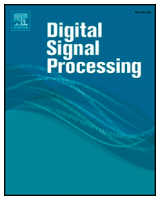
Digit. Signal Process. 144 (2024) 104283

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp



YOLO-PL: Helmet wearing detection algorithm based on improved YOLOv4

Haibin Li [a,b], Dengchao Wu [a,b], Wenming Zhang [a,b,*], Cunjun Xiao [a,b]

[a] School of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China
[b] Key Laboratory of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao, Hebei 066004, China

ARTICLE INFO

Keywords:
YOLOv4
Safety helmet-wearing detection
Small object detection
Lightweight

ABSTRACT

Workplace safety accidents are a pervasive issue worldwide. According to the National Work Safety Supervision Administration, a striking 67.95 % of construction accidents occur due to workers not wearing helmets. Existing helmet-wearing detection algorithms, however, tend to underperform in real-world scenarios where challenges such as smaller helmet areas in images, complex backgrounds, and object occlusions are present. Additionally, these models have a considerable amount of parameters, which impedes their practical deployment. This study proposes a novel, lightweight helmet detection algorithm, YOLO-PL, based on YOLOv4, to address these challenges. Initially, we designed the YOLO-P algorithms. YOLO-P algorithms optimize the network structure by refining its ability to detect small objects and improving the anchor assignment in the detection head. We design the Enhanced PAN (E-PAN) structure to merge the higher-layer, low-noise information with the lower-layer information based on the Path Aggregation Network (PAN). The YOLO-P algorithm improves detection accuracy by using the E-PAN structure. Subsequently, while preserving the performance of the YOLO-P algorithm, we enhanced its design for lightness. We proposed the Dilated Convolution Cross Stage Partial with X res units (DCSPX) module based on the Cross Stage Partial (CSP) structure, replacing the Spatial Pyramid Pooling (SPP) module with it. Additionally, we designed a Lightweight VoVNet (L-VoVN) structure based on the architecture of VoVNet, introduced a lightweight Max-Pooling (MP) down-sampling method, and fine-tuned the Swish activation function, which led to the final YOLO-PL algorithm. YOLO-PL significantly reduces the parameters in YOLO-P, thus achieving state-of-the-art performance that surpasses current object detectors like YOLOv5 and v7 in safety helmet detection. Moreover, our model exhibits substantial improvements in robustness and deployability, demonstrating considerable potential for practical implementations in industry.

Introduction

In numerous industrial and construction settings, on-site workers are mandated to wear safety helmets. Regrettably, the significance of wearing helmets is often overlooked due to a deficit in safety awareness and rule compliance. Conventional manual inspection methods are no longer adequate for the demands of contemporary construction safety management. As computer vision technology advances, both traditional visual algorithms and deep learning-based methods have been employed. However, the task of verifying helmet usage amongst workers continues to grapple with several challenges:

(1) Helmets are relatively small: Given the positioning of cameras, usually at a certain distance from the workers, existing algorithms display limitations in detecting small objects.
(2) The site's background is complex, and the objects are prone to occlusion: The environmental conditions on construction sites are intricate and variable. Industrial materials of various sizes and shapes, along with the work environment, can potentially interfere with the helmet detection. Furthermore, as workers engage in their tasks, other objects can easily obscure their heads, leading to incomplete information about the detected object and complicating the detection task.
(3) Diverse helmet colors and the presence of unworn helmets: A variety of helmet colors might be encountered on site, and the occurrence of unworn helmets can also introduce potential disturbances to the detection task.

Such challenges underscore the inherent difficulty and complexity of helmet wear detection tasks. The detection network requires a considerable receptive field and strong robustness. Developing algorithms that

* Corresponding author at: School of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China.
E-mail address: wudengchaoo@163.com (W. Zhang).

https://doi.org/10.1016/j.dsp.2023.104283

Available online 27 October 2023

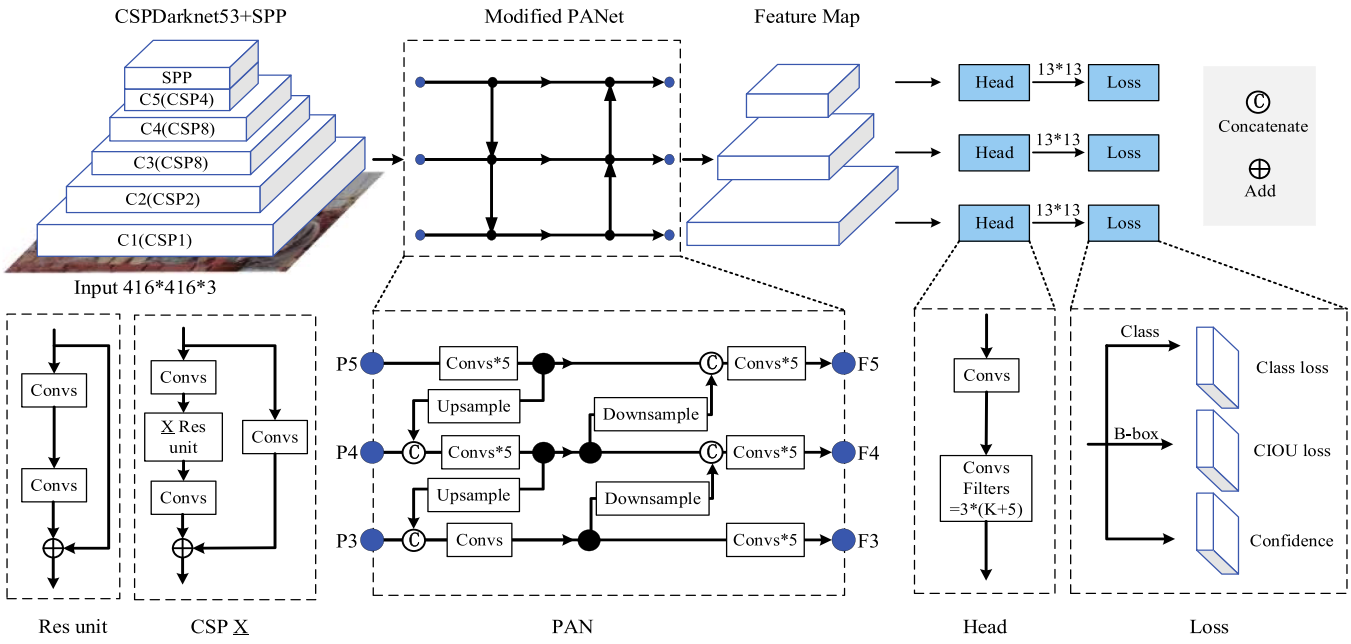


Fig. 1. YOLOv4 schematic diagram.

can effectively confront these challenges remains a pivotal area of contemporary research. For traditional machine vision algorithms, most methods do not need a powerful Graphics Processing Unit (GPU) to run smoothly on end devices. Park and Brilakis [1] Introduces a method that combines the Histogram of Oriented Gradient (HOG) with the HSV color histogram. This approach employs background subtraction to initially narrow down the detection to objects, then to humans, subsequently determining the presence of helmets. Rubaiyat et al. [2] presents a two-step detection mechanism. The first step merges the frequency domain information of an image with the widely-used Histogram of Oriented Gradient (HOG) for construction worker detection. Following this, the algorithm harnesses color features and the Circle Hough Transform (CHT) to ascertain whether the detected construction worker is donning a helmet. Liu and Ye [3] proposes a technique focused on facial localization using skin color as an indicator. Upon determining the face region, the method utilizes the Hu moment algorithm to process the information, achieving helmet recognition through SVM. The image feature extraction algorithm outlined above, specifically tailored to particular scenes, struggles to adapt to the intricate and dynamic conditions of the working environment, leading to diminished detection effectiveness for smaller helmets at greater distances. This challenge is further exacerbated within a multi-stage detection process, where instability at any given stage can have cascading effects, undermining the accuracy of subsequent detection stages. Such design characteristics indicate a lack of robustness in the algorithm, calling for further refinement and consideration. With advancements in computer hardware and technology, large deep neural networks, particularly convolutional neural networks (CNNs), can now be computed in parallel using GPUs [4]. Deep learning-based helmet-wearing detection methods are gaining popularity [5], and many researchers regard these methods to address industrial management challenges [6–8]. Due to the high accuracy of the early two-stage detectors, two-stage detectors were first used to detect safety helmets. Fang et al. [9] proposed a Faster R-CNN based helmet-wearing detector which achieves 90.1 % to 98.4 % accuracy in different scenarios. However, its practical application is challenging, as it requires approximately 0.2 s per image for detection. Due to the speed advantage of single-stage detectors and the fact that various state-of-the-art detectors have been proposed, more and more researchers in industrial applications are choosing single-stage detectors for practical application aspects of their designs. In [10], The DSFD (Dual shot face detector) algorithm [11] is used for face detection and localization, and then for helmet localization based on the return of the bounding box of the face, determine the presence of the helmet in the bounding box as well as to fine-tune the bounding box. Although the above algorithm achieves high accuracy, the inference speed is not mentioned in the paper, and the detection of helmets based on the candidate box generated by face detection is questionable for recognizing heavily obscured faces and for the particular shooting angles of workers. In [12], the confidence in helmet-wearing within a bounding box is determined by multiplying the pixels in the box by a set of weight coefficients, followed by an assessment through empirical thresholds to decide if a helmet is being worn. This method necessitates manually tuning both the weight coefficients and empirical thresholds. Such a process is laborious and poses a challenge in determining the optimal values. Chen et al. [13] presents four major methods for small object detection and their practical applications, yet it fails to address the algorithm's detection efficiency following the implementation of these methods. Wang et al. [14] by replacing the modules, including ghost and PixelShuffle lightweight in the neck to reduce the model size and computation, a slight increment in accuracy is achieved. However, the model still possesses a large number of parameters and does not improve detection for smaller helmet objects. Wang et al. [15] employs the Pyramid Split Attention (PSA) model to differentiate between background and foreground information, thereby filtering out noise interference [16]. Increased shallow detection head, using four detection heads at the same time. Although the detection performance has some improvement, the amount of modeling computation and the parameters are significantly larger, and the detection efficiency is reduced [17]. Based on YOLOv5 [18], a generalized method for detecting small objects is proposed, and it is concluded that increasing the network width while decreasing the network depth and increasing the detection head of the shallow layer can help detect small objects. However, the two approaches are not designed as lightweight, and the model inference speed becomes slower. The model agnostic matching network (MAMN) method proposed in Tao et al. [19] can mitigate the disparities among datasets resulting from equipment differences, demonstrating superior performance to traditional pre-training initialization parameter techniques across various datasets. However, the MAMN method primarily finds its application in the analysis of few-shot objects and for minimizing device differences during fault detection. Thus, it isn't ideal for

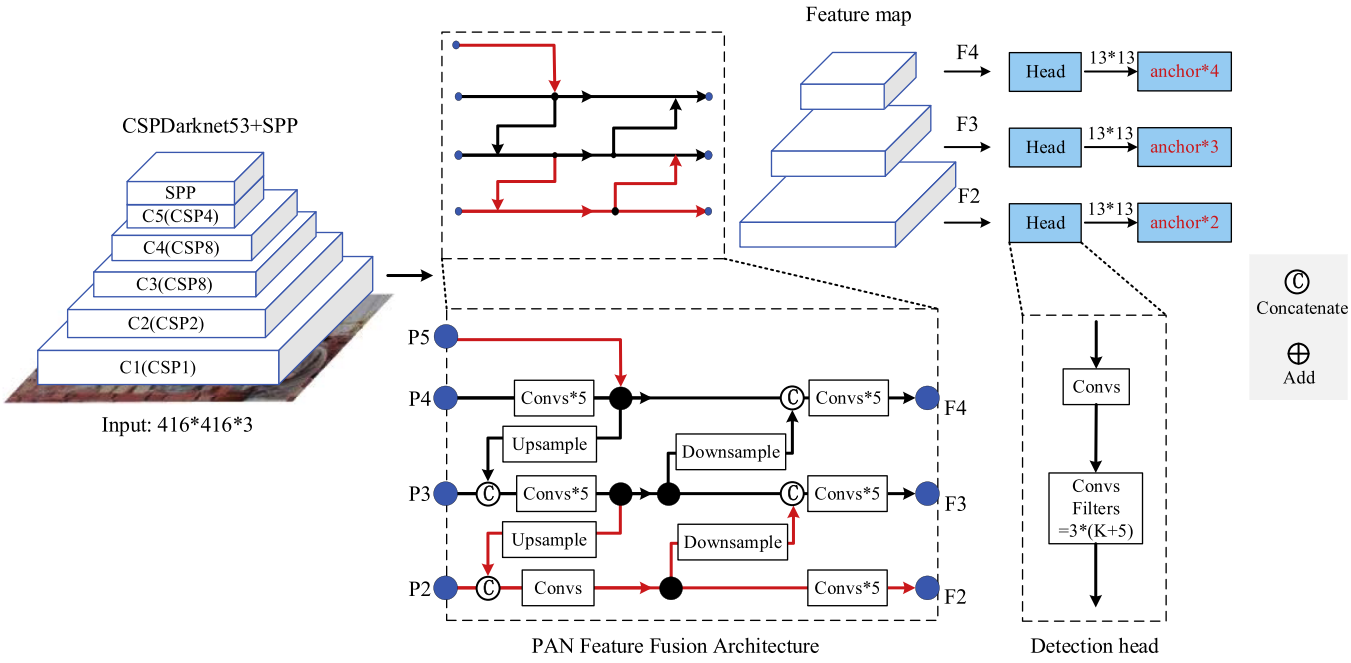


Fig. 2. Network structure improvement.

handling larger datasets like those associated with helmets. Additionally, the method cannot cope with situations where the algorithm's performance is degraded due to few-shot objects and significant occlusions.

In response to the above challenges, we proposed YOLO-P, an enhanced multi-scale feature fusion helmet detection algorithm based on YOLOv4 [20]. Considering the variance in the object sizes during helmet detection, with small objects being the predominant case, we specifically adapted the network structure for detecting small objects. Moreover, we optimize the anchor assignment in the detection head to thoroughly extract features from the C2 layer of the backbone. Owing to the intricate on-site background and prevalent occlusion between objects, we introduce the E-PAN module—an enhanced multi-scale feature fusion structure based on the PANet. The characteristic of the E-PAN module is that it effectively filters out complex background information through near up-sampling and residual operations between the same scales.

To address the challenges of high parameters and computational demands in the YOLO-P algorithm, which impede practical deployment, we propose a lightweight variant called YOLO-PL based on YOLO-P. We first suggest employing the DCSPX structure to broaden the receptive field as an alternative to the SPP module. Subsequently, a lightweight L-VoVN module is designed based on VoVNet [21], employing the one-shot aggregation architecture (OSA) and CSP structures to achieve a lightweight design for the original continuous convolution blocks. We enhance the conventional down-sampling method using the lightweight MP [22] down-sampling module. Lastly, to fine-tune the network, we integrate the Swish [23] activation function into the feature fusion structure. As a result, YOLO-PL maintains an equal equivalent Average Precision (AP) value to YOLO-P but reduces the parameters by 36.8 %, the computation is reduced by 33.1 %, and the inference speed is enhanced. We devised various experiments to confirm the improvements in performance that YOLO-PL offers within the field of safety helmet detection.

Related work

YOLOv4 is a successor to YOLOv3 and is developed on the Darknet framework. It is a fast end-to-end detection network that offers faster training speeds and easier deployment for practical applications than other networks. This end-to-end detection network exhibits enhanced training speeds and offers more convenient deployment in practical scenarios than its counterparts. As depicted in Fig. 1, the backbone of YOLOv4 employs CSPDarkNet53. The CSP structure [24] is used for gradient bifurcation, effectively reducing the parameters while

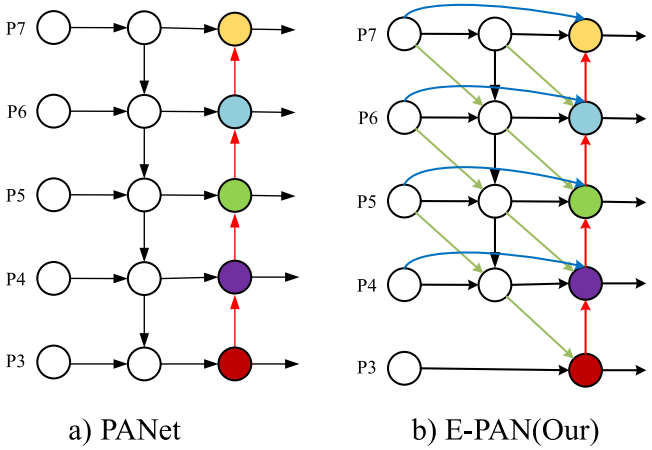


Fig. 3. Different feature fusion methods.

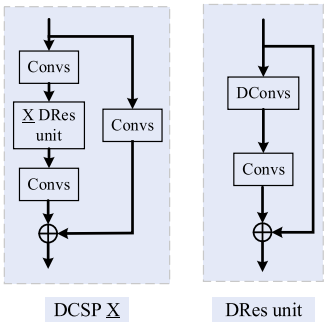
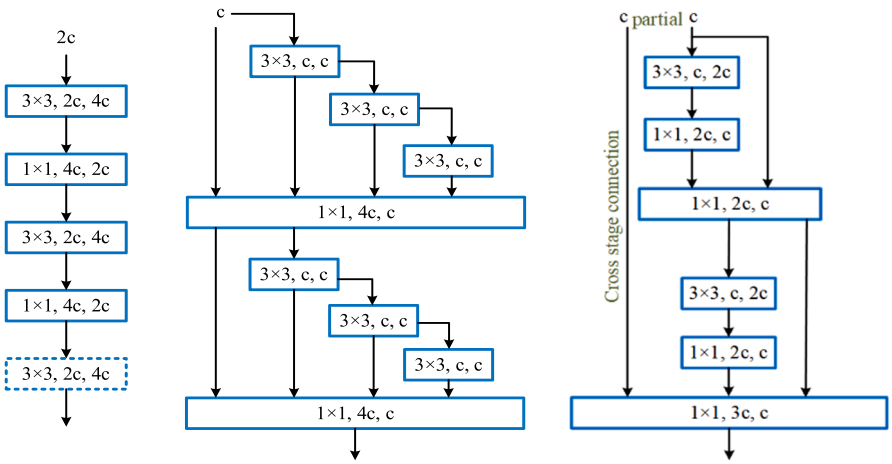


Fig. 4. DCSPXstructure.



a) Convolutional structure in YOLOV4 b) VoVNet(OSA structure) c) L-VoVN(Ours)

Fig. 5. L-VOVN designed analysis.

augmenting the expressive capacity of the backbone network. The neck of the model incorporates the PANet [25] for feature fusion, facilitating interactive representation of diverse information. Although the model retains the detection head from YOLOv3 [26], it refines the regression loss using the advanced CIoU [27] to better adjust the prediction bounding box. Multi-scale detection is performed at layers C3, C4, and C5, employing three detection heads. Each detection head is assigned three anchors.

Similarly, the YOLOv5 algorithm builds upon and enhances YOLOv3. Its overall architecture closely resembles YOLOv4, with the main difference in the backbone, which integrates the Focus structure. The feature fusion segment follows the same structure as YOLOv4, utilizing both the PAN and a Cross Stage Partial with X res units (CSPX) structure, similar to the backbone, for consecutive convolutions. This arrangement bolsters the capacity for network feature fusion.

YOLOv7 [22] also exhibits improvements in the feature fusion section. Its proposed E-ELAN, based on [28], employs expansion, shuffling, and merging bases to consistently boost the network's learning capability without compromising the original gradient paths. However, this design needs to be simplified and will increase the number of network parameters. The authors also propose a composite model scaling method in YOLOv7, which proposed composite model scaling method can continue the initial properties of the model to the maximum extent while ensuring that the structure is always optimal.

Proposed YOLO-PL architecture

In this section, we proposed the YOLO-P algorithm, which leverages multi-scale feature fusion to address the challenges posed by the presence of numerous small objects, complex detection backgrounds, and occlusions between individuals in helmet detection. In safety helmet-wearing detection, the YOLO-P algorithm exhibits better accuracy and recall compared to YOLOv5 and v7, but it has larger parameters. Therefore, we developed the YOLO-PL algorithm, a lightweight version of the YOLO-P, while ensuring it retains the performance metrics of the original YOLO-P.

YOLO-P algorithm designed based on multi-scale feature fusion

Improvements to network structure and detection head for helmet-wearing detection

In helmet-wearing detection scenarios, the considerable distance between workers and the camera often results in the presence of many small objects. During the convolution process, shallow feature maps,

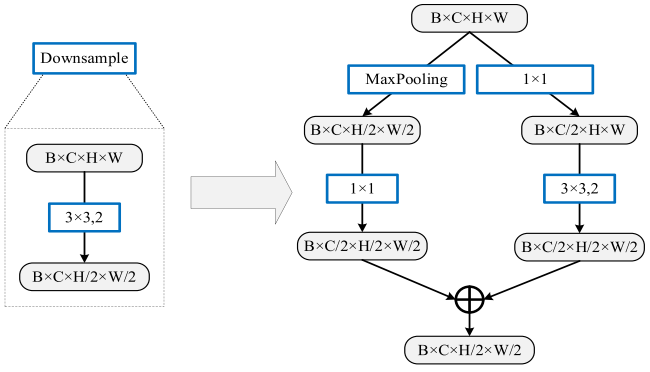


Fig. 6. Lightweight MP down-sampling structure.

which are larger, contain more detailed object information. However, as the convolution progresses through successive layers of down-sampling, the informative details of small objects tend to vanish with the progressive diminution of the feature maps. This effect is particularly pronounced for minuscule objects [29]. Furthermore, detecting such objects does not require an overly large receptive field. In contrast, an excessively deep network might not be as beneficial [17]. Consequently, we propose the following enhancements:

YOLOv4 employs the PAN approach to fuse features from P5, P4, and P3, conducting multi-scale detection sequentially. As depicted by the red line in Fig. 2, we propose the addition of a fusion operation on the P2 layer's output to improve the network performance for small object detection. Considering the global parameters volume and network simplification, the detection head corresponding to the P5 layer is eliminated, with its features being directly up-sampling for fusion.

We adjust the number of anchors assigned to each detection head to decrease the performance deficit in detecting large objects induced by the omission of the P5 layer. Instead of evenly distributing anchors among each detection head as we did previously, we configure the number of anchors corresponding to the P4, P3, and P2 layers to be 4, 3, and 2, respectively. Subsequent experiments confirm that such structural adjustments to the network enhance the detection of small objects despite an increase in the network's FLOPs. Following modifications in anchor assignment, the network's AP further improves while simultaneously reducing the FLOPs.

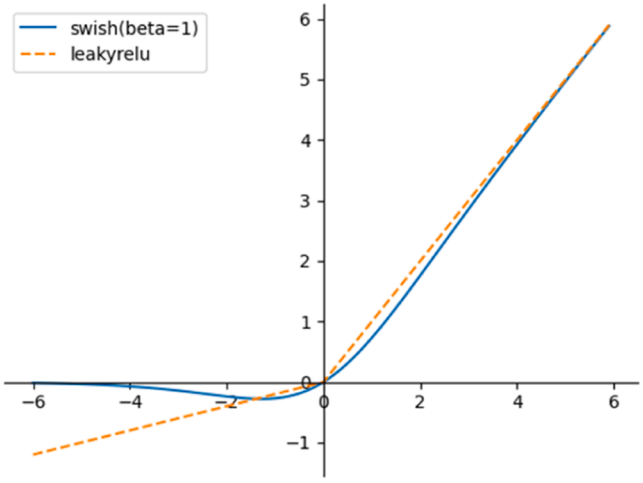


Fig. 7. Comparison of Swish activation function and LeakyReLU function.

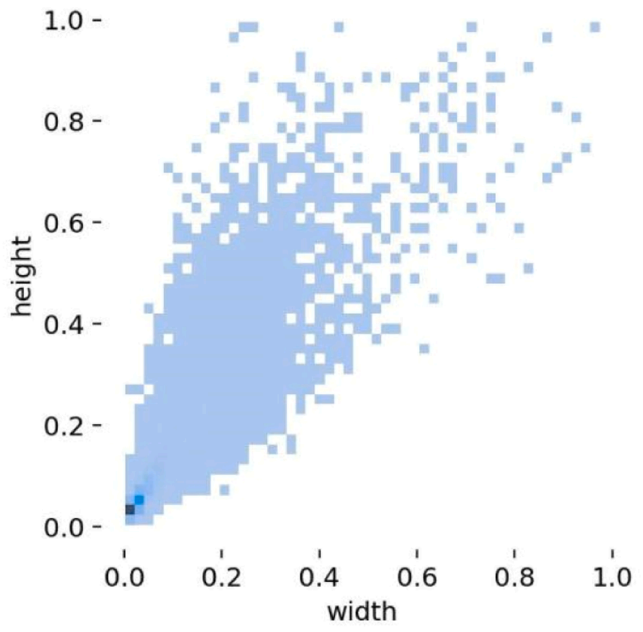


Fig. 8. Proportional distribution of object scales.

E-PAN design for multi-scale feature fusion

FPN (Feature Pyramid Network) [30] improves detection accuracy by integrating high and low-level features, especially for small objects. The PANet refines the FPN by introducing a bottom-up path enhancement. This structure emphasizes the significance of the network's shallow feature information, allowing the network to retain more of these specific features.

As shown in Fig. 3, we proposed E-PAN, based on PANet, which incorporates proximity up-sampling and residual connections at equivalent scales. Proximity up-sampling integrates information from the next higher scale, effectively reducing basal feature map noise, filtering complex background details, and emphasizing the feature map's foreground. The connection of residuals between the same scales is intended to fuse more features to enhances the expressiveness of feature maps. The above two operations without increasing too much computational cost. Moreover, this improvement can further enhance the network's feature fusion capability, effectively increasing the utilization of information within the low-level feature maps.

Table 1
Experimental environment settings.

Configuration	Parameter
CPU	Intel(R) Xeon(R) Gold 6240
GPU	GeForce RTX 3090
Operaing system	Ubuntu20.04
Accelerate environment	CUDA11.4 CUDNN8.2.4

Lightweight helmet wear detection YOLO-PL algorithm design

Though the YOLO-P algorithm boasts high AP and recall, its large parameters and FLOPs require significant resources, hindering real-time performance. To mitigate this, we introduced the lightweight YOLO-PL while retaining YOLO-P's performance.

DCSPX structure design

In this section, considering the prominence of small objects in helmet-wearing detection, adjustments are made to the CSP4 component within the backbone network. The DCSPX structure is proposed, as shown in Fig. 4. `DCSPX' stands for `Dilated Convolution Cross Stage Partial with X res units,' where `X' denotes the number of DRes units within the structure. Where DConvs are dilated convolutions, the DCSP structure extends the receptive field by replacing the standard convolutions with dilated convolutions within the CSP structure. Dilated convolutions can amplify the features of the network's receptive field without adding computational overhead. However, the continuous and excessive use of dilated convolutions can lead to the loss of local information. This loss is detrimental for pixel-level object prediction. To mitigate this, we apply dilated convolutions only to the network's deeper layers intended for large object-level prediction. As a result, the CSP4 component in the backbone network (as presented in Fig. 1) is substituted with DCSP2. This adjustment is used to expand the receptive field, acting as an alternative to the SPP module. This change permits the elimination of the SPP module and the convolution blocks that come before and after it. This improvement upholds the network's overall detection accuracy while diminishing parameters and floating point operations (FLOPs).

Lightweight L-VoVN structure design

Efficient network design often considers factors such as the number of parameters, computational load, and computational density. DenseNet [31] achieves exceptional performance by preserving intermediate features with various receptive fields through dense connectivity and concatenation, thus maintaining the original form of early information. However, it often results in slower speeds and lower energy efficiency due to the high memory access costs caused by the linear increase in input channels in dense connections. VoVNet addressed this by introducing the OSA, an energy and computationally efficient structure. As illustrated in Fig. 5(b), the OSA can capture diverse features using multiple receptive fields and only concatenates all prior features in the final feature map. This approach effectively circumvents the inefficiency associated with dense connectivity and notably improves the performance of small object detection.

As shown in Fig. 5(a), the convolution block connection structure in the PANet method during feature fusion in YOLOv4 is presented. The 3 $\times$ 3 convolution block, indicated by the dotted line, is not included in the continuous up-sampling process within PANet but in the continuous down-sampling process. Consecutive convolution can easily result in some deep features being lost, leading to overfitting. This overfitting often manifests as a decrease in detection accuracy.

With the design structure of VoVNet as a starting point, we propose a lightweight variant of VoVNet: L-VoVN, as shown in Fig. 5(c). L-VoVN retains the original OSA structure and introduces the Cross Stage Partial (CSP) structure. This enhancement allows the module to extract richer gradient combination information, and the diverse layers can learn a

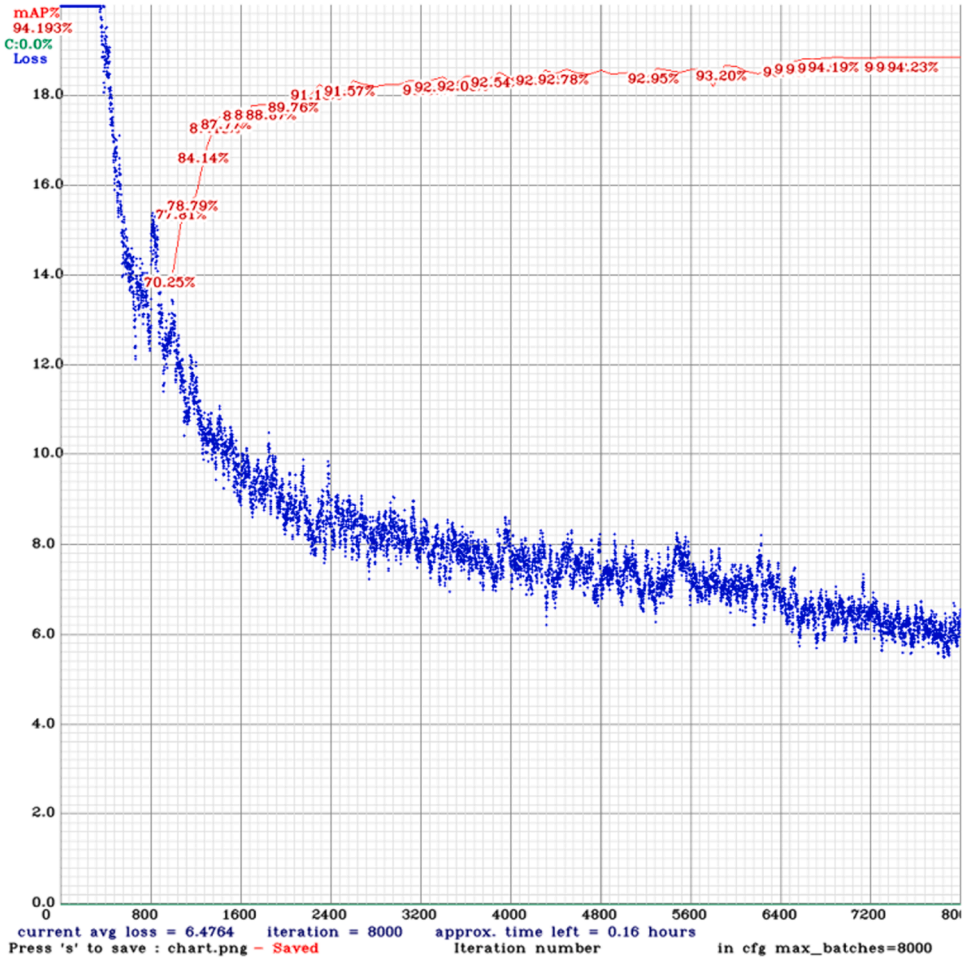


Fig. 9. Experimental results.

Table 2
YOLO-P ablation experiments.

#	ASHWD	E-PAN	AP_{50}/%	AP_{75}/%	Recall/%	BFLOPs	Params/MB
1			91.54	51.35	92.28	59.57	61.1
2	✓		93.88	58.34	94.04	63.95	45.9
3		✓	91.79	51.65	92.27	60.37	62.2
4	✓	✓	94.25	58.69	94.31	64.66	46.1

broader range of features. This configuration ensures that the structure can be used effectively for learning and convergence. Simultaneously, this structure increases the width of the 3 × 3 convolution in VoVNet and maintains the original convolution method used in YOLOv4. This broader network structure is more conducive to the detection of small objects [18]. This module significantly reduces the network's parameters and computational load. Through experimentation, it has been found that this structure outperforms other lightweight modules.

The number of convolutional block structure parameters used for down-sampling in the PANet of the original YOLOv4 is described in Eq. (1).

$$\begin{aligned} P_Y &= \sum_{l=1}^{N} \left(C_{l\_in} \times C_{l\_out} \times K_l^2\right) \\ &= 3 \times 3^2 \times 2C \times 4C + 2 \times 1^2 \times 4C \times 2C \\ &= 232C^2 \end{aligned} \tag{1}$$

where l denotes the convolutional layer in the network and denotes the number of input channels and output channels in the lth layer, respectively, and K denotes the size of the convolutional kernel. The amount of parameters of the L-VoVN module is calculated as shown in Eq. (2).

$$\begin{aligned} P &= \sum_{l=1}^{N} \left(C_{l\_in} \times C_{l\_out} \times K_l^2\right) \\ &= 2 \times 3^2 \times C \times C + 3 \times 1^2 \times 2C \times C + 1^2 \times 3C \times C \\ &= 45C^2 \end{aligned} \tag{2}$$

It can be seen that with the lightweight design, L-VoVN has up to 80.6 % fewer parameters than the convolution block structure in the original YOLOv4 PANet.

Lightweight MP down-sampling structure

In YOLOv4, traditional 3 × 3 convolution is directly employed to perform the down-sampling operation. In contrast, in the tiny-type network of the YOLO series, only the max-pooling layer is typically used for down-sampling. Max-pooling operates by retaining the maximum value within a specific region. This process may inadvertently lead to the loss of crucial information. Nevertheless, max-pooling offers several advantages. It requires fewer parameters and computational resources, enhances the invariance of image features, increases robustness to image offset and rotation, and helps maintain image texture information, thus reducing the risk of overfitting.

As illustrated in Fig. 6, the lightweight MP down-sampling structure is achieved by substituting certain convolutional layers with max-pooling and followed by concatenation. This modification can reduce the Params and FLOPs to nearly 50 % of their initial number.

Table 3
YOLO-PL ablation experiments.

#	DS	L-VoVN	MP	Swish	AP_{50}/%	AP_{75}/%	Recall/%	BFLOPs	Params/MB
1					94.25	58.69	94.31	64.66	46.4
2	✓				94.26	58.79	94.52	60.73	35.4
3	✓	✓			93.70	57.37	94.14	43.67	29.4
4			✓		94.28	58.72	94.29	64.19	46.2
5	✓	✓	✓		93.91	56.81	94.27	43.21	29.1
6	✓	✓	✓	✓	94.23	59.00	94.26	43.21	29.1

Table 4
Comparative experiments on structural improvement of adapted small objects.

#	Four heads	Three heads	Anchors assign	AP_{50}/%	AP_{75}/%	BFLOPs
1	✓			93.59	57.51	69.69
2		✓		93.78	57.96	63.96
3(Ours)		✓	✓	**93.88**	**58.34**	**63.95**

Additionally, it enhances detection performance while preserving the original number of output channels.

Swish activation function

Deep neural networks play a pivotal role in shaping the network's training dynamics and overall performance. Despite the LeakyReLU activation function mitigating the ``dying neuron'' problem inherent in ReLU, abundant empirical evidence suggests it can produce unstable results, limiting its practical application. Swish showcases superior performance on deep model architectures. The design of the Swish activation function is further improved, inspired by the LSTM and Sigmoid activation functions. As shown in Fig. 7, Swish leverages its impressive smoothness to optimize the algorithm effectively.

The Swish activation function is formulated as in Eq. (1), where σ is the Sigmoid activation function, which β is a constant or trainable parameter. The Swish activation function can effectively prevent the gradient from saturating during network training, and its derivative is constantly higher than 0.

$$\text{Swish}(x) = x \cdot \sigma(\beta x) \tag{4-9}$$

Experiment

Datasets

This study partially utilizes the SHWD (Safety Helmet Wear Detection) dataset [32], a public repository containing 7581 images, inclusive of 9044 instances of human heads wearing safety helmets and 111,514 standard head objects. To enrich our dataset's diversity, we added an 450 images from various scenes to the SHWD, resulting in a combined total of 8031 images.

Fig. 8 presents an analysis of the expanded SHWD dataset, with the x-axis and y-axis representing the ratios of the object's length and width to the image's length and width, respectively. The object sizes display significant diversity. However, a vast majority possess a ratio of less than 0.05, classifying them as small objects.

To demonstrate the robustness of our proposed model, we configured a comparative experiment using the ``Safety Helmet Detection'' dataset, published by Kaggle to enhance workplace safety. This dataset contains 5000 images, with our detection task specifically targeting the `helmet' and 'head' categories. This dataset will be uniformly referred to as the SHD dataset in ensuing discussions. Moreover, we introduce a Motorcycle Helmet Detection (MHD) dataset for comparative experiments. This dataset consists of 3052 images and is divided into two categories: Helmet and Without_Helmet.

Experiment settings

In the experiments, the batchsize is 64, and subdivisions are set to 16, which corresponds to 4 images per mini-batch, and the experiments are conducted for 8000 iterations. That equals 80 epochs. Parameters such as momentum, initial learning rate, and weight decay are the original parameters in the YOLOv4 network.

The pre-training phase utilizes the official CSPDarknet53 backbone network pre-training weights, with the anchor settings in the experiments deriving from the results of K-means++ algorithm clustering. To validate the algorithm's performance and serve as a practical deployment reference, we conducted experiments at a more challenging resolution of 416 × 416. To evaluate our algorithm's performance, we considered six metrics: AP50, AP75, Recall, floating point operations (FLOPs), Params, and Frames Per Second (FPS). AP_{50}, the Average Precision (AP) value when the Intersection over Union (IoU) of the prediction box and ground truth exceeds 0.5, it can adequately reflect whether the target has been detected. In contrast, AP_{75} leans more towards assessing the precision of the detected object's location than AP_{50}. Recall, or the rate of missed detection; FLOPs, which gauge the complexity of the algorithm/model; and Params, the total number of parameters requiring training during the model's training phase, along with Frames Per Second(FPS), which represents the efficiency of model inference, all collectively serve as comprehensive metrics for evaluating a model's overall performance in practical scenes. The specifics of the experimental environment settings are illustrated in Table 1.

Experimental results

We conducted a statistical analysis of the detection results on the validation set. The loss curves and AP_{50} curves of the training process are shown in Fig. 9, from which it can be seen that the loss value decreases faster before 2500 iterations, slowly oscillates down and

Table 5
Comparative experiments of different lightweight feature fusion modules.

#	L-VoVN	CSPVoVN	ELAN	AP_{50}/%	AP_{75}/%	Recall/%	BFLOPs	Params/MB
1				**94.26**	**58.79**	**94.52**	60.73	35.4
2		✓		93.61	57.08	93.67	45.05	30.0
3			✓	93.59	56.67	93.63	44.28	29.8
4 (Ours)	✓			93.70	57.37	94.14	**43.67**	**29.4**

Table 6
Performance comparison of different algorithms.

Algorithm	AP_{50}/%	AP_{75}/%	Recall/%	Params/MB	BFLOPs	FPS
YOLOv4	91.40	51.35	92.28	61.1	59.57	94.3
YOLOv5L	91.50	53.86	91.85	46.5	46.09	101.6
YOLOv7	92.48	**60.00**	90.00	34.7	43.62	104.2
YOLOv7 tiny	86.01	49.70	80.69	**5.6**	**5.52**	**187.9**
YOLO-PL	**94.23**	59.00	**94.26**	29.1	43.21	98.9

Table 7
SHD dataset comparison experiments.

Algorithms	AP_{50}/%	AP_{75}/%	Recall/%	BFLOPs	Params/MB
YOLOv4	92.59	55.07	91.86	59.57	61.1
YOLO-P	**93.71**	57.18	**92.22**	64.66	46.1
YOLOv7	92.95	**60.26**	89.55	43.62	34.7
YOLO-PL	93.54	59.43	92.07	**43.21**	**29.1**

stabilizes with the increase of the number of iterations, and finally, the loss value stabilizes at about 6.0. The model converges with the highest AP_{50} of 94.23 %.

Ablation experiments

YOLO-P algorithm ablation experiments

Table 2 presents the results of the ablation studies. ASHWD (Adaptive Safety Helmet Wearing Detection) denotes the enhanced network structure and the detector head reassignment for helmet-wearing detection. Notably, this method leads to a considerable enhancement in algorithm performance, with AP_{50} showing an improvement of 2.34 % and AP_{75} showing a boost of 6.99 %. When only the E-PAN structure is implemented in YOLOv4, AP_{50} sees a slight increase of 0.25 %. And AP_{75} increases by 0.30 %. Following the application of ASHWD, if the E-PAN structure is further implemented, there is a slight elevation in accuracy relative to the solo application of the E-PAN structure. AP_{50} escalates from 93.88 % to 94.25 %, showing a progress of 0.37 %; AP_{75} improves from 58.34 % to 58.69 %, demonstrating a gain of 0.35 %, and the recall rate also increases from 94.04 % to 94.31 %, marking an improvement of 0.27 %. The experimental data suggest that the E-PAN structure, a multi-scale feature fusion approach, is efficacious in enhancing network performance, particularly at the shallow layers.

YOLO-PL algorithm ablation experiments

This chapter's lightweight improvement of YOLO-P resulted in YOLO-PL, which can reduce the model's parameters and FLOPs significantly while retaining the AP and recall of the YOLO-P algorithm. To verify the effectiveness of each module proposed in this chapter, we establish the ablation experiments conducted on the expanded SHWD dataset, with the detailed results presented in Table 3.

In Table 3, entry 1 refers to the YOLO-P algorithm, while DS signifies the lightweight strategy employing the DCSP2 structure to replace the SPP module; L-VoVN denotes the lightweight design that builds upon VOVNet; MP represents the lightweight MP down-sampling structure; and Swish refers to the strategy introducing the Swish activation function. Notably, the YOLO-PL algorithm, represented by entry 6, necessitates 40 % fewer FLOPs than YOLO-P, represented by entry 1. Compared to the YOLOV4 network, it requires 30 % fewer FLOPs and halves the parameters. Additionally, it boosts the AP_{50} and AP_{75} metrics by 2.69 % and 7.65 %, respectively. This underscores the effectiveness of the algorithmic enhancements.

Comparison experiments

Comparison of different a priori methods

We compared various methods that are designed to enhance the small object detection network structures as discussed in Section 3.1.1. It contrasts the lightweight module proposed in Section 4.2.2 with other outstanding modules to validate its effectiveness. This serves to corroborate the superiority of the methods proposed in this research.

To fully exploit the information of small objects within lower-level feature maps, several algorithms use four detection heads to enhance the small object detection network structures [16]. These methods solely add the detection head in the P2 layer without making any adjustments to the one in the P5 layer. Moreover, many algorithms do not reassign the anchors of each detection head following an increase in the number of detection heads [33]. This section compares the above two cases and the proposed method in YOLO-P; the results are shown in Table 4.

Table 4 reveals that using three detection heads yields higher AP values and consumes fewer FLOPs than its four-headed counterpart. It is postulated that while increasing the quantity of detection heads can facilitate a more comprehensive detection of objects across different scales, it could potentially introduce more false detections. Moreover, an increased number of detection heads may inversely affect detection efficiency, thereby influencing the overall network's detection performance. When the anchors of the detection heads are reallocated, there is a slight improvement in precision coupled with a reduced computational load. The experiment substantiates that improving the structure to adapt to small objects, as proposed in YOLO-P, is more practical.

ELAN proposed in YOLOv7, along with CSPVoVN [34], is an improved feature fusion structure based on VoVNet. The lightweight L-VoVN proposed in this paper is also improved based on VoVNet. The results of the three improvement comparison experiments are shown in Table 5. The selected baseline algorithm, as indicated in Row 1, is the YOLO-P algorithm improved by the DSCPX proposed in this paper.

As indicated in Table 5, the lightweight L-VoVN feature fusion module proposed herein demonstrates lower parameters and FLOPs but a marginally higher AP value in comparison to CSPVoVN and ELAN, thereby evidencing the superiority of the L-VoVN module.

Performance comparison of different algorithms

To assess the performance of the YOLO-PL algorithm, we set up comparative experiments on the extended SHWD dataset, the results of which are shown in Table 6. It is observable that YOLO-PL falls slightly short of YOLOv7 in AP_{75}, However, both AP_{50} and the recall rate for YOLO-PL considerably outperform other algorithms. This indicates a marginally better localization accuracy for YOLOv7 compared to YOLO-PL. However, YOLOv7′s Recall is excessively low, leading to a significant amount of missed detection and a wide gap from YOLO-PL. For inference speed, the lightweight design allows YOLO-PL to achieve a 4.9 % speed boost over YOLOv4.

The improved network structure places a greater focus on low-level detailed information, enabling the detection of more small-sized safety helmets. Consequently, there is a significant enhancement in AP_{50} and Recall. However, the increased noise from the lower-level features impacts the localization precision, rendering AP_{75} less than optimal. While the multitude of connections introduced by E-PAN slightly hinders the overall inference speed of the model, it effectively increases the utilization of information in the feature maps. The subsequent lightweight

Table 8
MHD dataset comparison experiments.

Algorithm	AP_{50}/%	AP_{75}/%	Recall/%	Params/MB	BFLOPs
YOLOv4	92.26	72.00	92.32	61.1	59.57
YOLOv5L	91.93	78.56	92.25	46.5	46.09
YOLOv7	91.85	**81.88**	92.52	34.7	43.62
YOLO-PL	**93.31**	81.53	**92.82**	**29.1**	**43.21**

Fig. 10. Detection result of the expanded SHWD dataset.

design results in a substantial reduction in both parameter count and computational demands.

In practical applications, though still slightly slower than YOLOv5L and YOLOv7. Missed detection could result in severe accidents. Taking a generalized view, While YOLO-PL's inference speed and AP_{75} are slightly lower compared to YOLOv7, it surpasses in terms of AP_{50} and recall and also has a smaller size. Consequently, YOLOv7 is not the first choice for practical applications.

To verify the robustness of the YOLO-PL algorithm, we conduct a comparison experiment on the ``Safety Helmet Detection'' helmet detection dataset, and the results are shown in Table 7.

In the SHD dataset, the recall rate of YOLO-PL and AP_{50} are higher than YOLOv7, and only AP_{75} is slightly lower, but all metrics are higher than YOLOv4. Showing good robustness.

To verify the generalization capability of the YOLO-PL algorithm and further validate its potential, we designed comparative experiments using a dataset named MHD in the Motorcycle helmet detection domain. The experimental results are shown in Table 8.

From Table 8, it can be observed that YOLO-PL's AP_{50} and AP_{75} significantly outperform YOLOv4. Although its AP75 is slightly lower than that of YOLOv7, its AP_{50} remains notably higher than YOLOv7. Suggesting that YOLO-PL holds potential for practical applications.

To more clearly show the effectiveness of the YOLO-PL algorithm in helmet-wearing detection, images from the constructed dataset are presented in this section. Here, pink and green boxes mark correct detection, while red dashed boxes represent missed or false detections found by the algorithm after manual checking. Fig. 10 displays the detection results for the expanded SHWD dataset, while Fig. 11 shows the results for the SHD dataset.

For the SHWD dataset, Fig. 10(a) shows an image with a complex background, including occlusion situations and many small helmet objects. These factors led to five missed detections by the YOLOv4 algorithm and two by the YOLOv7 algorithm. In contrast, the YOLO-PL algorithm missed only one detection, with the other four objects successfully identified. In Fig. 10(b), the YOLOv4 and YOLOv7 algorithms missed a small object in the top right corner due to the reflective light; however, the YOLO-PL algorithm identified this object.

Regarding the SHD dataset, the YOLOv4 algorithm, in Fig. 11(a), wrongly infers the top object of the pole to be a safety helmet, showing a high level of false confidence in this incorrect detection. Two safety helmets on the left side of the image were also missed due to the camera angle causing facial occlusion; the YOLOv7 algorithm identified only one of these objects. In Fig. 11(b), a blurred and smaller safety helmet on the right side was missed, and the YOLOv7 algorithm also failed to detect this small helmet object. Nevertheless, the YOLO-PL algorithm identified all the objects mentioned above.

For analyzing the inference results of different algorithms on the two datasets, it is noticeable that the YOLOv4 algorithm shows substantial missed detections for objects with occlusion, indicating a lack of robustness and limited proficiency in recognizing small objects. While the more advanced YOLOv7 algorithm shows some improvement in detecting smaller objects, it still falls short of the YOLO-PL algorithm for detecting scenes with severe occlusion. Regarding the correctly identified helmet objects, the YOLO-PL algorithm inferred with significantly higher confidence than both YOLOv4 and YOLOv7 algorithms, demonstrating superior classification performance. It also effectively resolves the detection challenges associated with smaller helmet objects, with all small objects in the four images being successfully detected. Furthermore, the YOLO-PL algorithm outperforms YOLOv7 when the target is occluded, showcasing impressive robustness.

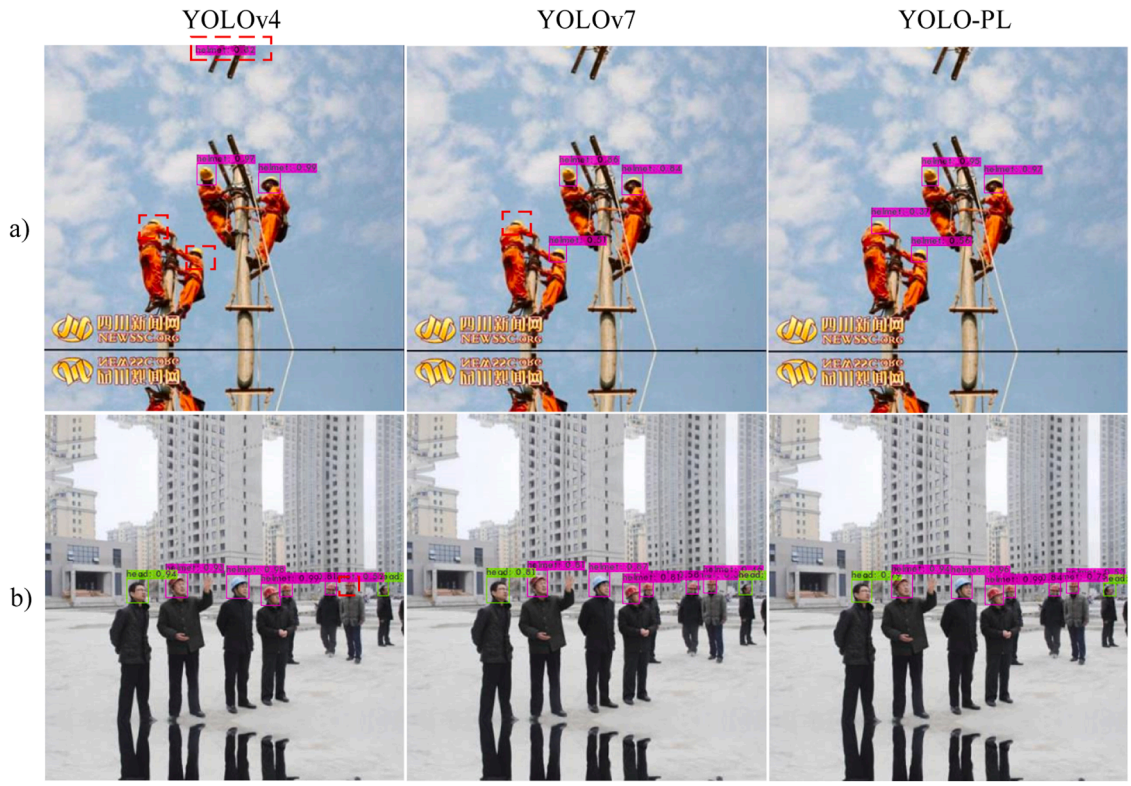


Fig. 11. Detection result of the expanded SHD dataset.

Conclusion

In this study, we propose YOLO-PL, an enhanced helmet-wearing detection algorithm that outperforms YOLOv4 in terms of higher AP value but smaller size. The algorithm presented in this study is specifically designed to address the detection of safety helmet-wearing in work environments characterized by small, severely occluded objects. By adjusting the network architecture and implementing a multi-scale feature fusion strategy, the YOLO-P model achieves enhanced precision in detection. Subsequently, by lightweighting YOLO-P without compromising its performance, we derived the YOLO-PL algorithm.

During extensive experimentation, YOLO-PL consistently demonstrated superior performance to YOLOv4 across several metrics, including AP_{50} and recall. Notably, its inference performance is on par with the YOLO-P. Moreover, YOLO-PL surpasses the advanced YOLOv7 algorithm in AP_{50} and recall while maintaining a smaller size, making it an ideal choice for deploying helmet detection tasks in real-world scenarios.

Despite the high performance exhibited by our algorithm in helmet detection experiments, we acknowledge that there remains space for enhancement. Firstly, future research could explore deeper into capitalizing on specific helmet characteristics to bolster the recognition of helmets against complex backdrops. Secondly, an combination of YOLO-PL with other vision algorithms, such as incorporating additional image augmentation algorithm, could be explored to attain heightened detection precision. Lastly, additional lightweight design strategies can be explored in the future to make the algorithm more suitable for deployment on resource-restricted devices while maintaining high performance.

Overall, we believe that the YOLO-PL algorithm presented in this research represents a novel and efficient solution for the challenge of helmet wear detection in practical scenarios, and it also gives a direction for future related research in this domain.

Funding

Natural Science Foundation of Hebei Province (F2019203195). National Natural Science Foundation of China (62106214).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Not applicable.

References

[1] M. Park, I. Brilakis, Construction worker detection in video boxes for initializing vision trackers, Autom. Constr. 28 (15) (2012) 15–25.
[2] A.H.M. Rubaiyat, T.T. Toma, M. Kalantari-Khandani, S.A. Rahman, L. Chen, Y. Ye, C.S. Pan, Automatic detection of helmet uses for construction safety, in: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence Workshops, 2016, pp. 135–142.
[3] X.H. Liu, X.N. Ye, Application of skin color detection and Hu moment in helmet recognition, J. East China Univ. Sci. Technol. 3 (2014) 365–370.
[4] L. Alzubaidi, J. Zhang, A.J. Humaidi, et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021) 53, https://doi.org/10.1186/s40537-021-00444-8.
[5] B.E. Mneymneh, M. Abbas, H. Khoury, Automated hardhat detection for construction safety applications, Proc. Eng. 196 (2017) 895–902.
[6] P. Chansik, L. Doyeop, K. Numan, An analysis on safety risk judgment patterns towards computer vision based construction safety management, in: Proc. Creative Construct. e-Conf, Budapest, Hungary, Budapest University of Technology and Economics, 2020, pp. 31–38.

[7] X. Wang, X. Jia, C. Jiang, et al., A wafer surface defect detection method built on generic object detection network, Digit. Signal Process. 130 (2022), 103718.
[8] H. Xia, J. Ma, J. Ou, et al., Pedestrian detection algorithm based on multi-scale feature extraction and attention feature fusion, Digit. Signal Process. 121 (2022), 103311.
[9] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2018) 1–9.
[10] J. Shen, X. Xiong, Y. Li, W. He, P. Li, X. Zheng, Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning, Computer-Aided Civ. Infrastruct. Eng. 36 (2) (2021) 180–196.
[11] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, F. Huang, DSFD: dual shot face detector, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun 2019, 2019, pp. 5055–5064.
[12] L. Huang, Q. Fu, M. He, D. Jiang, Z. Hao, Detection algorithm of safety helmet wearing based on deep learning, Concurr. Comput. 33 (13) (2021) e6234.
[13] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, A. Knoll, A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal. *IEEE Transactions on systems, man, and cybernetics*, systems 52 (2) (2020) 936–953.
[14] H. Wang, H. Ge, M. Li, PFG-YOLO: a safety helmet detection based on YOLOv4, in: 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), October 2021 5, IEEE, 2021, pp. 1242–1246.
[15] B. Wang, H. Xiong, L. Liu, Safety helmet wearing recognition based on improved YOLOv4 algorithm, in: 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC), March 2022 6, IEEE, 2022, pp. 1732–1736.
[16] L. Shen, H. Tao, Y. Ni, Y. Wang, V. Stojanovic, Improved YOLOv3 model with feature map cropping for multi-scale road object detection, Meas. Sci. Technol. 34 (4) (2023), 045406.
[17] A. Benjumea, I. Teeti, F. Cuzzolin, A. Bradley. (2021).YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles. arXiv:2112.11798.
[18] Ultralytics (2020) YOLOv5 2020 Available from: https://github.com/ultralytics/yolov5.
[19] H. Tao, L. Cheng, J. Qiu, V. Stojanovic, Few shot cross equipment fault diagnosis method based on parameter optimization and feature metric, Meas. Sci. Technol. 33 (11) (2022), 115005.
[20] Bochkovskiy A., Wang C.Y., Liao H.Y.M. YOLOv4: optimal speed and accuracy of object detection. arxiv preprint arXiv:2004.10934, 2020.
[21] Y. Lee, J. Hwang, S. Lee, et al., An energy and GPU-computation efficient backbone network for real-time object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–5.
[22] Wang, C.Y. Bochkovskiy, A., & Liao, H. (2022). YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv e-prints.
[23] Ramachandran P., Zoph B., Le Q.V. Searching for activation functions, arXiv:1710.05941, 2017.
[24] C.Y. Wang, H.Y.M. Liao, Y.H. Wu, et al., CSPNet: a new backbone that can enhance learning capability of CNN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391.
[25] Li H., Xiong P., An J., et al. Pyramid attention network for semantic segmentation. arxiv preprint arXiv:1805.10180, 2018.
[26] Redmon J., Farhadi A. YOLOv3: an incremental improvement. arxiv preprint arXiv:1804.02767, 2018.
[27] Z. Zheng, P. Wang, W. Liu, et al., Distance-IoU loss: faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence 34, 2020, pp. 12993–13000.
[28] anonymous. Designing network design strategies. anonymous submission, 2022. 3.
[29] B. Mahaur, K.K. Mishra, Small-object detection based on YOLOv5 in autonomous driving systems, Pattern Recognit. Lett. 168 (2023) 115–122.
[30] T.Y. Lin, P. Dollar, R. Girshick, et al., Feature Pyramid Networks for Object Detection, IEEE Computer Society, 2017.
[31] G. Huang, Z. Liu, L. Van Der Maaten, et al., Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[32] Njvisionpower. "NJVISIONPOWER/Safety-helmet-wearing-dataset: safety helmet wearing detect dataset, with pretrained model." GitHub. https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset. Accessed 18 Feb 2022.
[33] W. Sun, L. Dai, X. Zhang, P. Chang, X. He, RSOD: real-time small object detection algorithm in UAV-based traffic monitoring, Appl. Intell. (2021) 1–16.
[34] C.Y. Wang, A. Bochkovskiy, H.Y.M. Liao, Scaled-YOLOv4: scaling cross stage partial network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13029–13038.

Haibin Li is a professor at the Institute of Electrical Engineering at Yanshan University in Qinhuangdao, China. His current research interests lie in the area of computer vision.

Dengchao Wu was born in Shuozhou, Shanxi , China. He received Master's degree in Electronic and Information Engineering from Yanshan University, Qinhuangdao, China, in 2023. His research interest lie in computer vision.

Wenming Zhang is an associate professor at the Institute of Electrical Engineering at Yanshan University in Qinhuangdao, China. His current research interests are in the area of pattern recognition.

Cunjun Xiao is a doctoral student of Yanshan University. His research interest lie in computer vision.