TYPE Original Research
PUBLISHED 11 April 2025
DOI 10.3389/fbuil.2025.1563483

SD-YOLOv5: a rapid detection method for personal protective equipment on construction sites

ChunYa Li[1,2], Jianhua Wang[2,3]*, Bingfeng Luo[4], Tubing Yin[3], Baohua Liu[2,3] and Jianfei Lu[3]

[1]Wuhan University of Technology School of Management, Wuhan, China, [2]Shenzhen Yantian Port Real Estate Co., Ltd., Shenzhen, China, [3]Central South University School of Resources and Safety Engineering, Changsha, China, [4]Shenzhen Port Group Co., Ltd., Shenzhen, China

OPEN ACCESS

EDITED BY
Muhammad Junaid Munir,
RMIT University, Australia

REVIEWED BY
Pengfei Shan,
Xi'an University of Science and Technology, China
Syed Farhan Alam Zaidi,
Chung-Ang University, Republic of Korea
Pranali Kosamkar,
MIT World Peace University, India

*CORRESPONDENCE
Jianhua Wang,
✉ 393981086@qq.com

RECEIVED 20 January 2025
ACCEPTED 26 March 2025
PUBLISHED 11 April 2025

CITATION
Li C, Wang J, Luo B, Yin T, Liu B and Lu J (2025) SD-YOLOv5: a rapid detection method for personal protective equipment on construction sites.
Front. Built Environ. 11:1563483.
doi: 10.3389/fbuil.2025.1563483

With the rapid growth of urbanization, construction sites are increasingly confronted with severe safety hazards. Personal protective equipment (PPE), such as helmets and safety vests, plays a critical role in mitigating these risks; however, ensuring proper usage remains challenging. This paper presents SD (Small object detection and DilateFormer attention mechanism)-YOLOv5s, an improved PPE detection algorithm based on YOLOv5s, designed to enhance the detection accuracy of small objects, such as helmets, in complex construction environments. The proposed model incorporates a dedicated feature layer for small target detection and integrates the DilateFormer attention mechanism to balance detection performance and computational efficiency. Experimental results on the CHV dataset demonstrate that SD-YOLOv5s achieves an average precision (AP) of 93.7%, representing an improvement of 2.8 percentage points over the baseline YOLOv5s (AP = 90.9%), while reducing the model's parameter count by up to 14.6%. These quantitative improvements indicate that SD-YOLOv5s is a promising solution for real-time PPE monitoring on construction sites, although further validation on larger and more diverse datasets is warranted.

KEYWORDS

personal protective equipment (PPE) detection, YOLOv5, small object detection, DilateFormer attention mechanism, construction site safety

1 Introduction

In the process of urbanization, a significant number of workers are employed on construction sites, which often present numerous safety hazards (Cheng, 2024; Jia et al., 2025). Consequently, construction safety has consistently been a major concern, particularly in environments where workers frequently operate in dangerous conditions. According to statistics, the injury rate on construction sites exceeds 71% (Waehrer et al., 2007; Ahmed, 2019; Hwang et al., 2023; Soltanzadeh and Mohammadfam, 2022). However, the proper use of personal protective equipment (PPE) such as helmets, safety vests, and other items can mitigate these risks. For instance, correctly worn helmets not only protect against the impact of falling objects but also significantly reduce the severity of injuries from falls, potentially saving workers' lives (Hume et al., 1995). Safety vests, another essential form of PPE, enhance visibility, especially in low-light conditions, thereby helping to prevent accidents. Additionally, the colors of safety helmets indicate different job roles, facilitating the smooth execution of tasks (Wang et al., 2021). Despite the clear benefits, workers on construction sites often neglect to wear PPE, leading to increased dangers. Ensuring the correct use of PPE is crucial for reducing accidents. However, relying solely on manual inspections to enforce compliance is both time-consuming and inefficient (Yang et al., 2024).

In recent years, the rapid advancement of artificial intelligence technology and the significant improvement in computer processing power have made intelligent detection a prominent area of current research (Aradhya and Ravish, 2019; Masita et al., 2020; Chandan et al., 2018). Computer vision has emerged as the mainstream method for target detection, characterized by its contactless nature, high accuracy, and continuity. With the enhancement of computational power and the continuous optimization of algorithms, convolutional neural networks have become the leading approach in deep learning technology (Du, 2018; Chen et al., 2017). The development of target detection technology can be divided into two stages: the traditional target detection era before 2014 and the deep learning-based detection era post-2014. This technology has found widespread applications in fields such as security detection, intelligent perception, and autonomous driving. Target detection methods are primarily categorized into one-stage methods (e.g., the YOLO model) and two-stage methods (e.g., the R-CNN model). One-stage methods are known for their fast detection speed, making them suitable for real-time applications (Diwan et al., 2023), while two-stage methods excel in accurately detecting small objects, offering higher detection accuracy (Ren et al., 2017). Both approaches have distinct advantages and are extensively utilized in the detection of personal protective equipment.

The two-stage target detection method divides the entire detection process into two phases as shown in Figure 1: first, generating high-quality candidate regions, and then conducting further feature extraction on the selected windows, followed by classification and window regression based on the extracted features. In 2014, Girshick et al. (2014) proposed the R-CNN target detection algorithm, which identifies and locates targets by using selective search to generate candidate regions, employing SVM classifiers and bounding box regressors for target identification and localization, thereby significantly improving efficiency. Subsequently, in 2015, Girshick (2015) introduced Fast R-CNN, which integrates feature extraction, classification, and regression into a single network, accelerating both the training and testing processes while enhancing memory efficiency. Ren et al. (2017) advanced this further with the Faster R-CNN algorithm, introducing the Region Proposal Network (RPN) to replace traditional selective search, thereby achieving faster candidate region generation. In the context of worker safety, Madihah Saudi et al. (2020) developed an image detection model based on the R-CNN algorithm to assess PPE compliance, with experimental results showing an average accuracy of 70%. Riaz et al. (2023) proposed a novel method called PPE_Swin, which automatically detects PPE on construction sites by combining the Swin-Unet self-attention mechanism with global and local feature extraction, achieving a detection accuracy of 97%. Xiong and Tang (2021) introduced a scalable pose-guided anchoring framework for detecting multi-class PPE, utilizing a pose estimator and body-knowledge-based rule compliance, alongside a shallow CNN classifier to identify PPE classes, demonstrating high detection accuracy and scalability on the CPPE dataset. While the two-stage method excels in the accuracy of target detection, its slower detection speed limits its practical implementation.

In order to improve in the problem of a large number of small targets and a large Traditional one-stage methods primarily include the SSD and YOLO models. In 2016, Liu et al. (2016) introduced the SSD (Single Shot MultiBox Detector) target detection algorithm, which eliminated the proposal generation and resampling phases by integrating all computations into a single network, thereby simplifying the training process. Tests on the PASCAL VOC, COCO, and ILSVRC datasets demonstrated that SSD competes effectively with traditional methods in terms of speed and accuracy, making it widely used for the detection of safety and protective equipment.

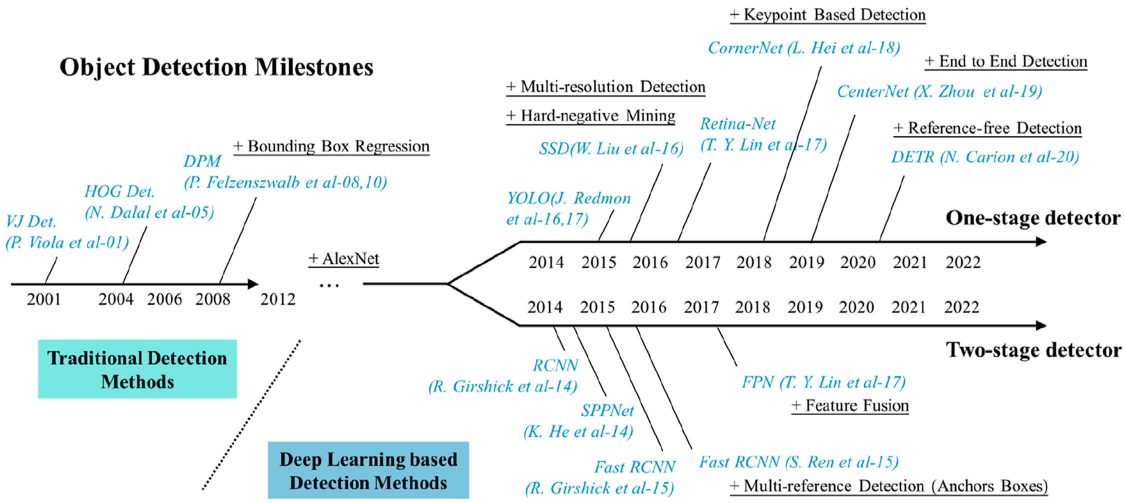
Object Detection Milestones
VJ Det. (P. Viola et al-01)
HOG Det. (N. Dalal et al-05)
DPM (P. Felzenszwalb et al-08,10)
+ Bounding Box Regression
2001
2004
2006
2008
2012
+ AlexNet
…
Traditional Detection Methods
Deep Learning based Detection Methods
+ Keypoint Based Detection
CornerNet (L. Hei et al-18)
+ End to End Detection
CenterNet (X. Zhou et al-19)
+ Multi-resolution Detection
+ Hard-negative Mining
SSD(W. Liu et al-16)
Retina-Net (T. Y. Lin et al-17)
+ Reference-free Detection
DETR (N. Carion et al-20)
YOLO(J. Redmon et al-16,17)
One-stage detector
2014 2015 2016 2017 2018 2019 2020 2021 2022
2014 2015 2016 2017 2018 2019 2020 2021 2022
Two-stage detector
RCNN (R. Girshick et al-14)
SPPNet (K. He et al-14)
Fast RCNN (R. Girshick et al-15)
FPN (T. Y. Lin et al-17)
+ Feature Fusion
Fast RCNN (S. Ren et al-15)
+ Multi-reference Detection (Anchors Boxes)

FIGURE 1
Development process of target detection.

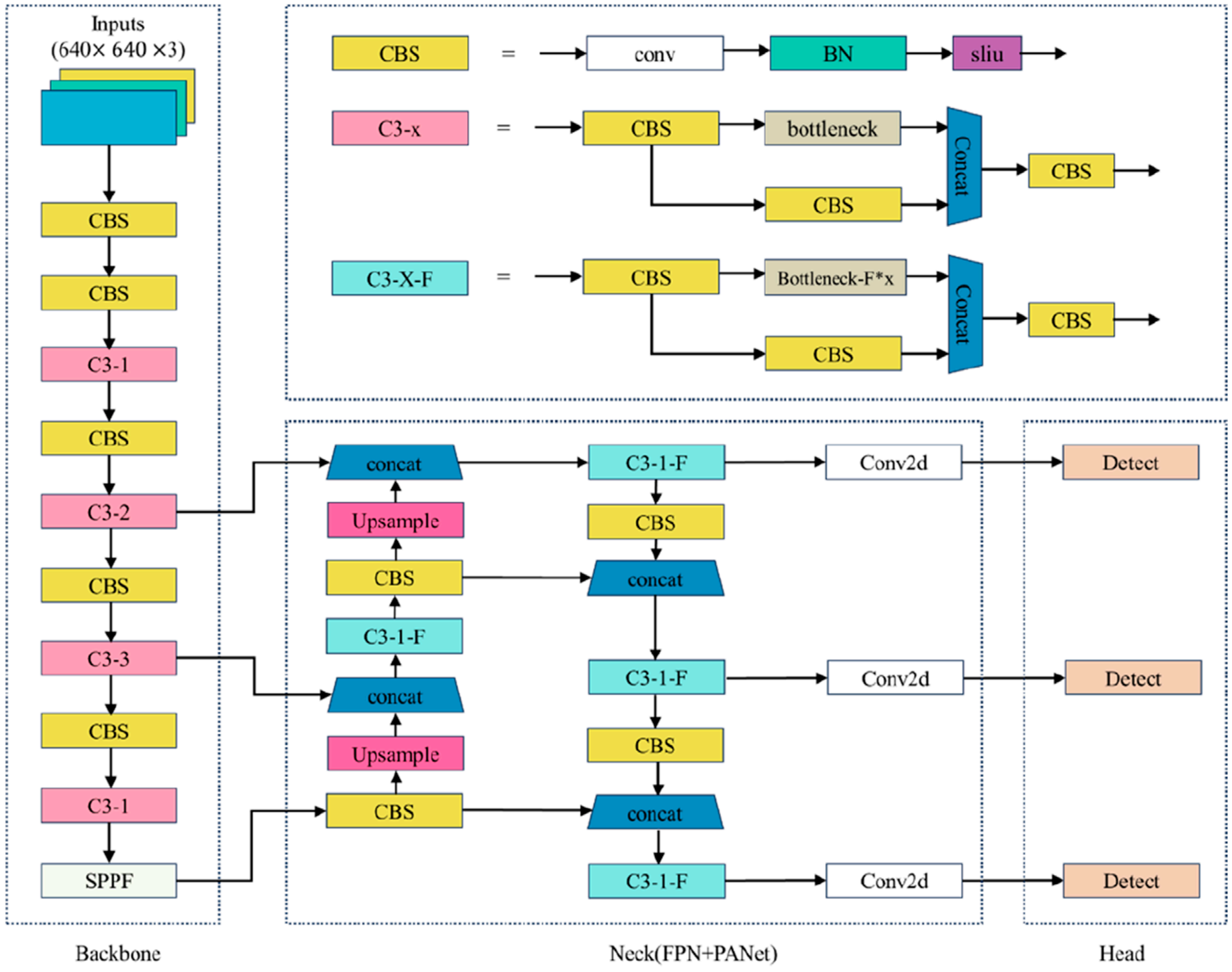

FIGURE 2
YOLOv5s model.

Wu et al. (2019) implemented color inspection of safety helmets using the SSD algorithm. In the same year, Redmon et al. (2016) proposed the YOLOv1 model at the IEEE International Conference on Computer Vision and Pattern Recognition. This model treated object detection as a regression problem involving spatially separated bounding boxes and correlated class probabilities, achieving a detection speed of up to 45 fps but with low detection accuracy. To address these issues, YOLOv2 was developed, which improved precision and recall (Redmon and Farhadi, 2017). YOLOv3 further enhanced YOLOv2 by utilizing multi-scale feature maps for detection and replacing the softmax function with an independent logistic regression classifier for prediction category classification, effectively improving prediction accuracy (Redmon and Farhadi, 2018). YOLOv4 introduced more advanced techniques such as CSPDarknet53, SPP, PANet, and Mish activation functions, along with more sophisticated data augmentation and automatic learning rate tuning strategies to enhance detection accuracy and speed (Bochkovskiy et al., 2020). YOLOv5 improved target detection performance and usability by incorporating a lightweight network architecture, advanced data augmentation techniques, and model export support (Zhu et al., 2021). Although the YOLO model demonstrated superior capabilities in detection speed and accuracy, it encountered challenges in detecting small targets. Numerous researchers have since conducted studies to address these shortcomings. Wang et al. (2023) propose an improved YOLOX method and a new dataset for detecting low light and small PPE. The ConvNeXt module is added to the backbone for deep feature extraction, a fourth YOLOX header is introduced to enhance multiscale prediction, and the CLAHE algorithm is employed to enhance the low-light images to achieve a higher performance detection. Yang and Wang (2022) proposed an improved helmet detection algorithm based on YOLOv4, which significantly enhances the detection accuracy of small and occluded targets by improving multi-scale feature extraction, introducing a channel attention module, and optimizing model training with the Eiou loss function and K-means clustering. Li et al. (2023) introduced a novel lightweight helmet detection algorithm, YOLO-PL, based on YOLOv4, which improves small target detection accuracy, reduces model parameters, and enhances the robustness and deployability

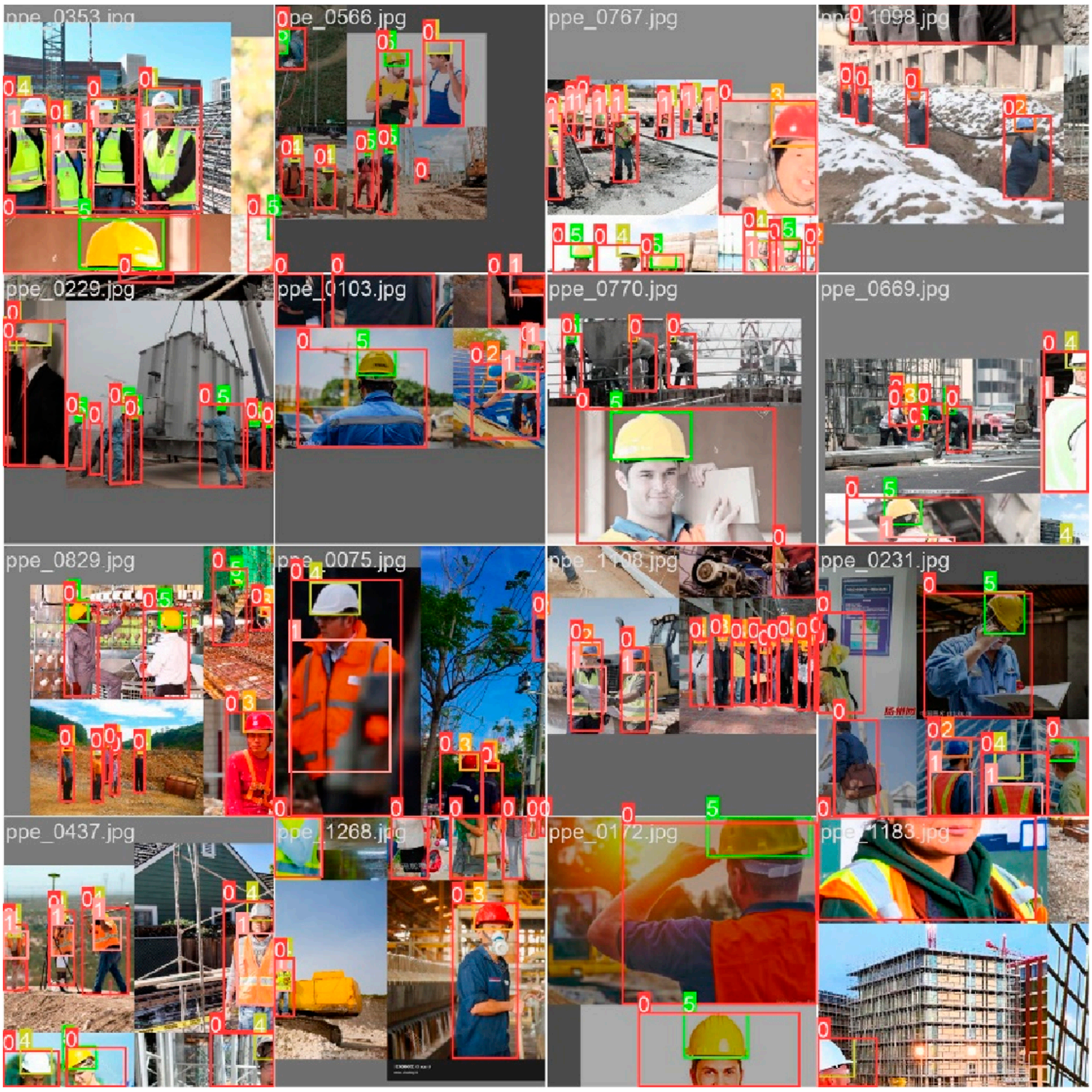

FIGURE 3
Mosaic algorithm for merging data.

of the model by optimizing the network structure, introducing E-PAN and DCSPX modules, and designing the lightweight VoVNet (L-VoVN) structure. This algorithm outperforms existing detectors in helmet detection and shows potential for practical applications in the industry. Qian and Yang (2023) proposed a lightweight model, YOLO CA, based on YOLOv5, which automates the detection of construction workers' helmet usage by incorporating enhancements such as coordinate attention (CA), depthwise separable convolution (DWConv), and the Ghost module. These improvements significantly boost detection accuracy and model efficiency in complex scenarios while reducing model parameters, making it suitable for lightweight embedding applications. Lian et al. (2024) introduced a novel deep learning framework called HR-YOLO, which substantially improves the accuracy of helmet detection, particularly in complex environments and small target detection, by combining helmet object features with human pose information and optimizing residual networks with the Laplace Perceptual Attention Model (LAAM). Chen et al. (2023) proposed an improved convolutional neural network model, YOLOv7-WFD, for detecting workers not wearing helmets, enhancing feature extraction capability, detail reconstruction, and model generalization through the introduction of the DBS module, Content-Aware ReAssembly of Features (CARAFE) module, and Wise-IoU loss function.

These studies have successfully applied deep learning computer vision to buildings and industries, although especially in the field of small target detection and fast detection, researchers have made relatively significant progress, however, often the introduction of small target detection will somewhat increase the computational cost of the model, therefore, we would like to

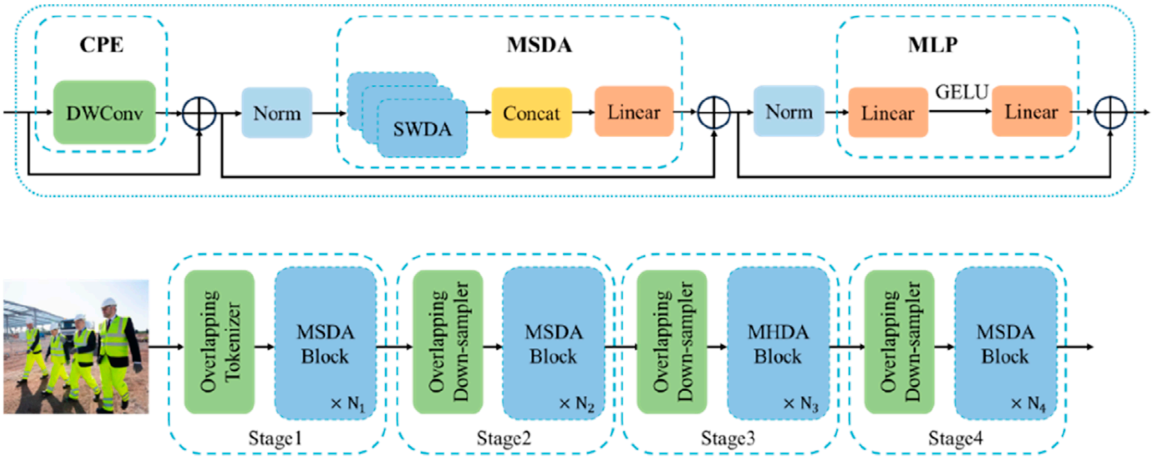

FIGURE 4
Overall architecture of DilateFormer.

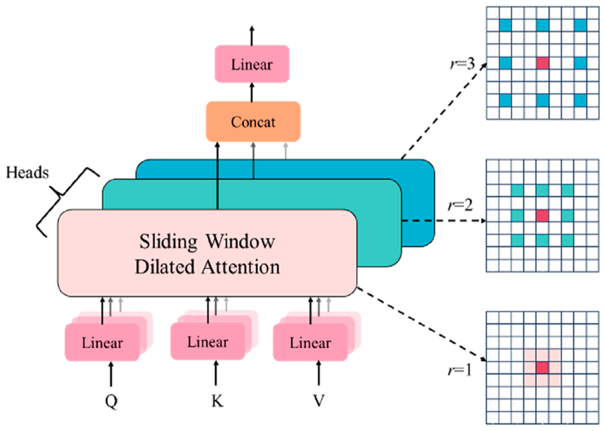

FIGURE 5
Graphical representation of Multi-Scale Dilated Attention (MSDA).

adopt a method that makes it possible to keep the computational cost of the model does not increase after adding the small target detection, i.e., to ensure that the real-timedetection to improve the performance of the overall model ultimately achieving faster detection coordinated with smaller computational cost, based on this to improve the detection performance of YOLOv5s for small targets (e.g., helmets) in complex architectural environments, this paper proposes the MSD-YOLOv5s algorithm, with the main contributions as follows.

1. To enhance the detection of small targets, such as safety helmets, in construction site environments, a dedicated feature layer for small target detection has been added to the YOLOv5s model, significantly improving detection accuracy.
2. The DilateFormer attention mechanism is introduced, reducing the redundancy of the self-attention mechanism, enhancing model accuracy, and lowering computational costs by integrating multi-scale dilation attention with multi-head self-attention.
3. Based on these improvements to the YOLOv5s model, termed SD-YOLOv5s, a PPE detection model is established, achieving an average precision (AP) of 93.7%.

2 Methods

YOLOv5 is currently the most widely used object detection model, as mentioned previously, due to its fast label detection algorithm (Wu et al., 2021; Zhang and Yin, 2022). Based on the depth of its network and the width of its feature maps, YOLOv5 is classified into several models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, all of which share the same network structure, consisting of a backbone, neck, and head (Zhou et al., 2021; Chen et al., 2022). Among these, YOLOv5s boasts the highest detection speed, while YOLOv5x achieves the highest detection accuracy. Compared to other models, YOLOv5s maintains a high level of accuracy with a more lightweight architecture, which has contributed to its widespread adoption. The YOLOv5s model is composed of four main components: input, trunk, neck, and prediction. YOLOv5 employs a standalone CNN model for end-to-end target detection. First, input images are standardized to a uniform size after data augmentation and then fed into the CNN network. The network outputs prediction results at three different scales, each corresponding to N channels containing prediction information. The network's prediction is then processed through network management operations to obtain the detection targets. Finally, the prediction results are refined using Non-Maximum Suppression (NMS) to finalize the detected targets. The structure of the network module is illustrated in Figure 2. Key components include Conv2d (2D convolution), BN (batch normalization), SiLU (activation function), Upsample (upsampling), and SPPF (a modified Spatial Pyramid Pooling module). Although the SPPF module offers fast computation, it remains computationally

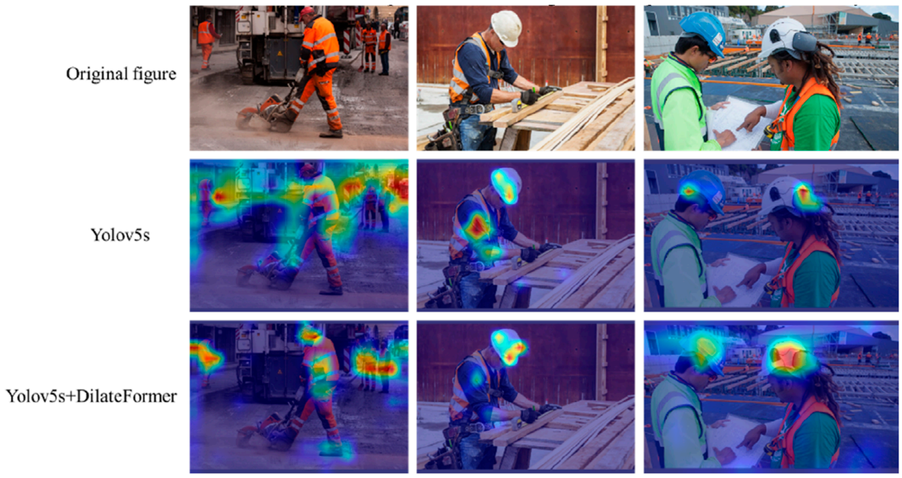

FIGURE 6
Comparison of DilateFormer before and after adding YOLOv5s.

intensive, necessitating modifications to the network's backbone to address this issue.

2.1 The mosaic algorithm

In the field of small target detection, optimizing the dataset through data augmentation is particularly important. First, data augmentation increases the diversity of the dataset, and second, it helps the model achieve better generalization. In this paper, the mosaic data augmentation method is employed, which enhances the dataset by randomly cropping, rotating, and joining any four images. This approach increases the variety of complex scenes and small targets, thereby improving the model's generalization ability and robustness. Figure 3 illustrates the enhanced image.

2.2 Backbone improvement

DilateFormer is a deep learning model based on a pyramid structure, primarily designed for processing fundamental visual tasks. The key concept behind its design is to utilize multi-scale dilated attention to capture multi-scale semantic information while reducing the redundancy of the attention mechanism (Jiao et al., 2023). As illustrated in Figure 4, the model consists of four main stages. In the first two stages, multi-scale dilated attention plays a crucial role, while the latter two stages employ standard multi-head self-attention. Upon image input, DilateFormer first applies an overlapping downsampler for patch embedding, where the resolution of input feature maps is adjusted by alternating the step size of the convolutional kernel. For initial patches, an overlapping downsampler with a kernel size of three and a step size of two is utilized. Conditional Positional Embedding (CPE) is employed throughout the model to adapt positional encoding to inputs of varying resolutions. The overall architecture of the model is as follows:

$$X = CPE(\hat{X}) + \hat{X} = DwConv(\hat{X}) + \hat{X}$$

$$Y = \begin{cases} MSDA(Norm(X)) + X, atlow - levelstages, \\ MHSA(Norm(X)) + X, atlow - levelstages \end{cases}$$

$$Z = MLP(Norm(Y)) + Y$$

Where $\hat{X}$ represents the input of the current block, either the image block or the output from the previous block. In practice, we implement the Conditional Positional Embedding (CPE) with zero padding and a 3 × 3 kernel size. We have also added a Multi-Layer Perceptron (MLP) to the previous work (Liu et al., 2021; Touvron et al., 2021), which consists of two linear layers with a channel expansion rate of four and a GELU activation function.

The core component of DilateFormer is its Multi-Scale Dilated Attention (MSDA) module. As illustrated in Figure 5, the MSDA module employs a multi-head design, where the channels of the feature map are divided into nnn different heads, and Sliding Window Dilated Attention (SWDA) is applied using different dilation rates for each head. This approach aggregates semantic information at various scales within the receptive field and effectively reduces the redundancy of the self-attention mechanism, all without the need for complex operations or additional computational costs. The specific operations are as follows: First, each head is assigned a distinct dilation rate. Next, slices are taken from the feature map, and SWDA is applied to obtain the output. Finally, the outputs from all heads are concatenated, and feature aggregation is performed through a linear layer.

DilateFormer effectively addresses the long-range dependency problem through its hybrid use of multi-scale dilated attention and multi-head self-attention, while maintaining computational

TABLE 1 Comparison of small target detection before and after incorporating YOLOv5s.

Models	mAP@0.5		
	Person	Vest	Helmet
YOLOv5s	92.6%	85.1%	93.8%
YOLOv5s + small	92.6%	92.6%	92.6%

efficiency and adaptability to inputs of varying scales and resolutions. As shown in Figure 6, the red regions indicate areas requiring high attention during feature extraction, with darker colors signifying greater significance. From the experimental results, it is evident that the inclusion of the DilateFormer module enhances the feature representation of targets in complex scenarios, as clearly illustrated in Figure 6.

2.3 Neck network improvement

The original YOLOv5 model includes three feature layers: 80 × 80, 40 × 40, and 20 × 20, corresponding to receptive fields of 32 × 32, 16 × 16, and 8 × 8, respectively. However, in practical applications, especially in dense scenes, some targets in the dataset are smaller than 8 × 8 pixels, such as helmets. The shallow feature information in these cases cannot be fully utilized, resulting in insufficient accuracy for small target recognition. To address this issue, a dedicated detection layer for small target detection is added to the YOLOv5s structure. This new feature layer has a size of 160 × 160, with a receptive field of 4 × 4, enabling the detection of targets as small as 4 × 4 pixels (Shan et al., 2024). In this dataset, helmets of different colors are considered smaller-scale objects. Table 1 lists the mean Average Precision (mAP) values for persons, vests, and helmets. The results show a significant improvement in the detection accuracy of helmets and vests after the addition of the small target detection layer. The improved SD-YOLOv5s model is shown in Figure 7.

2.4 Evaluation indicators

Detection accuracy and detection speed are critical indicators for evaluating model performance. Precision, recall, and average precision (AP) are key metrics used to assess the detection accuracy of a model, while frames per second (FPS) is an important metric for evaluating the speed of model detection. The relevant formulas are provided below (Jiang et al., 2022; Tian et al., 2019):

$$F = \frac{TP}{TP + FP} \times 100\%$$

$$R = \frac{TP}{TP + FN} \times 100\%$$

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

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$$

$$FPS = \frac{N}{t}$$

Where TP represents the number of correctly predicted positive samples, FP denotes the number of incorrectly predicted positive samples, and FN indicates the number of incorrectly predicted negative samples. The variable n refers to the number of target classes being tested, and $\mathrm{AP_i}$ is the Average Precision (AP) of the *i* target class.

3 Experimentation

3.1 Experimental environment

In this experiment, model construction, training, and validation were performed in a Windows environment using an NVIDIA Tesla T4 graphics card, with PyTorch version 1.8.2 and Python version 3.8.19. The input image size was set to 640 × 640, the weight decay coefficient was 0.0005, the initial learning rate was 0.01, and 300 epochs of optimization iterations were conducted.

3.2 Dataset characteristics

The dataset containing helmets of different colors was sourced from Wang et al. (Wang et al., 2021), and includes protective vests, four helmet colors (blue, red, white, and yellow), and images of individuals. This dataset is referred to as the CHV dataset. The majority of the images were captured at construction sites, providing an accurate reflection of real-world conditions. The dataset contains a total of 1,330 photographs with varying angles, distances, lighting conditions, and character states, encompassing a total of 9,209 instances. Figure 8 shows the distribution of images across the training, validation, and test sets.

The training set is used for the model's training process, the test set serves as a basis for evaluating the model's performance, and the validation set is utilized for model prediction. Figure 9 illustrates the percentage of each category label within the training, test, and validation sets.

The labeled visualization of the dataset distribution provides insights into the distribution of different categories of samples within the dataset. The top left image illustrates the dataset type, which has been described in detail in the previous section. The bottom left image shows the distribution of the objects' center of mass coordinates, emphasizing information about the focus and position of the samples. The top right image displays the bounding box coordinates in horizontal and vertical dimensions, aiding in understanding the dataset's skewness. The bottom right image shows the distribution of object sizes. This labeled visualization helps assess the reliability and usability of the dataset.

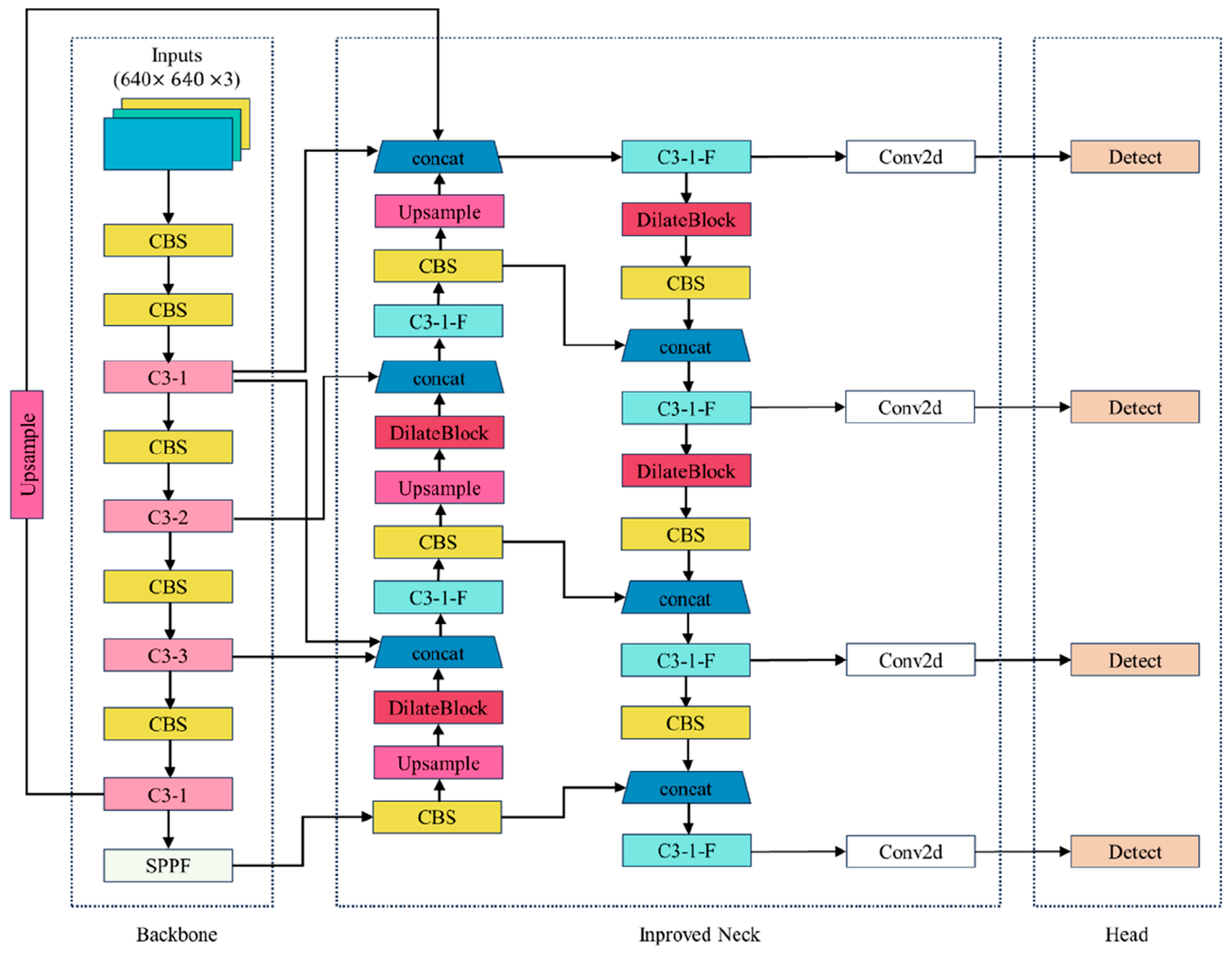

FIGURE 7
Improved YOLOv5s model.

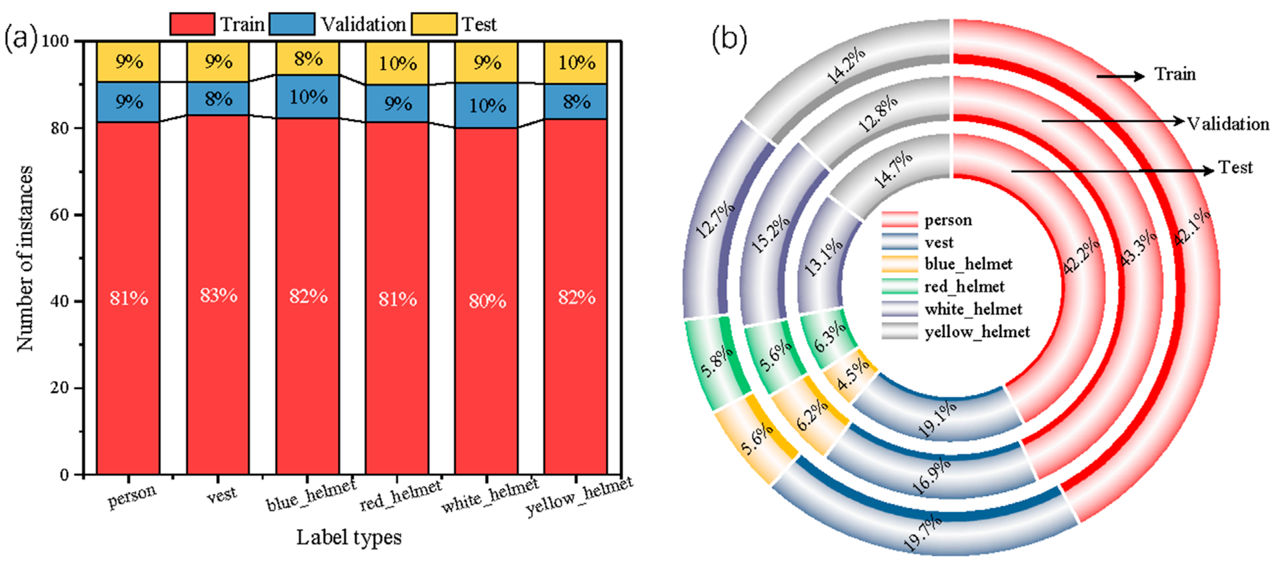

FIGURE 8
(a) Distribution of images in the training, test, and validation sets. **(b)** Percentage of each class in the training, test, and validation sets.

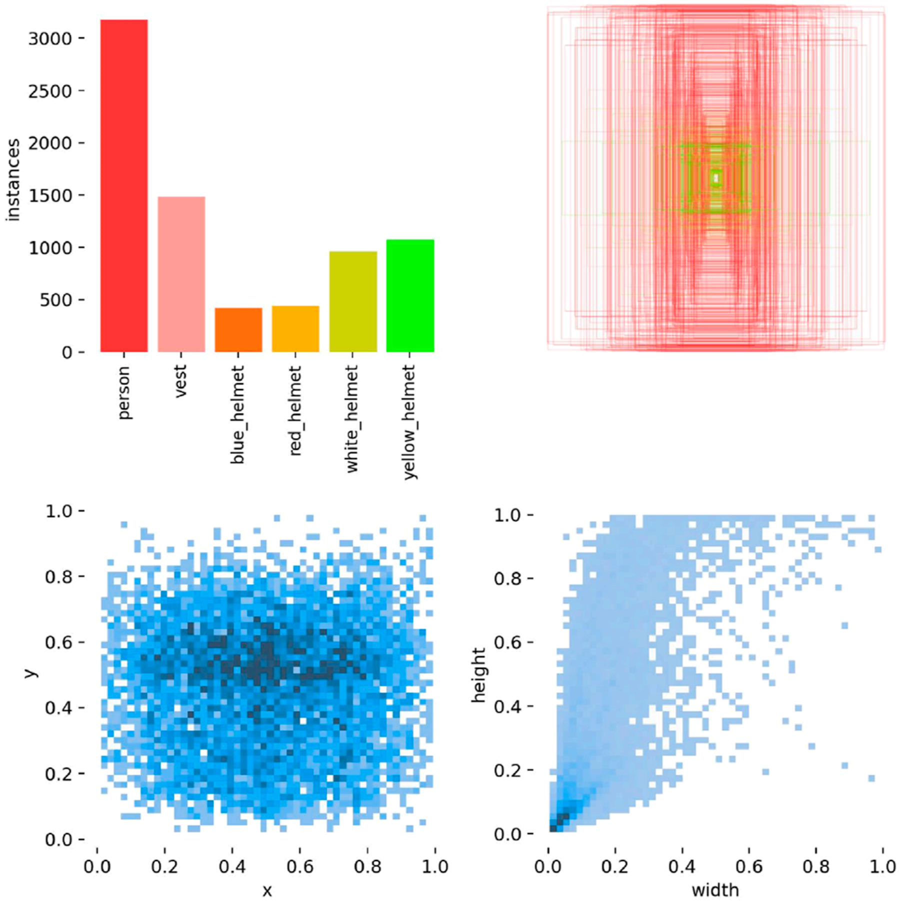

FIGURE 9
Distribution of label visualisations.

4 Results and discussion

4.1 Performance visualization

Figure 10 presents a visualization of the results from 300 rounds of training using the SD-YOLOv5s algorithm on the dataset. The images show minimal fluctuations in accuracy and recall metrics. The loss curves, including bounding box regression loss (box_loss), object confidence loss (obj_loss), and classification loss (cls_loss), for both the training and validation sets, converge gradually without significant fluctuations. This indicates that the model is neither overfitting nor underfitting during the training process. Additionally, the mAP_0.5 and mAP_0.5:0.95 accuracy metrics show a steady increase throughout the training, underscoring the robustness of the model.

Figure 11 shows the confusion matrix for both the improved model and the baseline model on the dataset. Compared to the baseline model, the improved model exhibits a reduced confusion rate and significantly enhanced classification accuracy across all categories.

The results of the image comparison before and after applying the SD-YOLOv5s algorithm are shown in Figure 12. The comparison clearly demonstrates that the improved model outperforms the YOLOv5 model, particularly in the detection of complex and small scenes. The improved model effectively reduces misdetections and omissions, exhibiting higher positional accuracy and robustness.

4.2 Results of improving the backbone network

To verify the effectiveness of the attention mechanism, various attention mechanisms, including CA (Wu et al., 2023), SE (Niu et al., 2023), and ECA (Zhang et al., 2023), were embedded into the algorithm. These attention mechanisms were integrated into the backbone network without any modifications to other parts of the model. The improved model was then tested and compared on the CHV dataset, with the experimental results presented in Table 2. Table 2 displays the detection accuracy and speed of YOLOv5s after integrating CA, SE, and ECA attention mechanisms. The results indicate that DilateFormer demonstrates the most superior performance under the same running environment and initial parameters. This superiority is attributed to the attention mechanism's ability to account for the locality and sparsity of the shallow self-attention mechanism, effectively aggregating multi-scale information

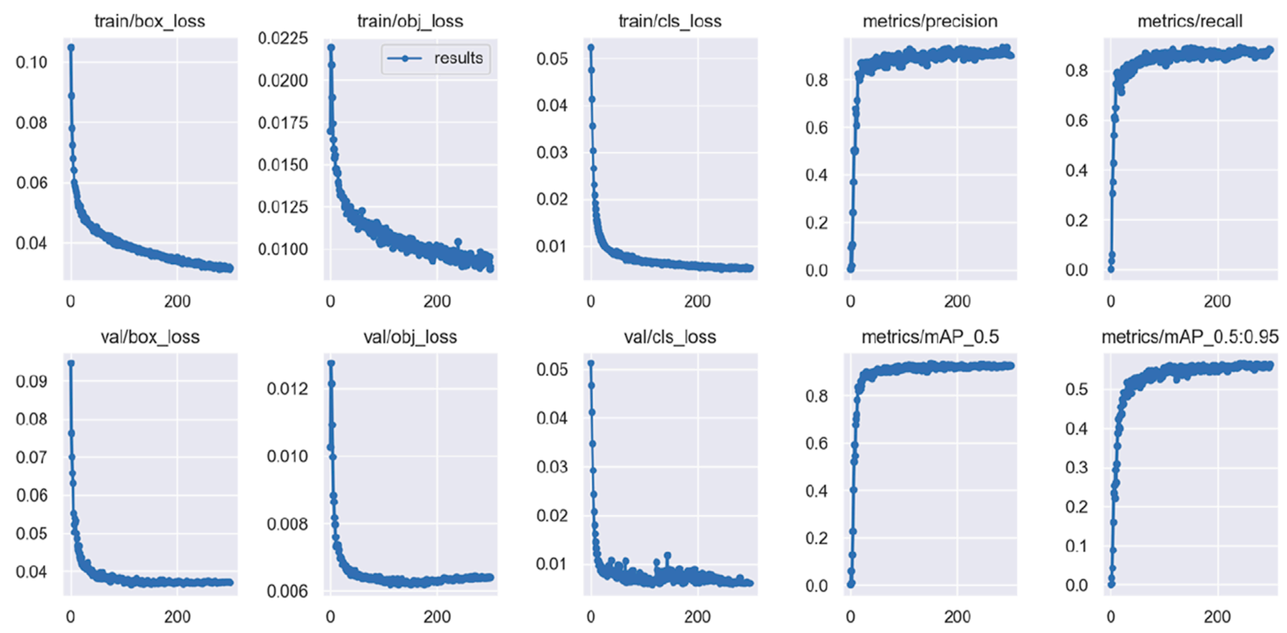

FIGURE 10
Visualisation of the improved YOLOv5s model training process.

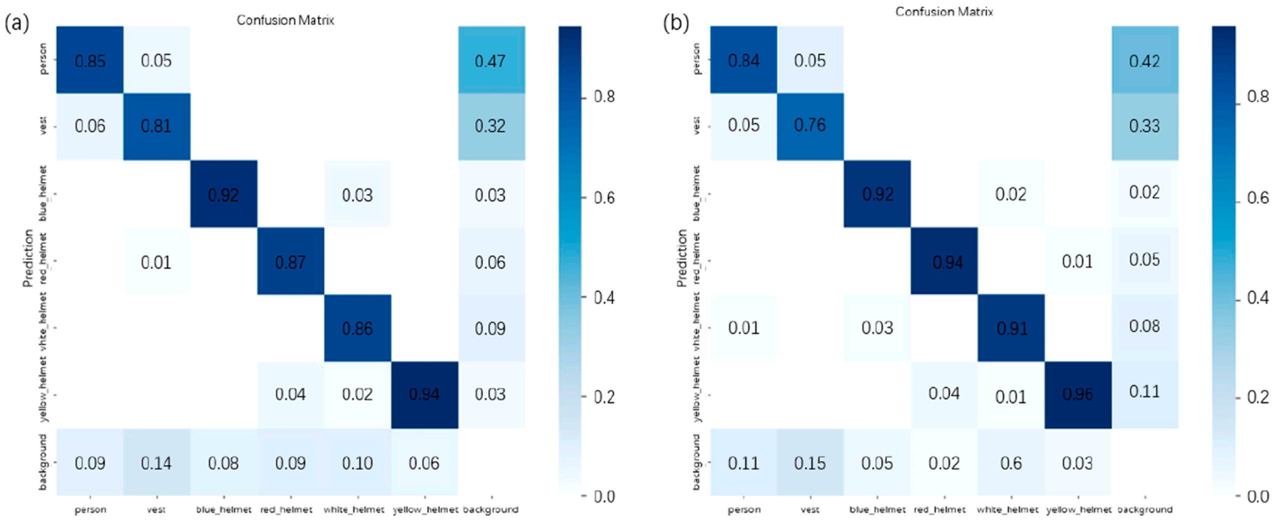

FIGURE 11
Confusion matrix for YOLOv5s and MSD-YOLOv5s. **(a)** YOLOv5s; **(b)** MSD-YOLOv5s.

while reducing redundancy. As a result, DilateFormer achieves leading computational performance with lower computational costs.

4.3 Ablation experiment

To explore the contribution of different components in our proposed model, an ablation test was conducted on the dataset as shown in Table 3. After applying the mosaic augmentation, the model's size and FLOPs remained unchanged, while accuracy improved by 0.8%, verifying that this improvement enhanced model accuracy by increasing the sample diversity. The inclusion of the small target detection layer led to a 1.1% improvement in the model's mAP. This enhancement allowed the model to learn more features related to small targets, significantly improving detection accuracy for these targets, albeit with a noticeable increase in model size. The introduction of DilatFormer contributed to a 1.7% increase in mAP, while reducing the model's parameters by 14.6% compared to the baseline model. Finally, when both DilateFormer and the small target detection layer were integrated into the model, the overall mAP increased by 2.3% compared to the baseline model, and the model size was also significantly reduced.

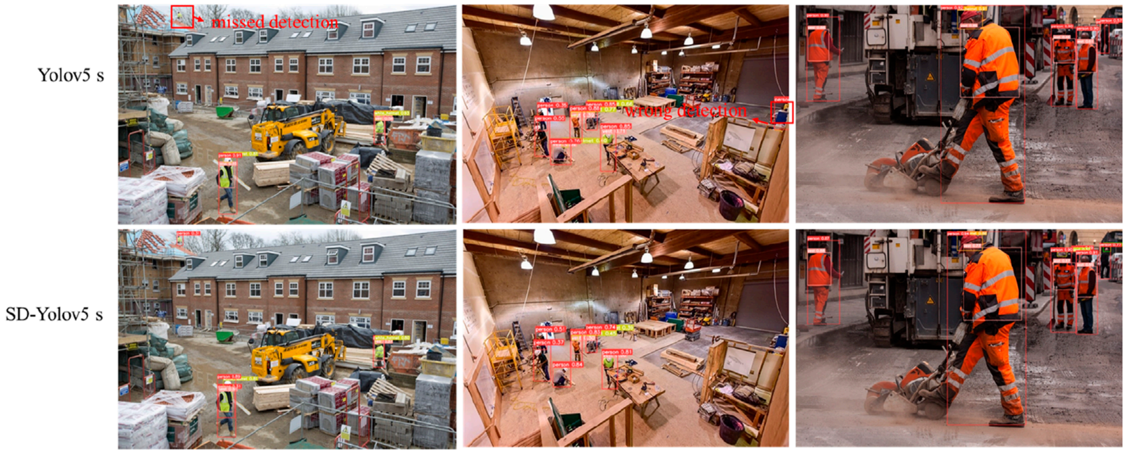

FIGURE 12
Comparison of detection effectiveness of improved algorithms.

TABLE 2 Comparison of different attention mechanisms.

Model	P	R	mAP@0.5	mAP@0.95	FPS
YOLOv5s	92.2%	85.8%	90.9%	53.0%	52.99
YOLOv5s + CA	92.7%	87.5%	91.5%	51.6%	50.26
YOLOv5s + SE	93.8%	87.0%	91.6%	51.5%	48.34
YOLOv5s + ECA	92.7%	87.9%	92.1%	53.7%	40.12
YOLOv5s + DF	92.9%	87.9%	92.6%	52.3%	58.25

4.4 Results of the comparative experiment

In this study, we analyze the YOLOv5, YOLOv8, YOLOv11 and YOLOv11 models in comparison with our proposed SD-YOLOv5s model. As shown in Figure 13, the SD-YOLOv5s model outperforms the other models in terms of mAP@0.50, demonstrating greater accuracy and stability, especially in the later stages of training. This improvement is attributed to the addition of a small target detection layer and the introduction of the DilateFormer attention mechanism, which effectively enhance the model's accuracy and robustness. In addition, in terms of model parameters, compared to the traditional YOLObv5 model, the improved model results in a significant reduction in model parameters, although its model parameters are higher compared to more advanced models such as YOLOv8, in that it is smaller and has optimal detection accuracy compared to the traditional model. Therefore, the SD-YOLOv5 model has superior detection accuracy and is suitable for use in complex construction scenarios that require real-time processing or resource constraints.

5 Conclusion and limitations

5.1 Conclusion

In conclusion, this study presents the SD-YOLOv5s model, an improved version of YOLOv5s specifically designed for detecting personal protective equipment in construction sites. By incorporating a dedicated small target detection layer, the method effectively addresses the challenges of identifying small and occluded objects in complex environments. Moreover, the integration of the Dilate-Former attention mechanism not only enhances detection accuracy but also reduces computational overhead, ensuring real-time performance. As a result, the SD-YOLOv5s model achieves an average precision (AP) of 93.7%, providing a highly effective solution for worker safety monitoring. Overall, the model demonstrates clear improvements in detection speed, accuracy, and reliability, underscoring its potential for practical deployment in safety monitoring systems and marking a significant advancement in intelligent detection technologies.

5.2 Limitations

While the SD-YOLOv5s model demonstrates significant improvements in detecting small and occluded targets, the study has several limitations. First, the dataset used for training and evaluation was relatively small, consisting mainly of images captured under controlled conditions. As a result, the model's generalization capability in varying real-world environments, such as nighttime construction or adverse weather conditions, remains uncertain. Second, the addition of the small target detection layer increases the computational complexity and inference time, which may hinder its deployment on low-power or edge devices. Lastly, while the DilateFormer attention mechanism helps reduce computational costs, it may not fully address the redundancy issues in more complex scenarios with highly dense objects. Future research should aim to incorporate additional datasets and further optimize the network structure to enhance detection robustness and computational efficiency.

TABLE 3 Ablation studies of different components in the improved model.

Model	mAP0.5 (%)	FPS	Parameters	FLOP(G)
YOLOv5s	91.6%	52.8	7.03	15.8
YOLOv5s + small	91.9%	41.90	7.17	18.6
YOLOv5s + DilateFormer	92.6%	58.26	6.00	13.7
YOLOv5s + small + DilateFormer	93.7%	45.14	6.09	15.4

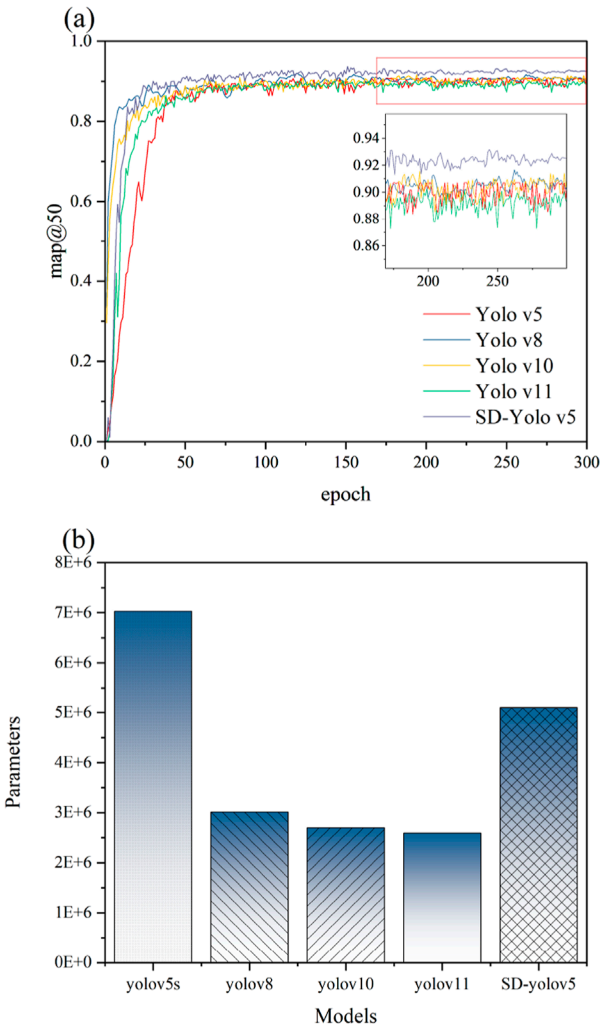

FIGURE 13
Comparison between different models **(a)** mAP values; **(b)** model parameters.

Data availability statement

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

Author contributions

CL: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing – original draft. JW: Investigation, Methodology, Software, Supervision, Validation, Writing – review and editing. BiL: Supervision, Validation, Writing – original draft. TY: Supervision, Validation, Writing – original draft. BaL: Supervision, Validation, Writing – original draft. JL: Supervision, Validation, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

Authors CL, JW, and BL were employed by Shenzhen Yantian Port Real Estate Co., Ltd. Author BL was employed by Shenzhen Port Group Co., Ltd.

The remaining 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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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

Ahmed, S. (2019). Causes of accident at construction sites in Bangladesh. *Organ. Technol. Manag. Constr.* 11 (1), 1933–1951. doi:10.2478/otmcj-2019-0003

Aradhya, H. R., and Ravish, H. (2019). Object detection and tracking using deep learning and artificial intelligence for video surveillance applications. *Int. J. Adv. Comput. Sci. Appl.* 10 (12). doi:10.14569/ijacsa.2019.0101269

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: optimal speed and accuracy of object detection. ArXiv.

Chandan, G., Jain, A., and Jain, H. (2018). "Real time object detection and tracking using Deep Learning and OpenCV," in *2018 International Conference on inventive research in computing applications (ICIRCA)*. IEEE, 1305–1308.

Chen, C., Liu, M.-Y., Tuzel, O., and Xiao, J. (2017). "R-CNN for small object detection," in *Computer vision–ACCV 2016: 13th asian conference on computer vision, Taipei, Taiwan, November 20-24, 2016, Revised selected papers, Part V 13*. Springer, 214–230.

Chen, J., Zhu, J., Li, Z., and Yang, X. (2023). YOLOv7-WFD: a novel convolutional neural network model for helmet detection in high-Risk Workplaces. *IEEE Access* 11, 113580–113592. doi:10.1109/access.2023.3323588

Chen, Z., Wu, R., Lin, Y., Li, C., Chen, S., Yuan, Z., et al. (2022). Plant disease recognition model based on improved YOLOv5. *Agronomy* 12 (2), 365. doi:10.3390/agronomy12020365

Cheng, L. (2024). A highly Robust helmet detection algorithm based on YOLO V8 and transformer. *IEEE Access* 12, 130693–130705. doi:10.1109/ACCESS.2024.3459591

Diwan, T., Anirudh, G., and Tembhurne, J. V. (2023). Object detection using YOLO: challenges, architectural successors, datasets and applications. *multimedia Tools Appl.* 82 (6), 9243–9275. doi:10.1007/s11042-022-13644-y

Du, J. (2018). "Understanding of object detection based on CNN family and YOLO [J]," in *Journal of Physics: conference Series* 1004 (1).012029

Girshick, R. (2015). Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision, 1440–1448. doi:10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition* 580–587. doi:10.48550/arXiv.1311.2524

Hume, A., Mills, N., and Gilchrist, A. (1995). "Industrial head injuries and the performance of the helmets," in *Proceedings of the 1995 international IRCOBI conference on the Biomechanics of Impact*, 13–15.

Hwang, J. M., Won, J. H., Jeong, H. J., and Shin, S. H. (2023). Identifying critical Factors and Trends leading to Fatal accidents in small-scale construction sites in Korea. *BUILDINGS* 13 (10), 2472. doi:10.3390/buildings13102472

Jia, X. J., Zhou, X. X., Shi, Z. H., Xu, Q., and Zhang, G. M. (2025). GeoIoU-SEA-YOLO: an advanced model for detecting Unsafe Behaviors on construction sites. *SENSORS* 25 (4), 1238. doi:10.3390/s25041238

Jiang, P., Ergu, D., Liu, F., Cai, Y., and Ma, B. (2022). A Review of Yolo algorithm developments. *Procedia Comput. Sci.* 199, 1066–1073. doi:10.1016/j.procs.2022.01.135

Jiao, J., Tang, Y.-M., Lin, K.-Y., Gao, Y., Ma, A. J., Wang, Y., et al. (2023). Dilateformer: multi-scale dilated transformer for visual recognition. *IEEE Trans. Multimedia* 25, 8906–8919. doi:10.1109/tmm.2023.3243616

Li, H., Wu, D., Zhang, W., and Xiao, C. (2023). YOLO-PL: helmet wearing detection algorithm based on improved YOLOv4. *Digit. Signal Process.* 144, 104283. doi:10.1016/j.dsp.2023.104283

Lian, Y., Li, J., Dong, S., and Li, X. (2024). HR-YOLO: a multi-Branch network model for helmet detection combined with high-resolution network and YOLOv5. *Electronics* 13 (12), 2271. doi:10.3390/electronics13122271

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*. Editors B. Leibe, J. Matas, N. Sebe, and M. Welling (Cham: Springer International Publishing), 21–37.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

Masita, K. L., Hasan, A. N., and Shongwe, T. (2020). "Deep learning in object detection: a review," in *2020 international conference on artificial intelligence, Big data, computing and data Communication systems (icABCD)*. IEEE, 1–11.

Niu, C. W., Song, Y. S., and Zhao, X. Y. (2023). SE-lightweight YOLO: higher accuracy in YOLO detection for Vehicle inspection. *Appl. SCIENCES-BASEL* 13 (24), 13052. doi:10.3390/app132413052

Qian, S., and Yang, M. (2023). Detection of safety helmet-wearing based on the YOLO_CA model. *Comput. Mater. and Continua* 77 (3). doi:10.32604/cmc.2023.043671

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only Look once: Unified, real-time object detection," in *Computer vision and Pattern Recognition*.

Redmon, J., and Farhadi, A. (2017). "YOLO9000: better, faster, Stronger," in *2017 IEEE conference on computer vision and Pattern Recognition (CVPR)*, 6517–6525.

Redmon, J., and Farhadi, A. (2018). YOLOv3: an Incremental improvement. ArXiv.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Analysis Mach. Intell.* 39 (6), 1137–1149. doi:10.1109/TPAMI.2016.2577031

Riaz, M., He, J., Xie, K., Alsagri, H. S., Moqurrab, S. A., Alhakbani, H. A. A., et al. (2023). Enhancing Workplace safety: PPE_Swin—a Robust Swin transformer approach for automated personal protective equipment detection. *Electronics* 12 (22), 4675. doi:10.3390/electronics12224675

Saudi, M. M., Ma'arof, A. H., Ahmad, A., Saudi, A. S. M., Ali, M. H., Narzullaev, A., et al. (2020). Image detection model for construction worker safety conditions using faster R-CNN. *Int. J. Adv. Comput. Sci. Appl.* 11 (6). doi:10.14569/ijacsa.2020.0110632

Shan, D., Yang, Z., Wang, X., Meng, X., and Zhang, G. (2024). An Aerial image detection algorithm based on improved YOLOv5. *Sensors* 24 (8), 2619. doi:10.3390/s24082619

Soltanzadeh, A., and Mohammadfam, I. (2022). Cause-consequence modeling of occupational accidents in construction sites: a Retrospective study in Iran. *J. HEALTH Saf. AT WORK* 12 (3), 446–458. doi:10.1093/schbul/sbq173

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., and Liang, Z. (2019). Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* 157, 417–426. doi:10.1016/j.compag.2019.01.012

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers and distillation through attention[C]," in *International conference on machine learning*. PMLR, 10347–10357.

Waehrer, G. M., Dong, X. S., Miller, T., Haile, E., and Men, Y. (2007). Costs of occupational injuries in construction in the United States. *Accid. Analysis and Prev.* 39 (6), 1258–1266. doi:10.1016/j.aap.2007.03.012

Wang, Z., Cai, Z., and Wu, Y. (2023). An improved YOLOX approach for low-light and small object detection: PPE on tunnel construction sites. *J. Comput. Des. Eng.* 10 (3), 1158–1175. doi:10.1093/jcde/qwad042

Wang, Z., Wu, Y., Yang, L., Thirunavukarasu, A., Evison, C., and Zhao, Y. (2021). Fast personal protective equipment detection for real construction sites using deep learning approaches. *Sensors* 21 (10), 3478. doi:10.3390/s21103478

Wu, J. X., Cai, N., Chen, W. J., Wang, H. H., and Wang, G. T. (2019). Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. *AUTOMATION Constr.* 106, 102894. doi:10.1016/j.autcon.2019.102894

Wu, W., Liu, H., Li, L., Long, Y., Wang, X., Wang, Z., et al. (2021). Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image. *PloS one* 16 (10), e0259283. doi:10.1371/journal.pone.0259283

Wu, X. Q., Qian, S. R., and Yang, M. (2023). Detection of safety helmet-wearing based on the YOLO_CA model. *CMC-COMPUTERS Mater. and CONTINUA* 77 (3), 3349–3366. doi:10.32604/cmc.2023.043671

Xiong, R., and Tang, P. (2021). Pose guided anchoring for detecting proper use of personal protective equipment. *Automation Constr.* 130, 103828. doi:10.1016/j.autcon.2021.103828

Yang, B., and Wang, J. (2022). An improved helmet detection algorithm based on YOLO V4. *Int. J. Found. Comput. Sci.* 33 (06n07), 887–902. doi:10.1142/s0129054122420205

Yang, X., Wang, J. Z., and Dong, M. G. (2024). SDCB-YOLO: a high-precision model for detecting safety helmets and reflective Clothing in complex environments. *Appl. SCIENCES-BASEL* 14 (16), 7267. doi:10.3390/app14167267

Zhang, J. R., Wei, X., Zhang, L. X., Yu, L. B., Chen, Y. N., and Tu, M. Q. (2023). YOLO v7-ECA-PConv-NWD detects Defective Insulators on Transmission Lines. *ELECTRONICS* 12 (18), 3969. doi:10.3390/electronics12183969

Zhang, M., and Yin, L. (2022). Solar cell surface defect detection based on improved YOLO v5. *IEEE Access* 10, 80804–80815. doi:10.1109/access.2022.3195901

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)*. IEEE, 6–11.

Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). "TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on Drone-captured scenarios," in *2021 IEEE/CVF international conference on computer vision Workshops (ICCVW)*, 2778–2788.