* 2017; <https://arxiv.org/abs/1710.05941v2>.
30. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. *IEEE Conf Comput Vision Patt Recog (CVPR)* **2017**, 936–944 (2017). <https://doi.org/10.1109/CVPR.2017.106>
31. Wang, C., Xu, C., Wang, C., Tao, D.: Perceptual adversarial networks for image-to-image transformation. *IEEE Trans. Image Process.* **27**(8), 4066–4079 (2018). <https://doi.org/10.1109/TIP.2018.2836316>
32. Gennari, M., Fawcett, R., Prisacariu, V.A.: DSConv: Efficient Convolution Operator. 2019a; <https://doi.org/10.48550/arXiv.1901.01928v1>.
33. Gennari, M., Fawcett, R., Prisacariu, V.A.: DSConv: Efficient Convolution Operator. 2019b; [arXiv.Org. https://arxiv.org/abs/1901.01928v2](https://arxiv.org/abs/1901.01928v2).
34. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018; pp. 7132–7141. <https://doi.org/10.48550/arXiv.1709.01507>.
35. Han, G., Zhu, M., Zhao, X., Gao, H.: Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection. *Comput Elect Eng* **95**, 107458 (2021)
36. Sannasi, G., Devansh, A.: An Intelligent Video Surveillance System for Detecting the Vehicles on Road Using Refined YOLOV4. *Comput Elect Eng* **113**, 109036 (2024)
37. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2019; 580–587. [arXiv.Org. https://arxiv.org/abs/1911.11907v2](https://arxiv.org/abs/1911.11907v2).
38. Gao, S.H., Cheng, M.M., Zhao, F., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans Patt Anal Mach Intell IEEE* **43**, 652–662 (2021)
39. Park, J., Woo, S., Lee, J.-Y., Kweon, I. S.: BAM: Bottleneck Attention Module. 2018; [arXiv:1807.06514](https://arxiv.org/abs/1807.06514). <https://doi.org/10.48550/arXiv.1807.06514>.
40. Ji, S.J., Ling, Q.H., Han, F.: An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information. *Comput Elect Eng* **105**, 108490 (2023)
41. Tan, L., Lv, X., Lian, X., Wang, G.: YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm. *Comput Elect Eng* **93**, 107261 (2021)
42. Wu, J., Cai, N., Chen, W., Wang, H., Wang, G.: Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. *Autom. Constr.* **106**, 102894 (2019). <https://doi.org/10.1016/j.autcon.2019.102894>
43. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S. CBAM: Convolutional Block Attention Module. 2018; [arXiv:1807.06521](https://arxiv.org/abs/1807.06521). <https://doi.org/10.48550/arXiv.1807.06521>.
44. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H. Expectation-Maximization Attention Networks for Semantic Segmentation. *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019; <https://arxiv.org/abs/1907.13426v2>.
45. Cai, H., Li, J., Hu, M., Gan, C., Han, S. EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction. *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022; <https://doi.org/10.48550/arXiv.2205.14756>.
46. Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. *IEEE/CVF Conference on Computer Vision and Pattern*

- Recognition (CVPR).2023; pp: 14420–14430. <https://arxiv.org/abs/2305.07027v1>.
47. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. YOLOX: Exceeding YOLO Series in 2021. [arXiv:2107.08430](https://arxiv.org/abs/2107.08430). <https://doi.org/10.48550/arXiv.2107.08430>.
48. Wang, C-Y., Bochkovskiy, A., Liao, H-Y. M.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023;7464–7475. <https://doi.org/10.1109/CVPR52729.2023.00721>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Hongge Ren received the PH.D degree in Pattern Recognition and Intelligent Systems from Beijing University of Technology. She is currently the professor of the School of Control and Mechanical Engineering, Tianjin Chengjian University. Her main research areas include

artificial intelligence, robot control, target tracking, image processing, and more.

Anni Fan is currently studying for the M.S. degree in Intelligent Construction Technology at the School of Control and Mechanical Engineering, Tianjin Chengjian University.

Jian Zhao received postdoctoral fellow in Mechanical Manufacturing and Automation at Tianjin University, currently serves as the Dean of the School of Control and Mechanical Engineering at Tianjin Chengjian University. Her main research areas include intelligent manufacturing technology, dynamic measurement and intelligent diagnosis, vibration and noise measurement and control technology.

Hairui Song is currently studying for the M.S. degree in Control and Mechanical Engineering at Tianjin Urban Construction University.

Xiuman Liang majoring in Control Theory and Control Engineering at North China Electric Power University, received the Master's degree in Engineering. Main research directions: detection technology and intelligent devices, advanced control theory and applications, motion control systems.