TYPE Original Research
PUBLISHED 10 November 2023
DOI 10.3389/fbuil.2023.1288445

Check for updates

OPEN ACCESS

EDITED BY
Andrew Agapiou,
University of Strathclyde, United Kingdom

REVIEWED BY
ZhiQiang Chen,
University of Missouri–Kansas City,
United States
Yang Yu,
University of New South Wales, Australia
Aibin Chen,
Central South University Forestry and
Technology, China

*CORRESPONDENCE
Jiaqi Ren,
✉ 210511051@sust.edu.cn

RECEIVED 04 September 2023
ACCEPTED 25 October 2023
PUBLISHED 10 November 2023

CITATION
Hu L and Ren J (2023), YOLO-LHD: an enhanced lightweight approach for helmet wearing detection in industrial environments.
Front. Built Environ. 9:1288445.
doi: 10.3389/fbuil.2023.1288445

YOLO-LHD: an enhanced lightweight approach for helmet wearing detection in industrial environments

Lianhua Hu and Jiaqi Ren*

College of Mechanical and Electrical Engineering, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China

Establishing a lightweight yet high-precision object detection algorithm is paramount for accurately assessing workers' helmet-wearing status in intricate industrial settings. Helmet detection is inherently challenging due to factors like the diminutive target size, intricate backgrounds, and the need to strike a balance between model compactness and detection accuracy. In this paper, we propose YOLO-LHD (You Only Look Once-Lightweight Helmet Detection), an efficient framework built upon the YOLOv8 object detection model. The proposed approach enhances the model's ability to detect small targets in complex scenes by incorporating the Coordinate attention mechanism and Focal loss function, which introduce high-resolution features and large-scale detection heads. Additionally, we integrate the improved Ghostv2 module into the backbone feature extraction network to further improve the balance between model accuracy and size. We evaluated our method on MHWD dataset established in this study and compared it with the baseline model YOLOv8n. The proposed YOLO-LHD model achieved a reduction of 66.1% in model size while attaining the best 94.3% mAP50 with only 0.86M parameters. This demonstrates the effectiveness of the proposed approach in achieving lightweight deployment and high-precision helmet detection.

KEYWORDS

safety helmet wearing detection, lightweight, high-resolution, Ghostnet, attention mechanism, loss function

1 Introduction

Industries with high accident rates, including construction, electrical power infrastructure, and coal mining, frequently face substantial casualties and economic losses. Ensuring workers' safety is, therefore, increasingly important. The construction industry's fatal accident analysis reveals that around 13.9% of fatal accidents occur due to individuals being struck by objects Shao et al. (2019). Wearing safety helmets is a mandatory preventive measure for anyone entering construction sites. Presently, methods for ensuring this compliance are predominantly based on safety training, manual on-site inspections, and monitoring surveillance videos. These methods are often limited by high costs and low efficiency. Consequently, intelligent real-time detection of safety helmet usage presents significant research value for enhancing worker safety.

In recent years, computer vision-based safety helmet detection methods have become increasingly prominent in construction sites, substantially improving worker safety Zhou et al. (2021); Li et al. (2022); Lee et al. (2023). However, traditional vision-based methods rely

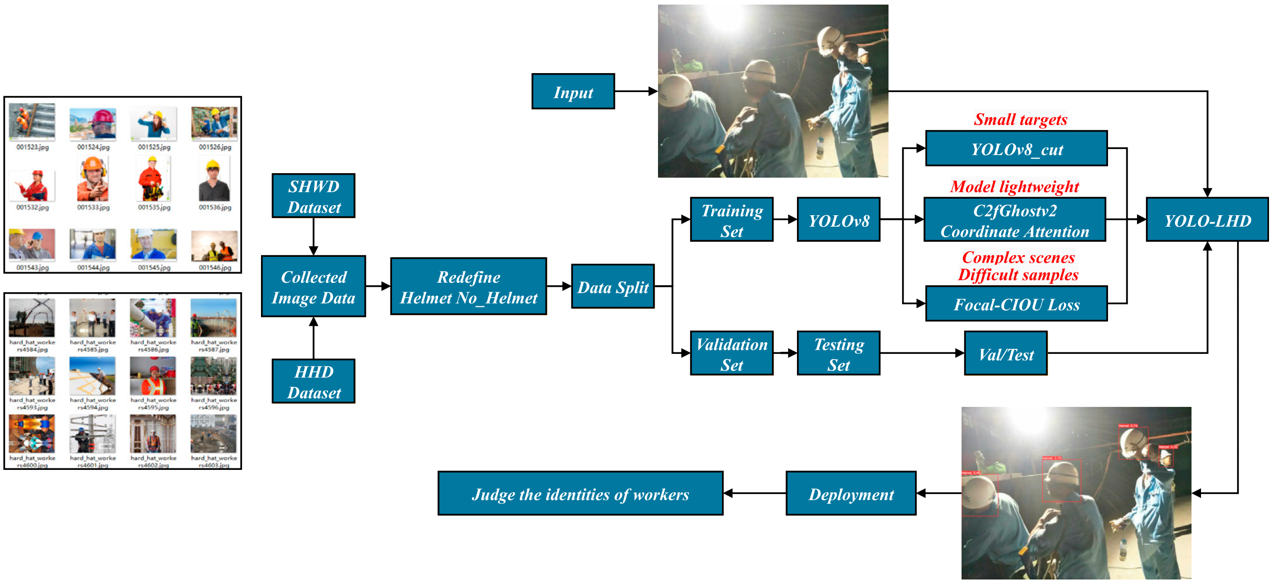

FIGURE 1
Overall flowchart of YOLO-LHD framework.

on manually designed feature extraction, such as the Histogram of Oriented Gradient (HOG) Cao et al. (2011); Wang et al. (2009), Hough transform Che-Yen Wen et al. (2003), and background subtraction algorithm Dar-Shyang Lee (2005). These extracted features are then classified to identify specific target categories using classifiers like Support Vector Machine (SVM) Shrestha et al. (2015) and AdaBoost Wu and Nagahashi (2014).

Traditional visual methods, which rely on manually designed features, are often limited by their low robustness and accuracy Ott and Everingham (2009). The advent of deep learning-based object detection methods has led to their application in industrial environments Yu et al., 2022, Yu et al., 2023, providing superior results compared to traditional machine learning methods. This includes methods such as SSD Fang et al. (2018a), Faster R-CNN Fang et al. (2018b), YOLO Wu et al. (2019). However, factors such as lighting conditions, complex backgrounds, occlusions, and small target sizes often result in issues like false and missed detections Yan et al. (2020).

Deep learning-based object detection models inherently feature complex networks, high computational resource costs, large parameter counts, and substantial model sizes. As a result, their deployment on embedded devices is challenging due to the extensive parameter and computation requirements Deng et al. (2022). To address these issues, simplified network versions, including “small” and “tiny” versions of YOLO Redmon et al. (2016) and SSD Liu et al. (2016), have been proposed. To lighten the model, lightweight backbone networks like MobileNet Kurdthongmee (2020) and ShuffleNet Cui et al. (2023) are employed. While these measures contribute to model lightening, they often result in reduced accuracy. Consequently, finding an equilibrium between model size and accuracy has emerged as a pressing issue in contemporary research.

In this study, we categorize a worker’s condition into two states: “Helmet” and “No_Helmet”, representing whether the worker is wearing a safety helmet or not. We aim to address the issues present in the existing algorithms, such as a large number of network parameters, low detection performance, and difficulties in application deployment, by proposing a lightweight detection algorithm, YOLO-LHD, as shown in Figure 1. This algorithm is intended for helmet detection at construction sites, leveraging the improved YOLOv8n object detection architecture. The main contributions of this paper can be outlined as follows:

1) We proposed a lightweight helmet detection method with superior average precision and minimal model size, enhancing its suitability for deployment on mobile or embedded devices. Specifically, we optimized the YOLOv8n network architecture by reducing the downsampling rate of the backbone network and eliminating deep structures detrimental to the detection of small targets. Based on high-resolution information, we developed a new lightweight backbone, YOLOv8_cut, which significantly reduced the parameter count. In the feature fusion network, we introduced high-resolution features containing shallow layer information and added a large-scale detection head, thereby enhancing the model’s capability to detect small targets.

2) To address the computational burden posed by high-resolution information and enhance the model’s feature extraction capability, we employed the improved GhostNetv2 Bottleneck to replace the existing Bottleneck. This change streamlined model complexity and further minimized its size.

3) To offset the accuracy loss induced by the lightweight backbone, we integrated a downsampling convolutional module embedded with CA (Coordinate Attention) into the feature fusion network. Additionally, by incorporating Focal-CIOU Loss, we enhanced the model’s performance and its detection capability for challenging samples.

4) We conducted experiments on the MHWD (Merged Helmet Wearing Detection dataset) dataset and implemented different module improvement measures, thereby validating the effectiveness of our

improved method. Comparative experiments with mainstream advanced algorithms demonstrated that YOLO-LHD excels in terms of model accuracy and detection efficacy.

The remainder of this paper is organized as follows. Section 2 provides a review of pertinent research in the realms of lightweight object detection and safety helmet detection. Section 3 elaborates on our proposed safety helmet detection model, detailing the network architecture, the enhanced modules, and the Loss function. Section 4 presents the experimental results, comparing our methodology to other state-of-the-art detection algorithms, and examines the factors influencing the performance of the detection model. Lastly, Section 5 recaps the contributions of this study.

2 Related work

To deliver a comprehensive understanding of the safety helmet algorithm proposed herein, this section offers a thorough review of existing algorithms for lightweight object detection and safety helmet detection.

2.1 Lightweight object detection method

In convolutional neural network-based deep learning algorithms, developing an adequately deep network can enhance its performance. However, this introduces significant computational costs. MobileNet Howard et al. (2017) curtails computational overhead via depth-wise separable convolutions. MobileNetv2 Sandler et al. (2018) introduces inverted residual blocks (IRBs) to mitigate information loss from non-linear transformations. MobileNetv3 Howard et al. (2019) further refines and simplifies the network structure while integrating the SE attention mechanism, thus improving accuracy and reducing latency. ShuffleNet Zhang et al. (2018) organizes the channels of feature maps and conducts channel shuffling through depth-wise separable convolutions. ShuffleNetv2 Ma et al. (2018) adds a 1 × 1 convolution layer before global average pooling to mix features. FasterNet Chen et al. (2023) proposes Partial Conv, applying regular convolutions to a portion of input channels while leaving the remaining channels intact. With a partial ratio = 1/4, the floating-point operations (FLOPs) are only 1/16 of regular convolutions, effectively diminishing memory access while preserving feature extraction efficacy. GhostNet Han et al. (2020) generates features Y via 1 × 1 convolutions, applies depth-wise separable convolutions on Y to generate features, and eventually concatenates these two features. GhostNetv2 Tang et al. (2022) introduces a novel DFC attention mechanism on the basis of Ghostnetv1 to improve the feature extraction ability of the model for spatial information.

Recently, Vision Transformer (ViT)-based models such as MobileFormer Chen et al. (2022), EfficientFormer Li et al. (2022c) and EfficientViT Liu et al. (2023) have gained traction in the field. Despite their unique advantages, it is crucial to highlight that these ViTs do not always surpass traditional CNNs when considering metrics like parameter count and computational complexity. Given the swift advancements in lightweight Transformer networks, there's a sustained effort from researchers to enhance both their efficiency and performance metrics. With these considerations in mind, our decision to enhance the YOLOv8 backbone with the improved C2fGhostv2 was made after careful evaluation, focusing on both the lightweight of the architecture and its detection efficacy.

2.2 Safety helmet wearing detection

Safety helmet detection refers to the process of employing computer vision techniques to ascertain whether individuals in areas like construction sites are wearing safety helmets. This task is pivotal for ensuring the safety of construction workers during their duties. In recent years, with advancements in deep learning and computer vision technologies, safety helmet detection based on deep learning has gained increasing popularity.

Currently, safety helmet detection primarily involves two methods: traditional image processing techniques and deep learning-based approaches. Traditional image processing methods utilize techniques such as color segmentation and morphological operations for safety helmet recognition. Che-Yen Wen et al. (2003) enhanced the Hough transform for safety helmet detection. Li et al. (2018) used head positioning, HSV transformation, adaptive thresholding, and other image processing techniques to detect safety helmets. Nevertheless, these methods require manual setting of specific thresholds and rules and impose stringent requirements for lighting, background, and adaptability to complex scenes.

Deep learning-based methods are widely used for safety helmet detection. They involve building deep neural networks that input images for feature extraction, target localization and classification, and output conclusions about whether safety helmets are being worn. Popular deep learning methods include SSD Liu et al. (2016), Faster R-CNN Ren et al. (2015), YOLO Redmon et al. (2016); Redmon and Farhadi 2017, Redmon and Farhadi 2018, among others. These methods, based on Convolutional Neural Networks (CNN) and object detection algorithms, allow for precise and rapid detection of safety helmets, with relatively lower requirements for image lighting and background. Wu et al. (2019) improved the YOLOv3 algorithm for effective safety helmet detection. Zhou et al. (2021) used the YOLOv5 model for safety helmet detection, demonstrating the effectiveness of helmet detection based on YOLOv5. Tai et al. (2023) improved the YOLOv5 algorithm using attention mechanisms and dynamic anchor boxes, enhancing the detection accuracy and speed for occluded targets.

Building upon the aforementioned studies, this paper proposes a lightweight safety helmet detection method based on an improved YOLOv8 to address the shortcomings of existing detection algorithms, such as large model size, low detection accuracy, and high false positive rates.

3 Methods

In this section, we provide a comprehensive description of both the baseline YOLOv8 model and our improved YOLO-LHD model.

3.1 YOLOv8 network model

You Only Look Once (YOLO) is the first one-stage object detector proposed by Redmon et al. (2016) at CVPR 2016.

YOLO divides the input image into a predefined number of grids and predicts a certain number of bounding boxes and their corresponding class probabilities for each grid. Each bounding box outputs four coordinate values and an objectness score, exhibiting high real-time performance and strong background discrimination capability. Subsequent versions of YOLO Redmon and Farhadi 2017, Redmon and Farhadi 2018; Bochkovskiy et al., 2020; Jocher 2020; Li C. et al., 2022; Wang et al., 2023 have continuously improved both detection accuracy and speed. Among them, YOLOv8, the latest version of the YOLO series object detection algorithm proposed by Jocher et al. (2023) in 2023, is a state-of-the-art, single-stage object detection algorithm. Building upon previous versions of YOLO, YOLOv8 further enhances performance by incorporating a lightweight network architecture.

The model architecture of YOLOv8 consists of three main structures: Backbone, Neck, and Head. The YOLOv8 algorithm includes different scaling factors, namely, N/S/M/L/X versions, with decreasing speed and increasing accuracy. The backbone network and the Neck component of YOLOv8 have undergone a redesign of the C3 module. The new C2f module connects multiple feature maps and leverages convolutional operations to fuse them, resulting in a richer gradient flow. The C2f module efficiently maintains computational efficiency while enhancing the expressive power of feature representation, thereby contributing to improved network performance.

For the Neck component, YOLOv8 employs the widely used PANet structure found in the YOLO series for feature fusion. PANet combines bottom-up and top-down approaches, effectively extracting multi-scale information. This facilitates the fusion of information across different feature layers, enabling the detector to better adapt to objects of varying sizes and shapes and improving localization accuracy. Moreover, YOLOv8 eliminates the 1 × 1 convolution before upsampling in the Neck component, reducing computational overhead.

In the Head component, YOLOv8 separates the classification and detection heads and removes the objectness branch. The model transitions from an Anchor-based approach to an Anchor-Free one, permitting dense prediction directly on the entire image without the necessity for predefined candidate boxes. It also enables the model to adapt to objects of different sizes and background conditions. YOLOv8 employs BCE Loss for Classification Loss, while the regression loss includes Distribution Focal Loss (DFL Loss) and CIOU Loss.

As the most recent version of the YOLO model, YOLOv8 presents the best performance on target detection tasks. However, its size remains relatively large for lightweight devices. In this paper, we implement corresponding improvements based on the YOLOv8n model to enhance safety helmet detection performance and create a lightweight network model.

3.2 The proposed YOLO-LHD algorithm

Our proposed YOLO-LHD safety helmet detection algorithm's overall architecture and implementation details are shown in Figure 2. The main components involved include the backbone network constructed with the C2fGhostv2 module, the network structure that reduces downsampling ratio and introduces high-resolution features, and the Regression loss that incorporates Focal Loss.

The improved C2fGhostv2 serves as the backbone of YOLO-LHD to extract features and outputs feature maps with resolutions of 160 × 160, 80 × 80, and 40 × 40 to the feature fusion network. Compared with the original YOLOv8 architecture, YOLO-LHD removes the feature map with a resolution of 20 × 20 and introduces higher-resolution features. After fusion with high-level features and combined with the CA attention mechanism and Focal-CIOU Loss, it further enhances the model's detection performance for small-sized targets. The specific details of these three components will be introduced in the following sections.

3.2.1 Optimize the network structure of YOLOv8n

In the feature extraction network of YOLO, the resolution of the feature maps decreases from the lower layers to the higher layers, while the depth increases. The feature maps output by the low-level modules usually have larger spatial dimensions and contain complex and more detailed features. In the context of small object detection, as the depth of the network increases, the feature information of small objects can be lost due to pixel aggregation, and a deep network structure is not conducive to small object detection. Based on this, we have optimized the network architecture of YOLOv8 in this paper.

We use a shallower network for feature extraction, by eliminating the deepest convolution and C2f module from the standard YOLOv8 backbone. Our approach introduces a low-level feature map with a resolution of 160 × 160 into the feature fusion stage, which shares the feature information of small targets more effectively. Subsequently, We add a detection head with a resolution of 160 × 160, removing the original 20 × 20 detection head. This inclusion of lower-level information on small targets effectively diminishes the false negative rate for small target detection.

The high-resolution feature fusion neck network expands the resolution of the feature map from 40 × 40 to 160 × 160 through two linear interpolations, extending the network width and feature resolution, generating a feature map that contains richer spatial position information and semantic information and is more suitable for detecting small objects. According to the aspect ratio distribution of labels in the dataset shown in Figure 3, we find that the size ratio of the targets is mainly distributed within 20%, most of which are medium and small targets. Through the aforementioned enhancement measures, we ultimately arrive at a novel model architecture, YOLOv8n_cut, as shown in Figure 4.

The reconstructed YOLOv8_cut, through targeted improvements to the backbone network, feature enhancement network, and detection head, reduces the number of parameters of the network from the original 3.1M–0.96M, thus preserving more feature detail information while facilitating model deployment. Eventually, compared to the baseline network, it retains more detail features and semantic features, improving the model's detection accuracy.

In the feature fusion network, this paper introduces the CA attention mechanism Hou et al. (2021) into the downsampling convolution to compensate for the feature extraction loss caused by the reduction in the number of downsampling times. Drawing on

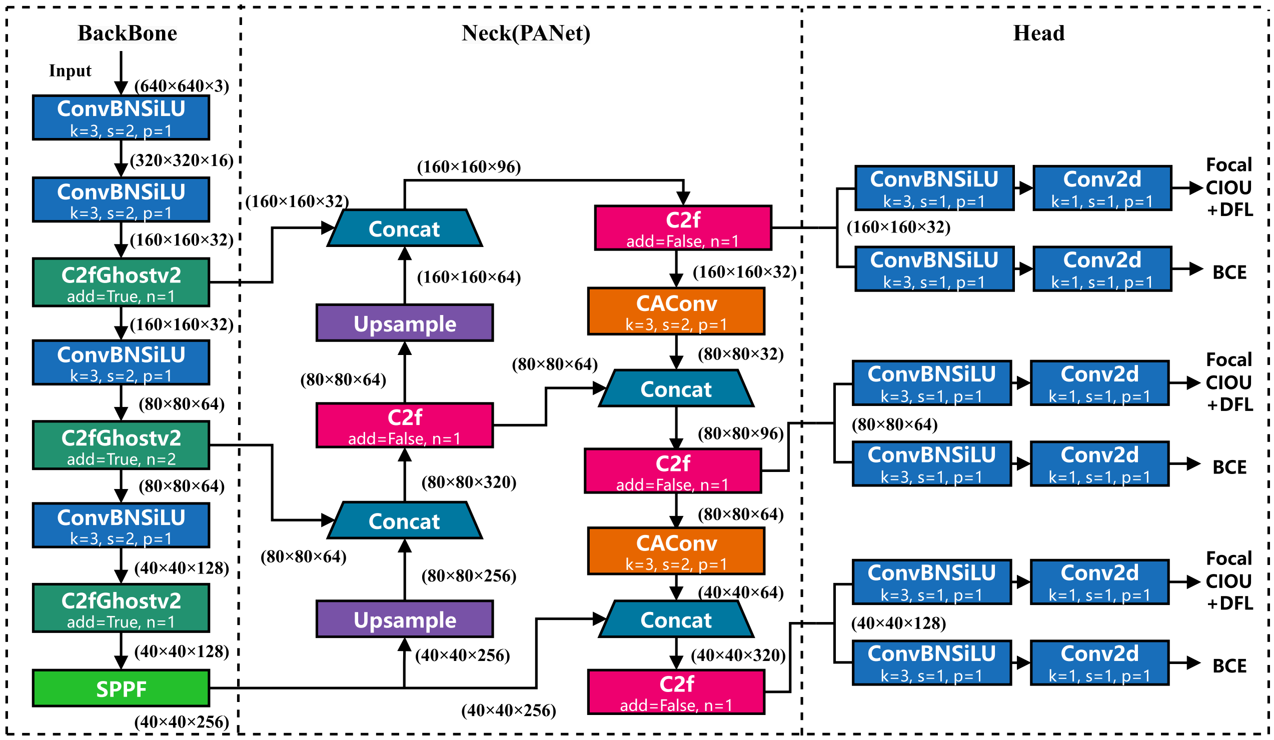

FIGURE 2
The architecture and network specifics of YOLO-LHD.

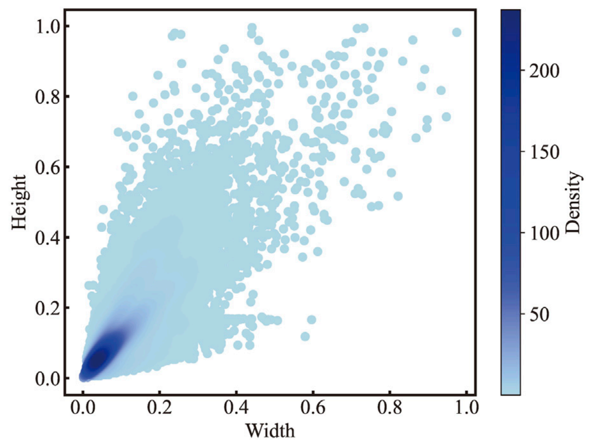

FIGURE 3
The distribution of ground-truth dimensions.

the ideas of Zhang et al. (2023), we embed the CA attention mechanism into the downsampling convolution module of YOLO-LHD's feature fusion network to get CAConv, that is, convolution operation based on attention. Embedding the CA attention mechanism module into the feature fusion network pays more attention to the helmet target of interest, effectively reduces the impact of the complex background, and improves the detection ability of small targets. The structure of the CAConv module is shown in Figure 5.

The primary principle of the CA mechanism is as follows: First, the input feature is globally average pooled separately across the width (W) and height (H) dimensions. Subsequently, the features mapped to W and H directions are merged, stacking the width and height features. This is followed by feature processing via convolution, normalization, and activation functions. The process is then divided into H and W parallel stages. After adjusting the number of channels, the attention weights are acquired using a sigmoid function. Finally, these weights are multiplied with the original features, yielding the final feature set imbued with attention weights.

This paper embeds CA attention into the downsampling convolution module, that is, a 3 × 3 convolution layer is introduced at the end of CA attention. This attention-based convolution operation replaces the original standard downsampling convolution module. Compared with directly adding the attention module in the network, the attention-based convolution operation reduces the storage and transmission steps of the intermediate feature map, which helps to improve the performance and representation ability of the model.

3.2.2 Improvement of the backbone network

Incorporating high-resolution features aids in extracting detailed information from the target objects. However, it also escalates computational complexity and memory usage as the network processes higher resolution feature maps, necessitating more convolution operations to achieve higher-resolution outputs. To enhance feature extraction capabilities and lighten the model concurrently, this study introduces an improved

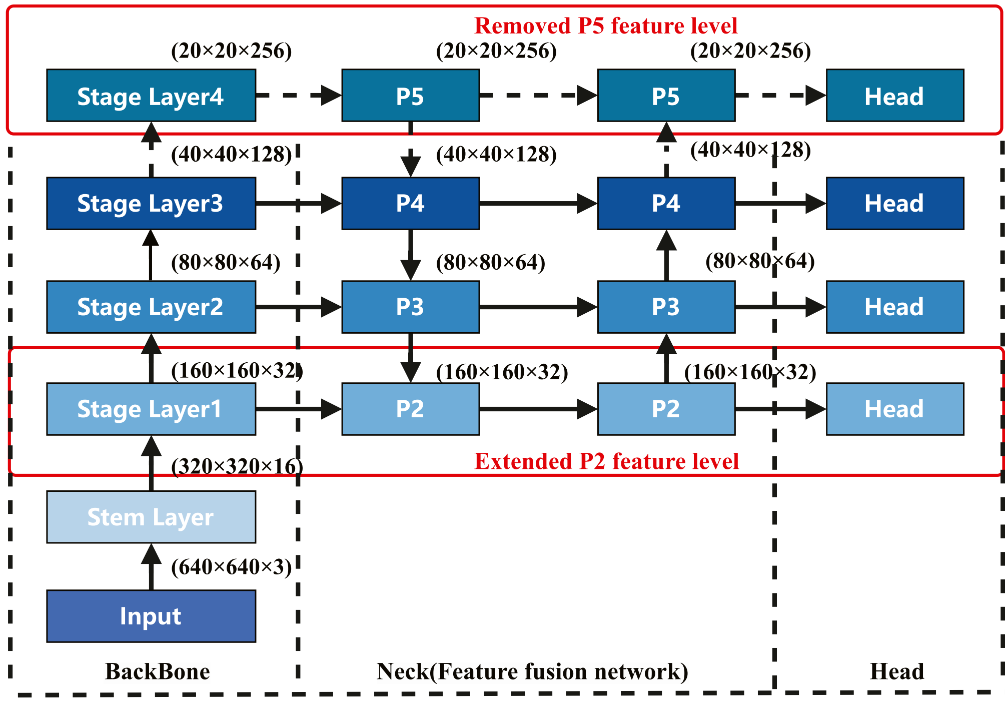

FIGURE 4
Reconstructed YOLOv8_cut network.

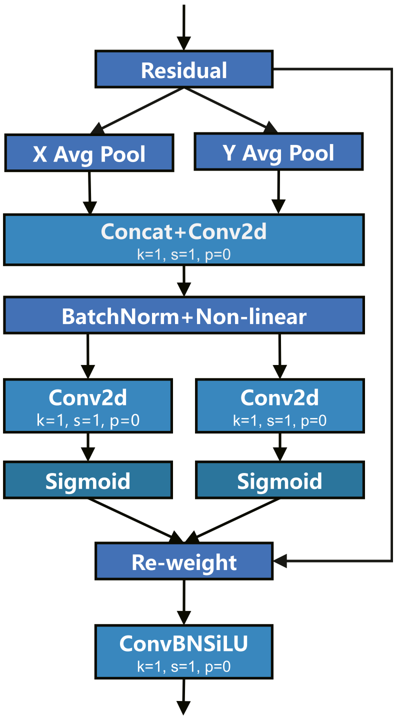

FIGURE 5
The structure of the coordinate attention mechanism.

GhostNetv2 Bottleneck to replace the existing Bottleneck, thereby bolstering the network model's feature extraction capacity, reducing model complexity, and minimizing model size. The specific process schematic is shown in Figure 6. We optimized the YOLOv8n model, adapting it to complex construction site scenarios, and achieved a more lightweight YOLO-LHD model that boasts high-precision detection accuracy.

In an effort to both lighten the YOLOv8n_cut network model and enhance its feature extraction capabilities, the backbone network was initially modified by integrating the module C2f from the original YOLOv8n_cut backbone with the improved GhostNetv2 Bottleneck to augment model performance. GhostNet is a lightweight feature extraction model consisting of two stacked Ghost Modules. The primary principle of a Ghost Module is as follows: initially, a 1 × 1 convolution extracts preliminary features from the input features $X \in \mathbb{R}^{H\times W\times C}$. Then, a depthwise separable convolution applies a linear transformation to the preliminary features, enriching the information in the feature map. Finally, these two feature components are concatenated. The main computation process of the GhostModule is as follows:

$$Y' = X * F_{1\times 1} \tag{1}$$

$$Y'' = Y' * F_{dp} \tag{2}$$

$$Y = \text{Concat}\left(\left[Y', Y''\right]\right) \tag{3}$$

In the equation, $*$ represents convolution operation, $F_{1\ \times\ 1}$ signifies a 1 × 1 standard convolution, and $Y' \in \mathbb{R}^{H\times W\times C_{out}}$

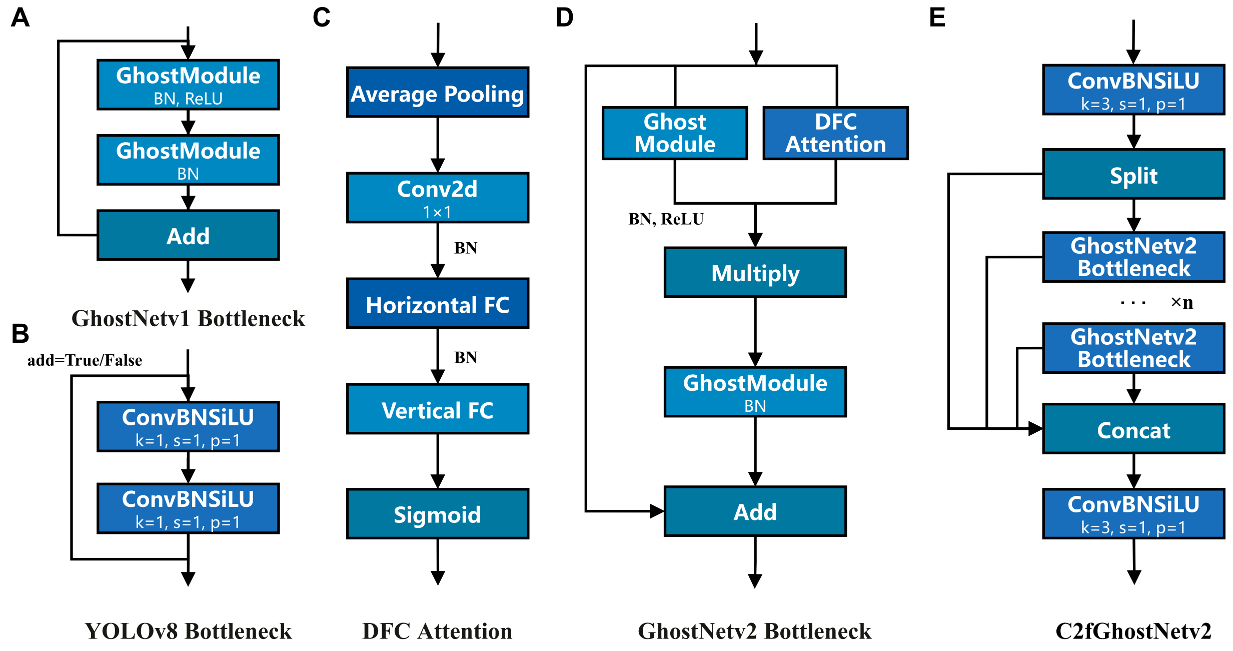

FIGURE 6
Schematic representations of the improved C2fGhostv2 and GhostNet: **(A)** GhostNetv1 Bottleneck, **(B)** the original YOLOv8 BottleNeck, **(C)** DFC Attention, **(D)** GhostNetv2 Bottleneck, **(E)** the improved C2fGhostNetv2.

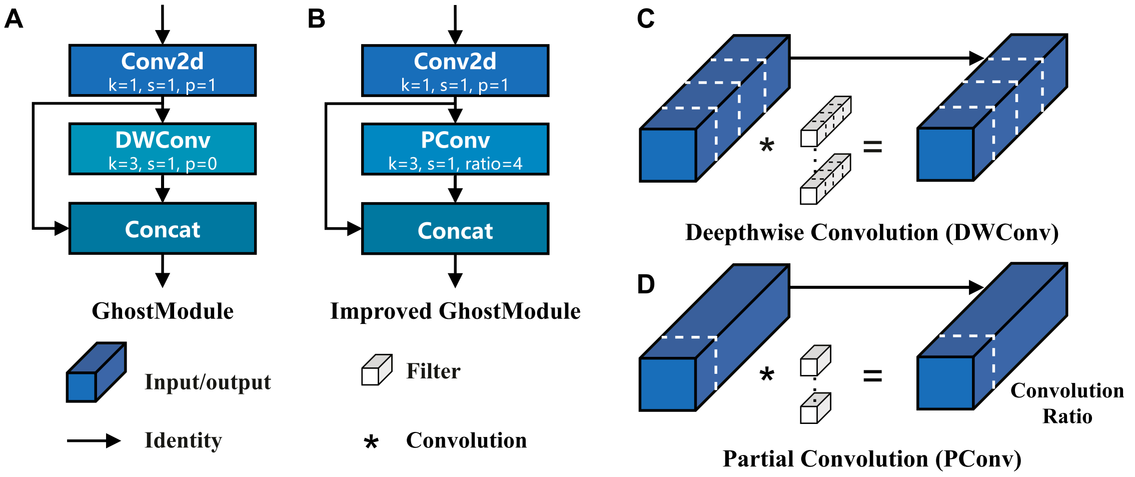

FIGURE 7
The improved GhostModule structure and the calculation schematic diagram of Partial Convolution (PConv): **(A)** GhostModule, **(B)** Imporved GhostModule, **(C)** Deepthwise Convolution (DWConv), **(D)** Partial Convolution (PConv).

denotes the output preliminary features. Subsequently, a depthwise convolution is applied to Y', further extracting features and generating s features denoted as Y''. Ultimately, Y' and Y'' are concatenated to generate the final output features $Y \in \mathbb{R}^{H\times W\times C_{out}}$, where Y possesses the same number of output channels as a standard convolution. The floating-point operations in the GhostModule operation are merely $\frac{1}{s}$ of the standard convolution.

In the linear transformation process within the GhostModule, this is realized through F_{dp} (depthwise convolution). While DWConv is extensively used in constructing low-FLOPs neural networks, its simple replacement with regular convolution could lead to a significant decrease in accuracy. In MobileNet Howard et al. (2017), an approach to compensate for the decline in accuracy is adopted by expanding the network width. Chen et al. (2023) proposed partial convolution (PConv), the feature maps are highly similar in different channels. To minimize computational redundancy, a regular convolution is applied to part of the input feature's channels, leaving the remaining channels unchanged, treating either the first or the last contiguous channels as representatives for feature computation.

The calculation diagram of PConv is shown in Figure 7, with the partial ratio $r = \frac{1}{4}$, the floating-point operations (FLOPs) count of PConv is only $\frac{1}{16}$ of a regular convolution. On the device, compared to DWConv, it presents a higher floating-point operations per second (FLOPS), exhibiting superior performance.

GhostNetv2 is a lightweight network architecture proposed Tang et al. (2022) at NeurIPS 2022. Building upon the original GhostNet model, they introduced a Decoupled Fully Connected (DFC) attention mechanism that enhances the model's representational capacity by further capturing long-distance spatial information. The primary structure of DFC, as depicted in Figure 6C, undergoes the following computation process: Firstly, to reduce additional computational cost, average pooling is used in the first step of DFC implementation to halve the feature size. Next, 1 × 1 convolution generates features, which are then aggregated in the horizontal and vertical directions over a long range of pixels via depthwise separable convolution with a kernel size of $1 \times k_H$ and $1 \times k_W$. After superposition, a global receptive field is achieved. Bilinear interpolation is used for upsampling, returning to the original dimensions. The features generated by the Ghost Module are then element-wise multiplied to achieve more information-rich features. The overall structure is shown in Figure 6D.

In the GhostNetv2Bottleneck of GhostNetv2, the DFC module and Ghost Module process input features in parallel. Then, the multiplied features serve as the input to the second Ghost Module, finally producing the output features. GhostNetv2Bottleneck captures the long-term dependencies between pixels at different spatial positions, thus enhancing the model's expressiveness. In this study, to further improve the model's feature extraction capability while keeping the network model lightweight, we utilize Partial Convolution (PConv) to replace DWConv in the Ghost Module of the GhostNetv2Bottleneck to complete Cheap Operation for linear transformation, as shown in Figure 7B. Subsequent experiments confirm the effectiveness of PConv in improving model accuracy.

The improved GhostNetv2Bottleneck achieves a better balance in model size, the number of model parameters, and FLOPs after replacing the original C2f Bottleneck. In the YOLOv8 backbone network, the output of the C2f module does not change the size of feature map, and the convolution stride of the original C2f Bottleneck is consistently 1. Correspondingly, the stride of the GhostNetv2Bottleneck used in this study is also 1.

3.2.3 Improvement of the loss function

In YOLOv8, the regression loss is composed of Distribution Focal Loss (DFL) and CIOU loss. CIOU loss was proposed by Zheng et al. (2020) based on geometric factors in bounding box regression, i.e., overlapping area, center point distance, and aspect ratio. It achieved better convergence speed and performance. Given a predicted box B and a real box B^{gt}, the definition of CIOU Loss is as follows:

$$L_{CIoU} = 1 - \mathrm{IOU} + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{v^2}{(1 - \mathrm{IOU}) + v} \tag{4}$$

where IOU represents the intersection over union of the predicted box and the real box, b and b^{gt} respectively represent the center points of the predicted box B and real box B^{gt}, $\rho(\cdot) = \|b - b^{gt}\|_2$ represents the Euclidean distance, c is the diagonal length of the smallest enclosing rectangle of the predicted and real boxes, and $v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$ represents the difference in aspect ratio between the predicted box and the real box.

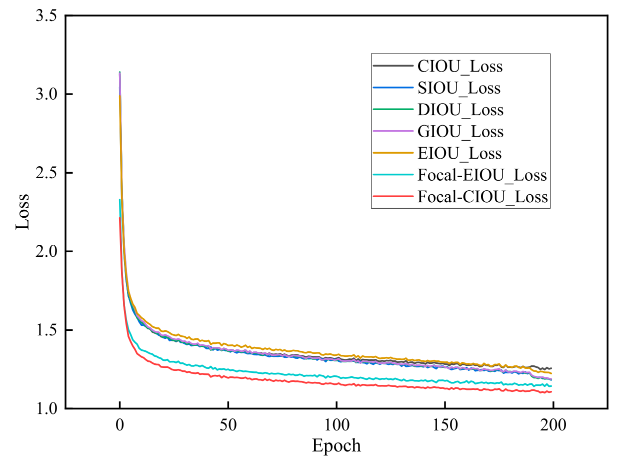

FIGURE 8
The regression error and distribution curves during training with different loss functions.

The idea of Focal-EIOU loss function Zhang et al. (2022) is introduced into CIOU loss to compensate for the overall gradient becoming smaller when the loss value tends to zero due to multiplication, setting a higher gradient where the error rate is high and paying more attention to the identification of difficult samples. To reduce the influence of low-quality samples on model performance, by using IOU to reweight CIOU loss. The definition of Focal-CIOU Loss is as follows, where γ is the adjustment parameter of the Focal loss:

$$L_{Focal-CIOU} = \mathrm{IOU}^{\gamma} L_{CIOU} \tag{5}$$

Figure 8 shows the performance of the YOLO-LHD model, where we use Focal-CIOU loss in place of the original CIOU loss. The figure demonstrates that the Focal-CIOU loss achieves better performance, as well as the performance of several other loss functions.

4 Experimental results

In this section, we evaluate our proposed YOLO-LHD on the MHWD dataset, discussing and contrasting the details of our improvements. First, we introduce the helmet detection dataset established for this study, along with our experimental environment and model evaluation metrics. After analyzing the experimental results, we provide evidence of the effectiveness of our improvements, including ablation studies and comparison experiments, and compare with other advanced object detection algorithms to demonstrate the detection performance.

4.1 Experimental datasets

In this paper, the dataset consists of two parts: safety helmet wearing detect dataset (SHWD) and Hard Hat Dataset (HHD). The

TABLE 1 Information on MHWD dataset.

Datasets	Helmet	No_Helmet	Images
Training	22,871	13,656	7,065
Validation	2,435	1,363	785
Testing	2,707	1,571	873

first part comprises 7,581 publicly available images from the SHWD dataset, containing the class labels “hat” and “person”. The second part consists of 5,000 images from the Kaggle platform, which includes the class labels “helmet”, “person”, and “head”.

In the original SHWD dataset, the ratio of images showing safety helmet wearing to not wearing is 1:12, making the categories extremely unbalanced. Therefore, we modified and enhanced the SHWD dataset to create the dataset used in this paper. We combined the SHWD and HHD datasets, unifying the category labels to “Helmet” and “No_Helmet”. Ultimately resulting in our final dataset, which we refer to as the ‘MHWD’ (Merged Helmet Wearing Detection dataset). The MHWD is made publicly available at https://drive.google.com/file/d/1GbiDTSdeKEB-BxuH4G54gtX-gDfC_uSF/view?usp=drive_link.

4.2 Experiment configuration and evaluation metrics

To validate the effectiveness of the proposed YOLO-LHD safety helmet detection model, all experiments in this study were conducted using the PyTorch 1.12.1 framework on a Windows 10 Professional 64-bit system with an NVIDIA RTX 3080 GPU and CUDA 11.3 environment. The model construction, training, and validation were performed under these settings.

During training, the following operations were applied to the training images: Mosaic data augmentation, adaptive padding, and scaling to 640 × 640. The training was performed for 300 epochs, with a batch size of 8. The initial learning rate was set to 0.01, the momentum parameter to 0.937, and the weight decay parameter to 0.0005. Regarding loss adjustments, the box loss gain was fixed at 7.5, the class loss gain at 0.5, and the DFL loss gain at 1.5, and the SGD optimizer was used for loss optimization. The dataset was divided into training, validation, and testing sets in an 8:1:1 ratio. The label distribution of the divided dataset is shown in Table 1. The table encompasses the number of labels about the wearing of safety helmets and the corresponding number of images.

The algorithm’s recognition speed and detection capability were evaluated in this study. The recognition speed was determined by measuring the inference time (in milliseconds) of the model on test images. The detection capability was assessed using precision, recall, and mean Average Precision (mAP). Lower inference times indicate faster network detection, precision represents the ratio of correctly detected boxes to predicted boxes and measures the accuracy of the network’s predictions. Recall represents the ratio of correctly detected boxes to the actual annotated boxes and measures the network’s ability to detect the labeled boxes. mAP is the average precision across all classes and serves as an indicator of the network’s overall recognition capability. In this study, predicted bounding boxes with an Intersection over Union (IoU) greater than 50% with the annotated boxes were considered as correctly predicted results.

We utilize the parameters count (Params) and computational complexity, quantified in GFLOPs (Giga Floating-Point Operations), as metrics for assessing the computational costs of the models. For a thorough evaluation, we incorporated COCO metrics, giving insights into performance across various IoU thresholds and object sizes: AP_{50-95}, AP_{50}, AP_{75}, AP_{small}, AP_{medium}, and AP_{large}. Figure 9 showcases the training mAP50 and loss curves of both YOLOv8n and YOLO-LHD models. Notably, the model’s mAP50 stabilizes around the 150-epoch mark, indicating progressive convergence as the iteration count rises. Figure 10 presents the confusion matrices for our model on the validation and testing sets. Our model demonstrates a strong capability in distinguishing between positive and negative cases. Moreover, the confusion matrix also indicates that the model has a relatively low rate of misclassifying the background, emphasizing its robustness in distinguishing safety helmets in varied scenarios.

4.3 Comparative experiment of improved backbone network

To validate the effectiveness of the improved backbone network structure in reducing model parameters and improving model accuracy, training and validation of six different lightweight backbone networks were conducted under the same experimental environment. The backbone network of the YOLOv8n model was replaced by MobileNetv3, ShuffleNetv2, FasterNet, GhostNetv2, EfficientViT_M0, EfficientformerV2_S0 and C2fGhostv2 networks. The improved models were compared in the experiments, and the experimental results are shown in Table 2.

The experimental results show that compared to the smaller computational cost of the ShuffleNetv2 and GhostNetv2 backbones, mAP50 increased by 4% under slightly increased computational overhead, and achieved higher detection accuracy than the pre-improved GhostNetv2 and other ViT-based lightweight backbones under smaller computational costs. The scheme in this paper surpasses other models in mAP50 and mAP50-95, demonstrating its advantage in extracting effective features and overall performance superior to the other six mainstream lightweight backbone networks.

4.4 Comparative experiment of different loss functions

We validated the performance of the improved Focal-CIOU Loss in the YOLO-LHD model and conducted comparison experiments with the original CIOU Loss and four other different loss functions. Under the same experimental environment, the Focal-CIOU Loss in this paper achieves the best performance. Compared with the original CIOU Loss, the accuracy has improved by nearly 1%, which proves the effectiveness of Focal-CIOU Loss. The experimental results are shown in Table 3.

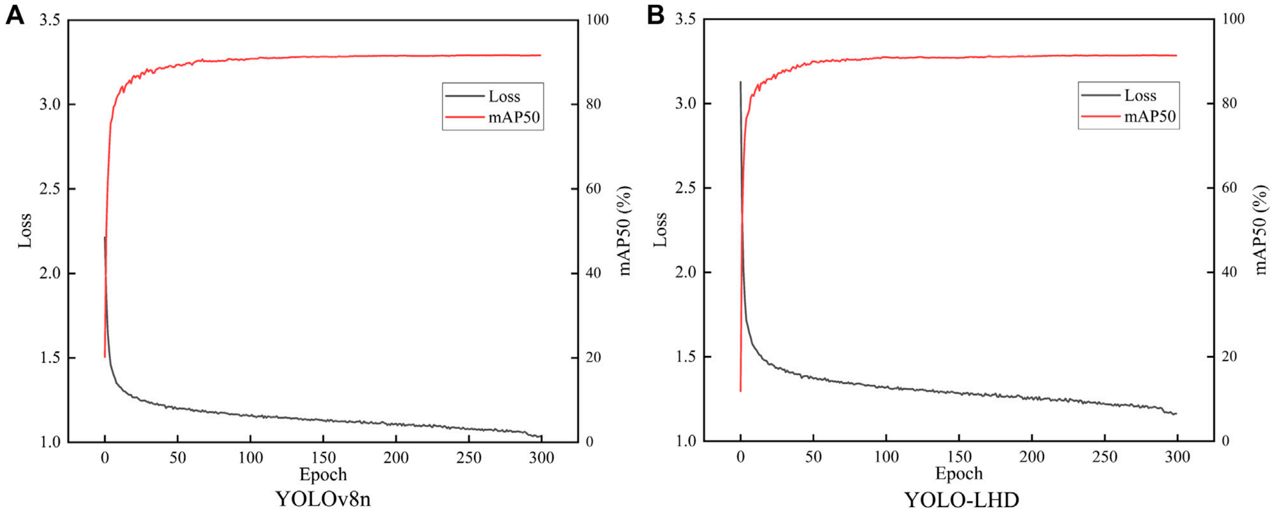

FIGURE 9
Training loss and mAP50 curves: **(A)** YOLOv8n, **(B)** proposed YOLO-LHD.

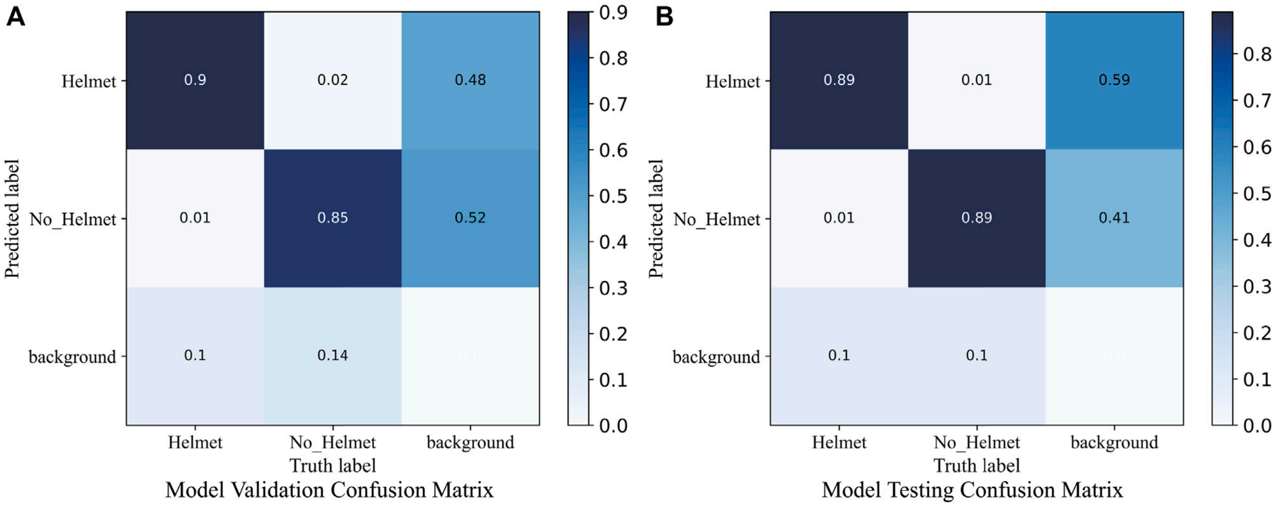

FIGURE 10
Confusion matrix for YOLO-LHD on validation and testing sets: **(A)** validation set, **(B)** testing set.

4.5 Performance of different attention mechanisms

To validate the accuracy improvement ability of our introduced downsampling convolution module based on CA attention mechanism, we introduced different attention mechanisms into the downsampling convolution module, such as SE (Hu et al., 2018), CBAM (Woo et al., 2018), CAM (Woo et al., 2018), and also directly added CA attention in the feature fusion network. Comparative experiments were conducted under the same experimental environment. Our improvement achieved a higher model Recall and average accuracy while maintaining relatively unchanged computational complexity and parameter count. In terms of Recall, the CAConv module reached 88.7%, and in mAP50, CAConv reached 94.3%, showing balanced performance in terms of accuracy. This proves the effectiveness and superiority of the CAConv module. The experimental results are shown in Table 4.

4.6 Ablation experiments

We validated the effectiveness of the proposed model through ablation experiments, analyzing the impact of the improved modules on the overall model by gradually adding each improvement to the original YOLOv8n model. The

TABLE 2 The performance comparison experiment results of improved backbone network.

Model	Model size (MB)	Params (M)	GFLOPs	mAP50 (%)	mAP50-95 (%)
YOLOv8n	6.2	3.01	8.2	93.3	58.4
+ MobileNetv3 Howard et al. (2019)	5.0	2.35	5.8	91.2	55.9
+ ShuffleNetv2 Ma et al. (2018)	3.6	1.86	**5.7**	89.4	54.3
+ FasterNet Chen et al. (2023)	4.7	2.25	6.7	93.1	57.8
+ EfficientViT-M0 Liu et al. (2023)	8.7	4.01	9.5	93.1	57.9
+ EfficientformerV2-S0 Li et al. (2022c)	39.5	5.11	11.7	91.2	57.1
+ GhostNetv2 Tang et al. (2022)	**3.3**	**1.42**	10.2	91.0	57.0
+ C2fGhostv2 (ours)	5.3	2.55	6.9	**93.4**	**58.1**

The bold values indicates the best results of each column.

TABLE 3 The performance comparison experiment results of different loss functions.

Loss function	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
CIOU Zheng et al. (2020)	92.3	87.6	93.4	58.1
SIOU Gevorgyan (2022)	91.4	87.1	93.3	58.2
DIOU Zheng et al. (2020)	91.8	87.9	93.5	58.0
GIOU Rezatofighi et al. (2019)	91.4	**88.8**	93.5	57.8
EIOU Zhang et al. (2022)	91.1	88.0	93.5	58.2
Focal-EIOU Zhang et al. (2022)	92.1	85.3	92.3	58.0
Focal-CIOU(ours)	**92.4**	**88.8**	**94.3**	**58.6**

The bold values indicates the best results of each column.

TABLE 4 The performance comparison experiment results of downsampling convolution modules combined with different attention mechanism.

Attention module	Params (M)	GFLOPs	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
Baseline	0.86	9.8	92.3	87.6	93.4	58.1
+ SEConv	**0.85**	**9.7**	92.8	86.3	93.3	58.4
+ CAMConv	0.87	9.8	91.4	87.7	93.5	58.3
+ CBAMConv	0.87	9.8	90.6	85.6	90.8	56.5
+ Coordinate Attention	0.86	**9.7**	**93.0**	87.1	93.6	**58.8**
+ CAConv(ours)	0.87	9.8	**92.8**	**88.7**	**94.3**	58.7

The bold values indicates the best results of each column.

performance of the YOLO-LHD model was evaluated, as shown in Table 5. ✓ indicates the use of that method. Reducing the downsampling rate significantly reduces the number of model parameters but increases the computational load. The use of C2fGhostv2 can reduce the computational overhead, but it also reduces accuracy. By introducing CAConv and Focal Loss, our proposed method reduced the size of the original YOLOv8n model by 66.1% and decreased the number of model parameters by 71.1%, achieving an average accuracy of 94.3%. While maintaining high accuracy, our method effectively reduces the size and parameter count of the original YOLOv8 model, showing a good lightweight effect.

4.7 Comparative experiment with different object detection algorithms

To verify the effectiveness of our algorithm, we conducted training and verification of mainstream algorithms under the same experimental environment and compared them with YOLO-LHD. As shown in Table 6, the comparison algorithms include SSD300 Liu et al. (2016), Faster-RCNN Ren et al. (2015), EfficientNet Tan and Le (2019), YOLOv3-tiny Redmon and Farhadi (2018), YOLOv5n Jocher (2020), YOLOv5s Jocher (2020), YOLOX-tiny Ge et al. (2021), YOLOv6n Li et al. (2022a), YOLOv7n Wang et al. (2023), and Rtmdet-tiny Lyu et al. (2022). Among them, SSD,

TABLE 5 The ablation experimental results of YOLO-LHD model.

Model	Cut	Ghostv2	CAConv	Focal-CIOU	Model size (MB)	Params (M)	GFLOPs	mAP50 (%)
YOLOv8n					6.2	3.01	**8.9**	93.3
M1	✓				2.2	0.96	10.5	93.8
M2	✓	✓			**2.0**	0.87	9.8	93.4
M3	✓		✓		2.2	0.97	10.6	94.0
M4	✓			✓	2.2	0.96	10.5	93.4
M5	✓	✓	✓		**2.0**	**0.86**	9.8	93.6
Ours	✓	✓	✓	✓	2.1	0.87	9.7	**94.3**

The bold values indicates the best results of each column.

TABLE 6 The performance comparison of different algorithms.

Model	Backbone	Params (M)	GFLOPs	COCOAP$^{test}_{50-95}$(%)	mAP50 (%)	Infer/ms
SSD300	VGG16	23.88	30.47	42.6	—	11.7
Faster-RCNN	ResNet50	41.35	69.53	47.8	—	21.9
EfficientNet	EfficientNet-b3	18.36	39.33	49.9	—	34.9
YOLOv3-tiny	Darknet53	8.67	13.0	46.5	86.2	**3.4**
YOLOv5n	CSPDarknet53	1.76	**4.1**	51.3	90.7	8.2
YOLOv5s	CSPDarknet53	7.03	16.0	**54.1**	92.5	9.8
YOLOX-tiny	DarkNet53	5.03	7.57	52.0	—	11.9
YOLOv6n	EffifienRep	4.63	11.34	53.9	—	10.1
YOLOv7n	ELAN + MP	6.02	13.2	51.7	92.4	7.2
Rtmdet-tiny	CSPNeXt	4.87	8.02	51.0	—	18.0
YOLOv8s	CSPDarknet53	11.12	28.4	53.4	94.1	10.7
YOLOv8n	CSPDarknet53	3.01	8.2	52.3	93.3	9.7
YOLO-LHD (ours)	CSPDarknet53-G	**0.86**	9.7	54.0	**94.3**	10.4

The bold values indicates the best results of each column.

Faster-RCNN, EfficientNet were conducted under the MMDetection Chen et al. (2019) framework, YOLOX-tiny, Rtmdet-tiny were conducted under the MMYOLO Contributors (2022) framework.

The results show that the number of parameters of our model is only 0.86M, which is far lower than other models, requiring smaller model storage. Our model's accuracy reached the best 94.3%, and the inference speed was 10.4 m. Although it was not the fastest, it showed higher efficiency compared to other models with similar or higher accuracy. With fewer parameters, although the inference delay has increased slightly, the overall goal of a lightweight helmet detection model has been achieved. It balances computational cost and detection accuracy well.

We compared our method with others based on the COCO metrics. As shown in Table 7, our YOLO-LHD achieved the best performance in both $\mathbf{AP}_{50}$ and $\mathbf{AP}_{small}$. Although our method may not be the top performer in some metrics, overall, it is on par with the best methods and demonstrates outstanding results, especially in detecting small-sized objects.

The detection effects of some typical complex backgrounds and poor lighting conditions are shown in Figure 11. It can be seen that our model performs well on small-size targets, obscured targets, and targets under poor lighting conditions. The model has high accuracy and does not produce false and missed detections. The experimental results demonstrate the practical significance of our YOLO-LHD model for helmet-wearing detection.

5 Conclusion

In this study, we propose a lightweight safety helmet detection model called YOLO-LHD. To enhance the model's detection and feature extraction capabilities for small objects, we decrease the downsampling rate of the backbone network and introduce high-resolution feature maps and the CA attention mechanism in the

TABLE 7 The performance comparison of different algorithms with COCO metrics.

Model	AP_{50-95}^{test}(%)	AP_{50}^{test}(%)	AP_{75}^{test}(%)	AP_{small}^{test}(%)	AP_{medium}^{test}(%)	AP_{large}^{test}(%)
SSD300	42.6	74.8	44.4	32.4	56.1	39.6
Faster-RCNN	47.8	82.2	51.1	41.1	58.6	42.2
EfficientNet	49.9	83.0	54.3	42.9	60.2	44.4
YOLOv3-tiny	46.5	85.7	45.2	37.4	54.9	53.2
YOLOv5n	51.3	91.4	54.8	43.4	59.9	62.0
YOLOv5s	**54.1**	92.3	57.7	45.4	61.4	62.8
YOLOX-tiny	52.0	86.1	56.9	43.2	60.8	41.8
YOLOv6n	53.9	90.2	**58.8**	41.8	**63.3**	**73.7**
YOLOv7n	51.7	92.5	50.4	42.2	59.6	61.5
Rtmdet-tiny	51.0	84.9	56.6	44.2	61.5	43.9
YOLOv8s	53.4	92.3	56.0	44.7	60.7	61.4
YOLOv8n	52.3	91.7	54.6	43.8	60.2	58.3
YOLO-LHD (ours)	54.0	**92.8**	54.4	**46.1**	60.7	61.4

The bold values indicates the best results of each column.

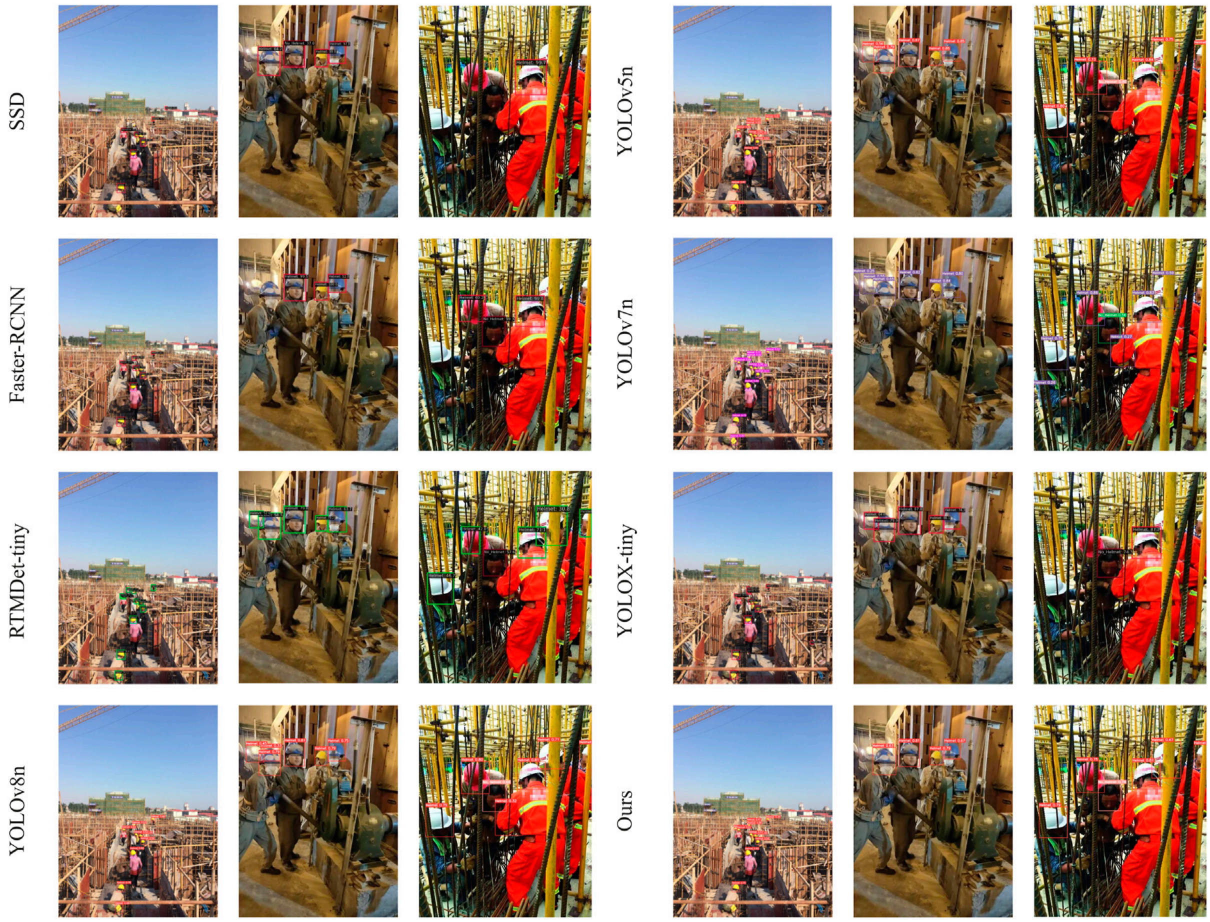

FIGURE 11
The results of helmet detection by different algorithms.

feature enhancement network, and add a large-scale detection head, thereby improving the detection performance of the model. Subsequently, we improve the C2f Bottleneck module by incorporating the GhostNetv2 Bottleneck module fused with PConv as the backbone network of YOLO-LHD, aiming to reduce the model's parameter and computational complexity. Finally, we introduce Focal Loss into the CIOU Loss and optimize the object detection model using Focal-CIOU Loss, thus improving the model's accuracy and robustness.

On the dataset used in this study, our YOLO-LHD safety helmet detection model has only 0.86M parameters and a model size of 2.1MB, achieving a mAP50 of 94.3%. While reaching the optimal detection accuracy, it also exhibits characteristics friendly to practical deployment. Through validation of actual detection results and comparisons with other existing algorithms, we demonstrated that the YOLO-LHD model possesses superior performance and practical application value in the domain of construction site safety helmet detection. Moving forward, we plan to explore additional avenues, such as further improving the model's noise robustness, collecting more real scene data to improve the generalization ability of the model, investigating real-time implementation possibilities, and extending its applicability to other domains. These potential directions will contribute to the continued advancement of safety helmet detection systems, making them even more effective and versatile in complex industrial environments.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article because the merged and optimized dataset utilized in this study is derived from publicly available datasets, which are described in the data acquisition section of the manuscript.

Author contributions

LH: Conceptualization, Resources, Supervision, Writing–review and editing. JR: Conceptualization, Methodology, Software, Writing–original draft, Validation, Visualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported by grants from the Key R&D Plan of Shaanxi Province (2023-YBGY-331) and Scientific Research Fund of Shaanxi University of Science and Technology (Doctoral Fund, Research on Data Fusion Technology in Paper Industry, BJ15-19).

Acknowledgments

The authors gratefully acknowledge the Youth Innovation Team of Shaanxi Universities.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). *Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934*. doi:10.48550/arXiv.2004.10934

Cao, X., Wu, C., Yan, P., and Li, X. (2011). "Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos," in *2011 18th IEEE international conference on image processing(ICIP)*, 2421–2424. doi:10.1109/ICIP.2011.6116132

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., et al. (2023). "Run, don't walk: chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (CVPR), 12021–12031. doi:10.48550/arXiv.2303.03667

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. (2019). *MMDetection: open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906*.07155. doi:10.48550/arXiv.1906.07155

Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., et al. (2022). "Mobile-former: bridging MobileNet and transformer," in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 5260–5269. doi:10.1109/CVPR52688.2022.00520

Contributors, M. (2022). *MMYOLO: OpenMMLab YOLO series toolbox and benchmark*.

Cui, M., Lou, Y., Ge, Y., and Wang, K. (2023). LES-YOLO: a lightweight pinecone detection algorithm based on improved YOLOv4-Tiny network. *Comput. Electron. Agric.* 205, 107613. doi:10.1016/j.compag.2023.107613

Deng, L., Li, H., Liu, H., and Gu, J. (2022). A lightweight YOLOv3 algorithm used for safety helmet detection. *Sci. Rep.* 12, 10981. doi:10.1038/s41598-022-15272-w

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., and Li, C. (2018a). Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Automation Constr.* 93, 148–164. doi:10.1016/j.autcon.2018.05.022

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., et al. (2018b). Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Automation Constr.* 85, 1–9. doi:10.1016/j.autcon.2017.09.018

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). *Yolox: exceeding yolo series in 2021. arXiv preprint arXiv:2107*.08430. doi:10.48550/arXiv.2107.08430

Gevorgyan, Z. (2022). *Siou loss: more powerful learning for bounding box regression. arXiv preprint arXiv:2205.12740*. doi:10.48550/arXiv.2205.12740

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2020). "Ghostnet: more features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 1580–1589. doi:10.1109/CVPR42600.2020.00165

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 13708–13717. doi:10.1109/CVPR46437.2021.01350

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 1314–1324. doi:10.1109/ICCV.2019.00140

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). *Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704*.04861. doi:10.48550/arXiv.1704.04861

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 7132–7141. doi:10.1109/CVPR.2018.00745

Jocher, G. (2020). *YOLOv5 by ultralytics*.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). *YOLO by ultralytics*.

Kurdthongmee, W. (2020). A comparative study of the effectiveness of using popular DNN object detection algorithms for pith detection in cross-sectional images of parawood. *Heliyon* 6, e03480. doi:10.1016/j.heliyon.2020.e03480

Lee, D.-S. (2005). Effective Gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Analysis Mach. Intell.* 27, 827–832. doi:10.1109/TPAMI.2005.102

Lee, J.-Y., Choi, W.-S., and Choi, S.-H. (2023). Verification and performance comparison of CNN-based algorithms for two-step helmet-wearing detection. *Expert Syst. Appl.* 225, 120096. doi:10.1016/j.eswa.2023.120096

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., et al. (2022a). *Yolov6: a single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209*.02976. doi:10.48550/arXiv.2209.02976

Li, J., Zhao, X., Zhou, G., and Zhang, M. (2022b). Standardized use inspection of workers' personal protective equipment based on deep learning. *Saf. Sci.* 150, 105689. doi:10.1016/j.ssci.2022.105689

Li, K., Zhao, X., Bian, J., and Tan, M. (2018). *Automatic safety helmet wearing detection. arXiv preprint arXiv:1802*.00264. doi:10.48550/arXiv.1802.00264

Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., et al. (2022c). Efficientformer: vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* 35, 12934–12949. doi:10.48550/arXiv.2206.01191

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "Ssd: single shot multibox detector," in *Computer vision–ECCV 2016: 14th European conference(ECCV)*, 21–37. doi:10.1007/978-3-319-46448-0_2

Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., and Yuan, Y. (2023). "EfficientViT: memory efficient vision transformer with cascaded group attention," in *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 14420–14430. doi:10.1109/CVPR52729.2023.01386

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., et al. (2022). *Rtmdet: an empirical study of designing real-time object detectors. arXiv preprint arXiv:2212*.07784. doi:10.48550/arXiv.2212.07784

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 116–131. doi:10.1007/978-3-030-01264-9_8

Ott, P., and Everingham, M. (2009). "Implicit color segmentation features for pedestrian and object detection," in *2009 IEEE 12th international conference on computer vision(ICCV)*, 723–730. doi:10.1109/ICCV.2009.5459238

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only Look once: unified, real-time object detection," in *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, 779–788. doi:10.1109/CVPR.2016.91

Redmon, J., and Farhadi, A. (2017). "YOLO9000: better, faster, stronger," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, 6517–6525. doi:10.1109/CVPR.2017.690

Redmon, J., and Farhadi, A. (2018). *Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767*. doi:10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: towards real-time object detection with region proposal networks. *Adv. neural Inf. Process. Syst.* 28, 1137–1149. doi:10.1109/tpami.2016.2577031

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 658–666. doi:10.1109/CVPR.2019.00075

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "MobileNetV2: inverted residuals and linear bottlenecks," in *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 4510–4520. doi:10.1109/CVPR.2018.00474

Shao, B., Hu, Z., Liu, Q., Chen, S., and He, W. (2019). Fatal accident patterns of building construction activities in China. *Saf. Sci.* 111, 253–263. doi:10.1016/j.ssci.2018.07.019

Shrestha, K., Shrestha, P. P., Bajracharya, D., and Yfantis, E. A. (2015). Hard-hat detection for construction safety visualization. *J. Constr. Eng.* 2015, 1–8. doi:10.1155/2015/721380

Tai, W., Wang, Z., Li, W., Cheng, J., and Hong, X. (2023). DAAM-YOLOV5: a helmet detection algorithm combined with dynamic anchor box and attention mechanism. *Electronics* 12, 2094. doi:10.3390/electronics12092094

Tan, M., and Le, Q. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 6105–6114. doi:10.48550/arXiv.1905.11946

Tang, Y., Han, K., Guo, J., Xu, C., Xu, C., and Wang, Y. (2022). *Ghostnetv2: enhance cheap operation with long-range attention. arXiv preprint arXiv:2211*.12905. doi:10.48550/arXiv.2211.12905

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). "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)*, 7464–7475. doi:10.48550/arXiv.2207.02696

Wang, X., Han, T. X., and Yan, S. (2009). "An HOG-LBP human detector with partial occlusion handling," in *2009 IEEE 12th international conference on computer vision(ICCV)*, 32–39. doi:10.1109/ICCV.2009.5459207

Wen, C.-Y., Chiu, S.-H., Liaw, J.-J., and Lu, C.-P. (2003). The safety helmet detection for ATM's surveillance system via the modified Hough transform. In *IEEE 37th annual 2003 international carnahan conference onSecurity Technology*. 364–369. doi:10.1109/CCST.2003.1297588

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 3–19. doi:10.48550/arXiv.1807.06521

Wu, F., Jin, G., Gao, M., Zhiwei, H., and Yang, Y. (2019). Helmet detection based on improved YOLO V3 deep model. In *2019 IEEE 16th Int. Conf. Netw. Sens. Control (ICNSC)*. 363–368. doi:10.1109/ICNSC.2019.8743246

Wu, S., and Nagahashi, H. (2014). Parameterized AdaBoost: introducing a parameter to speed up the training of real AdaBoost. *IEEE Signal Process. Lett.* 21, 687–691. doi:10.1109/LSP.2014.2313570

Yan, X., Zhang, H., and Li, H. (2020). Computer vision-based recognition of 3D relationship between construction entities for monitoring struck-by accidents. *Computer-Aided Civ. Infrastructure Eng.* 35, 1023–1038. doi:10.1111/mice.12536

Yu, Y., Hoshyar, A. N., Samali, B., Zhang, G., Rashidi, M., and Mohammadi, M. (2023). Corrosion and coating defect assessment of coal handling and preparation plants (CHPP) using an ensemble of deep convolutional neural networks and decision-level data fusion. *Neural Comput. Appl.* 35, 18697–18718. doi:10.1007/s00521-023-08699-3

Yu, Y., Samali, B., Rashidi, M., Mohammadi, M., Nguyen, T. N., and Zhang, G. (2022). Vision-based concrete crack detection using a hybrid framework considering noise effect. *J. Build. Eng.* 61, 105246. doi:10.1016/j.jobe.2022.105246

Zhang, X., Liu, C., Yang, D., Song, T., Ye, Y., Li, K., et al. (2023). *Rfaconv: innovating spatial attention and standard convolutional operation. arXiv preprint arXiv: 2304*.03198. doi:10.48550/arXiv.2304.03198

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018). "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," in *2018 IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 6848–6856. doi:10.1109/CVPR.2018.00716

Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., and Tan, T. (2022). Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* 506, 146–157. doi:10.1016/j.neucom.2022.07.042

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). Distance-IoU loss: faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell. (AAAI)* 34, 12993–13000. doi:10.1609/aaai.v34i07.6999

Zhou, F., Zhao, H., and Nie, Z. (2021). "Safety helmet detection based on YOLOv5," in *2021 IEEE international conference on power electronics, computer applications (ICPECA)*, 6–11. doi:10.1109/ICPECA51329.2021.9362711