Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research on Multi-object Detection in Security Video Surveillance Based on Improved YOLOv8n

Jincheng Huang
University of Science and Technology Beijing
Weiyi Min
University of Science and Technology Beijing
Ze Tian
University of Science and Technology Beijing
Jianing Liu
University of Science and Technology Beijing
Huiling Jiang
87398930@qq.com

University of Science and Technology Beijing

Research Article

Keywords: YOLOv8n, SEAttention, WIoU V3 loss function, small object detection head, multi-class object detection

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8559020/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Research on Multi-object Detection in Security Video Surveillance Based on Improved YOLOv8n

Huang Jincheng[1], Min Weiyi[1], Tian Ze[1], Liu Jianing[1], Jiang Huiling[1]

Abstract

To address the issues of slow processing efficiency, low accuracy, and the difficulty in multi-object recognition of large video streams in the cloud, this paper optimizes the security video detection algorithm based on the YOLOv8n model. It introduces the SEAttention mechanism into the backbone network of the model to enhance the perception of key area features, adds a small object detection head (named P2) to provide higher-resolution feature maps for better detection of small objects, and adopts the WIoU V3 loss function to optimize the bounding box regression. To test the effect of multi-object recognition, the network data and on-site video slices are integrated, and after preprocessing, unified annotation, quality verification, and multi-dimensional enhancement, a multi-task dedicated dataset for fire smoke and safety helmets is constructed. The experimental results show that the improved YOLOv8n model has a precision rate of 86.5%, a recall rate of 77.5%, and an mAP50 of 84.3%. The core indicators have all improved compared to the original model, and the detection effect of small objects such as flames and safety helmets has been significantly improved. Moreover, the number of parameters and computational requirements are suitable for deployment on edge computing devices. This model effectively solves the real-time detection problem of fire smoke and safety helmets in complex environments, verifies its applicability in industrial security scenarios, and provides technical support for the intelligent security of industrial facilities.

Keywords: YOLOv8n, SEAttention, WIoU V3 loss function, small object detection head, multi-class object detection

Funding Declaration：

This work was supported by the National Key R&D Program of China (Grant 2023YFC3009800) on the R&D and Application of Intelligent Monitoring and Early Warning System for Early Signs of Electrical Fires.

Huang Jincheng
3296703306@qq.com
Min Weiyi
m202510118@xs.ustb.edu.cn
Tian Ze
m202320033@xs.ustb.edu.cn
Liu Jianing
m202420016@xs.ustb.edu.cn
Jiang Huiling（corresponding author）
87398930@qq.com
University of Science and Technology Beijing, Beijing 100083, China

1. Introduction

With the acceleration of industrial intelligence, security monitoring has evolved from traditional passive manual duty to a modern system based on intelligent perception and active early warning [1], becoming the core technical support for ensuring industrial production safety and preventing accidents [2].

Although deep learning, especially YOLO series real-time target detection algorithms, has greatly improved the real-time recognition ability of security systems [3] for illegal intrusion, abnormal behavior and risk targets, and has become the core technology of intelligent security [4-6]. However, the current mainstream YOLO improved models generally have problems such as high computational overhead and high model parameters, which make it difficult to deploy efficiently on edge devices with limited resources, and still rely heavily on centralized cloud processing. This architecture leads to full upload of monitoring data, which not only causes bandwidth resource tension and redundancy of alarm information, but also seriously restricts the rapid identification and response timeliness of security operation and maintenance personnel to risks [7].

To effectively address these challenges, cloud edge collaboration architecture has become an important development direction in the field of intelligent security. This architecture significantly reduces data transmission pressure

and improves system real-time performance by pre-processing and preliminary risk screening of front-end image data on edge computing devices and uploading only high-value alarm information to the cloud. However, edge devices have strict limitations in terms of computing power, power consumption and deployment space, which puts forward extremely high requirements on the number of parameters and calculations of the model. Therefore, it is urgent to develop a highly lightweight YOLO model for edge end characteristics.

In the field of security monitoring, relying on video networking cameras to identify fire smoke and personnel violations without wearing helmets have become the two most common types of applications. However, different scenarios have different equipment arrangements, insufficient light, occlusion and environmental interference seriously affect the accuracy of identification, the flow of smoke spread, the frequency of personnel access, etc., which increases the difficulty of identifying such small targets again [8-10].

At the technical research level, researchers have made many improvement attempts based on YOLO in the field of fire smoke detection and helmet recognition: for example, FG-YOLO studied by JialeYao[8] reduces feature loss through FSCNet backbone network and GScELAN module, and improves the accuracy of flame smoke recognition; Penglei Wang et al.[9] introduced coordinate attention mechanism (CA) to enhance feature extraction ability, B.Lin[10] introduced CA attention mechanism and Mosaic data enhancement based on YOLOv8n improved model, which improved the recognition ability of helmet on construction site, Z.Xu et al.[11] studied helmet detection method based on semantic guidance and feature selection fusion based on YOLOv8m, but the above improved multi-focus single task [12, 13], and are mostly designed for general scenarios, failing to adapt to special environments with dense equipment and complex backgrounds; in addition, the model parameters are still high, making it difficult to deploy to edge devices. In addition, there is a lack of lightweight fusion research on fire and helmet multitasking. If multiple single-task models are deployed separately, the resource consumption of edge devices will be further aggravated. Single-task models will aggravate the resource consumption of edge devices.

Combined with the existing research status and engineering background, it is found that there are three shortcomings in the existing target detection technology: one is that the model is not lightweight enough to adapt to the computational requirements of edge devices; the other is that the multi-task fusion is poor, and the deployment of multiple models alone will increase the system complexity; and the third is that the scene adaptability is weak, and the equipment layout and environmental characteristics of complex scenes are not optimized.

Therefore, based on the cloud edge collaborative architecture, the improvement and application research of the fire smoke and helmet wearing multi-task lightweight YOLO model are carried out. Through the optimization of the model structure and the improvement of the multi-task loss function, it can realize efficient reasoning on the edge equipment. Meanwhile, the fire smoke and helmet multi-task dedicated data set is made according to the characteristics of complex layout of production equipment and large mobility of personnel, providing technical support and engineering examples for industrial scenario applications.

2. YOLOV8 Model Improvement Technology

YOLO series belongs to single stage target detection algorithm, the fire source carried by personnel and initial fire characteristics are not obvious, smoke diffusion is fast, high sensitivity detection is needed, personnel helmet belongs to small target, and there are many interference factors such as occlusion and long distance. The standard YOLO v 8n model has insufficient detection accuracy for small target and weak characteristic target in complex industrial environment. Therefore, this paper adds attention mechanism in feature extraction layer to The loss function is changed according to the detection target characteristics and the auxiliary detection training head (p2) is added to enhance the recognition of small targets.

2.1 SEAttention

SE (short for "Squeeze-and-Excitation") is an attention mechanism used to enhance Convolutional Neural Networks (CNNs). The SE network structure was proposed by Jie Hu et al. [14] in 2018, and its model structure is shown in Fig. 1. Its core idea is to introduce a global attention mechanism into CNNs to adaptively learn the importance of each channel. The

SE network implements the attention mechanism through two steps: squeeze and excitation.

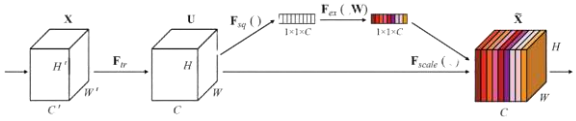


Fig. 1 SEAttention Attention Mechanism Model

In the squeeze step, the SE network performs global pooling on the feature map of each channel, compressing it into a scalar. In the excitation step, the SE network converts the compressed feature vector into a weight vector via a fully connected layer; this weight vector is then used to weight the feature map of each channel. Finally, the learned channel weights are applied to the original feature maps to enhance useful features and suppress irrelevant ones.

A large number of YOLO models integrated with the SEAttention mechanism have been developed, and they have achieved promising improvement effects [15, 16].

2.2 **P2 Layer 160×160 Small Object Detection Head**

For small target objects, there exist issues such as few features, unobvious semantic features, and susceptibility to feature submersion due to continuous convolution operations. Additionally, small targets usually occupy a very small pixel area in images and are prone to being submerged by large-scale features in high-level feature maps.

To address the problems in small target detection, by adding an additional 160×160 detection head to the P2 shallow feature map, more local information can be utilized (as shown in Fig. 2), thereby improving the detection accuracy of small targets. The 160×160 detection head is constructed based on the P2-layer feature map, which undergoes only one downsampling operation without excessive convolution or downsampling operations.

This detection head adopts a lightweight structure of "3×3 convolution + 1×1 convolution":

- The 3×3 convolutional layer is used for feature enhancement (with 256 output channels) to retain detailed information such as edges and textures;
- The 1×1 convolutional layer is responsible for the output of the classification and regression branches. Specifically, the classification branch outputs the object confidence score, while the regression branch outputs the object's coordinate offsets and aspect ratio.

The specific structure can be expressed as Formula (1):

$$F_{detect(x)} = Conv_{1\times1}(Conv_{3\times3}(F_{P2}(x))) \quad (1)$$

where $F_{P2}(x)$ denotes the input feature of the P2 layer, $Conv_{k\times k}$ represents the k×k convolution operation, and $F_{detect(x)}$ stands for the output feature of the detection head.

Although P3 has medium resolution, stride convolution can still cause critical local details to be lost. In contrast, P2 retains finer spatial granularity, captures low-level visual features, and is better suited for small target recognition, especially in complex backgrounds and occlusion scenes.

In order to effectively utilize these high-resolution features, this paper first upsamples the selected feature maps in the backbone network and stitches them with shallow outputs to form a rich feature set that fuses spatial details and context depth. The composite features are further refined by C2f module-C2f as an enhanced residual block, which can preserve key edge and shape information while achieving efficient feature fusion [17].

Subsequently, the processed P2 feature map is passed into the scale sequence (Scale Sequence, ScalSeq) module, which enhances small target feature representation by aggregating multi-scale information. Finally, the enhanced P2 output is integrated with existing P3-P5 detector heads to construct a multi-resolution detector head set. This design ensures comprehensive coverage of target scales, especially improves the detection sensitivity and accuracy of the model for small targets and densely distributed targets. In summary, the introduction of the P2 detection layer endows the model with early spatial cues and enhances its robustness in small target detection tasks typical of UAV vision applications [18].

2.3WIOUV3 Loss Function

Object detection is a core problem of computer vision, and its detection performance depends on the design of loss function. The default loss function in YOLOV8 uses the complete cross-merge ratio CIoU[19] to reflect the geometric relationship between bounding boxes. The aspect ratio of CIoU loss function describes the relative value, which may affect the performance of model detection when the aspect ratio of real frame is very close to that of predicted frame.

$$LossCIoU = 1 - IoU + \frac{d^2}{e^2} + \alpha v \quad (2)$$

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

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

In the formula: $LossCIoU$ is the calculation function of *CIoU*, and IoU is the Intersection over Union (IoU) between the prediction box and the ground truth box; *d* denotes the Euclidean distance between the centers of the prediction box and the ground truth box; *e* represents the diagonal length of the minimum enclosing box that surrounds the two boxes; α and υ are adjustment parameters used to balance the aspect ratio; *w* and *h* are the width and height of the prediction box, respectively; w^g and h^g are the width and height of the ground truth box, respectively.

Bounding box loss function is very important to target detection model performance. Existing studies have multiple default training samples with high quality, focusing on strengthening the loss fitting ability, but there are a large number of low quality samples in the actual data, blind regression will damage the performance, which is the starting point of this study.Focal-EIoUv1 [20] is designed to mitigate the interference of low-quality samples by focusing mechanism [21], but the effect is limited due to static focusing, which fails to give full play to non-monotonic focusing efficiency.

ZanjiaTong[22] et al. proposed a dynamic nonmonotonic focusing mechanism and designed Wise-IoU(WIoU). The dynamic nonmonotonic focusing mechanism uses outlier instead of IoU to evaluate the quality of anchor boxes and provides a wise gradient gain allocation strategy. This strategy reduces the competition of high quality anchor boxes while reducing the harmful gradient generated by low quality examples, which allows WIoU to focus on ordinary quality anchor boxes and improve the overall performance of the detector.

$$L_{WIoUv1} = R_{WIoU}L_{IoU} \tag{5}$$

$$RWIoU = \exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{((w^g)^2+(h^g)^2)^*}\right) \tag{6}$$

In the formula: (x, y) corresponds to the position of the ground truth frame (x_{gt}, y_{gt}); R_{WIoU} is the loss of the high-quality anchor frame; L_{IoU} denotes the base IoU loss; and * represents the separation operation.

WIoUv1, however, does not introduce aspect ratios and does not eliminate factors that hinder convergence, so this study will use the WIoUv3 loss function to attach a focusing mechanism by constructing gradient gains, which can avoid large harmful gradients for low-quality samples [23,24]. WIoUv3 is constructed from β and WIoUv1 as follows

$$L_{WIoUv3} = rL_{WIoUv1} \tag{7}$$

$$r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{8}$$

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \tag{9}$$

In the formula: *r* is the non-monotonic focusing coefficient; β is the outlier; and δ is the gradient gain.

1.4 Improved model architecture

The improved YOLOv8n model structure is still divided into input and backbone networks as shown in Fig. 2. The four core modules, Backbone, Neck and Head, improve the detection performance through multi-dimensional structure optimization on the basis of retaining the lightweight advantage of the original model, as follows: the input end of the model adaptively scales the original image to the default input size of 640×640, maintains the aspect ratio of the image during the scaling process, and fills gray pixels at the edges to avoid feature offset; The backbone network continues the C2f module as the core feature extraction unit, strengthens the feature fusion effect by means of multi-branch residual connection, and integrates SEAttention channel attention mechanism after the fast pyramid pooling (SPPF) module. This mechanism adaptively calibrates the feature channel weights through Squeeze (global average pooling) and Excitation (two-level full connection layer) operations, enhances the effective feature response, suppresses redundant information, and further improves the representation ability of high-level features. The neck network (PAN-FPN) adds an independent detection head at the shallow high-resolution feature map (160×160) to supplement the original multi-scale detection system's ability to capture small targets. At the same time, the CIoU loss function of the original model is replaced by the WIoUV3 loss function, which optimizes the gradient distribution in the process of bounding box regression through dynamic non-monotonic mechanism and reasonable gradient gain allocation strategy, effectively improves the difference between positive and negative

samples, and finally improves the accuracy and stability of the overall target detection.

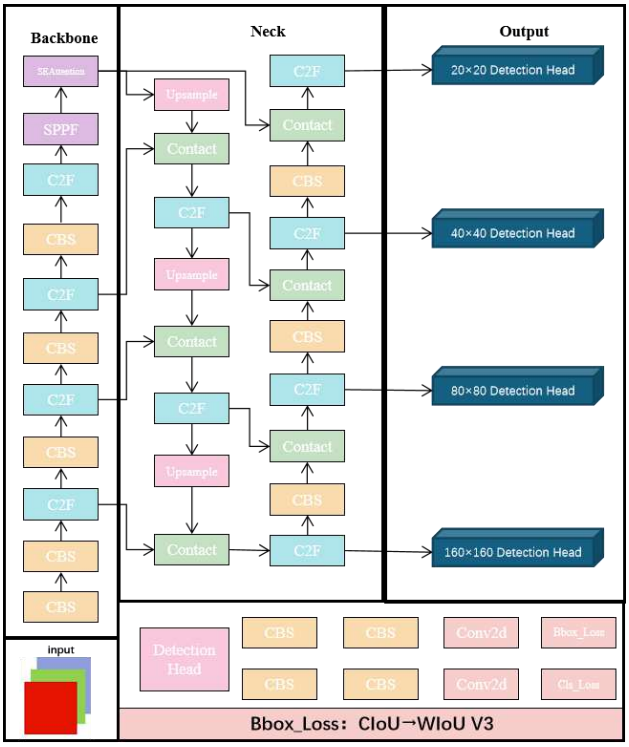


Fig 2 Overall Model Structure Diagra

2 Construction and preprocessing of datasets

2.1 model test environment

The experiment is based on Windows 10 operating system, graphics processor GeForce RTX 5090 TI (32 G). Deep learning framework PyTorch 2.7.0, Python 3.9.19, computing unified device architecture CUDA12.8, training process set batch-size 64, round 200.

2.2 Source of data set

For fire smoke recognition and personnel helmet recognition, the original data are obtained by means of network open data set integration, network crawling and monitoring video frame extraction.

The fire smoke data set comes from the multi-scene smoke flame target detection data set released by the Yangtze River Delta Science Data Center. This data set meets the security monitoring requirements and integrates public data sets and self-made video frames, including 10,641 high-resolution images in total.(Resolution ⩾640×480 pixels), covering a variety of indoor and outdoor scenes. The dataset focuses on smoke and flame core targets, while covering detection requirements under complex conditions such as low light at night, strong backlight, dynamic occlusion, and smoke with different colors, improving the adaptability of the model in actual monitoring scenes. In terms of labeling, high accuracy and consistency are guaranteed. The intersection ratio (IoU) between the bounding box and the target area is ⩾0.85, and the category label error rate is controlled within 1%. In addition, the dataset included 2,667 background images (including natural landscapes and urban street scenes) to reduce the risk of misjudgment of similar textures by the algorithm [25].

The helmet dataset is composed of helmet related images and personnel related images on the network and scene video collection. The acquisition process focuses on the activity scenes of personnel in the station, and captures the wearing status of helmets under specific working conditions such as equipment inspection and operation, so as to ensure that the data set is highly consistent with the actual complex scene of the heat exchange station, and can meet the requirements of subsequent detection tasks for complex environment.

2.3 multi-task data set construction

First, the unified annotation standard for the dual tasks was clarified: the LabelImg tool was used to redefine the annotation categories into 4 classes ("smoke": smoke, "fire": flame, "hat": safety helmet, "person": person), ensuring consistent annotation semantics between the two datasets.

Currently, there is an overlap in data types between the two datasets: the original fire and smoke dataset contains "person" and "hat" categories, while the original safety helmet dataset partially includes "fire" and "smoke" categories. Both cases are reasonable scenarios of multi-object coexistence in real-world settings; therefore, during dataset merging, partial images in the original datasets need to be re-annotated.

After re-annotating the above two datasets, quality verification was conducted on the merged dataset via code. It was found that some samples in the dataset had low clarity (including blurriness, strong noise, and resolution failing to meet the 640×480 standard) and an obvious class imbalance. Specifically, some samples were "pure person" or "pure safety helmet" samples, and a single sample contained too many categories, leading to uneven class distribution.

To address the above issues, an image quality assessment script was developed after importing OpenCV. This script

automatically filters out blurry samples with substandard clarity or insufficient resolution by calculating the image clarity metric. Additionally, a sample category judgment script was created to traverse all annotation files. It defines the "pure person/pure safety helmet scenario" as "annotated boxes containing only the 'person' or 'hat' category, with no 'smoke' or 'fire' annotations", and automatically counts the number of such redundant samples. Based on the random sampling principle, some redundant samples were deleted while a portion was retained to ensure scenario diversity, avoiding reduced generalization ability caused by over-cleaning.

To simulate "background interference" (i.e., images without targets) in real-world scenarios, all samples with empty annotations in the dataset were retained. This allows the model to learn to "distinguish between images with targets and those without" during training, thereby improving its robustness to complex scenarios. Meanwhile, to avoid inefficient training caused by excessive data volume, code was used to perform category-balanced reduction on samples containing targets, ensuring consistent learning weights for each category.

Finally, manual review and cleaning were conducted on the screening and deletion results. The final cleaned dataset retained a total of 13,862 samples, including 7,635 samples in the training set, 2,212 in the validation set, and 1,180 in the test set. The classes are evenly distributed, and the dataset covers complex scenarios such as indoor/outdoor environments, low-light nighttime conditions, and strong backlight conditions. Both the annotation quality and data diversity meet the training requirements of the dual-task model.

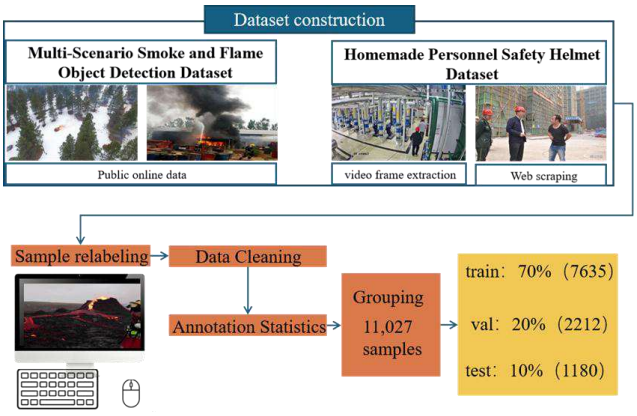


Fig 3 Dataset Creation Process

2.4Data augmentation strategy

To enhance the generalization and robustness of the model, two types of data augmentation were implemented in this study. First, for on-site inspection data, the augmentation focused on screening scenes where both safety helmets and personnel coexist—this enables the model to better learn the differences between the two categories. Specifically, data augmentation techniques such as brightness enhancement, Gaussian noise addition, random cropping, and brightness reduction were applied to images extracted from on-site inspection videos.

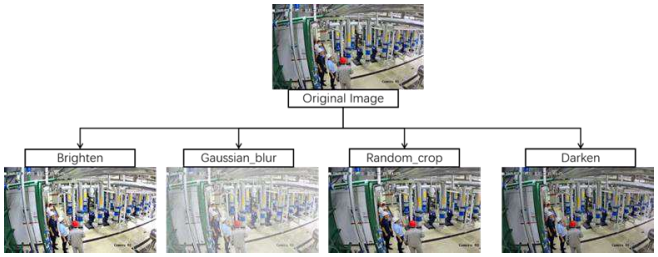


Fig 4 Surveillance Video Data Augmentation Strategies

To further improve the model's robustness to complex scenarios, a multi-dimensional data augmentation strategy was adopted for the training set, covering both general augmentation and category-specific augmentation. This significantly increased the diversity of samples.

(1)General Augmentation: The overall diversity of samples was improved by means of HSV adjustment, rotation, horizontal flipping, Mosaic tiling, and Mixup blending.

(2)Category-Specific Augmentation: Targeted optimizations were made for individual categories: the adjustment range of hue, saturation, and brightness was expanded for the "fire" category; brightness adaptability was optimized for the "smoke" category; shear transformation was added for the "hat" category; and translation transformation was incorporated for the "person" category. These measures achieved precise optimization for single categories.

This data augmentation was only applied to the training set, while the validation set and test set retained their original data. This ensures the objectivity of model performance evaluation.

2.5 Model evaluation metrics

To clarify the matching relationship between different models and improvement modules, and to facilitate linking to their improvement effects, Table 1 (Model Improvement Strategy) summarizes the adoption of key improvement modules—SEAttention, small object detection head, and Wolu V3—by

each model. This table lets you easily view the improvement module configuration of each model, supporting the subsequent analysis of effect differences across different improvement combinations.

Table 1Model Improvement Strategy

Improve module model name	SEAttention	**Small object detection head**	WIoU V3
Model1	×	×	×
Model2	√	×	×
Model3	×	√	×
Model4	×	×	√
Model5	√	√	×
Model6	√	×	√
Model7	×	√	√
Model8	√	√	√

Note: ×: Improved model not selected; √: Improved model selected. Model 1 refers to the original YOLOv8n model, while Models 2–8 correspond to new models developed based on each improved strategy.

To accurately evaluate the accuracy and real-time performance of the detection algorithm, the following model detection performance and evaluation metrics were selected in this study: mean Average Precision (mAP), Precision (P), Recall (R), model computational complexity (GFLOPS), and detection speed (average detection time per frame). Among them, the calculations of Precision, Recall, Average Precision, and mean Average Precision are shown in the formulas:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$P_{AP} = \int_0^1 P(R)dR \tag{12}$$

$$P_{Map} = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{13}$$

Where: True Positive (TP): The number of positive classes correctly predicted as positive; False Negative (FN): The number of positive classes incorrectly predicted as negative; False Positive (FP): The number of negative classes incorrectly predicted as positive; True Negative (TN): The number of negative classes correctly predicted as negative; N: Number of detection classes;P: Precision, %;R: Recall, %;PAP: Average Precision, %;PmAP: mean Average Precision, %;APi: Average Precision of the i-th class.

3 Analysis of Experimental Results

3.1 Analysis of Ablation Experiment Results

Ablation experiments were conducted based on the model improvement strategies in Table 1. The evaluation metrics included Precision (P), Recall (R), mean Average Precision at IoU=0.5 (mAP50), mean Average Precision at IoUs from 0.5 to 0.95 (mAP50-95), computational complexity, parameter count, and detection speed. The experimental results are presented in Table 2.

Table 2 Ablation Experiment Results

Improve module model name	P/%	R/%	mAP50/%	mAP50-95/%	Computational workload/GFLOPS	Number of parameters × 10^5/M	Detection speed（ms）
Model1	84.2	72.8	80.4	52.7	8.2	30.12	3.13
Model2	84.7	73.8	80.5	52.7	8.2	30.20	3.20
Model3	85.6	77.2	83.6	54.7	12.4	29.27	3.84
Model4	85.4	74.8	80.9	52.8	8.2	30.12	3.13
Model5	86.3	76.8	83.8	54.7	12.4	29.35	3.95
Model6	85.4	74.3	81.1	53.1	8.2	30.20	3.20
Model7	85.3	78.5	84.1	54.9	12.4	29.27	3.91
Model8	86.5	77.5	84.3	55.1	12.4	29.35	3.96

As can be seen from Table 2:

(1) Model 1 is the original YOLOv8n model, whose recognition accuracy for the two tasks was 84.2%, with mAP50 and mAP50-95 reaching 80.4% and 52.7%, respectively.

(2) Models 2, 3, and 4 were modified by adding the SEAttention attention mechanism, integrating a small-object detection head, and adopting the WIoU V3 loss function (each based on the original model). Compared with the original model, all three models showed improvements in various metrics to different degrees. Among them, Model 2 showed the smallest improvement, while Model 3 (with the added small-object detection head) achieved the most significant improvement.

(3) Model 5 was developed by adding the SEAttention attention mechanism to Model 3. Compared with Model 3, Model 5 exhibited higher P and mAP values. Models 6 and 7 were modified by adopting the WIoU V3 loss function based on Models 2 and 3, respectively; it can be observed that both models showed enhanced P and mAP values compared with their respective base models (Models 2 and 3). Model 8 is the improved YOLOv8 model, which integrated all three modifications and achieved the highest P and mAP values.

(4) The ablation experimental results indicate that Model 8 had the optimal comprehensive performance: its P reached 86.5%, R reached 77.5%, mAP50 reached 84.3%, and mAP50-95 reached 55.1%. These values represent increases of 2.3%, 4.7%, 3.9%, and 2.4%, respectively, compared with the original model. Since the selected baseline model YOLOv8n is a small-sized model, the increased computational complexity and the corresponding change in detection speed after modification do not affect the requirements of practical security monitoring.

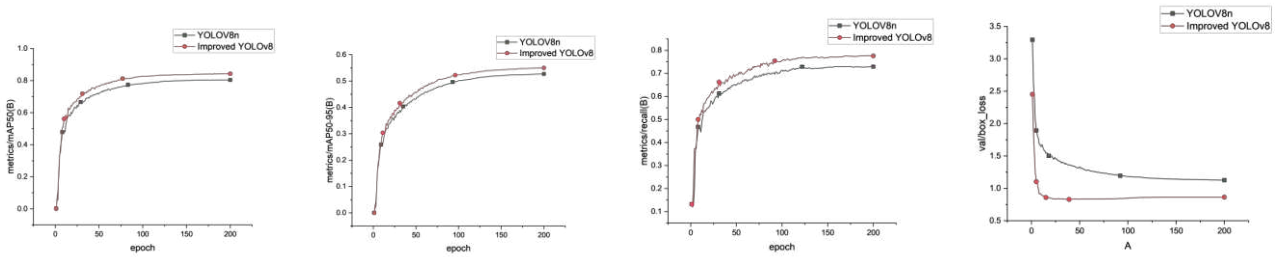


Fig 5 Comparison of Training Results Between Improved YOLOv8n and Original YOLOv8n

A further comparative analysis of mAP50, mAP50-90, R, and boundary loss during 200 training epochs shows the training results of the improved model and the original model as shown in Fig 5. It can be observed that: the improved model achieved significant enhancements in localization accuracy, detection precision, and comprehensiveness; it also accelerated the training process without oversampling issues, and significantly improved the practical effectiveness of security monitoring.

3.2 Comparative Analysis Table of AP Values for Each Category

Since security monitoring belongs to a multi-object detection task, it is necessary to evaluate the performance of each category after relevant improvements. By analyzing the AP value (i.e., the area under the Precision-Recall (P-R) curve) of each category, the improvement in detection performance for each category can be reflected

Table 3Improvement Effects of Each Model Type

Model Type	AP(smoke)/%	AP(fire) /%	AP(hat) /%	AP(person) /%	mAP50/%
Model1	86.4	69.2	88.8	77.1	80.4
Model2	85.7	69.2	88.9	78.4	80.5
Model3	86.1	74.1	90.2	84.3	83.7
Model4	86.2	69.4	89.3	78.6	80.9
Model5	86.5	74.1	90.4	84.3	83.8
Model6	86.2	70.5	89.4	78.4	81.1
Model7	85.9	75.3	90.8	84.6	84.1
Model8	86.7	75.5	89.9	84.9	84.3

Analysis of Table 3 reveals that Model 8 achieved the highest values among all improved models in terms of AP (smoke), AP (fire), AP (person), and mAP50. It was only slightly lower than that of Models 3, 5, and 7 in terms of AP (person), but this metric still remained at a high level, fully meeting the practical application requirements of security monitoring. This further verifies that Model 8 is the optimal model.

Compared with the original YOLOv8n model, the improved YOLOv8n model showed increases in AP (smoke), AP (fire), AP (hat), and AP (person) by 0.3%, 6.3%, 1.1%, and 7.8%,

respectively. Considering that flame and person targets are mostly classified as small-scale targets in the self-developed dataset, the above metric changes fully indicate that the proposed improvement strategy can significantly enhance small-object detection performance. Although multi-task and multi-category detection tasks themselves tend to cause a certain degree of detection accuracy degradation, the improvement scheme proposed in this study provides new insights and practical references for the lightweight implementation of multi-category object detection in industrial scenarios. Combined with the current evolutionary trend of security monitoring towards a cloud-edge-end collaboration architecture, the development of lightweight detection models with higher generalization ability for edge computing devices will become an important direction for future research in the field of industrial security.

3.3 Comparative Experiments with Other Models

To verify the detection performance of the improved model in security monitoring, comparative experiments were conducted in this study by selecting common lightweight detection models: YOLOv5n, YOLOv7-tiny, and YOLOv8n. The results are presented in Table 4.

Model Type	mAP50	mAP50-95	P	R	GFLOPs（@640×640）
YOLOV5n	74.9	46.2	78.4	68.3	4.5
YOLOV7-tiny	81.5	53.0	84.9	73.1	13.2
YOLOV8n	80.4	52.7	84.2	72.8	8.2
Improved YOLOV8n	84.3	55.1	86.5	77.5	12.4

(1) The YOLOv5n model showed relatively low performance across all metrics, especially in terms of mAP50 and mAP50-95 (reaching 74.9% and 46.2%, respectively). This indicates that with its low computational complexity, YOLOv5n cannot meet the requirements of multi-task detection, and its detection capability has certain limitations.

(2) By contrast, YOLOv7-tiny had the highest computational complexity, with mAP50 and mAP50-95 reaching 81.5% and 53.0%, respectively—showing strong detection capability under different IoU thresholds. Its Precision and Recall were 84.9% and 73.1%, respectively, demonstrating excellent overall performance. With a significantly reduced computational complexity, YOLOv8n showed slightly lower performance across all metrics than YOLOv7-tiny.

(3) The improved YOLOv8n model outperformed all other models across all metrics: its mAP50 and mAP50-95 reached 84.3% and 55.1%, respectively. This means the model can maintain high precision under different IoU thresholds and accurately detect all types of targets. Meanwhile, its Precision and Recall reached 86.5% and 77.5%, respectively—indicating that it can identify more true targets during multi-category detection, reduce the possibility of missing detections, and ensure that all types of targets are detected in a timely manner. Although its computational complexity is higher than that of the baseline YOLOv8n model, the overall computational complexity remains low, which meets the requirements for deployment on edge computing devices. Furthermore, its overall computational complexity is also lower than that of YOLOv7-tiny.

4 Conclusions

(1) Considering the transformation trend of security monitoring from centralized cloud processing to localized deployment of edge computing, this study aimed to address the issues of low detection accuracy and poor practicality caused by insufficient lightweight design and weak scenario adaptability of existing mainstream detection models. Focusing on two core security tasks—personnel safety helmet wearing detection and fire/smoke recognition—a scenario-specific detection dataset was constructed. This dataset integrated public multi-scenario fire/smoke datasets and on-site collected personnel safety helmet samples. Through unified annotation, image quality verification, and class-balanced cleaning, general augmentation and category-specific augmentation strategies were adopted to improve sample diversity. This not only meets the data adaptation requirements of edge computing devices for lightweight model training but also solves the problems of insufficient scenario complexity in general datasets and difficulty in deploying original models (with large parameter counts) on edge devices.

(2) The SEAttention attention mechanism and a 160×160 small-object detection head were integrated into the

YOLOv8n model, and the WIoU V3 loss function was adopted. The improved YOLOv8n model achieved a Precision (P) of 86.5%, a Recall (R) of 77.5%, and a mean Average Precision at IoU=0.5 (mAP50) of 84.3%. Compared with the original model, P, R, and mAP50 increased by 2.3%, 4.7%, and 3.9%, respectively. In the multi-category object detection task, the improved model showed increases in the Average Precision (AP) of each category—AP (smoke), AP (fire), AP (hat), and AP (person)—by 0.3%, 6.3%, 1.1%, and 7.8%, respectively, compared with the original model. Notably, the AP values of small targets (e.g., flames and personnel) showed particularly significant improvements. These results fully verify that the improvement strategy proposed in this study can effectively enhance the overall performance of object detection and achieve a significant optimization effect on small-object detection, providing strong support for the reliable application of the model in complex scenarios.

(3) The improved model was optimized based on the small-sized YOLOv8n architecture; the introduction of the attention mechanism enhanced its ability to focus on features of important information. In terms of performance metrics, the improved model has a parameter count of 29.35×10^5, which is a 2.6% decrease compared with the original YOLOv8n. Its computational complexity is 12.4 GFLOPS, and its detection speed is 3.96 ms. While achieving lightweight compression of the parameter count, both computational efficiency and detection response speed remain at a high level. Owing to the lightweight design of the small-sized model, combined with the feature enhancement effect of the attention mechanism and the balanced performance of metrics, the model can still fully meet the core requirements of security monitoring scenarios for real-time performance and target detection accuracy.

References

[1] BOUROU S, MANIATIS A, KONTOPOULOS D, et al. Smart Detection System of Safety Hazards in Industry 5.0 [J]. Telecom, 2023.

[2] SHARMA R. Enhancing Industrial Automation and Safety Through Real-Time Monitoring and Control Systems [J]. International Journal on Smart & Sustainable Intelligent Computing, 2024, 1(2): 1-20.

[3] BAJPAI M. YOLO Models for Security and Surveillance Applications [J]. International Journal for Research in Applied Science and Engineering Technology, 2024, 12(6): 2513-8.

[4] KANG S, HU Z, LIU L, et al. Object Detection YOLO Algorithms and Their Industrial Applications: Overview and Comparative Analysis [J]. Electronics, 2025, 14(6): 1104.

[5] BHATTACHARJEE S. A Literature-based Performance Assessment of the YOLO (You Only Look Once) CNN Approach for Real-time Object Detection [J]. Computology: Journal of Applied Computer Science and Intelligent Technologies, 2025, 4(2): 18-40.

[6] KUKRETI S, PRAVEEN R V S, BANSAL S, et al. Object Detection in Real-Time Surveillance Using Deep Learning-Based YOLO Framework [J]. 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), 2025: 601-6.

[7] YOU L, XIAO S, PENG Q, et al. ST-Seq2Seq: A Spatio-Temporal Feature-Optimized Seq2Seq Model for Short-Term Vessel Trajectory Prediction [J]. IEEE Access, 2020, 8: 218565-74.

[8] YAO J, LEI J, ZHOU J, et al. FG-YOLO: an improved YOLOv8 algorithm for real-time fire and smoke detection [J]. Signal, Image and Video Processing, 2025, 19(5).

[9] WANG P, FAN X, ZHU Z, et al. A Smoke and Fire Detection Model Based on Improved YOLO11 [J]. 2024 4th International Conference on Communication Technology and Information Technology (ICCTIT), 2024: 521-5.

[10] LIN B. Safety Helmet Detection Based on Improved YOLOv8 [J]. IEEE Access, 2024, 12: 28260-72.

[11] XU Z, LI Y, ZHU H. Safety helmet detection method based on semantic guidance and feature selection fusion [J]. Signal, Image and Video Processing, 2023, 17(7): 3683-91.

[12] SUN L, WANG L. An improved YOLO V5-based algorithm of safety helmet wearing detection [J]. 2022 34th Chinese Control and Decision Conference (CCDC), 2022: 2030-5.

[13] WANG L, ZHANG X, YANG H. Safety Helmet Wearing Detection Model Based on Improved YOLO-M [J]. IEEE Access, 2023, 11: 26247-57.

[14] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks [Z]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 7132-41.10.1109/cvpr.2018.00745

[15] DONG W, NIE S, WANG Y. Remote Sensing Object Detection Based on Improved YOLOv3 [J]. 2022 IEEE 4th International Conference on Civil Aviation Safety and …, 2022, 2022.

[16] WANG Y, GU D, ZHU Y, et al. Reflective clothing detection based on YOLOv5s fused with SE attention mechanism [J]. International Journal of Reasoning-based Intelligent Systems, 2025.

[17] WU Y-H, CHEN P-Y, HSIEH J-W, et al. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN [J]. Conference on Computer Vision and Pattern Recognition Workshops IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[18] WANG P, ZHAO J. SOD-YOLO: Enhancing YOLO-Based Detection of Small Objects in UAV Imagery [J]. Computing Research Repository, 2025, abs/2507.12727.

[19] HUANG P, TIAN S, SU Y, et al. IA-CIOU: An Improved IOU Bounding Box Loss Function for SAR Ship Target Detection Methods [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 10569-82.

[20] ZHANG Y-F, REN W, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression [J]. Neurocomputing, 2022, 506: 146-57.
[21] QIN G, PAN R, DENG Y, et al. Road pedestrian detection and tracking algorithm based on improved YOLOv5s and DeepSORT [J]. PLoS ONE, 2025.
[22] TONG Z, CHEN Y, XU Z, et al. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism [J]. Computing Research Repository, 2023, abs/2301.10051.
[23] YANG X, LIU C, HAN J. Reparameterized underwater object detection network improved by cone-rod cell module and WIOU loss [J]. Complex & Intelligent Systems, 2024, 10(5): 7183-98.
[24] LI S, CAO X, ZHOU Z. Research on inshore ship detection under nighttime low-visibility environment for maritime surveillance [J]. Computers and Electrical Engineering, 2024, 118: 109310.
[25] ZHAO C, ZHAO L, ZHANG K, et al. Smoke and Fire-You Only Look Once: A Lightweight Deep Learning Model for Video Smoke and Flame Detection in Natural Scenes [J]. Fire, 2025, 8(3).