Journal of Engineering, Project, and Production Management
2026, 16(4), 2026-147

TransSPCW-Net: A Personal Protective Equipment Detection Model

Jinhao Da
Undergraduate Student, School of Automation, Beijing Institute of Technology, Beijing 102488, P. R. China, E-mail: dajinhao.da@outlook.com

Project Management
Received January 28, 2026; revised April 15, 2026; accepted May 10, 2025
Available online May 29, 2026

Abstract: Construction site safety is an important issue for protecting workers health and reducing safety accidents. However, the traditional manual monitoring method often suffers from incomplete coverage and delayed response. To better overcome these problems, it is crucial to adopt real-time detection of both humans and Personal Protective Equipment (PPE) using intelligent recognition systems. To overcome these problems, this paper designs a model integrating the C3TR module, the SPP module, the CBAM attention mechanism, and the Wise-IoU (WIoU) loss function based on YOLOv8, named TransSPCW-Net. The TransSPCW-Net integrates these key innovations: the C3 Transformer (C3TR) module is used in conjunction with Spatial Pyramid Pooling (SPP). The joint module enhances the ability to extract features of all objections and reduce the impact of occlusion. The Convolutional Block Attention Module (CBAM) attention mechanism is designed to reduce background noise and focus more attention on the human body. Then the WIoU loss function is adopted to enhance the attention to the ordinary quality sample, thereby strengthening the precision of the prediction box. Experimental results show the improvements. The precision of the new model increased 1.377% from 94.932%, and recall increased 2.686% from 90.976%. Furthermore, it means Average Precision (mAP) increased 1.45% from 96.073%. These results demonstrate the model's robust performance in detecting PPE and humans in complex construction environments.

Keywords: Construction site safety, TransSPCW-Net, PPE detection, CBAM, WIoU-v3, C3TR.

DOI 10.32738/JEPPM-2026-147

1. Introduction

Construction settings, such as building sites and resource exploration sites, are inherently multifaceted and pose numerous safety concerns. Statistical analyses have shown that most construction accidents are associated with improper use of PPE (Zakariah and Alnuaim, 2024). Thus, it is important to have a method to detect whether workers wear PPE in compliance. However, traditional safety oversight relies on manual inspection. This method often causes problems such as incomplete coverage and delayed responses, which cannot effectively ensure workers safety. Therefore, it is crucial to introduce an object detection algorithm to detect PPE on construction sites (Al Khiami and ElHadad, 2024).

In recent years, Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) have matured in image understanding. One of them is two-stage detectors, such as Region-CNN (R-CNN) and Faster R-CNN, which focus on high precision but require substantial computational resources. Another method is one-stage detectors, such as You Only Look Once (YOLO), which are well suited for real-time operation on edge devices. YOLO has improved in speed and small-object detection after multiple iterations, making it a good choice for online monitoring on construction sites (Omar et al., 2024; Edozie et al., 2025).

Many researchers have applied deep learning methods to detect human bodies and their PPE in construction sites. For instance, Zhang et al. (2025) changed the CSP Bottleneck with 2 Convolutions (C2f) modules into Partial Convolution (PConv) modules and applied Gated Spatial-Temporal Attention mechanisms (GSTA) and the WIoU loss function. These enhancements improve feature extraction for safety helmets with differing shapes and sizes. The improved model achieved a mAP increase of 2.3% and a precision increase of 1.2%. Wang et al. (2025) introduced the EfficientViT backbone network, integrated Multi-Scale Dilated Attention (MSDA) and Dynamic Snake Convolution (DSConv) modules in the neck layer. These improvements enhance both detection precision and real-time performance for various PPE. The results demonstrated that the improved model achieved 4.1% increase in precision for detecting humans and six types of PPE.

Tong et al. (2025) designed a Multi-scale Feature Diffusion Pyramid Network (MFDPN) that dynamically selects

weights to improve classification precision. Experimental results indicated that improvements of 3.9% and 4.6% in mean Average Precision over IoU thresholds from 0.5 to 0.95 (mAP50-95) and mean Average Precision, IoU=0.5 (mAP50), for detecting PPE in construction sites. Chen et al. (2024) replaced all C2f modules with Spatial and Channel reconstruction Convolution (SCConv). In the head layer, they designed a lightweight decoupled head (PC-Head) to replace the detection modules. Ultimately, their improved models achieved a mAP of 93.8% for helmet detection.

Significant progress has been made in PPE detection models. However, construction sites still present several unresolved challenges. First, lighting conditions are highly variable, leading to unstable color and texture features. Under low-light conditions, the difference between PPE and the background is very low. Second, PPE and the human body will obstruct each other, making localization and classification difficult. Third, different PPE categories vary significantly in image scale. Helmets are relatively large, while boots are near the ground and occupy fewer pixels. These challenges indicate that the detection model should have both high precision for various detection targets and robustness in the complex construction environment (Al Khiami and ElHadad, 2024).

To address these challenges, this paper proposes the TransSPCW-Net model, built on YOLOv8. At the network input, the focus module is introduced into the backbone's first layer. It involves down-sampling and mapping spatial details into richer channel features. This preserves more details, which is beneficial for subsequent feature extraction. For mid-to-high-level feature fusion, the CBAM attention mechanism is injected into the sixth layer. This improves the model's sensitivity to the region of interest. In this paper, the model is trained to focus more on the human body, improving its precision. To address Spatial Pyramid Pooling-Fast (SPPFs) limited multi-scale receptive field, the paper replaces it with the combination of SPP coupled with the C3TR block. In this combination, the SPP expands pooling scales to extend both global and local receptive fields. At the same time, the C3TR uses a transform module to enhance the modeling ability of long-distance dependence. It is well suited for construction sites with occlusion and multi-target detection. Together, these improve the model's ability to detect various types of PPE and occluded PPE in complex construction environments. To reduce computational redundancy in the head layer, TransSPCW-Net replaces the C2f module with the C2 module. This reduces the number of parameters and calculation time without significantly affecting feature extraction. The CIoU loss is replaced with the WIoU-v3 loss function. This new loss function reduces the impact of low quality annotation on model training. It improves the objectness of the predicted box for all objects, especially the small objects such as boots.

These improvements provide a comprehensive approach, focusing on preserving input detail and on attention-guided feature refinement. They also emphasize multi-scale and global semantic integration, along with output optimization using reliable regression loss. Under complex construction scenarios, TransSPCW-Net better captures small-scale objects and reduces the impact of occlusion and background misdetections. It also achieves a more reasonable and focused distribution of attention to human bodies. This provides a feasible and efficient engineering solution for real-time detection of PPE and personnel compliance on edge devices at construction sites.

2. YOLOv8 Network Model

YOLO, as a deep learning algorithm for real-time object detection, was first proposed by Joseph Redmon and colleagues in 2016. Its core idea is to achieve object localization and classification through a single forward pass of a neural network. YOLOv8, a prominent model in the YOLO family, balances strong detection accuracy and real-time speed while reducing the parameter budget (Lin et al., 2025; Ali and Zhang, 2024; Sapkota et al., 2025).

YOLOv8 has three parts: the backbone for feature extraction, the neck for fusing features, and the head for predicting outputs. Through a series of convolutional blocks, the backbone abstracts local textures and edges into foundational feature maps. In this layer, YOLOv8 introduces the C2f module as a crucial component for efficient feature fusion. It uses a two-way fusion scheme: channels are split and then concatenated, while spatial paths keep information flowing (Varghese and Sambath, 2024). SPPF acts on the backbone's output feature maps to enhance multi-scale representation. This module enhances the receptive field and enables the network to better adapt to multi-resolution inputs. In the head layer, YOLOv8 incorporates C2f modules to improve the performance of classification and regression tasks. Finally, the head layer outputs object detection results by applying convolutional operations on high-dimensional features to precisely identify objects of different categories and scales (López et al., 2025; Raditya et al., 2024).

3. Improved TransSPCW-Net Network Model

To further expand the application scope of YOLOv8 and more accurately identify workers and their compliance at construction sites, this paper proposes an improved YOLOv8-based model, TransSPCW-Net. The improved network structure is shown in Fig. 1, and Table 1 shows the detailed parameters of each layer. In Table 1, the 'layer' represents the sequence of each layer. The 'from' marks the input of the current module, while the 'n' means the number of repeated modules. The 'params' means the learnable parameters, the 'module' represents the abbreviation of the current layer, and the 'arguments' means the key parameters inside the module. The specific improvements are as follows.

1) The Focus module is introduced in the first layer of the Backbone network to perform spatial reconstruction and down sampling on the original input image. The module halves the width and height of the input image by a slicing operation and centralizes the information into the channel space. Thus, it can increase the number of channels without losing information. This is profit to the extraction of small object features in the next convolution.

2) The CBAM is integrated into the sixth layer of the backbone. By analyzing the YOLOv8 attention heat map, the model's attention is consistently on the background under complex lighting conditions. After introducing the CBAM mechanism, the model's attention becomes more focused on human bodies in complex environments. At the same time, it also improves sensitivity to regions of interest, enhancing the ability to recognize small objects like boots.

3）The paper noticed that mid-level activations were fragmented. When strong light appeared, the precision of small PPE items, such as helmets, was poor. Introducing an SPP module stabilizes those responses by widening context. Furthermore, the C3TR module has strong global modeling capabilities. This helps network link a human's body to nearby PPE, even when partially occluded by construction materials. By reconstructing occluded objects, such as a human, the SPP-C3TR joint module reduces the negative impact of occlusion on detection accuracy.

4）Although the C2f module has a stronger feature extraction capability, it also has a higher computational cost. Given that the paper has already introduced the high-computational SPP-C3TR joint module, the use of C2f may not only affect the model's speed but also lead to redundant calculation. This, in turn, may lead to overfitting and reduce detection precision. This paper replaces all the C2f modules in the neck layer with C2 modules with lower computational cost.

5）Because of factors such as complex light conditions, the predicted box bounds by the Complete-IoU (CIoU) loss function in YOLOv8 are often inaccurate. Therefore, this paper replaces the CIoU loss function with the WIoU-v3 loss function, which employs the dispersion concept to emphasize mid-quality predictions and improve localization. Experimental results show that after replacing with WIoU-v3, the final classification loss (cls_loss) and bounding-box regression loss (box_loss) decrease by 0.07 and 0.19, respectively.

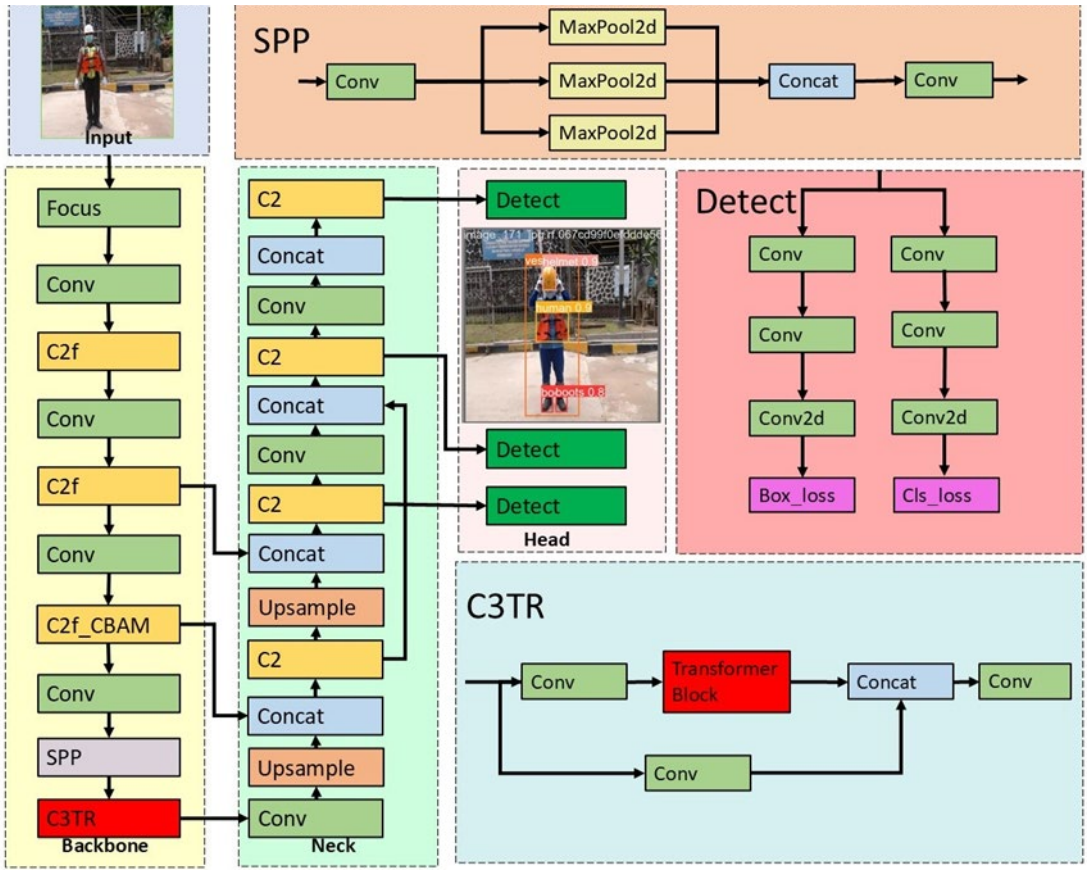


Fig. 1. TransSPCW-Net network model

3.1. Focus Module

The main purpose of the Focus module is to apply a 2×2 space-to-depth rearrangement, making an image from (height(H), width(W), channel(C)) into ($\frac{H}{2}$, $\frac{W}{2}$, $4C$) by concatenating four spatial slices into the channel dimension. Compared to direct down sampling, this module preserves more image details, which will help extract features of small PPE like boots in the next C2f module. Moreover, reducing special dimensions reduces computation in subsequent layers. The computational flow diagram is illustrated in Fig. 2.

3.2. CBAM Improvement Module

CBAM consists of two parts, one of which is the Channel Attention Module (CAM) and the other is the Spatial Attention Module (SAM). CAM can reweigh the importance of each channel using a Multi-Layer Perceptron (MLP), while SAM highlights regions of interest using global Average Pooling (AvgPool) and global Max Pooling (MaxPool). In this paper, CBAM is used to shift the model's attention from the complex background to the human body. The detailed flowchart is illustrated in Fig. 3 (Woo et al., 2018; Lv and Su, 2023).

Table 1. Specific parameters of each layer in TransSPCW-Net

layer	from	n	params	module	arguments
0	-1	1	3520	Focus	[3,32,3]
1	-1	1	18560	Conv	[32,64,3,2]
2	-1	1	29056	C2f	[64,64,1,True]
3	-1	1	73984	Conv	[64,128,3,2]
4	-1	2	197632	C2f	[128,128,2,True]
5	-1	1	295424	Conv	[128,256,3,2]
6	-1	2	788676	C2f_CBAM	[256,256,True]
7	-1	1	1180672	Conv	[256,512,3,2]
8	-1	1	656896	SPP	[512,512,[5,9,13]]
9	-1	1	1182976	C3TR	[512,512,1,False]
10	-1	1	131584	Conv	[512,256,1,1]
11	-1	1	0	Upsample	[None,2,'nearest']
12	[-1,6]	1	0	Concat	[1]
13	-1	1	493056	C2	[512,256,1]
14	-1	1	0	Upsample	[None,2,'nearest']
15	[-1,4]	1	0	Concat	[1]
16	-1	1	140032	C2	[384,128,1]
17	-1	1	147712	Conv	[128,128,3,2]
18	[-1,12]	1	0	Concat	[1]
19	-1	1	525824	C2	[640,256,1]
20	-1	1	590336	Conv	[256,256,3,2]
21	[-1,9]	1	0	Concat	[1]
22	-1	1	1838080	C2	[768,512,1]
23	[15,18,21]	1	12000604	Detect	[4,[384,640,768]]

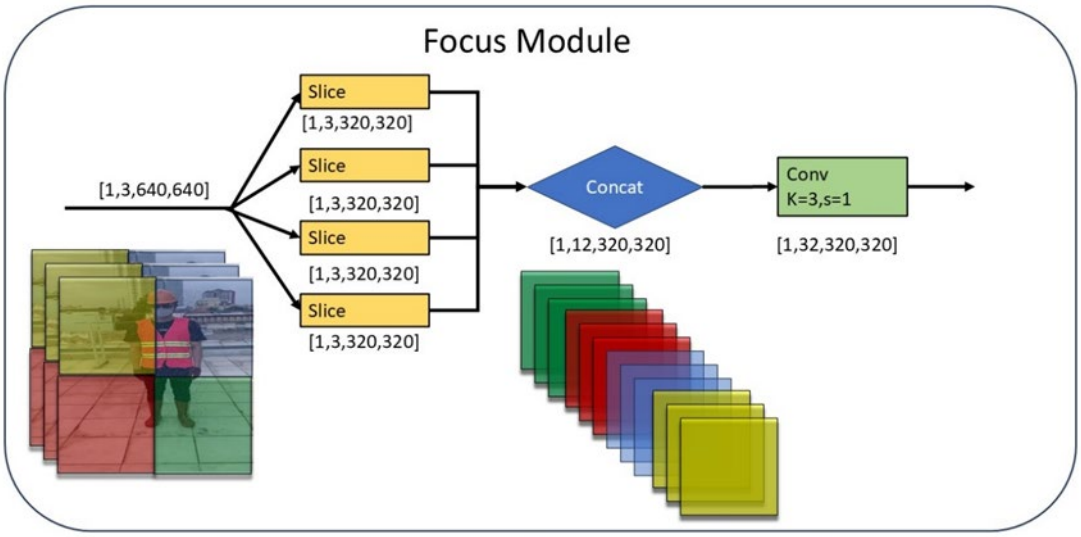


Fig. 2. Focus module operation flowchart

CAM will put the feature maps to the MaxPool and the AvgPool separately. The MaxPool captures the strongest activation of each channel, while the AvgPool reflects the overall average activation. In this paper, the MaxPool is used to capture distinctive textures, such as human body edges, helping to depict object boundaries. Moreover, the AvgPool emphasizes large-area patterns like vest. Then, these two descriptors are passed through MLP. After that, they are summed

and activated by a sigmoid function and finally get the channel attention map. This module not only suppresses the model's attention to the construction background but also enhances the recognition of human bodies and their PPE. The specific computational flow is shown in Fig. 4.

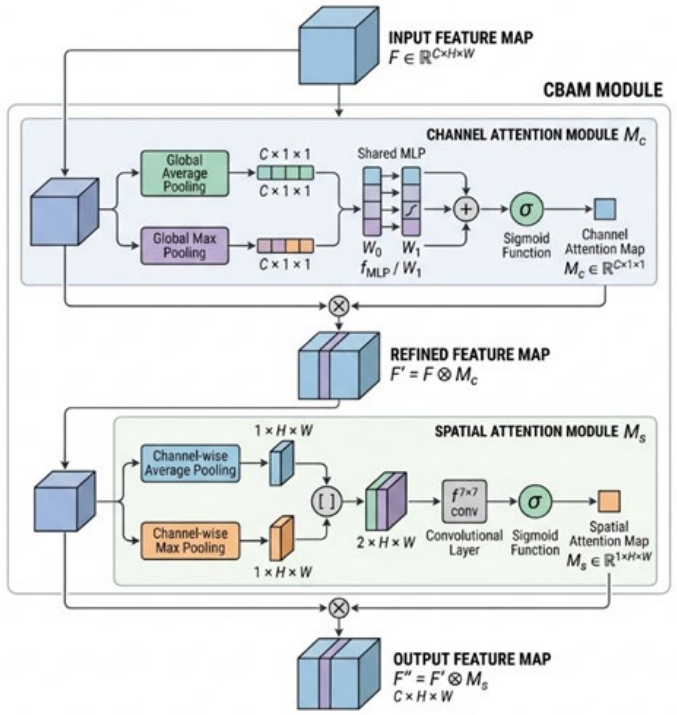


Fig. 3. Overall flowchart of the CBAM module

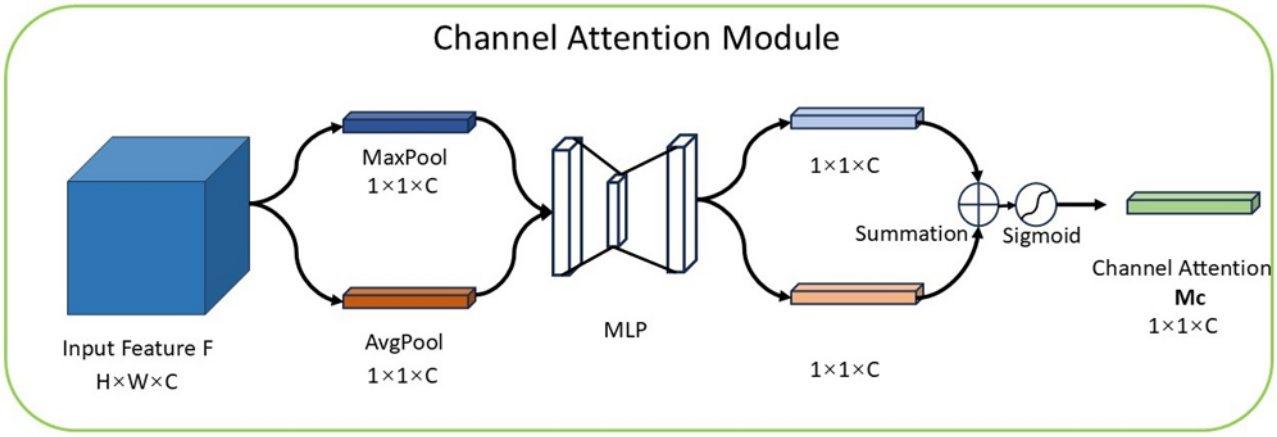


Fig. 4. Flowchart of channel attention

The output channel attention map $\mathbf{M}_c \in \mathbb{R}^{C\times1\times1}$ can be calculated using Eq. (1).

$$\begin{aligned}\mathbf{M}_c(\mathbf{F}) &= \sigma(MLP(AvgPool(\mathbf{F})) + MLP(\mathrm{Max}Pool(\mathbf{F}))) \\ &= \sigma(\mathbf{W}_1(\mathbf{W}_0(\mathbf{F}^c_{avg})) + \mathbf{W}_1(\mathbf{W}_0(\mathbf{F}^c_{max}))\end{aligned} \tag{1}$$

Here, **F** denotes the input feature map. AvgPool and MaxPool produce the channel descriptors $\mathbf{F}^c_{avg}$ and $\mathbf{F}^c_{max}$. MLP denotes a multi-layer perceptron. The two weights of MLP are $\mathbf{W}_0 \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{C\times C/r}$. Finally, σ means sigmoid activation function.

After CAM performs global pooling across each channel in the spatial dimension, SAM applies channel-wise AvgPool and MaxPool to the input feature map. These operations respectively obtain the mean response and the most salient response at each spatial location. The two resulting 2D maps $[AvgPool(\mathbf{F}'); MaxPool(\mathbf{F}')]$ denote channel-wise concatenation of the average-pooled and max-pooled 2D maps, which are then passed through a convolutional layer (or a fully connected layer applied spatially) to produce a spatial weight distribution. Then, these spatial weights are used to

reweigh the original feature map so that the model's attention can focus on important regions, such as the human body. The SAM flowchart is shown in Fig. 5.

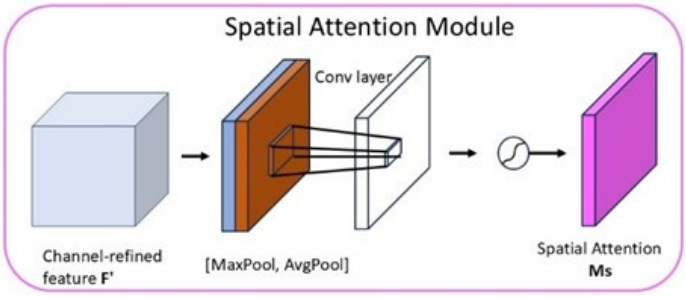


Fig. 5. Flowchart of spatial attention

The output spatial attention map $\mathbf{M}_s(\mathbf{F}) \in \mathbb{R}^{H\times W}$ of the spatial-attention module is shown in Eq. (2).

$$\begin{aligned}\mathbf{M}_s(\mathbf{F}) &= \sigma(f^{7\times7}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})])) \\ &= \sigma(f^{7\times7}([\mathbf{F}^s_{avg}; \mathbf{F}^s_{max}]))\end{aligned} \tag{2}$$

Here, input $\mathbf{F}$ means the channel-refined feature obtained in the last step. $f^{7\times7}$ denotes a 7×7 kernel convolution. $\mathbf{F}^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{F}^s_{max} \in \mathbb{R}^{1\times H\times W}$ are two 2D maps after pooling operation, and σ means sigmoid activation function.

The CBAM attention module combines CAM and SAM attention mechanisms to capture key information across both channel and spatial dimensions. This can filter the influence of complex backgrounds, such as cement in the construction scenes, and make attention more focused on the human body and PPE.

3.3. SPP+C3TR Collaborative Improvement Module

At the deepest layer of the backbone, the paper replaces the SPPF module with the SPP-C3TR joint module. The SPP module primarily enhances the model's feature extraction for objects of different sizes. The C3tr module enhances the model's global modeling capabilities to address occlusion, ensuring the human body is correctly detected even when covered by construction materials.

SPP was first proposed by He et al. (2015). SPP is used to obtain multi-scale feature information without changing the input size. Specifically, SPP uses max pooling with kernel sizes 5×5, 9×9, and 13×13 on the input feature map. To maintain the output size, the padding stride is 1. Finally, these feature representations of different scales are concatenated in the channel dimension. Because the texture and size of small objects, such as helmets and boots, are different from those of large objects, such as humans, the features of the two can be distinguished after concatenation.

C3TR replaces the bottleneck with a transformer block based on C3. Although this increases the computational cost, it improves the detection precision of the occluded human body by 1.82% from 94.93%. During the experiment, the paper also tried to replace c3tr with other layers of the backbone but found that CUDA Out Of Memory (CUDA OOM) occurred in all layers except the last one. This is because C3TR has a large calculation. If the module is used in layers with large feature maps, it can easily cause insufficient memory. After balancing performance and resources, the paper believes that putting C3TR at the last layer of the backbone is the best choice (Lv and Su, 2023).

Although the SPP-C3TR module improves the model's recognition of objects of different sizes and occluded objects, it also requires more computation. Therefore, this paper will proceed to address the problem of excessive computational load in the following section.

3.4. C2 Improvement Module

Although the C2f module shows stronger feature extraction capability, it also has a higher computational cost. Because the SPP-C3TR module has already expanded the subsequent feature fusion receptive field, if the model continues to use C2f blocks, it will introduce redundant computation. Redundant computation will increase the risk of overfitting, which is harmful for detection accuracy. Therefore, after SPP-C3TR, the paper replaces all C2f modules in the neck with C2 modules. Compared to C2f, C2 removes the bidirectional multi-stage fusion, thus it can reduce the redundant information. This replacement reduces computational cost while improving the model's detection performance when SPP–C3TR is introduced.

3.5. WIoU-v3 Loss Function

YOLOv8 uses CIoU for bounding-box regression. CIoU consists of three components: the vanilla IoU term, a center-distance penalty, and an aspect-ratio penalty. However, in a construction environment, image quality often degrades due to various lighting conditions and occlusion caused by building materials. As a result, geometric measures such as distance and aspect ratio tend to over-penalize low-quality samples, thereby harming generalization. This makes the CIoU loss less capable of capturing the object's shape. The limitations are alleviated by replacing CIoU with WIoU-v3, whose formulation

embeds a dynamically varying, non-monotonic focus term. When the anchor box overlaps well with the ground-truth box, WIoU-v3 weakens the geometric penalties, avoiding excessive intervention during training and thereby improving generalization.

WIoU-v1 introduced an attention-based bounding-box loss. WIoU-v3 introduces an outlier degree(β) based on WIoU-v1. By reducing the contribution of high-quality samples to the loss and dynamically assigning gradient gains to the bounding boxes, the new loss function enhances the model's localization capability. The loss calculation is as follows in Eqs. (3) through (8) (Wang et al., 2025; Khalil et al., 2024).

$$\beta = \frac{\mathbf{L}_{IoU}}{mean(\mathbf{L}_{IoU})} \tag{3}$$

$$\mathbf{L}_{IoU} = 1 - IoU = 1 - \frac{\mathbf{W}_i\mathbf{H}_i}{\mathbf{S}_u} \tag{4}$$

$$\mathbf{R}_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(\mathbf{W}_g^2 + \mathbf{H}_g^2)}\right) \tag{5}$$

$$\mathbf{L}_{WIoUv1} = \mathbf{R}_{WIoU}\mathbf{L}_{IoU} \tag{6}$$

$$r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{7}$$

$$\mathbf{L}_{WIoUv3} = r\mathbf{L}_{WIoUv1} \tag{8}$$

In Eq. (3), $\mathbf{L}_{IoU}$ represents the IoU loss, and *mean* ($\mathbf{L}_{IoU}$) denotes the moving average. $\mathbf{W}_i\mathbf{H}_i$ in Eq. (4) represents the area of the prediction box. $\mathbf{S}_u$ represents the summed area of the prediction box and the real box. In Eq. (5), $\mathbf{R}_{WIoU}$ means $\mathbf{L}_{IoU}$ of the ordinary-quality anchor box, and *exp* denotes the natural exponential function. In Eq. (6), the loss of WIoU-v1 is $\mathbf{L}_{WIoUv1}$. Then, r (Eq. (7)) is a nonmonotonic focusing coefficient constructed by β. Finally, in Eq. (8), WIoU-v3 $\boldsymbol{L}_{\mathrm{WIoUv3}}$ can then be calculated by multiplying $\mathbf{L}_{WIoUv1}$ by r.

The model attains better localization through attenuating the effect of high-quality instances and dynamically redistributing gradient gains over its boxes. The specific loss value comparison will be presented later in the text.

4. Experimental Design and Result Analysis

4.1. Experimental Design

4.1.1. Experimental parameters and experimental environment

To ensure the reproducibility of the experiments, here is the training configuration for the model. The optimization follows SGD with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. The paper uses a batch size of 4 and trains for 160 epochs on images resized to 864×864. The hardware/software stack consists of Windows 11, Python 3.9.23, PyTorch 2.0.0 (CUDA 12.7), a Ryzen 7 7840H processor, and an NVIDIA GeForce RTX 4060 GPU.

To verify the rationality of selecting the SGD optimizer with an initial learning rate of 0.01, a controlled experiment was designed in this study. Only the optimizers and initial learning rates were replaced to compare the model's detection accuracy, inference speed (FPS) and convergence stability. Experimental results show that SGD with a learning rate of 0.01 is the best setting.

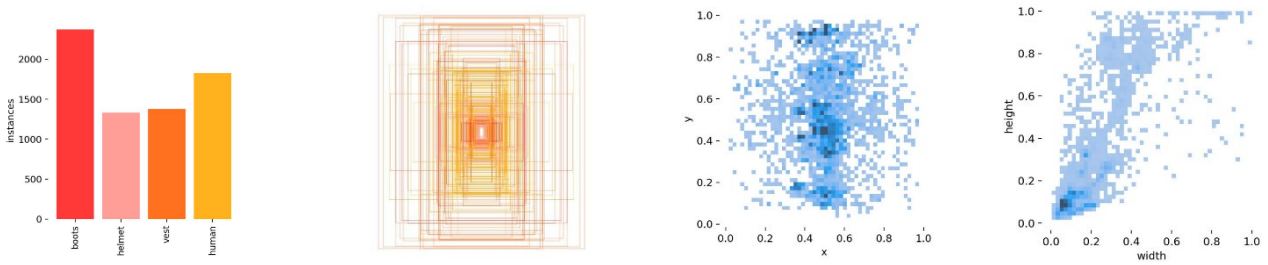
Table 2. Comparative Experimental Results of Different Optimizers and Initial Learning Rates

Optimizer	Initial Learning Rate	Precision (%)	Recall (%)	mAP50 (%)	Inference Speed (FPS)
SGD	0.001	95.126	92.058	96.214	28.5
SGD	0.01	96.309	93.662	97.517	31.2
SGD	0.1	92.087	88.534	93.106	29.8
Adam	0.001	42.106	35.729	38.514	18.7
Adam	0.01	29.635	21.087	25.306	15.2
AdamW	0.001	28.587	10.555	2.691	20.3
AdamW	0.01	21.308	8.742	1.956	16.5

4.1.2. Dataset and preprocessing

The paper uses a publicly available dataset from the Roboflow platform. It contains 2197 JPG images and four detection classes: human, helmet, vest, and boots. It is notable that the original dataset included the class gloves. However, the number of glove samples was very limited, so the paper preprocessed and removed the glove class. More detailed labeling

information is presented in Fig 6.



(a) number of labels (b) box length and width (c) center point distribution (d) ratio of label width to height

(e) example images for training

Fig. 6. Dataset annotation file statistics and visualization

Fig. 6(a) represents the number of each label. Fig. 6(b) superimposes the bounding boxes of all labels, and Fig. 6(c) shows the spatial position distribution of the bounding box center. Finally, Fig. 6(d) shows the ratio of In Fig. 6, the image (a) represents the number of each label. Fig. 6(b) superimposes the bounding boxes of all labels, and Fig. 6(c) shows the spatial position distribution of the bounding box center. Finally, Fig. 6(d) shows the ratio of label width to height, which is concentrated within the area from (0,0) to (0.2, 0.2). These 4 images show that the small size of labeled accounts for a large proportion of the dataset. Thus, enhancing the ability to detect small PPE items is a key goal of the paper. Fig. 6(e) shows example training images.

Due to practical constraints, we were unable to collect real images of construction sites under various low-visibility weather conditions. To address this issue, this paper adopts a fogging algorithm and processes the dataset by mixing original clear images and fog-enhanced images at a 2:1 ratio.

The specific algorithm process is as follows: First, a fog mask with the same size as the original image is constructed, using a light gray-white color that matches the characteristics of dust and haze at construction sites as the base color for the fog effect. Then, based on the principle of weighted image fusion, Eq. (9) is used to achieve the fogging effect.

$$\mathrm{I}_{fog}(x,y)=\mathrm{I}_{o}(x,y)\times(1-\alpha)+\mathrm{M}(x,y)\times\alpha \tag{9}$$

Herein, α means the fog density weight coefficient, which is randomly set within the range of 0.3 to 0.7. I_{fog} means the pixel values of the generated foggy image, I_{fog} means pixel values of the original clear construction site image, and M means the fog mask image. The fogging effect is shown in Fig. 7. Model training is performed based on this fog-

enhanced dataset.

4.1.3. Model evaluation metrics

To better evaluate the model's performance, the paper employs the following four metrics: Precision (P), Recall (R), and mAP50. Additionally, bounding Box regression loss (box_loss) and Classification loss (cls_loss) on the validation set are also used as auxiliary criteria. These metrics can reflect the model's detection capabilities and optimization performance of the human and PPE in the construction sites (Gallo et al., 2022; Barlybayev et al., 2024). P denotes the fraction of samples the model labels as positive. R represents the proportion of samples that are actually positive, and the model correctly identifies them as positive categories. mAP50 represents the average precision value across all categories. Box_loss represents the difference in shape and position between the predicted bounding box and the actual bounding box, and cls_loss represents the difference between the predicted category probability and the true category label.

Fig. 7. The original image and fog image

Note: (a), (b) and (c) are the original images, and (d), (e) and (f) are the same rendered images with the fogging mask based on the original images.

4.2. Analysis of Experimental Results

4.2.1. Experimental results

To better evaluate the improved model's efficiency in recognizing humans and PPE, this paper uses the same dataset to train YOLOv8 and TransSPCW-Net with identical training parameters. By comparing and analyzing the training results to see whether the improvements are effective. The test results are presented in Table 3.

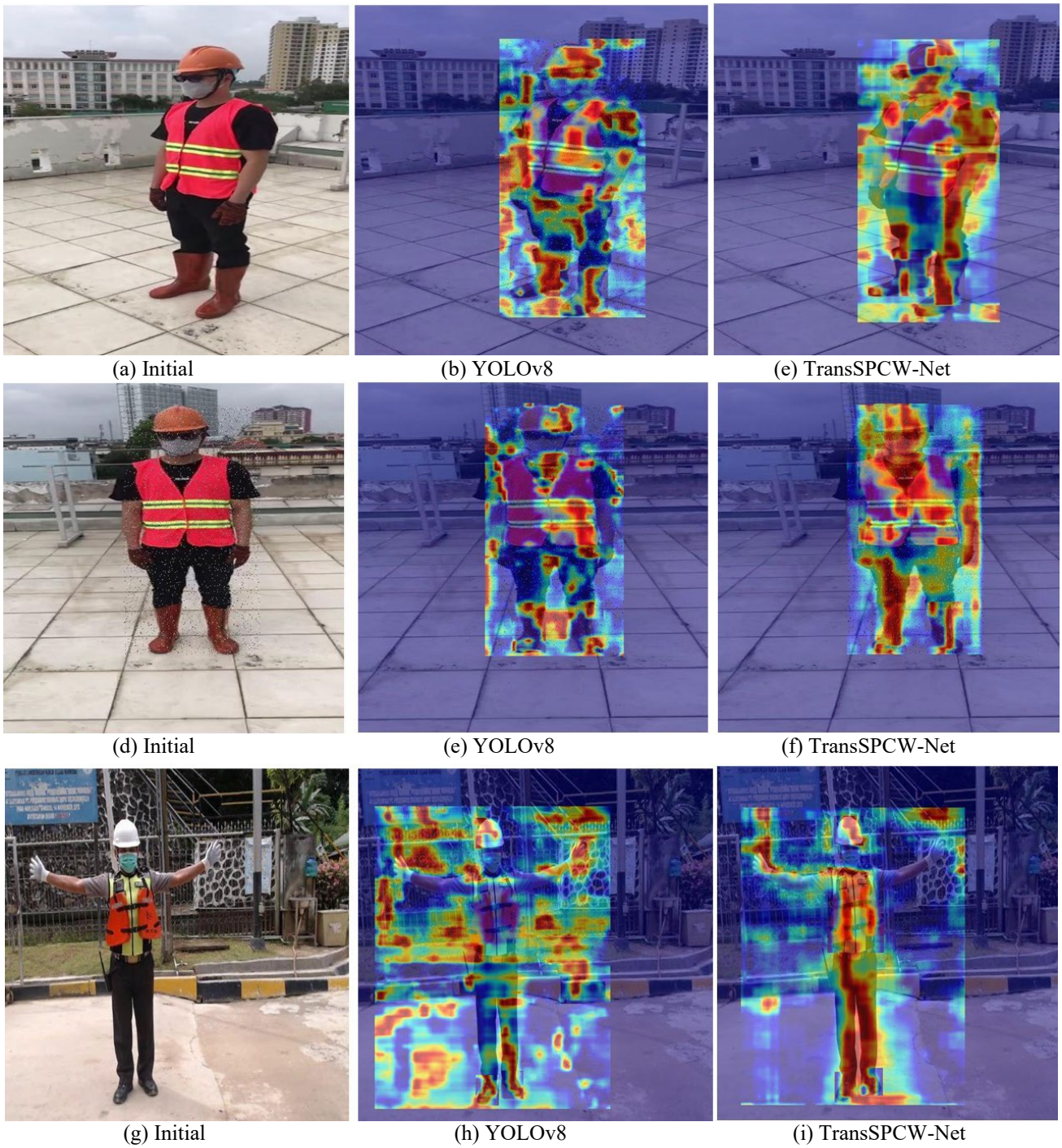
Table 3. Comparison of the algorithm before and after improvement

Model	Precision/%	Recall/%	mAP50/%	val	
				box_loss	cls_loss
YOLOv8	94.932	90.976	96.073	0.86123	0.6959
TransSPCW-Net	96.309	93.662	97.517	0.7907	0.50558

The evaluation of the experimental findings in Table 2 suggests that the TransSPCW-Net model outperforms YOLOv8 across all detection metrics. Compared with YOLOv8, TransSPCW-Net achieves improvements of 1.4% from 94.932% to 96.07%, 2.7% from 90.976% to 93.07%, and 1.45% from 96.07% to 97.45% on P, R, and mAP50. These gains demonstrate that the improved model delivers superior performance in human and PPE detection under construction-site conditions.

To better visualize the improvement, several representative test images were selected, and attention heatmaps were generated using the best-trained weights of YOLOv8 and TransSPCW-Net, as shown in Fig. 8. Darker regions indicate higher attention concentration. By analyzing images (b), (e), and (h), the paper finds that YOLOv8 tends to disperse attention across the construction background, especially in image (h). However, TransSPCW-Net focuses more on the human body and PPE, which is the desired outcome after analyzing images (c), (f), and (i). Then, take a closer comparison between (h) and (i), it is further revealed that background attention is repressed and human-focused attention is enhanced in TransSPCW-Net, just as the paper's design goals. These results confirm that TransSPCW-Net strengthens focus on human and PPE, represses background noise, and improves detection robustness under occlusion and multi-scale conditions.

To better observe the performance improvement of TransSPCW-Net, this paper presents the normalized confusion matrices of YOLOv8 and TransSPCW-Net on the test dataset, shown in Fig. 8. These matrices characterize each class's recognition behavior. A comparative analysis shows that TransSPCW-Net improves human recognition accuracy from 0.91 to 0.93, highlighting its enhanced capability for human detection.

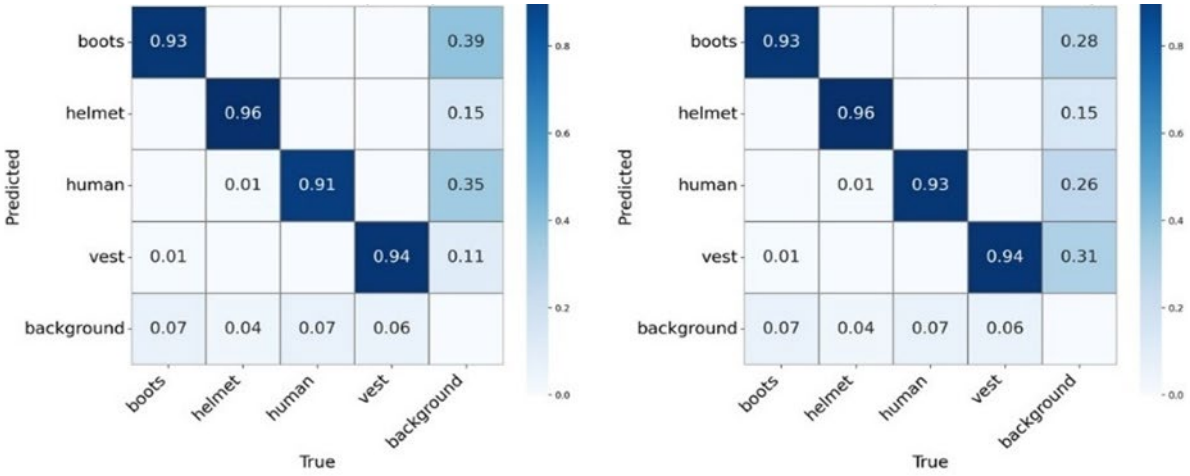


(a) Initial (b) YOLOv8 (e) TransSPCW-Net

(d) Initial (e) YOLOv8 (f) TransSPCW-Net

(g) Initial (h) YOLOv8 (i) TransSPCW-Net

Fig. 8. Heatmaps of some test set images generated using the weights obtained after training
Note: The (a), (d), and (g) are the original images; the (b), (e), and (h) are heatmaps drawn by YOLOv8; (c), (f), and (i) are heatmaps drawn by TransSPCW-Net.



(a) confusion matrix of YOLOv8 (b) confusion matrix of TransSPCW-Net

Fig. 9. Confusion matrix diagrams of the original model and the improved model

In terms of auxiliary detection metrics, TransSPCW-Net reduces box_loss and cls_loss by 0.07 and 0.19, respectively, compared with YOLOv8. It shows better convergence in bounding box regression and fewer classification errors. To show the impact of the loss function on model optimization. The paper assigns a weight of 0.5 to box_loss and cls_loss to create a weighted loss and plots the loss epoch curves for YOLOv8 and TransSPCW-Net, which is shown in Fig. 9. It shows that in the early training phase, the loss of TransSPCW-Net with WIoU is already much lower than that of YOLOv8 with CIoU. Throughout the entire training process, TransSPCW-Net's loss remains lower than YOLOv8's. The experimental results indicate that the use of WIoU loss function can improve the convergence speed and overall detection accuracy. It is a good choice for detecting humans and PPE in construction sizes.

4.2.2. Ablation experiment

Fig. 10 compares the convergence behavior of the loss functions used in TransSPCW-Net and YOLOv8 during training. To comprehensively validate the effectiveness of each improvement module and its ablation effect, this paper designs 5 ablation experiments on YOLOv8. These experiments are used to analyze their contribution to overall model performance. The paper employs P, R, and mAP50 as the evaluation metrics. The detailed ablation experiment schemes and results are presented in Table 4.

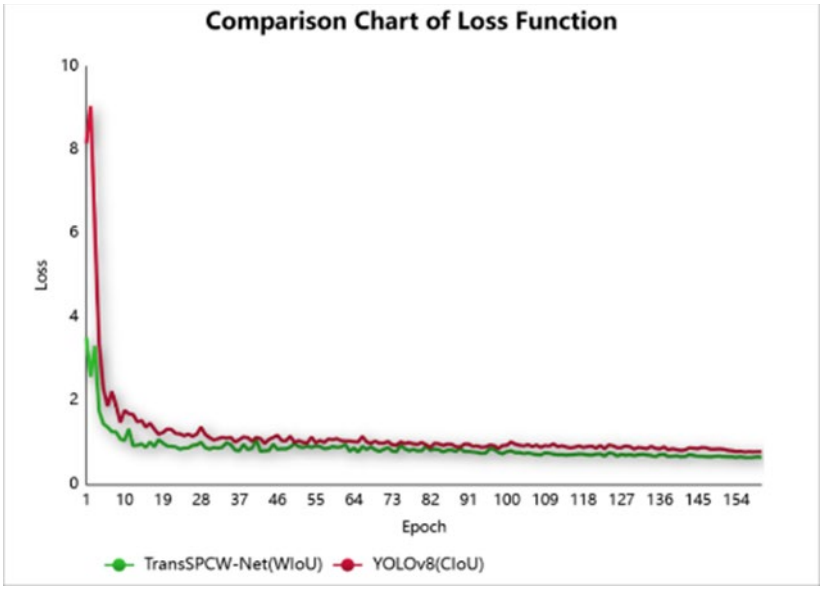


Fig. 10. Comparison chart of loss functions

Table 4. Results of ablation experiment

Model	SPP-C3TR	CBAM	WIoU-v3	Precision/%	Recall/%	mAP50/%
YOLOv8(Exp. 1)	×	×	×	94.93	90.97	96.07
SPP-C3TR(Exp. 2)	√	×	×	96.85	91.55	96.17
CBAM(Exp.3)	×	√	×	95.54	92.48	96.48
SPP-C3TR+CBAM(Exp.4)	√	√	×	95.73	93.00	96.72
CBAM+WIoU(Exp.5)	×	√	√	96.48	92.03	96.85
SPP-C3TR+WIoU(Exp.6)	√	×	√	96.87	92.50	97.23
TransSPCW-Net(Final)	√	√	√	96.30	93.66	97.51

As shown in Table 4, Exp. 1 is the YOLOv8 model, which has already achieved good results in P, R, and mAP50. Exp. 2 introduces the SPP-C3TR module, increasing P by 1.7% with slight gains in the other two metrics, proving that enhancing the receptive field and global modeling improves YOLOv8's performance. Exp. 3 adds the CBAM module, with P, R, and mAP50 rising by 0.6%, 1.5%, and 0.4%, respectively, indicating that the attention mechanism improves detection performance. Exp. 4 introduces CBAM together with SPP-C3TR, producing consistent improvements and indicating no conflict between the two modules that would degrade performance. Exp. 5 adds the WIoU-v3 loss function on top of CBAM, increasing P and R by 1.5% and 1.1%, with a slight rise in mAP50, demonstrating the crucial role of loss optimization in improving localization accuracy. Exp. 6 introduces the WIoU loss on top of SPP-C3TR and achieves increases of 1.9%, 1.5%, and 1.1% in P, R, and mAP50. Finally, all modules are combined to form the TransSPCW-Net model. The TransSPCW-Net model achieved the highest values across all metrics: precision of 96.30%, recall of 93.66%, and mAP50 of 97.51%. Although its P is not the highest among the ablations, it attains the highest R and mAP50 and is therefore adopted as the final model.

To clarify the balance between performance and computational efficiency and to demonstrate the rationality of the TransSPCW-Net model, ablation experiments were conducted by replacing the C2 modules in the neck layer with C2f modules, while keeping the backbone and detection head unchanged. The experimental variables are the combination modes of C2 and C2f modules in the neck layer, and the evaluation indicators include detection accuracy and computational efficiency, including parameters and Floating Point Operations (FLOPs). The detailed results are shown in Table 5.

Table 5. Ablation experiment results of c2/c2f module combination

Experiment Group	Neck Module Configuration	Precision /%	Recall /%	mAP50 /%	Parameters /M	FLOPs /G
1	All C2f	96.482	93.824	97.651	34.52	75.47
2	Layers 13/16/19: C2→C2f	96.456	93.801	97.632	33.06	72.13
3	Layers 13/16:C2→C2f	96.423	93.768	97.605	31.59	68.96
4	Layer 13: C2→C2f	96.387	93.715	97.562	30.14	65.82
5	All C2f(TransSPCW-Net)	96.309	93.662	97.517	28.67	62.35

The experimental results show that replacing C2 modules with C2f modules improves detection accuracy. When all C2 modules are replaced with C2f modules, the maximum increases in P, R, and mAP50 are only 0.173%, 0.162%, and 0.134%, respectively. But these can be ignored in the negligible fluctuation range in practical engineering applications. However, the introduction of C2f modules leads to a significant surge in computational costs. Compared with the all-C2 configuration,

the all-C2f configuration increases the number of parameters by 5.85M, rising 20.4%, and FLOPs by 13.12G, increasing 21.06%, resulting in substantial computational redundancy. Therefore, adopting the all-C2 configuration in the neck layer represents the optimal solution for TransSPCW-Net.

Evidence from the ablation analysis substantiates each module's contribution and the added benefit of collaborative optimization. It provides a more efficient and accurate solution for detecting humans and their PPE in construction sites.

4.2.3. Comparative experiment

For further validating the proposed model's performance, the paper selects several representative YOLOv8 structures as comparison models, including yolov8, yolov8x_DW_FOCUS2, yolov8x_DW_swin_FOCUS2_sppc, yolov8s_DW_CA_BOT3_FOCUS_C2, yolov8s_c2_DW, yolov8x_DW_FOCUS_sppc, and yolov8x_DW_swin_FOCUS. Each model shares the same dataset and training protocol for both training and evaluation. As for metrics, the paper chooses P, R, mAP50, box_loss, and cls_loss to evaluate the different models performance. The detailed information is shown in Table 4.

Experimental results show that TransSPCW-Net achieves the best overall performance. Its Precision (P), Recall (R), and mAP50 reach 96.3%, 93.7%, and 97.5%, respectively, which is significantly higher than the P of yolov8x_DW_FOCUS2, which is 95.5%, the R of YOLOv8, which is 90.9%, and the mAP50 of yolov8x_DW_swin_FOCUS2_sppc, which is 97.18. Although some models surpass YOLOv8 on individual metrics, they still underperform compared with TransSPCW-Net.

In terms of loss, apart from TransSPCW-Net, only yolov8x_DW_swin_FOCUS2_sppc has a lower box_loss, which is 0.8 lower than that of YOLOv8. All other models have higher losses. These results further prove that the final model exhibits stronger optimization capability and better convergence in both bounding box regression and classification.

In summary, TransSPCW-Net not only achieves comprehensive improvements in P, R, and mAP50 but also attains the lowest loss values. This demonstrates that, compared with other models, TransSPCW-Net offers superior detection accuracy and training convergence, providing a more reliable and efficient technical foundation for intelligent safety monitoring scenarios.

5. Conclusion

This study proposes TransSPCW-Net, a high-precision detection model for recognizing humans and their PPE in complex construction environments. By restructuring the YOLOv8 architecture, the model integrates the focus module to optimize the input feature representation and increase channel richness, providing sufficient support for subsequent feature extraction. The CBAM attention mechanism concentrates more attention on target regions like the human body and represses noise from complex construction backgrounds. The combination of the SPP structure and the C3TR module strengthens multi-scale feature extraction and global modeling, mitigating the impact of occlusion on detection. After introducing the SPP-C3TR joint model, the C2 module replaces C2f in the head layer to reduce redundant computation and avoid overfitting. To address inaccurate bounding box annotations and poor image quality in construction scenes, the WIoU-v3 loss function is adopted instead of CIoU and improves bounding box localization robustness. These improvements enhance the model's robustness and detection accuracy.

Ablation and comparative experiments indicate that TransSPCW-Net attains notable improvements over YOLOv8 and other enhanced baselines. Precision increases by 1.37% from 94.93%, recall by 2.68% from 90.97%, and mAP50 by 1.45% from 96.07%. The model performs well on small object detection and human recognition, while also reduces negative influence caused by occlusion and background noise. This indicates TransSPCW-Net is more stable and has higher detection precision in complex environments.

Based on the results of this paper, construction managers would replace traditional manual inspection with TransSPCW-Net to constantly detect whether the workers wear PPE correctly. This could reduce the occurrence of safety accidents on construction sites effectively. Furthermore, TransSPCW-Net features low computation resource consumption, which enables it to well adapt detection facilities on various construction sites.

Beyond direct safety detecting model, TransSPCW-Net can connect with alarming system to make real-time alerts to remind workers to wear PPE properly. Otherwise, it can also integrate weather monitoring system, so that the model can choose different pre-trained weights according to different weathers, such as normal weights, glare-prone weights, and low-visibility weights. Not just that, thanks to its strong ability to detect small objects and human bodies, TransSPCW-Net can also be used to detect the clothing or carried items of workers in indoor factory and office for safety prevention.

However, there are still some problems. Performance under extreme lighting conditions or dense occlusions can still be improved, and generalization to other construction site domains requires further verification. In the future, this work will expand the dataset size, continually track the latest YOLO developments to adapt and refine the framework. Strengthening robustness under nighttime and extreme weather conditions is also a future goal of this article.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author used DeepSeek solely for language editing and readability improvement. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

Table 6. Comparative experiment

	precision/%	recall/%	mAP50/%	val	
				box_loss	cls_loss
yolov8	94.932	90.976	96.073	0.86123	0.6959
yolov8x_DW_FOCUS2	95.515	87.742	96.056	0.96135	1.0496
yolov8x_DW_swin	94.491	88.212	97.18	0.76224	0.81849
yolov8s_DW_CA_BOT3	94.462	86.431	95.696	1.1764	1.0668
yolov8s_c2_DW	91.18	87.762	95.507	1.0156	1.228
yolov8x_DW_FOCUS_sppc	92.919	87.744	95.803	0.99989	1.0085
yolov8x_DW_swin_FOCUS	95.361	88.536	96.918	0.98488	0.96302
Faster R-CNN	95.627	91.843	96.429	0.82657	0.61834
EfficientDet-D4	95.986	92.751	97.105	0.79836	0.56241
TransSPCW-Net	96.309	93.662	97.517	0.7907	0.50558

References

Al Khiami, M. I., and ElHadad, M. M. (2024). Enhancing construction site safety using AI: The development of a custom YOLOv8 model for PPE compliance detection. In Proceedings of the *European Conference on Computing in Construction* (EC3), Chania, Crete, Greece, doi: 10.35490/EC3.2024.307.

Ali, M. L., and Zhang, Z. (2024). The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers,* 13(12), 336. doi: 10.3390/computers13120336.

Barlybayev, A., Krak, I., Amangeldy, N., Kurmetbek, B., Krak, I., Razakhova, B., Tursynova, N., & Turebayeva, R. (2024). Personal protective equipment detection using YOLOv8 architecture on object detection benchmark datasets: A comparative study. *Cogent Engineering,* 11(1), e2333209. doi: 10.1080/23311916.2024.2333209.

Chen, X., Jiao, Z., Liu, Y. (2024). Improved YOLOv8n-based helmet wearing inspection measurement method study. *Research Square* [Preprint]. doi: 10.21203/rs.3.rs-4733500/v1.

Edozie, E., Shuaibu, A. N., John, U. K., and Sadiq, B. O. (2025). Comprehensive review of recent developments in visual object detection based on deep learning. *Artificial Intelligence Review,* 58, 277. doi: 10.1007/s10462-025-11284-w.

Gallo, G., Di Rienzo, F., Garzelli, F., Ducange, P., and Vallati, C. (2022). A smart system for personal protective equipment detection in industrial environments based on deep learning at the edge. *Sensors,* 22(5), 1763. doi: 10.3390/s22051763.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904-1916. doi: 10.1109/TPAMI.2015.2389824.

Khalil, H. E., Abd El Aziz, M. R., Abdel Moneim, M. S. A., and Hassanien, A. E. (2024). A deep learning based approach for automated wind turbine detection using Sentinel 2 satellite imagery. *Remote Sensing,* 16(16), 2878. doi: 10.3390/rs16162878.

Lin, Y., Xiao, X., and Lin, H. (2025). YOLOv8-FDA: Lightweight wheat ear detection and counting in drone images based on improved YOLOv8. *Frontiers in Plant Science,* 16, 1682243. doi: 10.3389/fpls.2025.1682243.

López, L., Suárez-Ramírez, J., Alemán-Flores, M., and Monzón, N. (2025). Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation. *Automation in Construction*, 158, 105195. doi: 10.1016/j.autcon.2025.105195.

Lv, M., and Su, W.H. (2023). YOLOV5-CBAM-C3TR: An optimized model based on transformer module and attention mechanism for apple leaf disease detection. *Frontiers in Plant Science*, 14, 1323301. doi: 10.3389/fpls.2023.1323301.

Omar, M., Abidin, A. Z., and Filippini, D. (2024). Hardhat and safety vest detection in real-time using deep learning for safety compliance. Engineering, *Construction and Architectural Management*, 31(3), doi: 10.1108/ECAM-10-2022-0882.

Raditya, Y., Alunjati, F. A., Elviani, U., and Hidavat, F. (2024). Automated personal protective equipment (PPE) detection in construction environments: Enhancing safety through YOLOv8 object detection. In 2024 *IEEE International Conference on Information Systems Security* (ICISS), 1-6. doi: 10.1109/ICISS62896.2024.10751111.

Sapkota, R., Calero, M. F., Qureshi, R., Badgujar, C., Nepal, U., Poulose, A., Zeno, P., Vaddevolu, U. B. P., Khan, S., Shoman, M., Yan, H., and Karkee, M. (2025). YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series. *Artificial Intelligence Review*. doi: 10.1007/s10462-025-11253-3.

Tong, B., Li, G., Bu, X., Wang, Y., and Yu, X. (2025). A deep learning-based algorithm for the detection of personal protective equipment. *Plos one*, 20(5), e0322115. doi: 10.1371/journal.pone.0322115.

Varghese, R., and Sambath, M. (2024). YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In Proceedings of the *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems* (ADICS), 1-6. doi: 10.1109/ADICS58448.2024.10533619.

Wang, G., Zhao, X., Dang, D., Wang, J., and Chen, Y. (2025). Enhancing object detection with Shape-IoU and scale‐space‐task collaborative lightweight path aggregation. *Applied Sciences*, 15(22), 11976. doi: 10.3390/app152211976.

Wang, Z., Zhang, Y., and Zhang, S. (2025). Real-time personal protective equipment detection and classification with YOLOv8 multi-scale fusion. *Journal of Real-Time Image Processing*, 22, 131. doi: 10.1007/s11554-025-01715-w.

Woo, S., Park, J., Lee, J.Y. and Kweon, I.S., 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference On Computer Vision (ECCV)*, 3-19.

In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Tian (Eds.), Computer Vision ECCV 2018 (Lecture Notes in Computer Science, 11211, 3-19. Cham: Springer. doi: 10.1007/978-3-030-01234-2_1.

Zakariah, M., and Alnuaim, A. (2024). Recognizing human activities with the use of convolutional block attention module. *Egyptian Informatics Journal,* 27, 100536. doi: 10.1016/j.eij.2024.100536.

Zhang, L., Ma, H., Huang, J., Zhang, C., and Gao, X. (2025). An improved lightweight safety helmet detection algorithm for YOLOv8. *Computers, Materials and Continua*, 83(2), 2245-2265. doi: 10.32604/cmc.2025.061519.

Jinhao Da is an undergraduate student majoring in Automation at Beijing Institute of Technology, expected to graduate in 2027. His research interests include computer vision, object detection, and intelligent recognition algorithms. Jinhao Da has participated in several research and innovation projects, including a university-level innovation and entrepreneurship project on ship seam welding defect detection and target recognition for intelligent unmanned vehicles.