Electronics and
Electrical Engineering

https://www.aimspress.com/journal/ElectrEng

AIMS Electronics and Electrical Engineering, 9(3): 260–287.
DOI: 10.3934/electreng.2025013
Received: 07 March 2025
Revised: 27 April 2025
Accepted: 07 May 2025
Published: 16 May 2025

Research article

Real-Time personnel safety detection in chemical parks using YOLOv8-ARR

Zhong Wang[1,2,3], Lanfang Lei[2], Tong Li[1,]*and Peibei Shi[1]

[1] School of Computer and Artificial Intelligence, Hefei Normal University, Hefei 230601, China
[2] School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China
[3] Hefei Institute for Public Safety Research, Tsinghua University, Hefei, 230601, China

* **Correspondence:** Email: kamilt@foxmail.com; Tel: +86-135-156-44010.

Abstract: With the rapid development of the chemical industry, personnel safety has become a critical concern. Traditional monitoring systems, which rely heavily on human surveillance, are inefficient and often inaccurate. To address this issue, we proposed a real-time detection and identification system for personnel safety in chemical parks based on an advanced artificial intelligence algorithm named YOLOv8-ARR. The key contributions of this system include: (1) The introduction of Adaptive Powerful-IoU(APIoU) as a network optimization for bounding box regression loss, which effectively balances gradient gains between high-quality and low-quality samples, enhancing model localization; (2) a novel attention mechanism, Reinforced Channel Prioritized Contextual Attention(RCPCA), to improve background information extraction; (3) replacing traditional convolution with RFAConv to assign different weights to each receptive field position and feature channel, highlighting crucial details; (4) the use of a bidirectional feature pyramid network (BiFPN) for the weighted fusion of multi-scale feature maps; and (5) the addition of a small object detection layer in the YOLOv8 network to enhance the detection of small targets. Experimental results on a custom dataset of chemical park workers showed that our model improves the mean Average Precision (mAP@0.5) by 5.475% compared to the standard model. This system provides a more accurate solution for identifying abnormal behaviors and potential risks in chemical parks compared to traditional methods. Additionally, it significantly reduces dependency on human resources, minimizes false positives and negatives, and enhances monitoring efficiency and safety.

Keywords: safety behavior detection; chemical industrial park; YOLOv8-ARR algorithm; real-time monitoring system; attention mechanism RCPCA

1. Introduction

With the rapid development of the chemical industry, the safety management of chemical parks has become a crucial aspect of maintaining industrial ecological balance. Chemical parks, due to the presence of numerous hazardous chemicals and high-risk operational environments, demand strict management of personnel safety behaviors. Statistics show that frequent industrial accidents in China's manufacturing sector have resulted in significant casualties and economic losses each year, with most of these accidents caused by operator violations, such as smoking in work areas or failing to wear safety equipment. Consequently, enhancing safety management levels in chemical parks is imperative, prompting increasing attention from scholars and researchers toward exploring effective safety management strategies and technologies.

A significant portion of production accidents in the chemical industry is attributed to unsafe behaviors of personnel. For instance, behaviors such as smoking in hazardous areas or not wearing protective equipment are major causes of accidents. According to data from the State Administration of Work Safety, over 70% of accidents each year are due to unsafe behaviors. This situation not only threatens the lives of workers but also brings substantial economic losses and reputational damage to enterprises.

With the rapid advancements in computer technology and continuous breakthroughs in artificial intelligence, machine vision recognition technology based on image processing has emerged as a vital tool for enhancing monitoring efficiency. Machine vision technology spans from traditional target detection to deep learning-based object detection. Traditional methods typically use HOG, Haar features, or color information combined with classifiers like AdaBoost or Support Vector Machines (SVM) for target recognition. In the framework of deep learning, object detection algorithms have evolved into two major categories: Region-based two-stage detection algorithms, such as R-CNN, Fast R-CNN [1], and Faster R-CNN [2] and direct regression-based single-stage detection algorithms, including SSD [3] and the YOLO [4]series. These advanced machine vision technologies offer new solutions for safety management in chemical parks, significantly improving the accuracy and real-time performance of monitoring, thereby effectively preventing and reducing industrial accidents.

Although deep learning algorithms have shown great potential in detecting smoking behavior and helmet-wearing compliance in chemical industrial parks, they face numerous challenges in practical applications. On the one hand, the monitoring environment in chemical plants is often complex, with drastic lighting changes, high personnel density, and frequent occlusions, all of which pose significant obstacles to the detection of small targets. For instance, small objects such as distant workers or smoke are easily missed or misidentified, while complex background interference, such as occlusion or backlighting, can significantly reduce detection accuracy. On the other hand, traditional video image analysis methods rely on manual feature extraction, which is not only inefficient but also poorly adaptable to dynamic industrial environments. Although some industrial parks have deployed deep learning-based detection models in their equipment, these models often fail to meet practical demands, exhibiting low detection rates along with frequent false positives and false negatives. Therefore, improving the model's ability to detect small targets in complex environments, enhancing robustness, and effectively reducing false and missed detections remain critical research issues in this field.

To achieve accurate classification and precise localization of personnel safety in chemical parks, meet the precision requirements of industrial equipment, reduce false alarms and missed detections, and significantly decrease the workload of manual inspection, we propose a personnel safety behavior detection system based on an artificial intelligence framework. The major contributions of this algorithm include:

(1) **APIoU Bounding Box Regression Loss**: We introduce APIoU as the bounding box regression loss for network optimization. This modification effectively balances the gradient gains between high-quality and low-quality samples, improving the model's localization capabilities.

(2) **RCPCA Attention Mechanism**: A novel attention mechanism, RCPCA, is proposed, enabling the model to better extract background information, thereby enhancing detection performance.

(3) **RFAConv to Replace Traditional Convolution**: The use of RFAConv instead of traditional convolution (Conv) assigns different weights to each receptive field position and feature channel, highlighting important detail information.

(4) **Bidirectional Feature Pyramid Network (BiFPN)**: The neck design incorporates a Bidirectional Feature Pyramid Network (BiFPN) for the weighted fusion of multi-scale feature maps, enhancing the model's ability to detect targets of varying scales.

(5) **Small Object Detection Layer**: An additional small object detection layer is integrated into the YOLOv8 network, significantly improving the detection capability for small targets.

The structure of this paper is as follows: In Section 2, we provide an overview of related work and the current state of research; in Section 3, we introduce the methodology, detailing the proposed algorithm's framework and implementation specifics, including the loss function, RCPCA attention mechanism, RFAConv, and BiFPN; in Section 4, we discuss the experimental setup, results, and provide a discussion on the effectiveness of the method; finally, in Section 5, we conclude the research findings and discuss future research directions.

2. Related works

In the context of personnel behavior safety recognition in chemical plants, traditional target detection methods extract relevant features using descriptors and employ classifiers to detect worker safety based on category information. For instance, Rubaiyat et al. [5] utilized Discrete Cosine Transform (DCT) to extract frequency domain information and HOG features from images, employing SVM to identify candidate worker regions. They further used Circular Hough Transform and color combination features to detect whether workers were wearing safety helmets. Sun et al. [6] improved a visual background extractor to detect moving targets and determined the helmet position based on the head-to-body ratio. They applied Principal Component Analysis (PCA) for feature vector dimensionality reduction and used a Bayesian-optimized SVM model to recognize safety helmets, followed by the Mean Shift algorithm to track them. Additionally, Huang et al. [7] designed a smoking behavior detection method, using SVM to model and classify smoke features and smoking actions. Despite their effectiveness to some extent, these traditional methods face challenges in complex real-world scenarios, including high time complexity, low robustness, low accuracy, and susceptibility to false positives and false negatives.

Deep learning-based two-stage object detection methods involve initially identifying candidate regions, followed by classification and localization of these regions to detect worker safety. Park et al. [8] employed a Region-based Fully Convolutional Network (R-FCN) for object detection and classification, utilizing transfer learning techniques to train the model for safety helmet detection. Widiarsini et al. [9] constructed a smoking detection model based on Region-based Convolutional Neural Networks (R-CNN) and used Mask R-CNN for cigarette segmentation. While two-stage detection algorithms perform well in terms of accuracy, their detection speed is relatively slow, making it challenging to meet real-time requirements. Sond et al. [10] utilized Very Deep Residual Networks (VD-ResNet) to detect construction workers in image sequences with varying postures and backgrounds, providing efficient and accurate technical support for real-time detection.

Deep learning-based single-stage object detection methods achieve a balance between detection accuracy and real-time performance, making them widely applicable in engineering practices. These methods use global information to regress target detection bounding boxes and category information directly from the entire image. For instance, Wang et al. [11] replaced the residual convolutional layers in YOLOv3's Darknet-53 network with Inception-ResNet modules, modified the number of convolutional layers to enhance network performance, and added detection scales for small targets. They used the K-means algorithm to cluster anchor boxes, thereby improving the network's ability to detect small safety helmets and uniforms. Deng et al. [12] improved YOLOv4 for helmet detection, using K-means clustering to obtain optimal prior boxes and multi-scale training to enhance suitability across detection scales. Tan et al. [13] extended YOLOv5 by adding scales for small target detection and introduced the DIoU-NMS algorithm to replace the traditional non-maximum suppression, resulting in more accurate suppression of helmet prediction boxes. Fu et al. [14] proposed the GD-YOLO network based on YOLOv7, which includes an efficient feature extraction module, D-LAN, for detecting smoking and phone usage behaviors. Li et al. [15] developed an improved YOLO-PL algorithm for helmet detection in construction environments, aiming to enhance detection accuracy and real-time performance. Nath et al. [16] introduced a real-time detection method for construction worker personal protective equipment (PPE) compliance based on deep learning, which helps reduce construction site accidents and improve safety compliance. Li et al. [17] proposed the YOLOv5-SFE algorithm, integrating spatial and temporal features to improve the accuracy of detecting and recognizing worker behaviors in real-time.

Although traditional methods and deep learning approaches have made certain progress in personnel behavior safety recognition in chemical environments, challenges remain in terms of detection accuracy in complex scenarios. Traditional methods have high computational complexity and lack robustness, making it difficult to handle complex backgrounds and variable lighting conditions typically encountered in practical applications.

Two-stage detection models, such as Faster R-CNN and R-FCN, although showing significant advantages in detection accuracy, involve higher computational costs due to the separation of candidate region generation and subsequent classification, which makes them less efficient for high-performance processing. In recent years, Transformer-based models (such as DETR) have achieved end-to-end training and made significant progress, but their high computational complexity and limitations in processing large-scale datasets restrict their application in resource-constrained environments. In contrast, YOLO series models, by adopting a single-stage detection framework, directly regress the target bounding boxes and category information from the entire image, achieving a good balance between accuracy and efficiency. Especially, YOLOv8n, with its streamlined network structure and optimized training strategies, performs exceptionally well in multi-class object detection tasks in complex backgrounds, effectively addressing detection requirements in various safety monitoring scenarios. This gives YOLOv8n a clear advantage in applications that require high safety standards, such as construction sites and chemical plants.

In summary, considering the system's requirements for detection accuracy and efficiency, we choose YOLOv8n as the baseline model. It achieves an optimal balance between accuracy and efficiency, and can run stably in resource-constrained monitoring environments, meeting the practical needs of high-safety scenarios such as chemical plants.

3. Research method

The YOLO (You Only Look Once) algorithm, introduced in 2016, is renowned for its speed and real-time capabilities. However, as a single-stage algorithm, it often leads to higher false-positive and false-negative rates, especially with small and densely packed objects. To address these limitations, the YOLO algorithm has undergone continuous optimization, resulting in several iterations, including YOLOv2 [18], YOLOv3 [19], YOLOv4 [20], YOLOX [21], and YOLOv7 [22], as well as modifications like YOLO-Tiny. YOLOv8, one of the latest single-stage object detection algorithms, strikes a good balance between detection accuracy and speed. Given the practical application scenarios in chemical parks, YOLOv8n is particularly suitable due to its simpler network structure, minimal computational requirements, and fastest running speed, making it highly portable. Therefore, we focus on further improving the YOLOv8n model. The network structure of the YOLOv8n model comprises four main components: Input, Backbone, Neck, and Head. The Input stage involves Mosaic data augmentation to enrich the dataset and an anchor-free strategy to reduce the number of predicted boxes, thereby accelerating the Non-Maximum Suppression (NMS) process. The Backbone includes Conv, C2f, and SPPF (Spatial Pyramid Pooling-Fast) modules, which handle convolution, batch normalization, and feature extraction. The Neck combines the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) to ensure effective feature fusion across different stages. The Head uses a Decoupled Head strategy, separating the classification head from the detection head, and derives target class and location information from feature maps of three different scales. By refining these components, the improved YOLOv8n model is tailored to better address the challenges of detecting unsafe behaviors in chemical parks, thereby enhancing safety management.

In this paper, we design an improved target detection algorithm for chemical parks, named YOLOv8-ARR, using YOLOv8n as the baseline model. The specific contributions of this work are as follows: (1) We introduce APIoU as the network optimization for bounding box regression loss, effectively balancing the gradient gains between high-quality and low-quality samples and enhancing the model's localization capabilities. (2) We propose a novel attention mechanism, RCPCA, enabling the model to better extract background information and improve feature extraction capabilities. (3) We design the RFConv module to replace the original Conv module, assigning different weights to each receptive field position and feature channel, highlighting important detail information. Additionally, we modify the stride of P2 in the backbone network to 1, enhancing the backbone's extraction capabilities and reducing false positives and missed detections in chemical park images. (4) The neck design incorporates a Bidirectional Feature Pyramid Network (BiFPN) for the weighted fusion of multi-scale feature maps. (5) Considering the need for shallow feature maps for small targets, we introduce an additional P2 small object detection head in the head network to more effectively capture the details and local features of small targets. The YOLOv8-ARR structure diagram is shown in Figure 1.

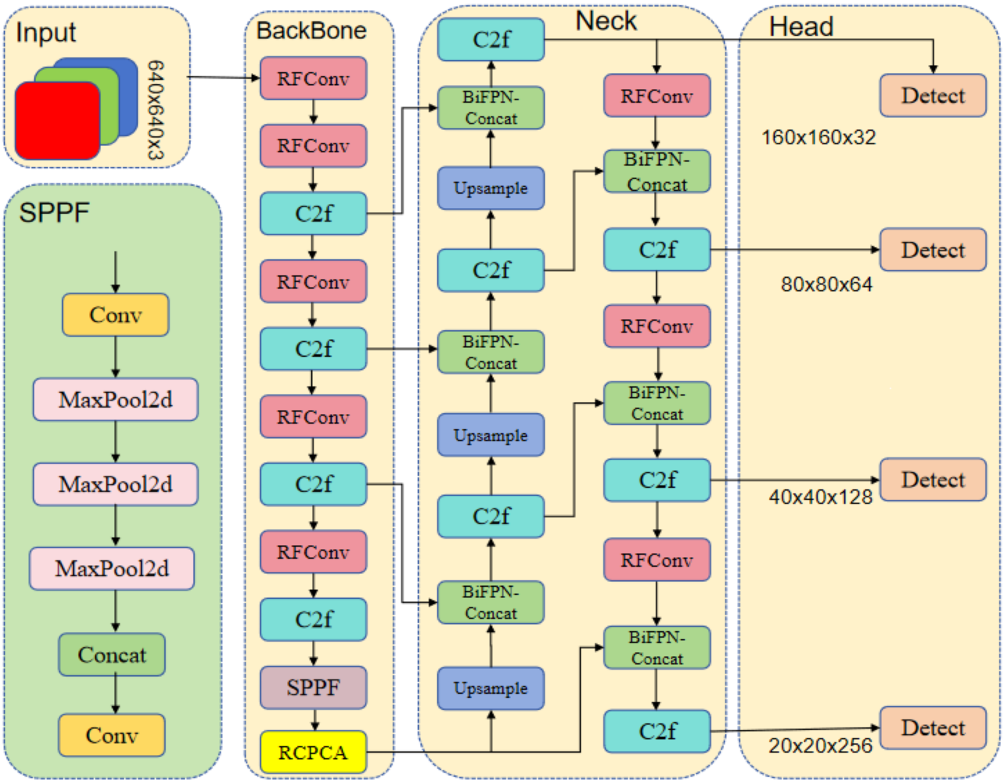

Figure 1. YOLOv8-ARR Structure Diagram. The RCPCA attention mechanism enables the model to better extract background information, enhancing its feature extraction capabilities. The RFConv module is designed to replace the original Conv module, assigning different weights to each receptive field position and feature channel, highlighting important detail information. A Bidirectional Feature Pyramid Network (BiFPN-Concat) is introduced in the neck design for the weighted fusion of multi-scale feature maps.

3.1. Adaptive powerful-IoU (APIoU)

The loss function is a crucial component in evaluating samples for deep neural networks. Choosing the appropriate loss function significantly impacts the model's convergence speed, thereby enhancing detection accuracy and minimizing false positives and false negatives. Accurate localization in chemical park detection is particularly essential. The YOLOv8 model employs Complete Intersection over Union (CIoU) for bounding box regression. The CIoU loss function is defined as follows:

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{d^2}{c^2} + \alpha v \tag{1}$$

where IoU represents the Intersection over Union, d is the distance between the center points of the predicted and ground truth boxes, and c is the diagonal length of the smallest enclosing box covering the predicted and ground truth boxes. The expressions for α and v are given by:

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

$$\alpha = \frac{v}{(1 - IoU + v)} \tag{3}$$

Here, h^{gt} and w^{gt} denote the height and width of the ground truth box, while h and w represent the height and width of the predicted box.

Although the CIoU loss function introduces two penalty terms to account for the differences in center point distance and aspect ratio, it does not directly capture the shape differences between the anchor box and the target box. This can lead to suboptimal solutions or unstable convergence during training. Moreover, these penalty terms do not reflect changes in the size of the target box, potentially affecting the model's performance in detecting objects of various sizes.

The IoU loss function is affected by unreasonable penalty factors, leading to the expansion of anchor boxes during regression, significantly slowing down the convergence rate. Some loss functions even cause the anchor boxes to increase in size. Therefore, we introduce and improve the PIoU loss function. The PIoU loss function incorporates a target size-adaptive penalty factor and a gradient adjustment function based on anchor box quality, guiding the anchor box regression along an efficient path to achieve faster convergence than the existing IoU loss. The PIoU loss function is given by Eq (4).

$$\mathcal{L}_{PIoU} = 1 - PIoU \tag{4}$$

where PIoU and p are defined in Eqs (5) and (6):

$$PIoU = IoU + e^{-p^2} - 1 \tag{5}$$

$$p = \frac{\left(\frac{dw_1}{w_{gt}} + \frac{dw_2}{w_{gt}} + \frac{dh_1}{h_{gt}} + \frac{dh_2}{h_{gt}}\right)}{4} \tag{6}$$

In the detection process within chemical industrial parks, the pixels occupied by objects in the field of view are influenced by lighting conditions and object types. Therefore, to balance the positioning and size of the target bounding boxes and enhance the model's generalization ability in different target scenarios, we have improved the PIoU loss function to obtain the APIoU loss. In the PIoU loss function, we introduce an area ratio, which is the ratio of the product of the predicted box area and the ground truth box area to the square of the area of the smallest enclosing box for the predicted and ground truth boxes. To better illustrate APIoU, we have drawn the structure diagram in Figure 2. The APIoU loss function is given by Eq (7):

$$\mathcal{L}_{PIoU} = 1 - APIoU \tag{7}$$

where APIoU and $area$ are defined in Eqs (8) and (9):

$$APIoU = IoU + e^{-p^2} + e^{-area^2} - 2 \tag{8}$$

$$area = \frac{(w * h) * (w_{gt} * h_{gt})}{(w_c * h_c)^2} \tag{9}$$

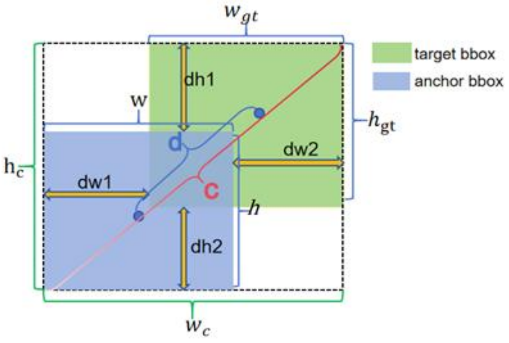

Figure 2. The structure of APIoU.

In complex industrial scenarios such as chemical parks, targets often exhibit significant variations in scale and are frequently subject to severe occlusion. To address these challenges, we have optimized the APIoU loss function by introducing an area ratio term, which more accurately captures the size variations of bounding boxes and enhances the model's ability to detect objects of different scales. Additionally, by integrating the regression task with area-related information, the loss function strengthens the model's feature learning capability, enabling it to better capture subtle differences in object shapes and improving adaptability to multi-scale object detection. This effectively overcomes the limitations of traditional PIoU in localization accuracy and better meets the demands of chemical parks for high-precision and robust object detection.

3.2. Reinforced channel prioritized contextual attention (RCPCA)

We propose a reinforced channel-priority contextual attention mechanism (RCPCA) to enhance the model's ability to extract background information, thereby improving object detection performance. In chemical park environments, objects are usually small or occluded, making traditional appearance-based detection methods ineffective. Therefore, extracting background information becomes crucial. The RCPCA mechanism enhances the model's ability to handle complex backgrounds and small objects by integrating three key components: A multi-scale feature extraction unit (MSFU), feature fusion with residual connections, and a channel-priority contextual attention mechanism (CPCA), as shown in Figure 3. MSFU extracts diverse feature representations from multiple scales, while feature fusion and residual connections retain important spatial information. CPCA dynamically adjusts the importance of different channels and guides the model to focus on the most relevant contextual information for accurate object detection. This mechanism significantly improves the robustness and accuracy of the model, especially in challenging scenarios such as occlusion and complex background.

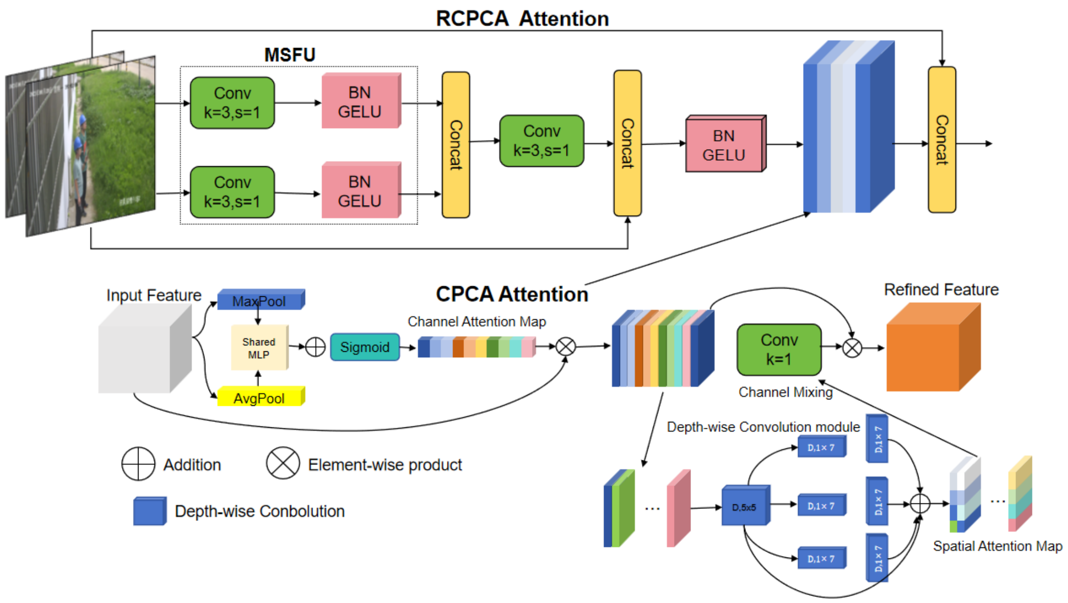

Figure 3. The structure of the RCPCA attention mechanism.

3.2.1. Multi-scale feature extraction unit (MSFU)

The core of the RCPCA consists of two key feature extraction modules: conv1 and conv2 of the MSFU. Conv1 utilizes a 3×3 convolutional kernel to capture local features, while conv2 uses a 5×5 convolutional kernel to capture broader contextual information. The outputs of these modules are normalized using BatchNorm2d and activated using the GELU activation function, introducing non-linear transformations that provide rich multi-scale feature representations for subsequent feature fusion. Given an input feature map $F \in R^{H \times W \times C}$, where H and W represent the height and width of the feature map, respectively, and C represents the number of channels, the mathematical expressions for MSFU are as follows:

$$F_{conv\,1} = \mathrm{GELU}\Big(\mathrm{BatchNorm2d}\big(\mathrm{Conv2d}(F, 3 \times 3,1,1)\big)\Big) \tag{10}$$

$$F_{conv\,2} = \mathrm{GELU}\Big(\mathrm{BatchNorm2d}\big(\mathrm{Conv2d}(F, 5 \times 5,1,2)\big)\Big) \tag{11}$$

3.2.2. Feature fusion and residual connection

Based on the multi-scale features generated by MSFU, channel fusion is performed using a 1×1 convolutional kernel, effectively integrating feature dimensions to form richer feature representations. This fusion strategy helps to combine information from different scales, providing more comprehensive input for the subsequent attention mechanism. Following this, a residual connection mechanism is introduced, adding the fused features to the original input to maintain the deep feature transmission of the network and enhance the model's learning ability. This design helps to alleviate the vanishing or exploding gradient problem in deep network training, while enabling the network to learn more complex feature representations. Finally, the attention features generated by CPCA are multiplied by the result of the feature fusion to obtain enhanced feature representations. The feature fusion strategy is implemented through the 1×1 convolution operation as follows:

$$F_{cat} = \text{Conv2d}(\text{Concatenate}(F_{conv\,1}, F_{conv\,2}), 1 \times 1,1,0) \tag{12}$$

The residual connection is:

$$F_{res} = F_{cat} + F \tag{13}$$

3.2.3. Channel prioritized contextual attention (CPCA)

The CPCA mechanism, proposed by Huang et al. [23], is a lightweight and high-performance convolutional neural network attention mechanism that dynamically allocates attention weights in both the channel and spatial dimensions. This mechanism more effectively utilizes the information in the input feature map, enhancing the network model's feature extraction capabilities. The CPCA attention mechanism consists of two major parts: The channel attention module and the spatial attention module.

The channel attention module enhances the feature representation of each channel by calculating the weights of each channel. Channels containing significant or important feature information are assigned larger weights, while channels with less important feature information are assigned smaller weights. First, the input feature map $F(H \times W \times C)$ undergoes global max pooling and global average pooling along the channel dimension to compute the maximum and average feature values for each channel, resulting in two feature vectors ($1 \times 1 \times C$) representing the global maximum and average features of each channel. These two feature vectors are then input into a two-layer shared multilayer perceptron (MLP), where the first layer has C/r neurons (being the reduction ratio), and the second layer has C neurons. This MLP is used to learn the attention weights for each channel. By learning these weight parameters, the network can adaptively determine which channels are more important for the current task. The outputs of the MLP are then element-wise summed and processed through a Sigmoid function to obtain the final channel attention weight vector $CA(F)$. The calculation formula is:

$$\text{CA}(F) = \sigma\Big(\text{MLP}\big(\text{AvgPool}(F)\big) + \text{MLP}\big(\text{MaxPool}(F)\big)\Big) \tag{14}$$

where AvgPool represents global average pooling, MaxPool represents global max pooling, and σ represents the Sigmoid activation function. The channel attention weight vector $\text{CA}(F)$ is element-wise multiplied with the input feature map F to generate the channel-refined feature map F_{cr}. The calculation formula is:

$$F_{cr} = \text{CA}(F) \otimes F \tag{15}$$

where $\otimes$ denotes element-wise multiplication.

The spatial attention module captures spatial structural information of the image by calculating spatial weights for each pixel. Complementing the channel attention module, the spatial attention module dynamically updates the feature information at each pixel location based on its spatial weight. First, the channel-refined feature map F_{cr}, generated by the channel attention module, is used as the input feature map for this module. F_{cr} is fed into a depthwise convolution layer with a kernel size of 5, producing an intermediate feature map F_m. Next, F_m is passed through three depthwise separable convolution paths L_1, L_2, and L_3 to obtain feature maps of different scales: F_{L1}, F_{L2}, and F_{L3}. Path L_1has convolution kernels of sizes (1,7) and (7,1); path L_2 has convolution kernels of size (1,11) and (11,1); and path L_3 has convolution kernels of sizes (1,21) and (21,1). These operations effectively capture multi-scale spatial information within the channels. Then, the feature maps F_m,

F_{L1}, F_{L2}, and F_{L3} are element-wise summed. The fused features are processed through a convolution layer with a kernel size of 1, ensuring the integration of channel information and the effective extraction of spatial information, resulting in the spatial attention feature map $SA(F_{cr})$. The calculation is as follows:

$$\mathrm{SA}(F_{cr}) = \mathrm{Conv}_{1\times1}\left(\sum_{i=0}^{3} \mathrm{Branch}_i\left(\mathrm{DwConv}(F)\right)\right) \tag{16}$$

where DwConv represents depthwise convolution, $Branch_i$, $(i \in \{0,1,2,3\})$ represents the i-th branch, and Branch_0 is the residual connection. $Conv_{n\times m}$ denotes a convolution operation with a kernel size of (n,m).

The spatial attention feature map $SA(F_{cr})$ is then element-wise multiplied with the channel-refined feature map F_{cr} to generate the CPCA attention-enhanced feature F_{cpca}. The calculation formula is:

$$F_{cpca} = \mathrm{SA}(F_{cr}) \otimes F_{cr} \tag{17}$$

3.3. RFAConv

By introducing RFAConv [24] to replace the Conv module, we improved the YOLOv8n model. The specific structure is shown in Figure 4. RFAConv uses a receptive field weight matrix to assign different weights to each receptive field position and feature channel, highlighting important detail information. Additionally, RFAConv dynamically generates receptive field spatial features and adapts the shape and range of the receptive field according to the size of the convolutional kernel to accommodate targets of different sizes. It generates smaller receptive fields for small targets to retain fine details, and larger receptive fields for larger targets to capture global features. This flexible adjustment of receptive field size improves detection accuracy for targets of various sizes.

The module implements a lightweight convolutional layer (group convolution), saving many network parameters. Additionally, the generated attention mechanism enables the network to focus on learning important information at each feature map level. Assuming $F \epsilon R^{C\times H\times W}$ and $F' \epsilon R^{C\times H\times W}$ are the input and output feature maps, respectively, the working principle of RFAConv can be expressed as follows:

$$F' = \mathrm{Conv2D}^{3\times3}\left(\mathrm{Reshape}(A_{RF} \times F_{RA})\right), \tag{18}$$

where A_{RF} is the receptive field attention map, F_{RA} is the receptive field spatial feature, $Conv2D^{3\times3}$ is a 3×3 standard convolution, and Reshape is a reshaping operation that changes the tensor dimensions. The formulas for A_{RF} and F_{RA} are as follows:

$$A\_RF = \mathrm{Softmax}(g^{1\times1}\ (\mathrm{AvgPool}(F))) \tag{19}$$

$$F_{RA} = \mathrm{ReLU}\left(\mathrm{BN}\left(g^{3\times3}(F)\right)\right) \tag{20}$$

where $g^{1\times1}$ represents a group convolution operation with a kernel size of 1×1, AvgPool is an average pooling layer, and BN is batch normalization. Softmax and ReLU are activation functions.

By integrating the receptive field attention convolution (RFAConv) into the YOLOv8n network, detection performance is significantly enhanced. RFAConv emphasizes the detailed features of targets, reduces information loss, and more accurately locates targets, providing reliable safety support and assurance for personnel safety in chemical industrial parks.

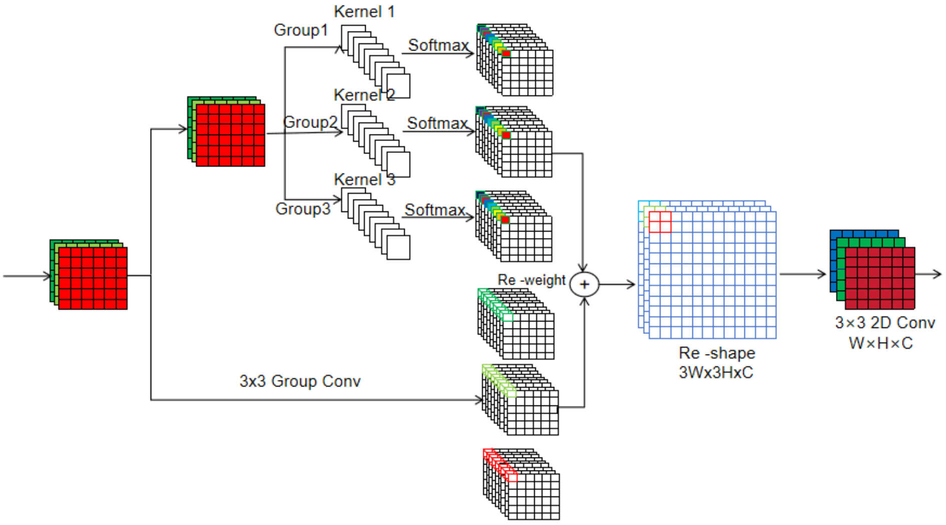

Figure 4. The structure of RFAConv.

3.4. BiFPN-Concat

To solve the problem of detecting targets with various scales in the complex environment of the chemical industrial park, we optimized the feature fusion method. The traditional Concat structure treats high-level semantic features and low-level detail features equally during fusion, resulting in insufficient detection accuracy of the model for multi-scale targets such as helmet wearers under occlusion and smoking behavior at a distance.

To this end, we designed the BiFPN-Concat structure and introduced the weighted fusion strategy of BiFPN to achieve adaptive integration of features: for small targets, low-level detail features are strengthened to retain position information, and for large targets, high-level semantic features are emphasized to enhance category discrimination, significantly improving the multi-scale detection performance in complex scenarios. The core advantage of BiFPN lies in the bidirectional feature pyramid architecture and dynamic weighting mechanism; the bidirectional path supports the bidirectional flow of features between high and low layers, retaining the position details of low-level features and integrating the semantic information of high-level features; and the weighted strategy avoids the "equal treatment" defect of traditional fusion methods by learning the importance weights of feature maps, effectively reducing information redundancy and loss. Compared with unidirectional fusion structures such as FPN, BiFPN demonstrates stronger feature expression capabilities and detection efficiency in multi-scale target detection in chemical parks. Its dynamic weights are continuously optimized during training, enabling the model to adaptively adjust the fusion strategy according to the importance of the input features, providing an efficient solution for accurate detection in complex scenarios. The formula for adjusting the fusion weights is as follows:

$$O = \Sigma_i \frac{w_i I_i}{\epsilon + \Sigma_j w_j} \tag{21}$$

where w_i represents the learnable weights, with ReLU ensuring $w_i \geq 0$, I_i represents the features of the i-th layer, and a small constant is introduced to avoid numerical instability.

By integrating the BiFPN structure into the neck network of the model and replacing the original Concat module, we not only improved the detection accuracy of the model but also enhanced its robustness in complex environments. The BiFPN-Concat structure is particularly suitable for target detection in the complex scene of the chemical industrial park, which can meet the needs of precise control of the feature fusion process, thereby further improving the practical application performance of the model.

3.5. Improved spatial pyramid pooling

SPP (Spatial Pyramid Pooling), proposed by He et al. [25], is a pooling structure used for image processing and computer vision tasks. This structure can perform standard pooling on images of different sizes and ultimately combine them into feature vectors of the same size as the input to the fully connected layer. Considering that some targets in chemical industrial parks are small and require high accuracy in target detection networks, we adopted the SPPCSPC module to replace the original SPPF module in YOLOv8 to improve the model. The SPPCSPC module integrates the CSP (Cross Stage Partial) structure on the basis of the SPP module.

In SPPCSPC, the overall input is divided into two different branches. The 3 × 3 convolution in the middle is not grouped and remains a standard convolution, while the right side uses pointwise convolution. Finally, the information streams output by all branches are concatenated. The SPPCSPC module provides significant improvements in target detection networks compared to the original SPP module and the SPPF module used in YOLOv8, particularly for smaller targets. The structure of the SPPF module is shown in Figure 5.

The SPPCSPC structure mainly consists of two substructures: The SPP structure and the CSPC (Cross Stage Partial Connections) structure. The main idea is to introduce cross-stage partial connections into the network to replace the traditional serial connection method in convolutional neural networks for feature propagation. This addresses the bottleneck problem in information transmission, improves feature propagation efficiency, and better utilizes information between low-level and high-level features. Adopting the SPPCSPC structure is beneficial for recognizing objects in chemical industrial parks, as the model can better extract features related to lighting, texture, and other target characteristics.

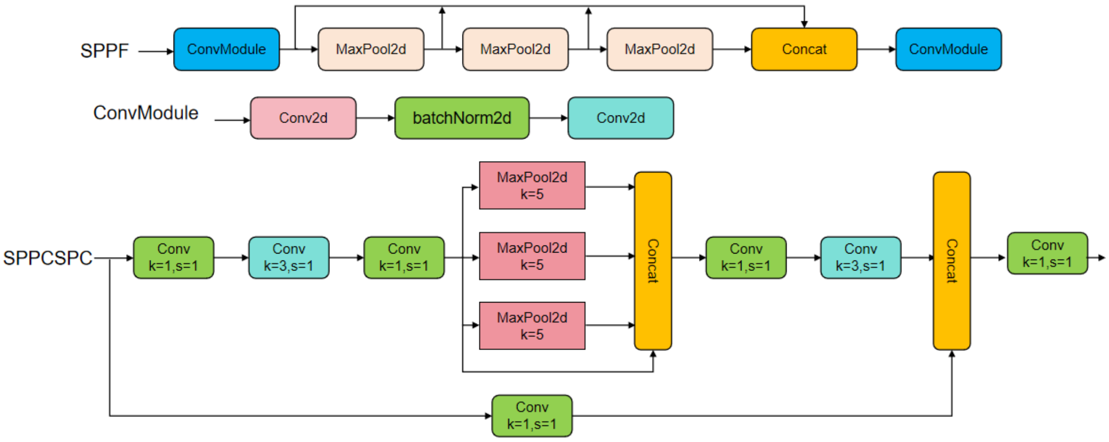

Figure 5. SPPF module.

4. Experimental results

4.1. Experimental environment and performance metrics

To validate the effectiveness of this method, an experimental platform was established using Ubuntu 18.04 as the operating system and PyTorch as the deep learning framework. YOLOv8n was used as the baseline network model. The specific configuration of the experimental environment is shown in Table 1.

Table 1. Experimental environment configuration table.

Environmental Parameter	Value
Operating system	Ubuntu18.04
Deep learning framework	PyTorch
programming language	Python3.8
CPU	Intel Xeon Scale 8358
GPU	NVIDIA A100(SXM4, 80GB)
RAM	256 GB

Consistent hyperparameters were applied throughout the training process for all experiments. Table 2 shows the exact hyperparameters used during training.

Table 2. Training hyperparameters.

Hyperparameters	Value
Learning Rate	0.01
Image Size	640 × 640
Momentum	SGD
Batch Size	16
Epoch	200
Weight Decay	0.0005

We selected Precision (P), Recall (R), F1 Score, and mean Average Precision (mAP@0.5) as evaluation metrics to assess the effectiveness of the improved network. The calculation formulas are as follows:

$$IoU = \frac{A \cup B}{A \cap B} \tag{22}$$

$$Precision = \frac{TP}{TP+FP} \tag{23}$$

$$Recall = \frac{TP}{TP+FN} \tag{24}$$

$$AP = \sum_{i=1}^{n-1}(r_{i+1} - r_i)\, P_{inter}\, (r_i + 1) \tag{25}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{26}$$

$$F1\, Score\ = \frac{Precision\ *Recall}{Precision\ +Recall} \tag{27}$$

In Eq (22), A represents the ground truth box, B represents the predicted box, $A \cap B$ represents the intersection area of A and B, and $A \cup B$ represents the union area of A and B. In Eqs. (23) and (24), TP denotes true positives, FP denotes false positives, and FN denotes false negatives. In Eq. (25), $r_{1,}r_2, \cdots, r_n$ are the recall values corresponding to the first interpolation segment sorted in ascending order of precision P. In Eq (26), k denotes the number of classes, which is 2 in this study. In Eq (27) is the F1 score formula.

4.2. Dataset description

There are no publicly available datasets on the internet specifically describing workers in chemical industrial parks. Therefore, the dataset for this study was collected manually and from online sources, comprising 12,048 images. Sample images are shown in Figure 6. The data was split into training and testing sets in a ratio of 0.8 to 0.2, using the labeling tool "labelImg" for annotation. The annotated data includes image size, the location of objects to be detected within the images, and the types of behaviors to be recognized. The label types are divided into two categories: "smoke" and "hat".

Figure 6. Sample images.

4.3. Experimental results

4.3.1. Comparison of loss functions

To evaluate the impact of different loss functions, we used the YOLOv8 model as a baseline and selected eleven loss functions, CIoU, SIoU, GIoU, DIoU, EIoU, InnerIoU, ShapeIoU, MDPIoU, NWD, WIoU, and PIoU, along with our newly developed APIoU for experimental comparison based on mAP50 and mAP50-95 metrics. Table 3 presents the experimental results. mAP50 is an important metric for assessing the performance of object detection models, reflecting the model's ability to accurately detect target objects. The experimental results show that the APIoU loss function significantly outperforms the other loss functions, achieving the highest mAP50 value of 0.84703. Notably, compared to the initial CIoU, the mAP50 value of APIoU improved by 1.2777%, and by approximately 0.7% compared to PIoU. This indicates that APIoU not only addresses the problem of anchor box area expansion, making the positioning and size of target bounding boxes more balanced by introducing the area ratio, but also adapts to the complex environments of chemical industrial parks, achieving broader detection performance. Especially when dealing with smaller, dim, or feature-blurred targets, APIoU can achieve precise boundary box regression, thereby obtaining accurate target location and size information. This further validates the effectiveness of the APIoU loss function.

Table 3. Comparison of loss functions.

Loss Function	mAP50	mAP50-95	smoke	hat
CIoU [26]	0.83426	0.52213	0.806	0.865
SIoU [27]	0.83331	0.52313	0.802	0.863
GIoU [28]	0.83124	0.52351	0.802	0.858
DIoU [26]	0.83067	0.52584	0.797	0.865
EIoU [29]	0.82232	0.52337	0.785	0.859
InnerIoU [30]	0.83271	0.52227	0.803	0.861
ShapeIoU [31]	0.83105	0.52379	0.807	0.855
MDPIoU [32]	0.83427	0.52346	0.804	0.863
NWD [33]	0.83689	0.52234	0.808	0.864
WIoU [34]	0.84064	0.52526	0.810	0.869
PIoU [35]	0.84066	0.52671	0.811	0.869
APIoU	0.84703	0.53358	0.825	0.869

4.3.2. Comparison of attention mechanisms

We compared 23 different attention mechanisms, including TripletAttention, CBAM, SimAM, and others, as shown in Table 4. The results indicate that the performance of the RCPCA attention mechanism is significantly better than the other attention mechanisms in the experiments. The RCPCA achieved mAP50 and mAP50-95 values of 0.84343 and 0.53034, respectively, which are approximately 0.4% and 0.3% higher than those of CPCA. The smoke detection remained at 0.812, and the hat detection value was 0.864. This demonstrates that RCPCA performs better in complex environments of chemical industrial parks, particularly in detecting small and occluded objects and handling the need for background information.

The data shows that RCPCA, due to its unique MSFU, feature fusion with residual connections, and CPCA mechanism, significantly outperforms other competing attention models in handling image tasks in chemical industrial parks. This validates the effectiveness and optimization of the proposed RCPCA attention mechanism.

Table 4. Comparison of attention mechanisms.

Model	mAP50	mAP50-95	Smoke	hat
TripletAttention [36]	0.83715	0.52758	0.812	0.862
CBAM [37]	0.83864	0.52787	0.806	0.871
SimAM [38]	0.83184	0.52682	0.804	0.860
PolarizedAttention [39]	0.83837	0.52649	0.813	0.863
BiLevelnchwAttention [40]	0.83887	0.53097	0.809	0.865
BiLevelRoutingAttention [40]	0.83065	0.52524	0.799	0.860
SpatialGroupEnhance [41]	0.83065	0.52463	0.801	0.859
SpatialAttention [42]	0.84051	0.5285	0.812	0.869
FocalModulation [43]	0.83374	0.52634	0.812	0.854
MLCA [44]	0.8399	0.52907	0.812	0.867
LSKblock [45]	0.83585	0.53055	0.803	0.867
deformableLKA [46]	0.83893	0.53302	0.807	0.869
SKAttention [47]	0.83816	0.52944	0.809	0.865
SEAttention [48]	0.83828	0.52597	0.813	0.860
ParNetAttention [49]	0.83816	0.52906	0.809	0.862
MHSA [42]	0.8371	0.52793	0.810	0.862
EfficientChannelAttention [50]	0.8362	0.52864	0.809	0.863
DoubleAttention [51]	0.63624	0.38685	0.520	0.749
CoTAttention [52]	0.8361	0.52482	0.808	0.863
EffectiveSEModule [53]	0.83551	0.52637	0.807	0.863
DAttention [54]	0.83733	0.52683	0.808	0.866
EMA [55]	0.83778	0.52699	0.807	0.867
CPCA [23]	0.83861	0.52747	0.812	0.864
RCPCA	0.84343	0.53034	0.824	0.863

4.3.3. Model comparison experiments

To validate the performance of the YOLOv8-ARR model in recognizing personnel safety behaviors in complex scenarios, we compared it with several mainstream object detection models. The experiment used a self-built dataset to evaluate the performance of each model in the complex environment of a chemical industrial park. As shown in Table 5, YOLOv8-ARR performed outstandingly in key metrics such as mAP50, smoke detection accuracy, and hat detection accuracy, especially showing a clear advantage in personnel safety behavior recognition tasks.

YOLOv8-ARR achieved an mAP50 of 0.8890, a 5.45% improvement over YOLOv8n (0.8345). Compared to other models, YOLOv8-ARR demonstrated significantly higher detection accuracy than YOLOv3-tiny (0.7907) and YOLOv5n (0.8224), further confirming its precision advantage in complex scenarios. In smoke detection, YOLOv8-ARR achieved an accuracy of 0.8620, improving by 5.6% over YOLOv8n (0.8060) and performing significantly better than YOLOv3-tiny (0.7690) and YOLOv5n (0.7920), highlighting its exceptional performance in complex environments. In hat detection, YOLOv8-ARR achieved an accuracy of 0.9160, a 5.3% improvement over YOLOv8n (0.8630), surpassing YOLOv3-tiny (0.8120) and YOLOv5n (0.8520), demonstrating its clear advantage in head protection detection.

Although YOLOv8-ARR has higher FLOPs (92.0 G) and Params (16.1 M), this increase is justified by the improvement in accuracy. YOLOv8-ARR incorporates several enhanced modules, such as the APIoU loss function and RCPCA attention mechanism, which significantly enhance its feature extraction capability. Therefore, the increase in FLOPs and Params is necessary to support higher accuracy detection, and this increase is essential for the performance improvement.

Table 5. Comparison of different models.

Model	FLOPs/G	Params/M	mAP50	smoke	hat
Faster RCNN	13.1	41.3	0.6687	0.7627	0.5747
YOLOv3-tiny [56]	13.0	8.7	0.7907	0.7690	0.8120
YOLOv5n	4.2	1.8	0.8224	0.7920	0.8520
YOLOv6 [57]	11.9	4.2	0.8221	0.7990	0.8520
YOLOv7 [22]	103.2	37.2	0.7729	0.6890	0.8568
YOLOX [21]	**15.4**	10.6	0.8344	0.7852	0.8837
DETR [58]	187.0	41.0	0.5508	0.3948	0.7068
SSD [3]	87.6	26.3	0.7389	0.6770	0.8007
CenterNet [59]	70.2	32.6	0.8243	0.7830	0.8656
YOLOv8n	8.1	3.0	0.8345	0.8060	0.8630
YOLOv8-ARR	92.0	16.1	0.8890	0.8620	0.9160

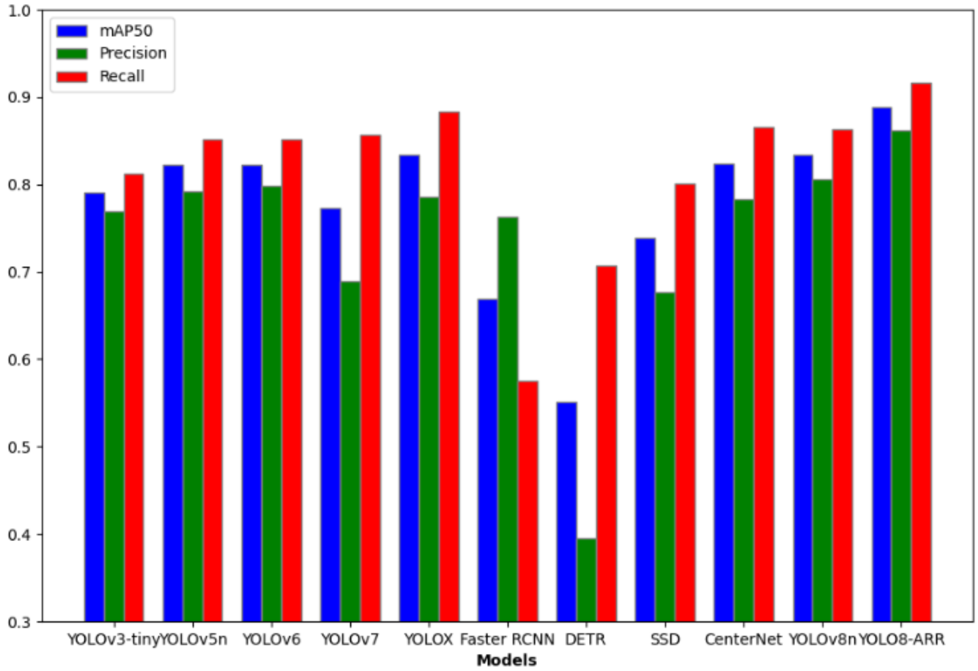

Figure 7. Detection results of different models.

Overall, YOLOv8-ARR's significant improvement in smoke and hat detection further validates its potential for application in complex environments. It demonstrates outstanding robustness and precise recognition capabilities, especially valuable for personnel safety behavior detection in high-risk environments. As shown in Figure 7, we can see the comparative metrics of each model. Therefore, it is evident that the YOLOv8-ARR model demonstrates superior recognition performance in identifying safety behaviors of personnel in complex scenes within chemical industrial parks compared to other models.

4.3.4. Ablation experiment

The proposed YOLOv8-ARR model is an optimized version of YOLOv8n, achieved by improving the loss function with APIoU, enhancing the attention mechanism with RCPCA, incorporating BiFPN_Concat, and introducing RFAConv and SPPCSPC. To evaluate the performance of each optimization module, an ablation study was conducted using a variable control method, with training and testing performed on the same dataset and training parameters. The results are shown in Table 6.

Table 6. Ablation experiment results.

Model	YOLOv8	APIoU	RC	RFA	BiFPN	SPPC	P2	Recall	Precious	mAP50	mAP50-95	Smoke	Hat
			PCA	Conv	_Concat	SPC							
1	√							**0.781**	**0.863**	**0.834**	**0.522**	**0.806**	**0.863**
2	√	√						**0.782**	**0.888**	**0.847**	**0.533**	**0.825**	**0.868**
3	√	√	√					**0.788**	**0.895**	**0.853**	**0.541**	**0.833**	**0.874**
4	√	√	√	√				**0.788**	**0.903**	**0.857**	**0.543**	**0.833**	**0.886**
5	√	√	√	√	√			**0.799**	**0.889**	**0.859**	**0.546**	**0.837**	**0.881**
6	√	√	√	√	√	√		**0.820**	**0.902**	**0.885**	**0.568**	**0.855**	**0.916**
Our	√	√	√	√	√	√	√	**0.820**	**0.903**	**0.889**	**0.570**	**0.862**	**0.916**

The results indicate that using the improved APIoU loss function significantly enhances detection performance, with mAP50 and mAP50-95 increasing by 1.3% and 1.1%, respectively, and other metrics also showing slight improvements. To ensure the model pays more attention to the background information required for object detection in chemical industrial parks, the improved RCPCA attention mechanism was introduced, resulting in a 0.6% improvement in mAP50, mAP50-95, and helmet detection, with a notable 0.8% improvement for small smoking targets. Both Recall and Precision show significant improvements.

To emphasize important detail information during the detection process in chemical industrial parks, RFAConv was introduced. This module uses a receptive field weight matrix to assign different weights to each receptive field position and feature channel, reducing information loss and more accurately locating targets, providing reliable safety support. The mAP50 and mAP50-95 increase by 0.4% and 0.2%, respectively, with a notable improvement of 1.2% for helmet targets.

To address the challenges of diverse scale variations in the complex environment of chemical industrial parks, BiFPN_Concat was introduced. This advanced feature fusion and transformation mechanism is better suited for detection applications in such environments. Smoke detection accuracy improves by 0.2%, with other metrics remaining stable, laying a foundation for future improvements.

Considering the small size of some targets and the high precision requirements of the target detection network in chemical industrial parks, the SPPCSPC structure was adopted. This structure enhances the recognition of objects in these environments, with improved extraction of features such as lighting and texture. Both mAP50 and mAP50-95 increased by 0.6%, and smoke and hat detection improve by 1.2% and 1.7%, respectively. Precision increases by 1.8%. Finally, the introduction of a small target layer (p2) significantly improves the detection rate of small smoking targets, reaching 0.862, an increase of 0.7%.

To effectively demonstrate the improvements of our model in complex environments of chemical industrial parks, we tested different scenarios within these environments. As shown in Figure 8. In the tests for group A, we can see that our proposed YOLOv8-ARR significantly improves the accuracy of detecting four helmets compared to YOLOv8n. The results were similarly positive for group B, where YOLOv8-ARR provides more stable detections. In group C, YOLOv8-ARR not only maintains high precision but also resolves false positive issues. For further validation, we conducted frame-by-frame detection using high-definition camera footage. In Figure D, we observe a significant improvement in the accuracy of detecting helmets and smoking. In Figure E, YOLOv8n produces a false positive for a person, which YOLOv8-ARR successfully resolves. In Figure F, YOLOv8n shows issues with missed detections and instability, whereas YOLOv8-ARR does not exhibit these problems. In summary, our algorithm can accurately classify and precisely locate personnel safety behaviors in chemical industrial parks. It meets the precision requirements of industrial equipment, reduces false and missed detections, significantly decreases the workload of manual inspections, and improves on-site management efficiency.

4.3.5. Experimental comparison based on public datasets

To comprehensively evaluate the generalization ability and robustness of the proposed YOLOv8-ARR model, we conducted additional validation experiments on widely recognized public benchmark datasets, Pascal VOC and VisDrone. The VOC dataset covers various object categories and real-world scenes, making it an ideal choice for assessing general object detection performance. In contrast, the VisDrone dataset contains aerial images with small objects, dense distributions, and complex backgrounds, presenting significant detection challenges. The experimental results are summarized in Table 7, demonstrating the outstanding detection accuracy and robustness of YOLOv8-ARR across scenarios.

On the VisDrone2019 dataset, characterized by complex object scales, orientations, and significant occlusions, YOLOv8-ARR significantly outperforms the baseline model YOLOv8n. Specifically, precision increases from 0.45032 to 0.47041 (+2.01%), showing an enhanced ability to correctly identify positive samples and reduce false positives. Recall improves by 0.56% (from 0.33466 to 0.34023), highlighting better performance in capturing true positive samples. mAP@50 increases significantly from 0.33459 to 0.47041 (+13.58%), demonstrating substantial improvement in both object localization and classification in challenging scenarios. mAP@50-95 also shows improvement, rising from 0.19337 to 0.29023 (+9.69%), further confirming enhanced robustness in detecting overlapping objects. The F1 score increases from 0.38397 to 0.39487 (+1.09%), indicating a better balance between precision and recall.

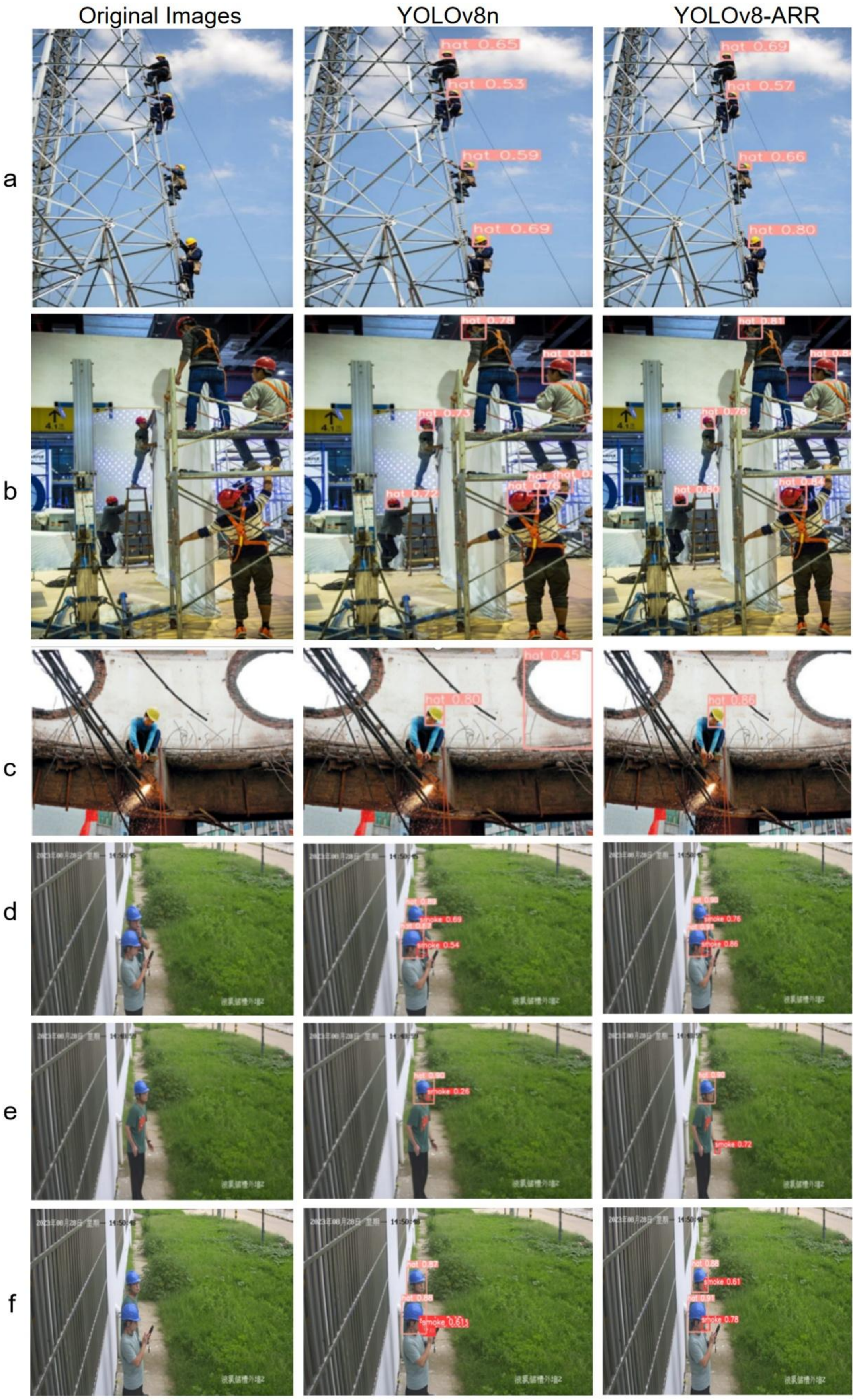

Figure 8. Comparison of model ablation experiment results.

On the VOC2007 dataset, YOLOv8-ARR maintains its competitive edge, with precision rising to 0.76342 (+0.72%), effectively reducing the false detection rate. Recall increases by 4.03% (from 0.62236 to 0.66270), further emphasizing the model’s reliability in detecting actual objects. mAP@50 increases from 0.69885 to 0.73531 (+3.78%), reflecting stronger object localization accuracy. mAP@50-95 improves slightly from 0.48257 to 0.48257 (+0.22%), further proving the model's precision in various scenarios. The F1 score increases from 0.68277 to 0.70950 (+2.67%), further demonstrating optimization in the balance between precision and recall.

For the VOC2012 dataset, YOLOv8-ARR shows more significant improvement compared to YOLOv8n. Precision increases from 0.62079 to 0.63006 (+0.38%), and recall rises by 4.88% (from 0.51201 to 0.56085). mAP@50 increases from 0.57222 to 0.59054 (+1.83%), clearly highlighting substantial progress in classification and object localization performance. mAP@50-95 improves from 0.41163 to 0.41163 (+1.49%), reflecting the model’s ability to maintain high-quality detection performance across a wider range of IoU thresholds. The F1 score rises from 0.56118 to 0.59344 (+3.23%), further indicating better optimization of the trade-off between precision and recall.

Overall, the experimental results strongly demonstrate that YOLOv8-ARR significantly improves detection accuracy in complex scenarios involving diverse scales, orientations, and occlusions, making it a robust and reliable model for practical deployment in various application environments.

Table 7. Model comparison study on public datasets.

Datasets	Model	Precision	Recall	mAP50	mAP50-95	F1
VisDrone2019	YOLOv8n	0.45032	0.33466	0.33459	0.19337	0.38397
	YOLOv8-ARR	0.47041 **(+2.01%)**	0.34023 **(+0.56%)**	0.47041 **(+13.58%)**	0.29023 **(9.69%)**	0.39487 **(1.09%)**
VOC2007	YOLOv8n	0.75618	0.62236	0.69885	0.48257	0.68277
	YOLOv8-ARR	0.76342 **(+0.72%)**	0.66270 **(+4.03%)**	0.73531 **(+3.78%)**	0.48257 **(+0.22%)**	0.70950 **(2.67%)**
VOC2012	YOLOv10n	0.62079	0.51201	0.57222	0.41163	0.56118
	YOLOv8-ARR	0.63006 **(+0.38%)**	0.56085 **(+4.88%)**	0.59054 **(+1.83%)**	0.41163 **(+1.49%)**	0.59344 **(3.23%)**

4.3.6. Grad-CAM visualization analysis

To better demonstrate the outstanding performance of the YOLOv8-ARR model in smoking detection and safety helmet detection, we conducted a detailed visualization analysis of the model’s detection results using Grad-CAM. The Grad-CAM images in Figure 9 clearly show how the model focuses on key regions in the image during detection. Through this image analysis, we can gain a more intuitive understanding of how the model locates and recognizes targets in real-world tasks.

In Group (a), the YOLOv8-ARR model successfully identified the smoke region, and compared to the YOLOv8n model, its confidence increased by 3%. The Grad-CAM image visually demonstrates how the YOLOv8-ARR model concentrates on the smoke source area, further proving its significant advantage in smoking detection tasks.

In Group (b), we show the model’s performance in detecting multiple safety helmets. For the first helmet on the left, the confidence of the YOLOv8-ARR model increased by 4% compared to YOLOv8n; for the first helmet on the right, the confidence increased by 1%. The Grad-CAM image clearly highlights the model’s focus on the head region, further emphasizing the superior performance of YOLOv8-ARR in multi-target detection and its efficient detection capability in complex environments.

The exceptional performance of YOLOv8-ARR not only offers significant advantages in improving detection accuracy but also demonstrates its robustness in complex environments and multi-target detection capabilities, making it highly valuable in real-world applications. Particularly in high-risk environments, such as personnel safety monitoring systems, YOLOv8-ARR provides efficient and real-time detection services, greatly enhancing security and reliability.

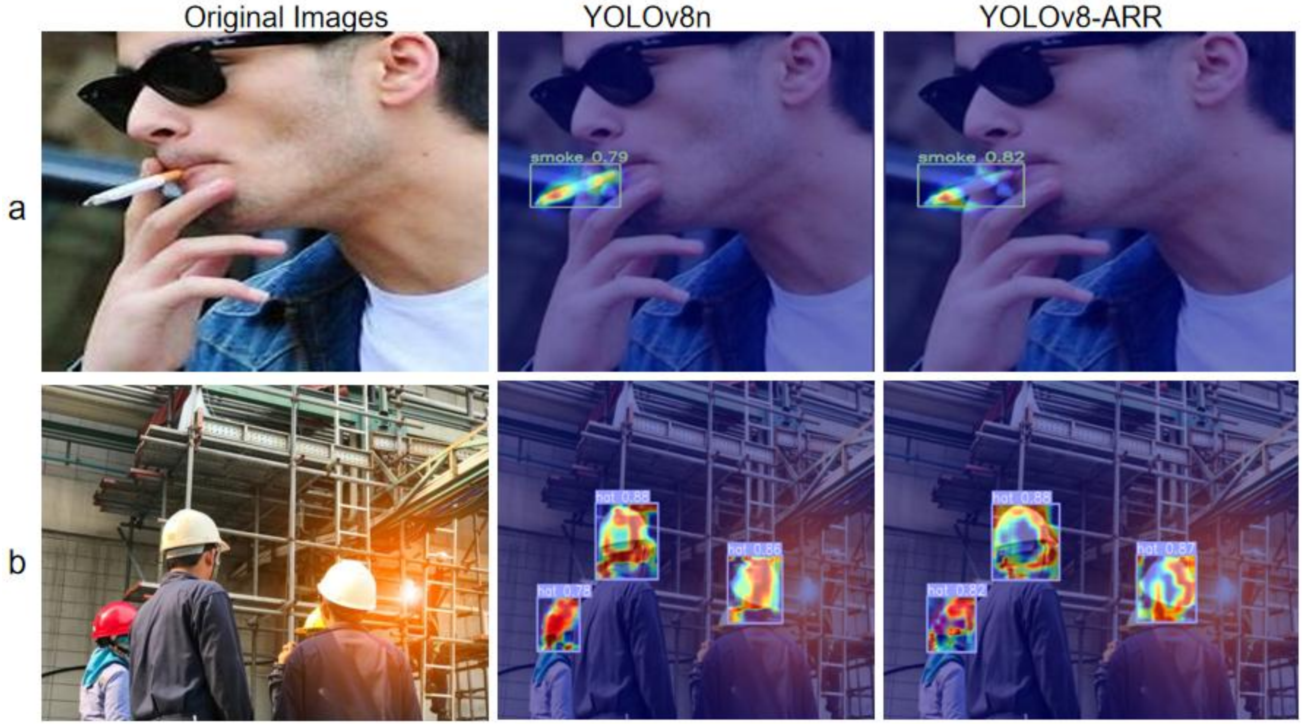

Figure 9. Grad-CAM visualization results.

5. Conclusions

In this paper, we propose a new algorithm for detecting personnel safety behaviors in the complex environments of chemical industrial parks, based on YOLOv8-ARR. The goal of this algorithm is to accurately classify and precisely locate safety behaviors, meeting the precision requirements of industrial equipment while reducing false positives and false negatives. This can significantly reduce the workload of manual inspections and improve on-site management efficiency. YOLOv8-ARR is optimized based on YOLOv8n. First, the algorithm introduces the enhanced Intersection over Union (APIoU) as an optimization method for bounding box regression loss, effectively balancing the gradient gain between high-quality and low-quality samples, thereby improving the accuracy of object localization. Second, a Reinforced Channel-Priority Contextual Attention (RCPCA) mechanism is proposed to improve the extraction of background information, enhancing the model's robustness in complex backgrounds. Next, the RFAConv convolution layer replaces traditional convolution layers, assigning different weights to each receptive field location and feature channel, which highlights critical details and enhances feature extraction ability. Then, the Bidirectional Feature Pyramid Network (BiFPN) is employed to perform weighted fusion of multi-scale feature maps, further strengthening the model’s ability to handle multi-scale objects. Finally, a small object detection layer is added, significantly improving the detection accuracy of small objects. Experimental results show that the YOLOv8-ARR model, on a custom chemical industrial park personnel dataset, improves the mean average precision (mAP@0.5) by 5.475% compared to the original YOLOv8n model, significantly increasing accuracy and effectively reducing false positives and false negatives. However, the model faces challenges in cases of extremely small targets or severe occlusion. Future research will introduce super-resolution technology, and integrate image segmentation and multi-view fusion methods to further enhance the model's robustness in complex environments. Future research will incorporate super-resolution technology and combine image segmentation with multi-view fusion methods to improve robustness in complex environments. Additionally, the focus will be on real-time deployment of the model, optimizing the graphical user interface (GUI), and developing mobile applications to enable on-the-go monitoring of personnel safety behaviors in chemical industrial parks. Last, infrared cameras will be used in conjunction with image enhancement algorithms for preprocessing, and spatiotemporal fusion methods will be employed to further enhance real-time detection capabilities in complex environments.

Author contributions

Zhong Wang: Writing – review & editing, Writing –original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization; Lanfang Lei: Writing – review &editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization; Tong Li: Methodology, Writing – review & editing, Investigation, Formal analysis; Peibei Shi: Validation, Supervision, Methodology.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61976198), and the Natural Science Research Key Project for Colleges and University of Anhui Province (Grant No. 2022AH052141 and 2022AH052142), and the 2023 Humanities and Social Science General Program sponsored by the Ministry of Education of the People’s Republic of China (Grant No. 23YJCZH067), and the Hefei Municipal Natural Science Foundation (Grant No. 202322).

Conflict of interest

The authors declare there is no conflict of interest in this paper.

References

1. Girshick R (2015) Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*.
2. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
3. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. (2016) Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37. Springer. https://doi.org/10.1007/978-3-319-46448-0_2
4. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788. https://doi.org/10.1109/CVPR.2016.91
5. Rubaiyat AH, Toma TT, Kalantari-Khandani M, Rahman SA, Chen L, Ye Y, et al. (2016) Automatic detection of helmet uses for construction safety. *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*, 135–142. IEEE. https://doi.org/10.1109/WIW.2016.045
6. Sun X, Xu K, Wang S, Wu C, Zhang W, Wu H (2021) Detection and tracking of safety helmet in factory environment. *Meas Sci Technol* 32: 105406. https://doi.org/10.1088/1361-6501/ac06ff
7. Huang X, Jia K, Liu P (2020) Automatic Detection of Taxi Driver Smoking Behavior Based on Traffic Monitoring. *Computer Simulation* 37: 337–344.
8. Park S, Yoon S, Heo J (2019) Image-based automatic detection of construction helmets using R-FCN and transfer learning. *KSCE Journal of Civil and Environmental Engineering Research* 39: 399–407.
9. Widiarsini KU, Khrisne DC, Suyadnya IMA (2021) Automatic Cigarette Object Concealment in Video using R-CNN.
10. Son H, Choi H, Seong H, Kim C (2019) Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks. *Automat Constr* 99: 27–38. https://doi.org/10.1016/j.autcon.2018.11.033
11. Wang X, Niu D, Luo P, Zhu C, Ding L, Huang K (2020) A safety helmet and protective clothing detection method based on improved-yolo v 3. *2020 Chinese automation congress (CAC)*, 5437–5441. IEEE. https://doi.org/10.1109/CAC51589.2020.9327187
12. Benyang D, Xiaochun L, Miao Y (2020) Safety helmet detection method based on YOLO v4. *2020 16th International conference on computational intelligence and security (CIS*), 155–158. IEEE. https://doi.org/10.1109/CIS52066.2020.00041
13. Tan S, Lu G, Jiang Z, Huang L (2021) Improved YOLOv5 network model and application in safety helmet detection. *2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*, 330–333. IEEE. https://doi.org/10.1109/ISR50024.2021.9419561
14. Fu Y, Ran T, Xiao W, Yuan L, Zhao J, He L, et al. (2024) GD-YOLO: An improved convolutional neural network architecture for real-time detection of smoking and phone use behaviors. *Digital Signal Processing* 151: 104554. https://doi.org/10.1016/j.dsp.2024.104554
15. Li H, Wu D, Zhang W, Xiao C (2024) YOLO-PL: Helmet wearing detection algorithm based on improved YOLOv4. *Digital Signal Processing* 144: 104283. https://doi.org/10.1016/j.dsp.2023.104283

16. Nath ND, Behzadan AH, Paal SG (2020) Deep learning for site safety: Real-time detection of personal protective equipment. *Automat Constr* 112: 103085. https://doi.org/10.1016/j.autcon.2020.103085
17. Li L, Zhang P, Yang S, Jiao W (2023) YOLOv5-SFE: An algorithm fusing spatio-temporal features for detecting and recognizing workers' operating behaviors. *Adv Eng Inform* 56: 101988. https://doi.org/10.1016/j.aei.2023.101988
18. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 7263–7271. https://doi.org/10.1109/CVPR.2017.690
19. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767.*
20. Bochkovskiy A, Wang CY, Liao HYM (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934.*
21. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) YOLOX: Exceeding YOLO Series in 2021. *arXiv preprint arXiv:2107.08430.*
22. Wang CY, Bochkovskiy A, Liao HYM (2022) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
23. Huang H, Chen Z, Zou Y, Lu M, Chen C, Song Y, et al. (2024) Channel prior convolutional attention for medical image segmentation. *Comput Biol Med* 178: 108784.
24. Zhang X, Liu C, Yang D, Song T, Ye Y, Li K, et al. (2023) Rfaconv: Innovating spatital attention and standard convolutional operation. *arXiv preprint arXiv:2304.03198.*
25. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE T Pattern Anal* 37: 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
26. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI conference on artificial intelligence,* 34: 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999
27. Gevorgyan Z (2022) SIoU loss: More powerful learning for bounding box regression. *arXiv preprint arXiv:2205.12740.*
28. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: A metric and a loss for bounding box regression. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 658–666. https://doi.org/10.1109/CVPR.2019.00075
29. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T (2022) Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *Neurocomputing* 506: 146–157. https://doi.org/10.1016/j.neucom.2022.07.042
30. Zhang H, Xu C, Zhang S (2023) Inner-IoU: more effective intersection over union loss with auxiliary bounding box. *arXiv preprint arXiv:2311.02877.*
31. Zhang H, Zhang S (2023) Shape-IoU: More Accurate Metric considering Bounding Box Shape and Scale. *arXiv preprint arXiv:2312.17663.*
32. Siliang M, Yong X (2023) Mpdiou: a loss for efficient and accurate bounding box regression. *arXiv preprint arXiv:2307.07662.*

33. Wang J, Xu C, Yang W, Yu L (2021) A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv preprint arXiv:2110.13389*.
34. Tong Z, Chen Y, Xu Z, Yu R (2023) Wise-IoU: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*.
35. Liu C, Wang K, Li Q, Zhao F, Zhao K, Ma H (2024) Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism. *Neural Networks* 170: 276–284. https://doi.org/10.1016/j.neunet.2023.11.041
36. Misra D, Nalamada T, Arasanipalai AU, Hou Q (2021) Rotate to Attend: Convolutional Triplet Attention Module. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3139–3148. https://doi.org/10.1109/WACV48630.2021.00318
37. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
38. Yang L, Zhang RY, Li L, Xie X (2021) SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. *International Conference on Machine Learning*, 11863–11874. PMLR.
39. Liu H, Liu F, Fan X, Huang D (2022) Polarized Self-Attention: Towards High-quality Pixel-wise Regression. *Neurocomputing* 506: 158–167. https://doi.org/10.1016/j.neucom.2022.07.054
40. Zhu L, Wang X, Ke Z, Zhang W, Lau RW (2023) Biformer: Vision transformer with bi-level routing attention. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10323–10333. https://doi.org/10.1109/CVPR52729.2023.00995
41. Li X, Hu X, Yang J (2019) Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*.
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention Is All You Need. *Advances in neural information processing systems* 30.
43. Yang J, Li C, Dai X, Gao J (2022) Focal modulation networks. *Advances in Neural Information Processing Systems* 35: 4203–4217.
44. Wang H, Guo P, Zhou P, Xie L (2024) MLCA-AVSR: Multi-Layer Cross Attention Fusion based Audio-Visual Speech Recognition. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8150–8154. https://doi.org/10.1109/ICASSP48485.2024.10446769
45. Li Y, Hou Q, Zheng Z, Cheng MM, Yang J, Li X (2023) Large selective kernel network for remote sensing object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16794–16805. https://doi.org/10.1109/ICCV51070.2023.01540
46. Azad R, Niggemeier L, Hüttemann M, Kazerouni A, Aghdam EK, Velichko Y, et al. (2024) Beyond self-attention: Deformable large kernel attention for medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1287–1297. https://doi.org/10.1109/WACV57701.2024.00132
47. Li X, Wang W, Hu X, Yang J (2019) Selective kernel networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 510–519. https://doi.org/10.1109/CVPR.2019.00060

48. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
49. Goyal A, Bochkovskiy A, Deng J, Koltun V (2022) Non-deep Networks. *Advances in neural information processing systems* 35: 6789–6801.
50. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542. https://doi.org/10.1109/CVPR42600.2020.01155
51. Chen Y, Kalantidis Y, Li J, Yan S, Feng J (2018) A^2-nets: Double attention networks. *Advances in neural information processing systems* 31.
52. Li Y, Yao T, Pan Y, Mei T (2022) Contextual Transformer Networks for Visual Recognition. *IEEE T Pattern Anal* 45: 1489–1500.
53. Lee Y, Park J (2020) Centermask: Real-time anchor-free instance segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13906–13915. https://doi.org/10.1109/CVPR42600.2020.01392
54. Xia Z, Pan X, Song S, Li LE, Huang G (2022) Vision transformer with deformable attention. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4794–4803. https://doi.org/10.1109/CVPR52688.2022.00475
55. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. (2023) Efficient multi-scale attention module with cross-spatial learning. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 1–5. IEEE. https://doi.org/10.1109/ICASSP49357.2023.10096516
56. Hu L, Li Y (2021) Micro-YOLO: Exploring Efficient Methods to Compress CNN based Object Detection Model. In *ICAART (2),* 151–158. https://doi.org/10.5220/0010234401510158
57. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. (2022) YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976.*
58. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-End Object Detection with Transformers. *European conference on computer vision*, 213–229. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58452-8_13
59. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: Keypoint triplets for object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, 6569–6578. https://doi.org/10.1109/ICCV.2019.00667