



Safety helmet detection based on improved YOLOv7-tiny with multiple feature enhancement

Shuqiang Wang¹ · Peiyang Wu¹ · Qingqing Wu¹

Received: 29 February 2024 / Accepted: 18 June 2024 / Published online: 25 June 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Safety helmets are vital protective gear for construction workers, effectively reducing head injuries and safeguarding lives. By identification of safety helmet usage, workers' unsafe behaviors can be detected and corrected in a timely manner, reducing the possibility of accidents. Target detection methods based on computer vision can achieve fast and accurate detection regarding the wearing habits of safety helmets of workers. In this study, we propose a real-time construction-site helmet detection algorithm that improves YOLOv7-tiny to address the problems associated with automatically identifying construction-site helmets. First, the Efficient Multi-scale Attention (EMA) module is introduced at the trunk to capture the detailed information; here, the model is more focused on training to recognize the helmet-related target features. Second, the detection head is replaced with a self-attentive Dynamic Head (DyHead) for stronger feature representation. Finally, Wise-IoU (WIoU) with a dynamic nonmonotonic focusing mechanism is used as a loss function to improve the model's ability to manage the situation of mutual occlusion between workers and enhance the detection performance. The experimental results show that the improved YOLOv7-tiny algorithm model yields 3.3, 1.5, and 5.6% improvements in the evaluation of indices of mAP@0.5, precision, and recall, respectively, while maintaining its lightweight features; this enables more accurate detection with a suitable detection speed and is more in conjunction with the needs of on-site-automated detection.

Keywords Construction workers · Helmet detection · YOLOv7-tiny · Attention mechanism · Loss function

1 Introduction

Construction workers face many dangers and safety risks in their work, including head injuries [1], and wearing a helmet is important for the prevention of head injuries. Many construction workers often disregard safety regulations by not wearing helmets at construction sites for convenience or comfort, causing increased safety risks. Therefore, reasonable constraints on the wearing of worker helmets at construction sites need to be implemented. At present, there are two main supervision methods, traditional manual supervision and automated detection based on image processing [2]. The traditional supervision method involves safety supervisors monitoring the violations through video surveillance equipment; however, due to the complexity of the construction site and the dim lighting conditions associated

with accessing many surveillance images, supervisors are unable to effectively oversee each scene, which can easily lead to omissions and safety accidents. To improve the safety of the construction workers, many construction sites have started to use the second type of supervision, which is based on image processing technology. However, the complexity of the production site video acquisition environment, target occlusions, uneven illuminations, and large target scale differences [3] pose challenges to the automatic detection and recognition of helmet wearing based on image processing.

Many scholars worldwide have carried out extensive research on helmet-wearing detection and recognition algorithms. The helmet-wearing detection and recognition algorithms can be divided into the traditional detection methods and deep learning-based methods. Traditional helmet-wearing detection methods can be divided into two categories, the sensor-based detection methods and computer vision-based detection methods. Sensor-based detection techniques focus on remote location and follow-up techniques such as radio frequency identification (RFID) [4] and wireless local area networks (WLANs) [5]. Kelm

✉ Peiyang Wu
102200821@hbut.edu.cn

¹ School of Civil Engineering, Architecture and Environment, Hubei University of Technology, Wuhan 430068, China

et al. [6] designed a mobile RFID portal for checking the correctness of personal protective equipment (PPE) worn by personnel. However, the RFID reader located at the entrance of the construction site could not monitor the non-entrance areas and could not identify whether the helmets were being worn by the construction workers. Li et al. [7] developed a real-time location system (RTLS) for tracking the position of workers by placing pressure sensors on their helmets; then the pressure information was transmitted via Bluetooth to monitor and determine whether safety helmets were needed, and warnings were sent when the helmets were deemed necessary. Zhang et al. [8] developed a smart helmet system using an IoT-based architecture. To determine the usage status of the helmet, an infrared beam detector and a thermal infrared sensor were placed inside the helmet. When both the infrared beam detector and the thermal infrared sensor were activated, helmet usage was confirmed. In general, the existing sensor-based methods have difficulty in accurately identifying whether people are wearing helmets at the construction sites. In addition, the use of sensors can result in significant production costs. Traditional computer vision-based detection methods usually utilize manually selected features or statistical features in using various steps such as background subtraction, human detection, and safety helmet detection and identification. Liu et al. [9] proposed the use of skin color to assist in determining the helmet location, extracting Hu moment feature vectors, and then using support vector machines (SVMs) to identify and classify helmet usage. Li et al. [10] used Vibe to segment motion backgrounds for motion targets in a surveillance scene at a fixed location, utilized the real-time human classification framework C4 to locate the human body, and finally achieved color feature discrimination for helmet detection. Traditional computer vision-based detection still relies on manual interventions, and problems such as low real-time performance and low robustness, make it difficult to meet the current requirements for automated helmet detection.

Traditional computer vision (CV)-based detection methods perform well in some specific scenes, but are often limited by complex backgrounds and variable lighting conditions, and their accuracy and robustness need to be improved. In recent years, with the rapid development of deep learning technology, deep learning-based methods have shown significant advantages in computer vision tasks. Compared to traditional CV methods, the deep learning methods are able to automatically learn feature representations in images rather than relying on hand-designed features. This enables

the deep learning models to have stronger generalization performance and accuracy in processing complex images.

Deep learning-based methods can be further classified into “two-stage” and “one-stage” methods. The “two-stage” approach consists of an algorithm that extracts features for candidate region generation and then uses a classifier to perform classification regression.

Yogameena et al. [11] used Faster R-CNN to detect motorbike targets with markers and then used a convolutional network model and spatial converter to identify helmets. Ferdous et al. [12] designed ResNet50 as the backbone fused feature pyramid network (FPN) to classify and localize helmets using classification and regression models. Wu et al. [13] improved the Faster R-CNN algorithm to fuse multiple feature layers and perform multiscale detection of helmets. The advantage of the “two-stage” method is that it can effectively improve the detection accuracy, but it has difficulty meeting the requirements of real-time detection. The “single-stage” approach uses an end-to-end strategy to detect and classify the target location in the image. The SSD (Single Shot MultiBox Detector) model [14] and the YOLO (You Only Look Once) model [15] are the most effective methods for detecting helmets in real time. The SSD and YOLO models are typical examples of “single-stage” algorithms. In recent years, many improved single-stage target detection algorithms for helmets have emerged. Redmon et al. [16] proposed introducing the MobileNet network into the SSD algorithm and applied their modified algorithm for helmet detection to improve the detection speed. Li et al. [17] increased the feature fusion role of the branch network in the SSD model and improved the default frame configuration to improve the accuracy of the algorithm for helmet detection in real application scenarios. Geng et al. [18] improved the detection accuracy using the Gaussian fuzzy method in YOLOv3 to address the problem of imbalanced data in the helmet dataset. Xiao et al. [19] significantly improved the accuracy of helmet-wearing detection via the YOLOv3 network by increasing the scale of the input image and reducing the loss of image features via depth separable convolution. Shen et al. [20] used bounding box regression and migration learning for helmet detection and improved the efficiency of the model by introducing DenseNet. Wang et al. [21] improved YOLOv5 by introducing a Convolutional Block Attention Module (CBAM), which significantly improved the detection accuracy, and the detection speed could meet the needs of real-time detection. Zhang et al. [22] combining both Bidirectional Feature Pyramid Network (BiFPN), additional detection layer and CBAM modules in YOLOv5 reduces the model’s false and missed detection rates. C.Geupta et al. [23] improved YOLOv8 for the fuzziness and invisibility of objects in images by introducing two

feature extraction methods and pruning operations on the model. While reducing the size of the model, the accuracy of detection is improved.

Based on the above literature review, it is clear that helmet detection techniques based on deep learning have been extensively researched. However, construction sites are usually complex environments with a variety of objects and interferences, such as mechanical equipment, obstacles, irregular light and shadows; these interferences can have an impact on helmet detection. On the other hand, some of the past algorithms fail to achieve a good balance between detection speed and detection accuracy.

In this study, YOLOv7-tiny [24], the network with the simplest structure and the fastest computational speed among the YOLOv7 series of algorithms, is used as the framework for the construction-site helmet detection algorithm. While the detection speed is fast, a certain level of detection accuracy is guaranteed in complex environments. On this basis, the YOLOv7-tiny network will be applied to improve upon the other possible problems in the detection process. The main contributions of this study are as follows:

First, screening for mainstream target detection algorithms, the YOLOv7-tiny algorithm, which performs better in helmet detection, is initially selected as the benchmark algorithm model for this paper.

Second, to address the problem of the influence of the complex environment of construction sites, the EMA attention mechanism [25] is introduced to focus more efficient attention to helmets against complex backgrounds, reduce the interference of irrelevant information, and improve the model detection performance.

Third, the IDetect Head is replaced with the self-attentive dynamic detection head, DyHead [26], which uses the attention mechanism on each feature dimension of the feature tensor. DyHead can further address the problems of complex environments and target size variance at construction sites and improve the robustness of target detection as well as the detection performance of the model.

Fourth, for the hard-to-classify samples of occluded and dense targets in the helmet dataset, the original loss function is replaced with the weighted out-of-union (WIoU) loss function [27]; this function effectively reduces the contribution of simple samples to the loss value and simultaneously enables the model to focus on difficult samples such as occlusions and enhances the generalizability of difficult samples.

Fifth, through experiments and results analysis, the improved YOLOv7-tiny algorithm was found to perform well in the task of safety helmet-wearing recognition for construction workers. While keeping the small size, the recognition accuracy and model stability are improved to meet the demand of real-time helmet detection in construction sites.

2 Preliminary work

2.1 Algorithm selection

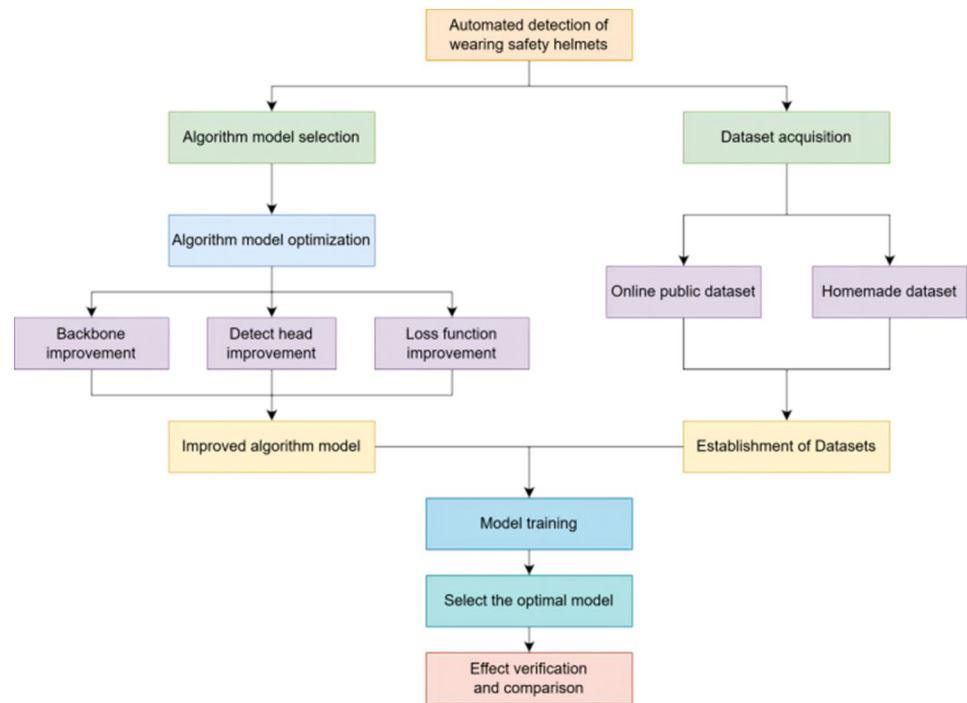
Automated testing of helmet wear requires a series of preparatory steps. First, we need to acquire the datasets. The dataset is then divided into a network public dataset and a homemade dataset. In this study, the combination of a network public dataset and a homemade dataset is used to establish the complete dataset. On the other hand, we need to select the algorithm with better performance from the current mainstream algorithms as the benchmark algorithm model. Moreover, the benchmark algorithm model needs to be improved accordingly with enhancements in the backbone network, detection head, and loss function, and an improved algorithm model can then be derived. Finally, the improved algorithm model is trained using the established dataset, the model with the best performance during training is derived, and its detection effect is verified and compared. The technology roadmap of this study is shown in Fig. 1.

To initially select the benchmark algorithm model with the best performance, we first recorded images at the construction site and transformed them into preliminary detection data through labeling and other operations. Moreover, several current mainstream target detection algorithm models are selected and applied to the prepared data for a preliminary verification of the actual detection effect. In Fig. 2a–d are the sample images of the detection effects from the Single Shot MultiBox Detector (SSD), Faster R-CNN [28], YOLOv5s, and YOLOv7-tiny algorithm models, respectively. Overall, the YOLOv7-tiny algorithm correctly identified all the targets, while the other algorithms yielded one false detection. In addition, the YOLOv7-tiny algorithm's confidence level for hat category recognition is also the highest among all the tested algorithms. The experimental results preliminarily show that the YOLOv7-tiny algorithm model outperforms the other mainstream algorithm models in helmet detection, and thus can be used as a benchmark algorithm model to study its potential improvement and evaluate its performance on various evaluation metrics.

2.2 YOLOv7-tiny algorithm

The YOLOv7-tiny algorithm is a deep learning-based target detection algorithm that is a lightweight version of the YOLO series of algorithms. Compared with the YOLOv7 algorithm, the YOLOv7-tiny algorithm reduces the number of parameters while increasing the detection speed. The algorithm consists of four parts: an input layer (Input), a feature extraction backbone network (Backbone), a feature fusion layer, and a detection head (Head). A fixed-size image is input and fed into a feature extraction backbone network

Fig. 1 Technology road map



consisting of ordinary convolutional layers and Mconv and ELAN convolutional layers. The feature maps extracted from the backbone network are fed into the SPPCSPC module, which refers to the combination of Spatial Pyramid Pooling (SPP) and Cross Stage Partial Connections (CSP), and then processed and subsequently fed into the Head network. Subsequently, the aggregated feature pyramid structure is used, convolution is used to adjust the channels of the features at different scales, and the confidence of the target frame is calculated with the aid of the Complete-IoU (CIoU) loss function. The YOLOv7 network introduces a multibranch stacking module, E-ELAN, in which left branch I and left branch II both contain one convolutional normalized activation function unit, while right branch I and right branch II contain three and five convolutional normalized activation function units, respectively. The features of these four branches are fused to perform one convolutional normalized activation operation. The YOLOv7-tiny network prunes this module while using the leaky ReLU activation function. Specifically, the two branches of the right branch are cut into 2 and 3 units of the convolutional normalized activation function, and a comparison is shown in Fig. 3.

The YOLOv7 network combines a convolution of size 3×3 with a step size of 2 and maximum pooling with a step size of 2 to act as a downsampling module. The YOLOv7-tiny network, on the other hand, uses only maximum pooling with a step size of 2 for the downsampling operation. In particular, due to the presence of the feature pyramid structure SPPCSPC at the tail of the network, residual operations on the SPP structure can be performed to assist in optimization

and feature extraction and to improve the sensory wildness of the network. The structure of the YOLOv7-tiny network is illustrated in Fig. 4.

The backbone network of YOLOv7-tiny is mainly responsible for extracting image features. Usually, a pre-trained model is used as the basis, and feature extraction is performed through upsampling or convolution operations; finally, three feature layers are output to the neck network, which further extracts the features and carries out the feature fusion. These layers usually include convolutional layers, pooling layers, and other operations; additionally, these layers are used to enhance the feature representation, after which the final detection results are generated through the head network.

In YOLO's network, the loss function usually consists of three parts: category loss, confidence loss, and position loss. The category loss usually adopts the cross-entropy loss; the confidence loss is mainly divided into the confidence loss with the target and without the target, which is also calculated using the cross-entropy loss; the position loss calculates the border loss between the prediction frame and the real frame; the loss is calculated by the intersection over union (IoU) ratio between the two frames; and the IoU is calculated by the following formula:

$$\text{IoU} = \frac{|b \cap b^{gt}|}{|b \cup b^{gt}|}, \quad (1)$$



Fig. 2 Comparison of the initial selection of algorithms

where b is the prediction frame region; b^{gt} is the real frame region; and $b \cap b^{gt}$ and $b \cup b^{gt}$ are the intersection and concatenation of the two regions, respectively.

The IoU has limitations in handling the intersection of two frames. To measure the intersection of two frames more accurately, the loss of generalized intersection over union (GIoU) is obtained by calculating the smallest outer rectangle of the two frames to obtain a proportion of the two frames in the rectangle; this more accurately reflects the degree of intersection of two frames. However, the computational speed and convergence speed of the GIoU metric are slightly affected by this process. To solve this problem, distance intersection over union (DIoU) regresses the Euclidean distances of the centroids of the two boxes on the basis of the IoU and helps to increase the convergence speed. Moreover, using the ratio of the centroid distance to the diagonal distance as a penalty term, DIoU effectively avoids the problem of optimization difficulty when the loss value is large. However, DIoU still suffers from the problems of centroid overlap and inconsistent aspect ratios. To obtain more accurate prediction frames, the complete intersection over union (CIoU) approach considers the consistency of the aspect ratio between two frames on the basis of the DIoU to measure the intersection of two frames from a more comprehensive perspective. Therefore, in the YOLOv7-tiny model network, both the confidence loss and category loss are calculated using the BCEWithLogitsLoss function, and the CIoU is adopted as the position loss function, accounting for the overlap area, centroid distance, aspect ratio, etc. The formula for the CIoU is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \lambda v, \tag{2}$$

$$\lambda = \frac{v}{1 - IoU + v}, \tag{3}$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{\omega^{gt}}{h^{gt}} - \arctan \frac{\omega}{h} \right)^2, \tag{4}$$

where λ is the parameter used to make the trade-off, v is the parameter measuring the consistency of the aspect ratios, ρ is the Euclidean distance between the centroids of the predicted and real frames, and c is the diagonal distance between the smallest outer rectangles of the two frames. To note, $v = 0$ when the aspect ratios are the same, at which point the partial penalty term loses its effect and is unstable.

On this basis, we have improved the YOLOv7-tiny network structure, and the improved YOLOv7-tiny network structure is shown in Fig. 5. The details of the improvement will be explained in detail in Sect. 3.

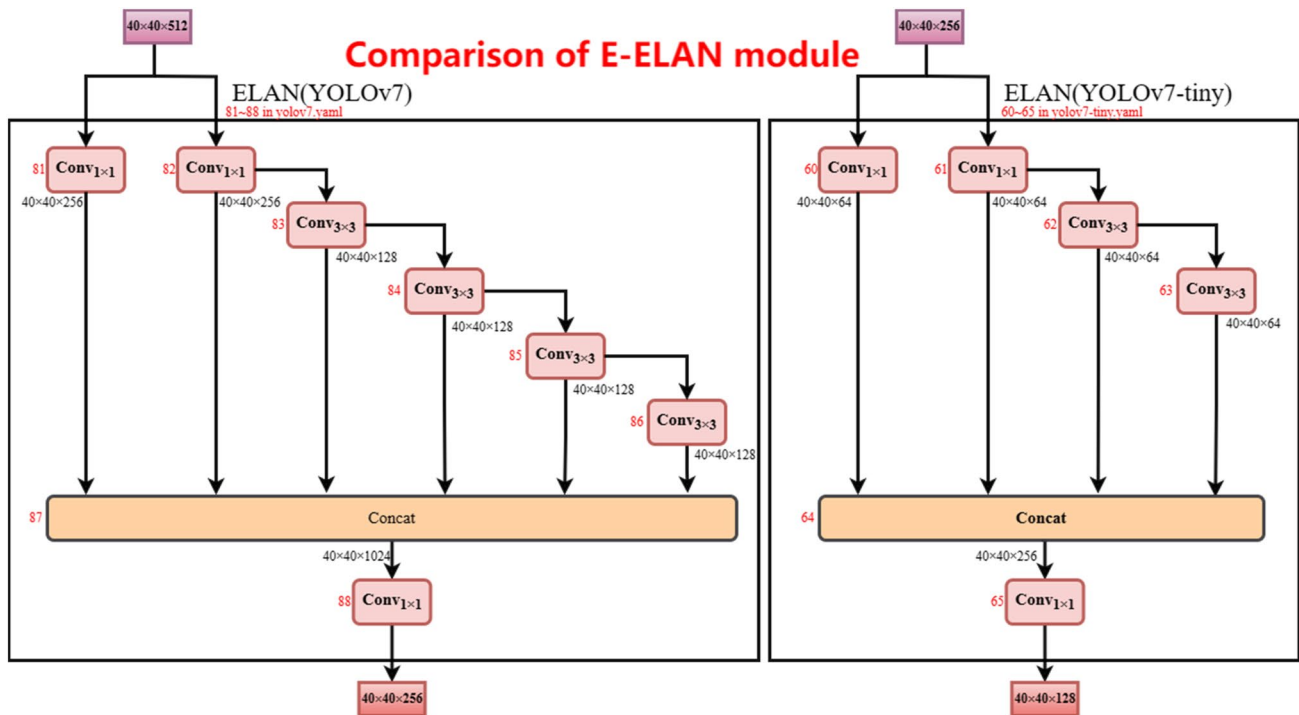


Fig. 3 Comparison of the E-ELAN modules of YOLOv7 and YOLOv7-tiny

3 An improved YOLOv7-tiny algorithm

3.1 Improvement of backbone

Attentional mechanisms are widely used in computer vision, especially in target detection tasks. However, modeling cross-channel relationships through channel dimensionality reduction may have side effects on extracting deep visual representations. There are two main types of current attention mechanisms, the channel attention mechanisms and spatial attention mechanisms. The SE module [29] models cross-dimensional interactions using a global average pooling operation to extract channel attention. The CA module [30] embeds spatial location information into the channel attention graph to enhance feature aggregation. The SGE module [31] groups channel dimensions into multiple sub-features to improve the spatial distribution of different semantic sub-feature representations. The CBAM module [32] exploits the semantic interdependencies between spatial and channel dimensions in the feature graph to construct cross-channel cross-space information. However, dimensionality reduction and grouping of channels yield better performance but inevitably reduce the processing efficiency of the detector, thereby increasing the latency.

With the goal of preserving the information in each channel and reducing the computational overhead, the Efficient Multi-scale Attention (EMA) module, which is based on cross-spatial learning, reshapes part of the channel into batch

dimensions and groups the channel dimensions into multiple sub-features, such that the spatial semantic features are uniformly distributed in each feature group. Specifically, in addition to encoding global information to recalibrate the channel weights in each parallel branch, the output features of two parallel branches are further aggregated through cross-dimensional interactions to capture pixel-level pairwise relationships. In this study, the EMA attention mechanism is added to the ELAN module in backbone. This improvement increases the ability of the model to extract features while being more efficient in terms of the required parameters. The overall structure of the EMA is shown in Fig. 6.

The EMA attention mechanism uses a parallel substructure to avoid performance degradation caused by complex sequential processing and deep convolution to extract pixel-level attention eigenvalues. The EMA aggregates the multiscale spatial structural information and uses the 1x1 convolution, naming it a 1x1 branch; in addition, the 3x3 convolution is placed in parallel with the 1x1 convolution to reduce the response latency.

First, for any given input feature map, $X \in R^{C \times H \times W}$ is used as the input to the EMA module. Then the EMA divides the channel dimension into G sub-features $X = [X_0, X_i, \dots, X_{G-1}], X_i \in R^{C // G \times H \times W}$. $G \ll C$, such that the learned attentional weight descriptors are used to enhance the feature representation of the region of interest in each sub-feature.

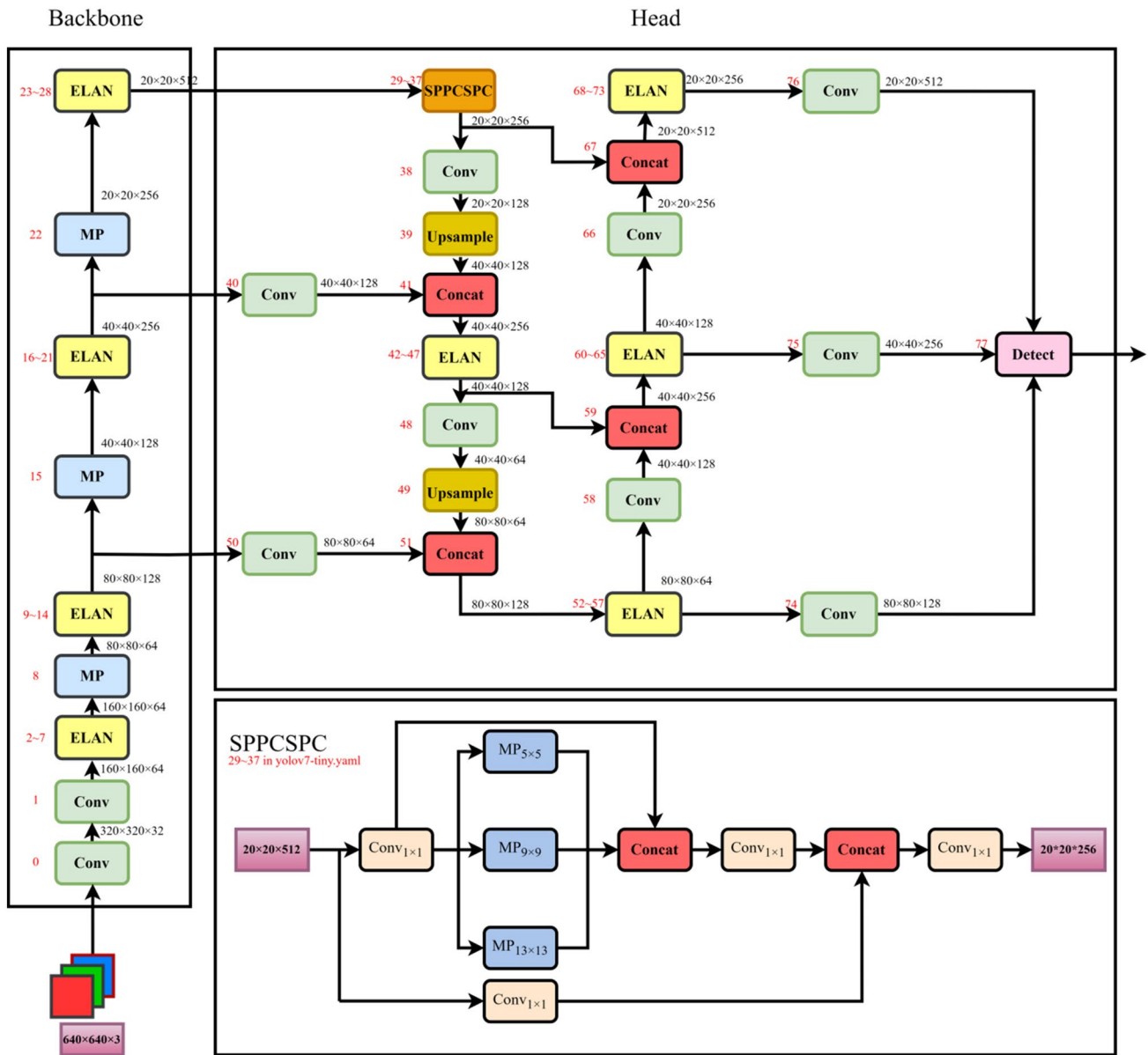


Fig. 4 YOLOv7-tiny network structure

Second, the EMA module proposes the use of three routes to extract the attention weights. In this case, the 1×1 convolution is located in the first two routes, and the 3×3 convolution is located in the third route. To reduce the computational overhead and to obtain the dependencies between all the channels, the EMA models the cross-channel interactive information interactions in both directions of the channels. Specifically, in the 1×1 convolutional branch, a 1D global average pooling operation is added for coding operations across channels in both directions of the channel, while the GN normalization and average pooling operations

are omitted in the 3×3 convolutional branch for extracting multiscale feature representations.

Finally, the EMA module also provides a cross-space information aggregation method in different spatial dimensional directions for richer feature aggregation. The EMA introduces two tensors for the output of the 1×1 branch and the output of the 3×3 branch. The global spatial information output from the 1×1 branch is subsequently encoded via 2D global average pooling, and the channel features at the output of the smallest branch are converted to the corresponding dimensional shape,

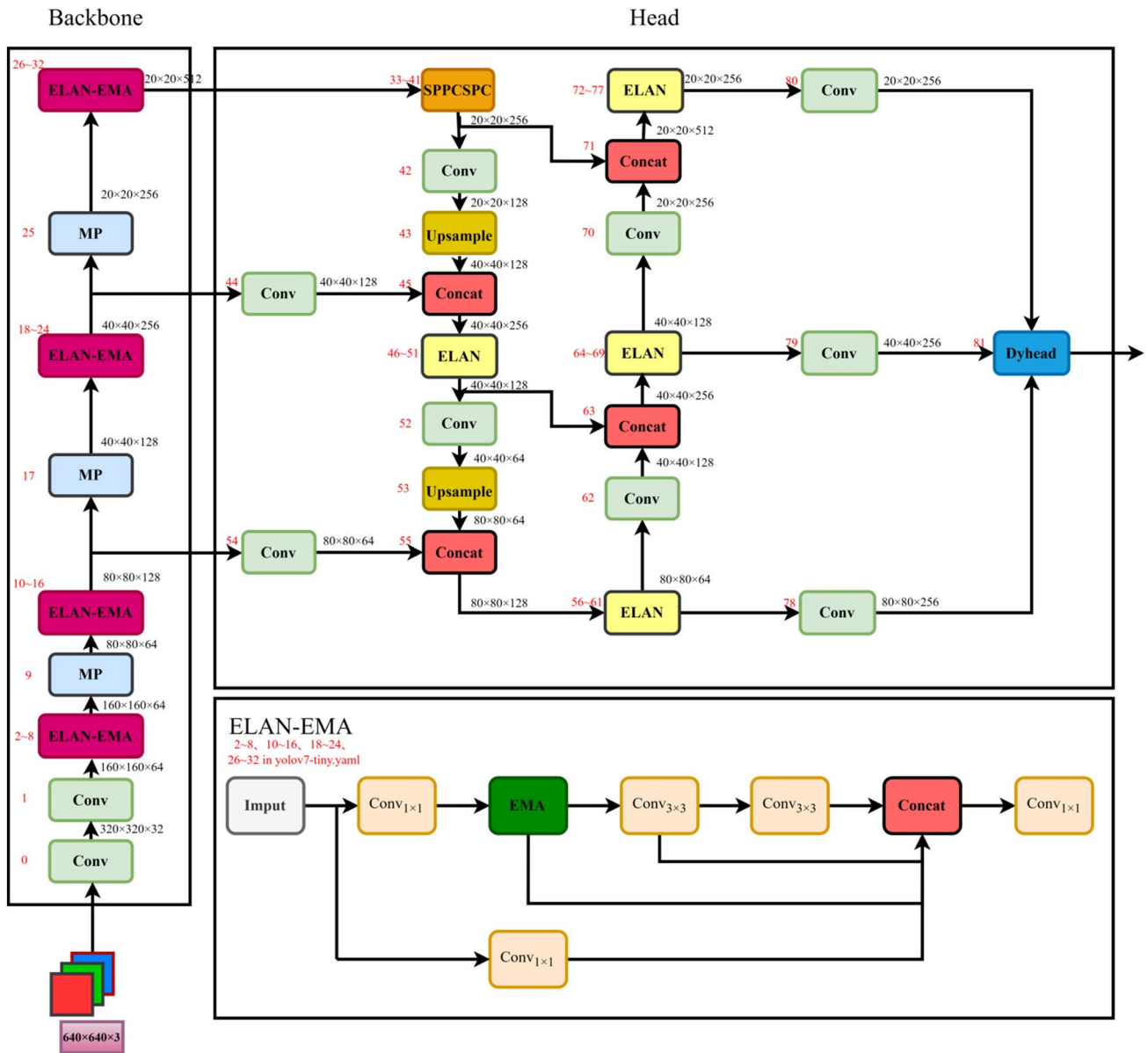


Fig. 5 Improved YOLOv7-tiny network structure

namely, $R_1^{1 \times C // G} \times R_3^{C // G \times HW}$. The formula for the pooling operation is given in Eq. (5):

$$Z_c = \frac{1}{H \times W} \sum_j^H \sum_i^W x_c(i, j), \tag{5}$$

Cross-space learning enlarges the feature space, efficiently extracts the dependencies between the three channels, preserves spatial structural information among the channels, and reduces computational overhead. EMA places a nonlinear SoftMax normalization function at the output of the 2D global average pooling to fit the linear

transformations. Finally, the output features of the three routes are computed as an aggregation of the two spatial attention weight values by a sigmoid activation function highlighting the context pixels of all the pixels, and the final output is of the same dimensional size as the input feature map $X \in R^{C \times H \times W}$.

The cross-spatial information aggregation method proposed in the EMA module models remote dependencies and stores precise location information in the EMA. Fusing contextual information at different scales enables the neural network to produce better pixel-level attention for the feature map. The parallelization of the convolutional kernel is then a more powerful structure for addressing short-term and

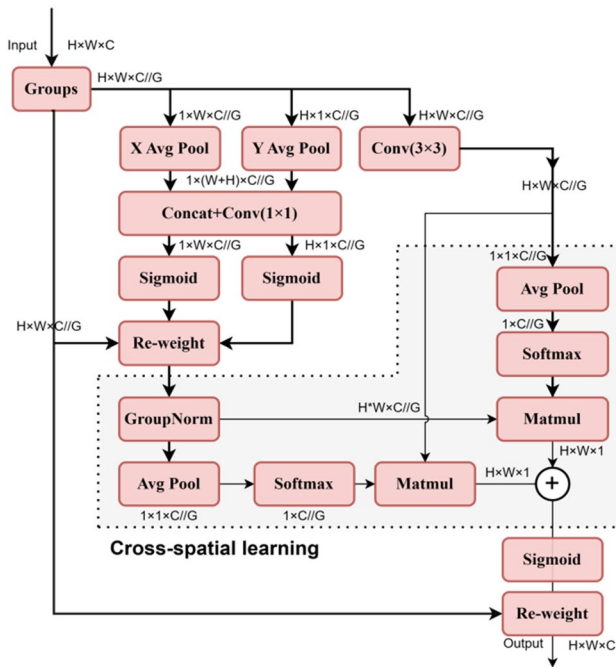


Fig. 6 Structure of the EMA attention mechanism

long-term dependencies using cross-space learning methods. In addition, the parallel use of 3×3 convolution and 1×1 convolution utilizes more contextual information in the intermediate features.

3.2 Improvement of detection head

In YOLOv7-tiny, the output of the backbone network is a three-dimensional tensor with dimensions of level, space, and channel. Therefore, for the model to achieve better scale awareness, spatial semantic feature learning and multitask adaptivity, the IDetect Head in the original structure of YOLOv7-tiny is replaced with a self-attentive dynamic detection head, DyHead, that unifies scale-aware attention, spatial-aware attention, and task-aware attention. As a generic detection head framework, DyHead uses the attention mechanism in each feature dimension of the feature tensor, and both can be applied to single- and two-stage target detection models.

DyHead converts the attention function into three sequential attentions, each of which focuses on only one feature dimension. Given a three-dimensional feature tensor $F \in R^{L \times S \times C}$, the following equation can be used:

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F, \tag{6}$$

where $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ are the different attention functions used for the three different dimensions L , S , and

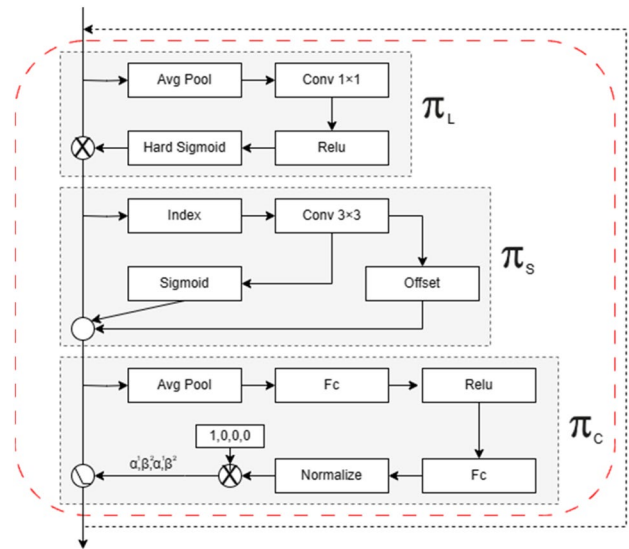


Fig. 7 DyHead structure

C , respectively. The expressions for the attention functions are in order, as follows:

$$\pi_L(F) \cdot F = \sigma \left(f \left(\frac{1}{SC} \sum_{S,C} F \right) \right) \cdot F, \tag{7}$$

$$\pi_S(F) \cdot F = \frac{1}{L} \sum_{l=1}^L \sum_{k=1}^K w_{l,k} \cdot F(l, p_k + \Delta p_k, c) \cdot \Delta m_k, \tag{8}$$

$$\pi_C(F) \cdot F = \max(\alpha^1(F) \cdot F_c + \beta_c(F), \alpha^2(F) \cdot F_c + \beta^2(F)), \tag{9}$$

where $f(\cdot)$ in Eq. (7) is a linear function approximated by a 1×1 convolutional layer, while $\sigma(\cdot)$ is a hard-sigmoid function. K in Eq. (8) is the number of sparsely sampled locations; $p_k + \Delta p_k$ denotes the transformed location where the Δp_k offset is increased by the self-learning space to focus on the discriminative region; and Δm_k is an important metric for self-learning at location p_k . F_c in Eq. (9) is the c -th channel feature slice, and $\max(\cdot)$ is a hyperfunction for global features that are first dimensionalized and then output using the fully connected and normalized layers. Finally, since the above three attentional mechanisms can be applied sequentially, the above three attentional modules can be utilized together multiple times by continuously stacking and combining them. The overall structure of the DyHead is shown in Fig. 7.

3.3 Improvement of the loss function

The loss function is an important part of the target detection model. The model detection performance depends on the design of the loss function, and a good bounding box loss calculation function can significantly improve the performance of the target detection model. The YOLOv7-tiny loss function consists of a localization loss function CIoU-Loss, a classification loss function, and a loss function of the target confidence BEC-Loss.

In the YOLOv7-tiny model, bounding box regression is performed using the CIoU loss function, which accounts for the aspect ratio of the predicted box to the real box during the calculation of the loss value, effectively solving the problem of providing a moving direction for the bounding box in the case of non-overlaps. However, since the CIoU loss function calculates all loss variables as a whole, it may lead to slow convergence and instability, and it also fails to consider the imbalance of difficult and easy samples.

The construction site environment is complex and variable, affecting the quality of the sample through many factors. Safety helmet-wearing samples taken within the construction site often are present in the shade and as dense targets and other difficult-to-classify situations; these images can cause considerable difficulty for helmet detection. Moreover, the CIoU-Loss function is highly sensitive to the positional deviation of the small target helmet, which is close to the distance of the large target helmet; thus, the frame of the small target helmet can be easily located. When large deviations occur, the direct use of the CIoU loss function detection effect is not good. Therefore, in this study, WIoUv3 is selected with a dynamic nonmonotonic focusing mechanism to replace CIoU-Loss as the bounding box loss calculation function of the improved algorithm model.

An imbalance of positive and negative samples is inevitable in the training dataset, which inevitably leads to the appearance of low-quality samples, and the previous loss function increases the penalty for low-quality samples, thus reducing the generalizability of the model. The dynamic nonmonotonic focusing mechanism in WIoUv3 can effectively avoid the negative impact of low-quality samples during the training process by balancing the ratio of high- and low-quality samples and focusing the bounding box regression results on the target object. The regression results focus on the target object, enabling the model to focus on complex samples such as occlusions and enhances the generalization performance for complex samples of target occlusions to overcome the difficulty in detecting signals between helmet samples due to occlusion. A schematic diagram of the WIoU as a whole is shown in Fig. 8.

There are three versions of the WIoU (Wise-IoU), i.e., WIoUv1, WIoUv2, and WIoUv3. WIoUv1 is a two-level attention mechanism constructed on the basis of the distance

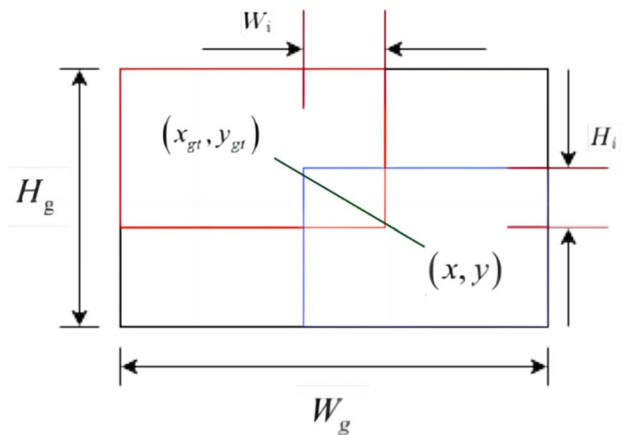


Fig. 8 Schematic diagram of WIoU parameters

metric that addresses the fact that low-quality datasets inevitably negatively affect the model, with the following formula:

$$L_{\text{WIoUv1}} = R_{\text{WIoU}} \mathcal{L}_{\text{IoU}}, \quad (10)$$

$$R_{\text{WIoU}} = \exp \left[\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*} \right], \quad (11)$$

where W_g and H_g are the width and height of the minimum closed area of the prediction box and the real box, respectively; * indicates that the area will be separated from the graph; and the positioning constants prevent the generation of the gradient that hinders convergence and can effectively improve the convergence efficiency.

WIoUv2 draws on the Focal-Loss design method and constructs the monotonic focusing coefficient $Y (Y > 0)$ on the basis of WIoUv1, which effectively reduces the contribution of simple samples to the loss value; then the model can focus on difficult samples and improve the classification performance, with the following formula:

$$\mathcal{L}_{\text{WIoUv2}} = \left(\frac{\mathcal{L}_{\text{IoU}}^*}{\mathcal{L}_{\text{IoU}}} \right)^Y \cdot \mathcal{L}_{\text{WIoUv1}}. \quad (12)$$

WIoUv3 constructs the nonmonotonic focusing coefficient r based on WIoUv1 by means of the outlier β with the following equation:

$$L_{\text{WIoUv3}} = r L_{\text{WIoUv1}}, \quad (13)$$

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}}, \quad (14)$$

where β is the outlier degree and represents the quality of the regression frame, and the hyperparameters α and δ control the mapping between the nonmonotonic focusing coefficient r and the outlier degree β . When the outlier degree of the regression frame satisfies $\beta = C$ (C is a preset value), the regression frame can obtain the highest gradient gain. Moreover, due to the existence of the sliding average, its dynamic update can adjust the quality classification criteria of the regression frames dynamically; this enables WIoUv3 to design the gradient gain allocation strategy that best meets the current situation at any time during training.

4 Experiments and analysis of results

4.1 Dataset

The quality of the safety helmet datasets open-sourced on the Internet are varied, and the efficiency of using all homemade datasets is very low; therefore, after a preliminary screening, we first selected the safety helmet-bearing dataset (SHWD) from GitHub. This dataset contains 7581 images from 9044 helmet-wearing subjects (positive samples) and 111,514 non-helmet-wearing subjects (negative samples). We went through each image in the dataset one by one by visual inspection and eliminated 1104 low-quality or non-conforming pictures. In particular, most of the removed photos were negative samples that were not suitable as helmet inspection images, such as images from classroom surveillance, etc., so we removed these invalid negative samples. Then using cameras and other mobile devices to collect images from offline construction scenes where photography is allowed and combining these images with crawler technology to collect web-related helmet images, we resupplemented the 1104 images such that we could initially develop the homemade helmet dataset. However, additional operations, such as format conversion and annotation, are still needed. On the one hand, the VOC format of the remaining annotated files of the SHWD dataset is batch converted to YOLO format by writing scripts, which is convenient for training. On the other hand, to unify the use of the YOLO dataset, the open-source software LabelImg was used to annotate the 1104 supplemented images according to the two annotation categories of the SHWD dataset: Helmet (wearing a helmet) and Head (not wearing a helmet). After the annotation is completed, the annotation file in the YOLO format is automatically generated. Finally, we first ensure that the images in the public and homemade datasets are evenly represented in the training set, the test set, and the validation set. We randomly divide this dataset into a training set, a test set, and a validation set at a 7:2:1 ratio. The training set contains 5457 images, the test set contains

1517 images, and the validation set contains 607 images. During the partitioning process, we ensured that each subset contained an appropriate proportion of public dataset images and homemade dataset images to maintain data consistency and representation. As a result, a helmet detection dataset is established that consists of a combination of a public dataset and homemade dataset.

4.2 Experimental environment and parameter settings

The operating system used in the experimental environment of this study was Windows 10, the CPU used was a 12th Gen Intel Core i9-12900KF 3.19 GHz, the GPU used was an NVIDIA GeForce RTX 3090, and the running memory used was 42 GB. In addition, the deep learning frameworks Python 3.8, PyTorch 1.11.0 and cuda 11.3 were used for computational acceleration.

To train with better results, the experiments in this study did not use pre-trained weights for migration learning, and the related parameter settings are shown in Table 1.

4.3 Evaluation indicators

Evaluation metrics are important criteria for measuring the performance and effectiveness of the improved algorithms [33]. The metrics used in the experiments of this study include precision (P), recall (R), mean average precision (mAP), and frames per second (FPS). P is the proportion of predicted true positive cases among all the predicted positive cases; recall R is the proportion of predicted true positive cases among all the true positive cases; average precision (AP) is the average of the precision values over the area enclosed between the precision–recall curve (PR curve) and the axes; mAP is the average of the AP values computed on top of the AP values for each detected category; and FPS denotes the number of images that can be detected by the model per second. Moreover, there is a category imbalance problem in our dataset, and we also use the $F1$ score as an evaluation indicator. $F1$ is the reconciled average of the precision and recall rates. The expression of each evaluation index is as follows:

Table 1 Experiment-related parameter settings

Parameter	Value
Lr0	0.01
Lrf	0.1
Momentum	0.937
Weight_decay	0.0005
Batch size	32
Epoch	150
Size	640×640

$$P = \frac{TP}{TP + FP}, \tag{15}$$

$$R = \frac{TP}{TP + FN}, \tag{16}$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP(i), \tag{17}$$

$$FPS = \frac{N}{t}, \tag{18}$$

$$F1 = \frac{2PR}{P + R}, \tag{19}$$

where TP is the number of true positive samples (number of positive samples correctly identified), FP is the number of false positives (number of negative samples misreported), and FN is the number of false negatives (number of positive samples missed). In this paper, the detection target is divided into two categories, so n is 2.

The higher the values of the above indicators are, the better the detection effect. By comprehensively analyzing the above evaluation indices, the performance and effectiveness of the improved YOLOv7-tiny helmet-wearing recognition algorithm for construction workers can be comprehensively assessed.

4.4 Ablation experiment

To verify the detection performance of the improved algorithm in this study and the effectiveness of the improvement, under the premise that each experimental parameter is the same, an ablation experiment is designed, and the impact of each improvement method on the model performance is analyzed. The results of the ablation experiment are shown in Table 2.

Based on the comparison of the ablation experiments in Table 2, in the three groups of single-improvement

experiments from A to C, the precision of experimental group C, in which the replacement loss function is the WIoU, has a very small decrease, but the recall and mAP@0.5 are both improved to some extent. The remaining two single-improvement experiments achieved an improvement of approximately 0.5–1% in precision, recall, and mAP@0.5; these results preliminarily verify the feasibility of each improvement. In the three groups of two-improvement combination experiments from D~F, all the evaluation indices are significantly improved. Among them, the combination of group D experiments has the best effect, and compared with those of the original model, the precision, recall, and mAP@0.5 are improved by 1.1, 1.9, and 2.1%, respectively; these results confirm that the combination of the two attention-boosting structures is very effective. Groups E and F perform their experiments by combining the two improvements on the basis of replacing the WIoU loss function, and the evaluation indices maintain the same magnitude of growth as the results from the group D experiments. This also confirms that the WIoU loss function can more accurately identify the target without increasing the computational cost. Group G is a combination of

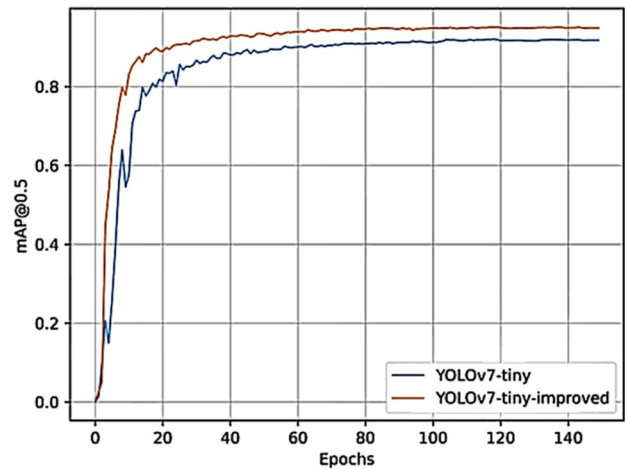


Fig. 9 mAP@0.5 comparison chart

Table 2 Comparison of the ablation experiments

Model	Proposed improvement			Precision/%	Recall/%	mAP0.5/%
	EMA	DyHead	WIoU			
YOLOv7-tiny	×	×	×	92.4	85.0	91.6
A	√	×	×	92.5	85.4	92.1
B	×	√	×	92.6	85.6	92.3
C	×	×	√	92.3	85.6	91.7
D	√	√	×	93.5	86.9	93.7
E	√	×	√	93.2	87.4	93.1
F	×	√	√	93.1	87.6	93.2
G	√	√	√	93.9	90.6	94.9

the three improvements. Figure 9 shows the comparison of the mAP@0.5 before and after the improvement. Overall, the best results are obtained from the combination of the three improvements (group G), and compared with the original YOLOv7-tiny, the improved YOLOv7-tiny models achieved improvements of 1.5, 5.6, and 4.6% in terms of precision, recall, and mAP@0.5, respectively. Figure 10 shows the comparison between the models before and after the improvement in terms of training loss. The improved YOLOv7-tiny model shows enhanced performance across all training losses. Notably, Box_Loss and Obj_Loss are significantly reduced, and fluctuations in the Obj_Loss curve are markedly smaller; these results confirm that the improved YOLOv7-tiny model is more stable and more robust. In general, from the ablation study of the loss function, we preliminarily proved that the improvement of the loss function is effective by comparing the base group and the C group. By analyzing the results of groups E, F, and G, it is proven that the network module combination based on the improved loss function is effective. From the ablation study of network module, we preliminarily proved that the improvement

of network module is effective by comparing the baseline group with group A and group B. Through the analysis of the results of group D and group G, it is proved that the improvement of network module combination is effective.

Through ablation experiments and comparisons of various indices, the improved model not only has better feature perception and extraction abilities but also has better overall stability; in addition, the improved model can adapt effectively to complex environments, such as construction sites, faster and better, further confirming the feasibility of the improved model.

4.5 Comparative experiments

To further verify the superiority of the improved algorithms in this study and to consider the timeliness of the related algorithms, comparison experiments are conducted with the current classical target detection algorithms under the same experimental equipment and dataset, and the selected target detection algorithms include SSD, Faster-RCNN, Cascade-RCNN [34], Libra-RCNN [35], YOLOv3 [36], YOLOv4

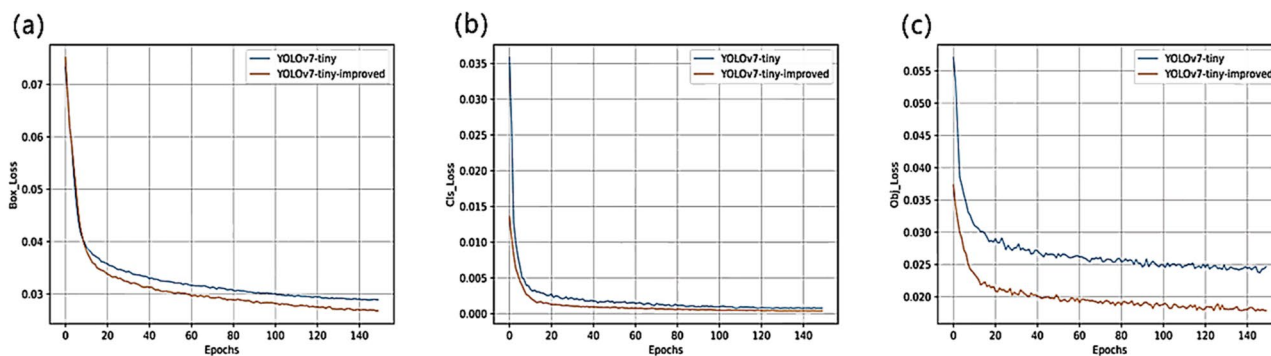


Fig. 10 Comparison of training loss

Table 3 Experimental comparison of the different algorithms

Model	Params/MB	Precision/%	Recall/%	mAP0.5/%	mAP0.5:0.95/%	FPS	F1%
SSD	40.3	86.3	79.5	80.3	/	99.7	82.76
Faster-RCNN	137.1	89.6	83.3	87.4	/	20.3	86.34
Cascade-RCNN	83.8	91.3	86.5	91.4	/	53.2	88.84
Libra-RCNN	68.5	90.9	85.7	91.1	/	64.7	88.22
YOLOv3	61.4	89.4	84.3	89.7	56.7	88.5	86.78
YOLOv4	60.8	90.3	83.4	88.3	57.4	92.4	86.71
YOLOv5s	7.0	89.7	84.5	90.3	58.6	243.3	87.02
YOLOv5l	46.5	92.3	87.4	92.3	58.4	124.4	89.78
YOLOv7-tiny	6.1	92.4	85.0	91.6	58.8	274.3	88.55
YOLOv7	37.2	94.1	88.7	93.8	59.7	126.5	91.32
Reference1 [21]	9.8	90.5	86.9	91.7	58.8	210.3	88.66
Reference2 [22]	10.5	93.7	89.6	94.1	59.8	203.2	91.60
Ours	6.2	93.9	90.6	94.9	60.0	263.1	92.22

[37], YOLOv5s, YOLOv5l, YOLOv7, the improved models mentioned in Reference 1 and the improved models mentioned in Reference 2. The comparative experimental results are shown in Table 3.

As shown in Table 3, the improved YOLOv7-tiny model not only retains the lightweight features of the original benchmark model, such as a small parameter count and fast detection speed, but also greatly improves the precision, recall, $F1$ score, and mapping metrics. First, the improved YOLOv7-tiny model substantially outperforms the more classical target detection algorithms, such as SSD, Faster-RCNN, Cascade-RCNN, and Libra-RCNN in terms of volume, precision, recall, mapping, and detection speed. On the other hand, when YOLOv3 and YOLOv4 are compared, in addition to the 5.2 and 6.6% improvements in the map (mAP@0.5), respectively, the number of parameters, FPS and $F1$ score also greatly increased. Finally, in comparison to experiments with several current mainstream target detection models, our improved model ensures faster detection speed and a smaller number of parameters as the map improves. At the same time, compared with some of the mainstream algorithms used to improve the model, the model proposed in this paper also has a certain degree of leadership in the evaluation indicators. Compared with YOLOv7, the improved model proposed in this study has insufficient precision; however, the number of parameters of YOLOv7 is much greater than that of the model proposed in this study, and the detection speed and $F1$ score are much lower than those of the improved YOLOv7-tiny model. Based on the comparative experimental results of this study, the improved YOLOv7-tiny model outperforms the original

YOLOv7-tiny model and other mainstream target detection algorithms in terms of the number of parameters, precision, recall, $F1$ score, mapping, and detection speed, and the advantages of the improved method are confirmed.

4.6 Proof of results

To further validate and more intuitively illustrate the effectiveness of the improved algorithm proposed in this study, several scenarios involving dense occlusions, small targets, and other characteristics were selected for comparison experiments. The experimental results show that our improved model exhibits significant advantages in challenging scenarios in the presence of occlusions, shadows, and small targets at long distances.

As shown in Fig. 11a, in the presence of an occlusion, the performance of the original model is not very good, and false detection occurs; in contrast, by replacing the WIoU loss function to better handle objects of different sizes and occlusions, our improved model can identify and locate occluded objects more accurately. As shown in Fig. 11b, when the environment is affected by shadows, the original model still suffers from misdetection. By introducing the EMA attention mechanism and replacing the original detector head with the DyHead detector head, the feature representation capability of the model is greatly enhanced, and the improved model is able to accurately identify all the helmets. As shown in Fig. 11c, for small targets at long distances, the original model results in missed detections, while the introduction of the EMA attention mechanism and the DyHead detection head gives the model a greater

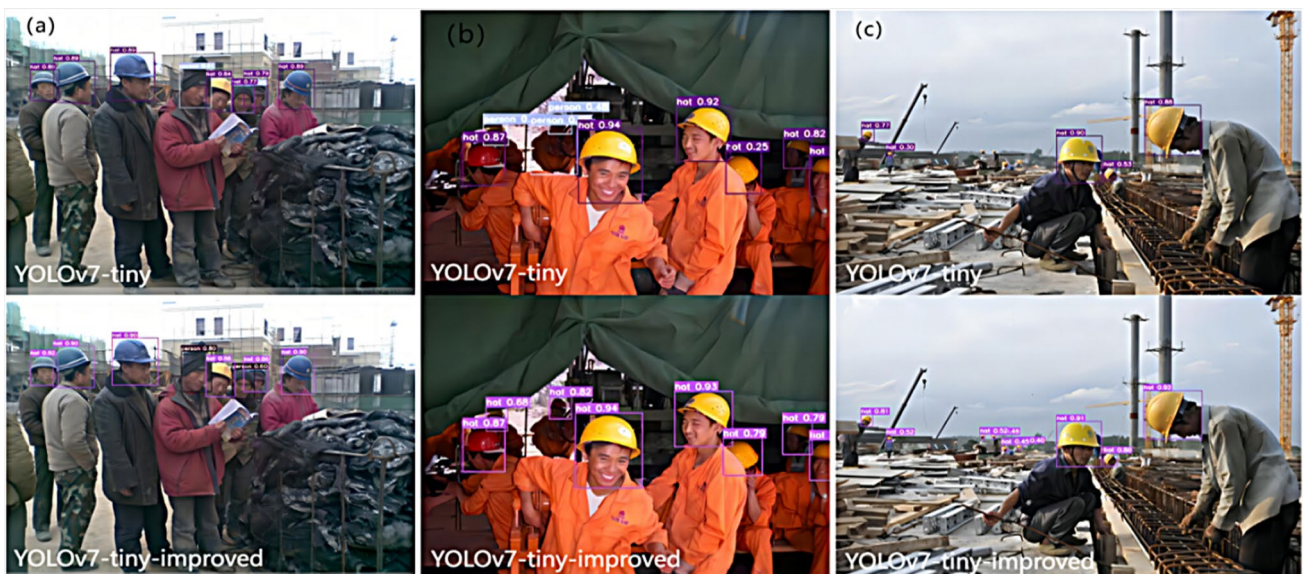


Fig. 11 Comparison of detection results in different scenes before and after algorithm improvement. **a** Occlusion environment, **b** shadow environment, **c** small target environment

ability to focus on areas and better focus on small targets. Overall, the improved model is able to adapt to the complex environments and the examination of long-distance small targets; in addition, it has a higher confidence level than the original model in terms of the targets that can be successfully recognized.

In summary, the original YOLOv7-tiny model poorly performs in the detection of scenes possessing features such as complex environments and small targets at long distances; however, the improved algorithm proposed in this study is able to maximally enhance the focusing on the featured region due to the introduction of the EMA attention mechanism and the self-attentive dynamic detection head DyHead, which can better complete the detection of targets in complex environments. Furthermore, the introduction of the WIoU loss function increases the accuracy of object detection under occlusion.

5 Summary

Aiming at current construction-site helmet-wearing detection algorithms that have problems such as low accuracy, poor real-time performance, and high influence from the environment, in this study, an improved YOLOv7-tiny algorithm is proposed. We first compared the mainstream target detection algorithms, and the YOLOv7-tiny algorithm model is selected as the benchmark model; the YOLOv7-tiny algorithm model initially shows excellent performance for helmet detection. Then by analyzing the problems of the YOLOv7-tiny algorithm model in actual detection, the model is improved accordingly. First, the EMA attention mechanism is introduced to dynamically adjust the weights of the different feature maps in both the spatial and channel dimensions, which increases the accuracy of the feature expression. Moreover, the detection head is replaced with a DyHead detection head, and the number of channels in the feature maps is dynamically adjusted in the channel dimension; these further enhance the channel feature expression capability. Finally, the WIoU loss function better adapts to different target sizes, can identify occluded targets, and yields better robustness. Through numerous experimental comparisons, while maintaining the advantages of small volume and fast detection speed of the original YOLOv7-tiny algorithm, the recognition accuracy and model stability are improved.

Despite the promising results of the proposed method, it also has some limitations. The improved algorithm in this paper is validated on specific datasets, and its performance may vary in different environments or on different distributed datasets. Therefore, the generality and generalization ability of the algorithm need to be further verified. In addition, in the future, we will further study the model compression techniques, such as pruning, quantization, etc.,

to further reduce the model size and improve the running speed while maintaining the performance.

Author contributions S.W. reviewed and edited the manuscript, P.W. conceptualized and wrote the main manuscript text. Q.W. checked the data. All authors reviewed the manuscript.

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

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Mo, S.: Improvement of construction worker's safety awareness and self-protection awareness. *Urban Constr. Theory Res.* 36, 130–131 (2011). <https://d.wanfangdata.com.cn/periodical/csjslyj201136328>
2. Dhillon, A., Verma, G.K.: Convolutional neural network: a review of models, methodologies and applications to object detection. *Prog. Artif. Intell.* 9(2), 85–112 (2020). <https://doi.org/10.1007/s13748-019-00203-0>
3. Zhang, Wu.: Niu: Summary of application research on helmet detection algorithm based on deep learning. *Comput. Eng. Appl.* 58(16), 1–17 (2022). <https://doi.org/10.3778/j.issn.1002-8331.2203-0580>
4. Ngai, E.W.T., Moon, K.K.L., Riggins, F.J., Candace, Y.Y.: RFID research: an academic literature review (1995–2005) and future research directions. *Int. J. Prod. Econ.* 112(2), 510–520 (2008). <https://doi.org/10.1016/j.ijpe.2007.05.004>
5. Sharma, K., Dhir, N.: A study of wireless networks: WLANs, WPANs, WMANs, and WWANs with comparison. *Int. J. Comput. Sci. Inform. Technol.* 5(6), 7810–7813 (2014). <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ff3e8a75932416553f16adf113245c1842a0f09b>
6. Kelm, A., Laußat, L., Meins-Becker, A., Platz, D., Khazaei, M.J., Costin, A.M., Helmus, M., et al.: Mobile passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites. *Autom. Constr.* 36, 38–52 (2013). <https://doi.org/10.1016/j.autcon.2013.08.009>
7. Li, H., Yang, X., Wang, F., Rose, T., Chan, G., Dong, S.: Stochastic state sequence model to predict construction site safety states through real-time location systems. *Saf. Sci.* 84, 78–87 (2016). <https://doi.org/10.1016/j.ssci.2015.11.025>
8. Zhang, T., Cheng, J.: The site management system with intelligent safety cap. *Internet Things Technol.* 4(1), 89–91 (2014). <https://doi.org/10.3969/j.issn.2095-1302.2014.01.040>
9. Liu, X., Ye, X.: Skin color detection and hu moments in helmet recognition research. *J. East China Univ. Sci. Technol.* 3, 365–370 (2014). <https://doi.org/10.3969/j.issn.1006-3080.2014.03.016>
10. Li, J., Liu, H., Wang, T., Jiang, M., Wang, S., Li, K., Zhao, X.: Safety helmet wearing detection based on image processing and machine learning. In 2017 ninth International Conference on advanced computational intelligence (ICACI), pp. 201–205. IEEE, Doha, Qatar (2017). <https://doi.org/10.1109/ICACI.2017.7974509>
11. Yogameena, B., Menaka, K., Saravana Perumaal, S.: Deep learning-based helmet wear analysis of a motorcycle rider for

- intelligent surveillance system. *IET Intell. Transp. Syst.* **13**(7), 1190–1198 (2019). <https://doi.org/10.1049/iet-its.2018.5241>
12. Ferdous, M., Ahsan, S.M.M.: Multi-scale safety hardhat wearing detection using deep learning: a top-down and bottom-up module. In 2021 International Conference on electrical, communication, and computer engineering (ICECCE), pp. 1–6. IEEE, Kuala Lumpur, Malaysia (2021). <https://doi.org/10.1109/ICECCE52056.2021.9514144>
 13. Wu, D., Wang, H., Li, J.: Safety helmet detection and identification based on improved faster RCNN. *Inform. Technol. Inform.* **1**, 17–20 (2020). <https://doi.org/10.3969/j.issn.1672-9528.2020.01.003>
 14. Xu, X., Zhao, W., Zou, H., Zhang, L., Pan, Z.: Detection algorithm of safety helmet wear based on MobileNet-SSD. *Comput. Eng.* **47**(10), 298–305 (2021). <https://doi.org/10.19678/j.issn.1000-3428.0058733>
 15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: single shot multibox detector. In computer vision–ECCV 2016: 14th European Conference, pp. 21–37. Springer, Amsterdam, The Netherlands (2016). https://doi.org/10.1007/978-3-319-46448-0_2
 16. 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 (CVPR), pp. 779–788. IEEE, Las Vegas, Nevada (2016). https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf
 17. Li, M., Han, Q., Zhang, T., Wang, D.: Safety helmet detection method of improved SSD. *J. Comput. Eng. Appl.* **57**(8), 192–197 (2021). <https://doi.org/10.3778/j.issn.1002-8331.2008-0155>
 18. Geng, R., Ma, Y., Huang, W.: An improved helmet detection method for YOLOv3 on an unbalanced dataset. In 2021 3rd International Conference on advances in computer technology, information science and communication (CTISC), pp. 328–332. IEEE, Shanghai, China (2021). <https://doi.org/10.1109/CTISC52352.2021.00066>
 19. Xiao, T., Cai, L., Gao, X., Huang, H., Zhang, C.: Improved YOLOv3 helmet wearing detection method. *J. Comput. Eng. Appl.* **57**(12), 216–223 (2021). <https://doi.org/10.3778/j.issn.1002-8331.2009-0175>
 20. Shen, J., Xiong, X., Li, Y., He, W., Li, P., Zheng, X.: Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning. *Comput. Aided Civil Infrastr. Eng.* **36**(2), 180–196 (2021). <https://doi.org/10.1111/mice.12579>
 21. Wang, L., Cao, Y., Wang, S., Song, X., Zhang, S., Zhang, J., Niu, J.: Investigation into recognition algorithm of helmet violation based on YOLOv5-CBAM-DCN. *IEEE Access* **10**, 60622–60632 (2022). <https://doi.org/10.1109/ACCESS.2022.3180796>
 22. Zhao, L., Tohti, T., Hamdulla, A.: BDC-YOLOv5: a helmet detection model employs improved YOLOv5. *Signal Image Video Process.* **17**, 4435–4445 (2023). <https://doi.org/10.1007/s11760-023-02677-x>
 23. Gupta, C., Gill, N.S., Gulia, P., Yadav, S., Chatterjee, J.M.: A novel finetuned YOLOv8 model for real-time underwater trash detection. *J. Real Time Image Process.* **21**, 48 (2024). <https://doi.org/10.1007/s11554-024-01439-3>
 24. 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. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 7464–7475. IEEE, Seattle WA, USA (2023). https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.pdf
 25. Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z.: Efficient multi-scale attention module with cross-spatial learning. In ICASSP 2023–2023 IEEE International Conference on acoustics, speech and signal processing (ICASSP), pp. 1–5. IEEE, Rhodes Island, Greece (2023). <https://doi.org/10.1109/ICASSP49357.2023.10096516>
 26. Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., Zhang, L.: Dynamic head: unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 7373–7382. IEEE, Virtual (2021). https://openaccess.thecvf.com/content/CVPR2021/papers/Dai_Dynamic_Head_Unifying_Object_Detection_Heads_With_Attentions_CVPR_2021_paper.pdf
 27. Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-IoU: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051* (2023). <https://doi.org/10.48550/arXiv.2301.10051>
 28. 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**, 1137–1149 (2017). <https://doi.org/10.1109/TPAMI.2016.2577031>
 29. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp. 7132–7141. IEEE, Salt Lake City, UT (2018). <https://doi.org/10.1109/CVPR.2018.00745>
 30. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 13708–13717. IEEE, Nashville, TN, USA (2021). <https://doi.org/10.1109/CVPR46437.2021.01350>
 31. Li, X., Hu, X., Yang, J.: Spatial group-wise enhance: improving semantic feature learning in convolutional networks. <http://arxiv.org/abs/1905.09646> (2019)
 32. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In Proceedings of the European Conference on computer vision (ECCV), pp. 3–19. IEEE, Munich, Germany (2018). https://openaccess.thecvf.com/content_ECCV_2018/papers/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.pdf
 33. Peng, G., Nourani, M., Harvey, J., Dave, H.: Personalized EEG feature selection for low-complexity seizure monitoring. *Int. J. Neural Syst.* **31**, 2150018 (2021). <https://doi.org/10.1142/S0129065721500180>
 34. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp. 6154–6162. IEEE, Salt Lake City, UT (2018). <https://doi.org/10.1109/CVPR.2018.00644>
 35. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: towards balanced learning for object detection. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 821–830. IEEE, Long Beach, CA, USA (2019). <https://doi.org/10.1109/CVPR.2019.00091>
 36. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. <http://arxiv.org/abs/1804.02767> (2018)
 37. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection. <http://arxiv.org/abs/2004.10934> (2020)

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.