Journal of Industrial Intelligence
https://www.acadlore.com/journals/JII

YOLOv8n-AM: Enhanced Real-Time Smoke Detection via Attention-Based Feature Interaction and Multi-Scale Downsampling

Zijun Yao, Lin Zhang, Ashim Khadka*

[1] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, 223003 Huaian, China
[2] Faculty of Management Engineering, Huaiyin Institute of Technology, 223003/Huaian, China
[3] Nepal College of Information Technology, Pokhara University, 44700 Lalitpur, Nepal

* Correspondence: Lin Zhang (zlmjl@hyit.edu.cn)

Received: 10-15-2024 **Revised:** 12-15-2024 **Accepted:** 12-31-2024

Citation: Z.J. Yao, L. Zhang, A. Khadka, "YOLOv8n-AM: Enhanced real-time smoke detection via attention-based feature interaction and multi-scale downsampling," *J. Ind Intell.*, vol. 2, no. 5, pp. 240–250, 2024. https://doi.org/10.56578/jii020405.

Abstract: Accurate smoke detection in complex industrial environments, such as chemical plants, remains a significant challenge due to the inherently low contrast, transparency, and weak texture features of smoke, which often exhibits blurred boundaries and diverse spatial scales. To address these limitations, YOLOv8n-AM, an enhanced lightweight detection framework belonging to the YOLO (You Only Look Once) series, was developed by integrating advanced architectural components into the baseline YOLOv8n model. Specifically, the conventional Spatial Pyramid Pooling-Fast (SPPF) module was replaced with an Attention-based Intra-scale Feature Interaction (AIFI) Convolution Synergistic Feature Processing Module (SFPM), i.e., AIFC-SFPM, enabling more effective semantic feature representation and an improvement in detection accuracy. In parallel, the original convolutional module was optimized using a Multi-Scale Downsampling (MSDown) module, which reduces model redundancy and computational overhead, increasing the detection speed. Experimental evaluations demonstrate that the YOLOv8n-AM model achieves a 1.7% improvement in mean Average Precision (mAP), accompanied by a 9.1% reduction in Giga Floating-point Operations Per Second (GFLOPs) and a 15.4% decrease in parameter count when compared to the original YOLOv8n framework. These improvements collectively underscore the model's suitability for real-time deployment in resource-constrained industrial settings where rapid and reliable smoke detection is critical. The proposed architecture thus provides a computationally efficient and high-precision solution for safety-critical visual monitoring applications.

Keywords: Smoke detection model; AIFC-SFPM; MSDown module; Real-time detection; Industrial safety

1 Introduction

In the modern industrial field, especially in large chemical plant areas, rapid detection and monitoring of smoke are crucial for ensuring safe production and preventing environmental pollution [1]. Chemical plants often involve the production, transportation, and storage of hazardous substances. Once a leak or accident occurs, the spread of smoke can pose a great threat to on-site workers and the surrounding environment. Therefore, developing efficient and accurate smoke detection methods has become a research focus. However, due to the complexity of the environment in the chemical plant area, the diversity of smoke forms, and the variability of meteorological conditions, traditional smoke detection technology [2] faces problems such as insufficient perception, a high false detection rate, and a slow response speed. With the continuous development of industrial automation and intelligence, smoke detection technology, as a key means of environmental monitoring and safety protection, has gradually received widespread attention. At present, research mainly focuses on two directions: traditional image processing-based detection methods and intelligent detection methods based on deep learning.

In terms of traditional video smoke detection technology, Pundir et al. [3] studied methods for extracting smoke features in different color spaces. In the Red, Green, and Blue (RGB) color space, the color features of smoke were extracted and converted into the YCbCr color space to calculate brightness values (Y) and chromaticity values (Cb and Cr). These features provide an effective basis for further smoke detection, especially under complex backgrounds and variable lighting conditions, which can significantly improve detection accuracy. However, it lacks sensitivity to low contrast or transparent smoke and is susceptible to interference from complex backgrounds such as cloud layers and vapors. Lin et al. [4] extracted suspicious areas in the foreground using Gaussian mixture models and background subtraction, and established smoke color models in the RGB and Hue, Saturation, and Intensity (HSI) color spaces. By combining contour features, a feature vector was formed and classified using Support Vector Machine (SVM) [5] to accurately identify smoke areas. Although it can extract contour features, its robustness to dynamic lighting changes is poor. Zhao et al. [6] classified smoke using the Cost-Sensitive Adaptive Boosting (CS-Adaboost) algorithm by extracting its movement, heat, and color features. However, it is difficult to detect static or slowly spreading smoke, and the computational complexity is high. Yuan et al. [7] proposed a smoke detection method based on fuzzy logic, which reduces negative effects such as sky clouds and lighting changes by taking the differences between RGB and HSI models as inputs. The use of the Extended Kalman filtering (EKF) to reshape the input-output of fuzzy rules improves the flexibility of the detection method. But the anti-interference ability is weak. Jia et al. [8] proposed a smoke pixel classification detection method based on saliency detection, which enhances smoke color nonlinearly, measures saliency by combining enhancement maps and motion maps, and uses motion energy and saliency maps for smoke detection. But its generalization ability is weak, making it difficult to adapt to the changing smoke patterns and environmental disturbances in chemical plants.

In summary, in recent years, traditional smoke detection research based on feature extraction has been continuously deepening and has achieved certain results. However, it can be observed that traditional methods have limitations. On the one hand, due to the single feature extraction of smoke recognition classification and the fact that manually set features such as color and texture cannot fully represent smoke, the false detection rate of smoke detection is relatively high. On the other hand, it performs poorly in dynamic lighting, noise interference, or diverse data.

With the rise of deep learning technology, especially the application of the convolutional neural network (CNN), smoke detection technology has been significantly improved. Deep learning can automatically learn and extract deep-level features of images, reducing reliance on manually designed features and significantly improving the model's generalization ability and detection accuracy. In this type of research, the most representative ones are object detection networks, such as YOLO series [9], Faster Region-based Convolutional Neural Network (Faster R-CNN) [10], Single Shot MultiBox Detector (SSD) [11], etc., which have excellent real-time performance and detection accuracy.

In terms of improving the application of the attention mechanism, Li et al. [12] proposed a Local Binary Pattern Silhouette Coefficient Variable (LBPSCV) for industrial smoke segmentation, which uses the variation of the silhouette coefficient as the weight to calculate the local binary pattern (LBP). The extracted texture features make it easier to detect smoke. But the segmentation effect is not good for low contrast and transparent smoke. He et al. [13] constructed a smoke dataset containing multiple positive and negative samples by combining online collection with offline shooting. In terms of model design, a module was embedded that combines the spatial attention mechanism and the channel attention mechanism into the second convolutional layer of Visual Geometry Group 16 (VGG16) to enhance the model's ability to detect smoke features and further improve the overall performance of the algorithm. But the contribution of high- and low-layer features has not been distinguished, resulting in background noise interference. Li et al. [14] introduced a lightweight framework for smoke video detection, which utilizes BFBlock to enhance feature extraction and considers the influence of weather factors to improve the robustness and accuracy of detection. Although lightweight design helps to improve detection speed, it sacrifices some detection accuracy while pursuing speed.

In terms of module optimization and lightweight design, Khan et al. [15] proposed a CNN-based smoke detection and segmentation framework aimed at effectively handling outdoor clear and hazy scenes. In this framework, EfficientNet was used for smoke detection, ensuring efficient feature extraction and accuracy. Meanwhile, DeepLabv3+ was applied to segment smoke areas to achieve optimal localization results. The proposal of this method provides a new solution for smoke detection. Although high-precision segmentation was achieved, the multi-scale feature fusion was not optimized, resulting in a high missed detection rate for small targets. Masoom et al. [16] improved the algorithm's ability to locate smoke by using smoke principal component analysis (PCA) as a preprocessing module to remove redundant features. Although PCA helps reduce computational complexity, the additional detection scale added may increase the computational cost of the improved network. Huo et al. [17] proposed a single-scene smoke detection algorithm based on the depthwise separable convolution, which effectively reduces model complexity. However, the depthwise separable convolution was performed in a two-dimensional plane, making it difficult to fully utilize the feature information at the same spatial position between channels.

In the improved method based on object detection networks, Sun et al. [18] proposed an improved CNN that achieves automatic feature extraction through optimization strategies in multi-convolutional kernels and batch normalization processes without manual intervention. This improvement significantly optimizes the loss function, thereby improving the smoke recognition rate and providing new ideas and methods for the development of smoke detection technology. But the feature fusion strategy is single. Wang et al. [19] used YOLOv5m [20] as the base model and improved the model's ability to extract smoke features by adding a spatial channel attention mechanism to the module composed of the convolutional layer, the Batch Normalization (BN) layer, and the LeakyReLU activation function in its backbone network. Although the feature extraction capability was improved, redundant calculations were not reduced. Shao et al. [21] introduced a new method for small target fire and smoke detection called YOLOv7scb, which enhances the model's ability to effectively extract features from small targets. However, the module is not lightweight, and the entire framework still requires a large amount of computing resources during runtime. Chen et al. [22] proposed a lightweight forest fire detection model for unmanned aerial vehicle applications based on the YOLOv7 model. The model adopts the GSConv convolution design and constructs the GSELAN and GSSPPFCSPC modules, which significantly reduce the number of parameters in the model and improve computational efficiency. In addition, the model introduces a coordinate attention mechanism, which further enhances the ability to extract smoke features and effectively improves the detection performance of the model in complex natural environments. Although the model is lightweight, high-level semantic representation has not been optimized for smoke transparency. Deng et al. [23] proposed an integrated attention mechanism for aircraft hangar fire detection based on the YOLOv8n framework. However, the module structure that is not lightweight is difficult to deploy widely.

In recent years, deep learning technology has significantly promoted the development of smoke detection, with its core advantage being the end-to-end feature learning achieved through CNNs, effectively breaking through the limitations of traditional manual features. Despite these achievements, several persistent challenges remain unresolved. On the one hand, it is difficult for existing methods to accurately detect small, localized smoke to large, diffuse smoke. On the other hand, the complexity of network models can waste computational resources for simple smoke recognition requirements.

This study proposes a new smoke detection model, YOLOv8n-AM, by combining AIFC-SFPM and the MSDown module based on the above-mentioned previous studies.

2 YOLOv8n-AM

The development of the YOLOv8n-AM model was guided by the need to address two critical challenges in smoke detection:

(a) Smoke usually manifests as low-contrast, transparent or semi-transparent areas with weak texture and edge information, and it may appear at different scales, ranging from small local smoke to large diffuse smoke, making it difficult to be accurately detected.

(b) The high computational complexity and resource consumption of detection models make it difficult to deploy on low-power mobile terminals, which limits the widespread application of smoke detection technology. Therefore, there is a higher demand for lightweight and real-time detection models.

In response to the above issues, this study mainly improves the YOLOv8n-based model and proposes a smoke detection model, YOLOv8n-AM. Replacing the SPPF module of YOLOv8n with AIFC-SFPM significantly increases the accuracy of model detection. By using the MSDown module to improve the Conv module, it greatly reduces redundancy, lowers model complexity, and improves detection speed. The YOLOv8n-AM network architecture is shown in Figure 1.

The working principle of the technique is shown in Figure 1.

2.1 YOLOv8 Object Detection Neural Network

The main structure of YOLOv8 [24] includes a backbone network, a neck network, and a head network. The backbone network is responsible for extracting meaningful features from input images, capturing simple patterns such as edges and textures in the image, and providing multi-scale feature representations at different levels, thereby providing rich information for subsequent object detection. As a bridge between the backbone network and the head network, the neck network mainly performs feature fusion operations, integrates different levels of feature information, constructs feature pyramids to ensure that the network can detect objects of different sizes, and improves detection accuracy by considering a wider range of scene context information, while reducing spatial resolution and resource dimensions to accelerate computation speed. As the final part, the head network is responsible for generating the final output of object detection, such as generating bounding boxes associated with objects that may exist in the image and assigning a reliability score to each bounding box to represent the likelihood of object existence. At the same time, objects in the bounding boxes are classified and sorted according to their categories.

YOLOv8 includes multiple versions, namely YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. The smallest version is YOLOv8n. Under the same dataset, the YOLOv8n model has the shortest training time, smallest size, and real-time efficiency. YOLOv8n was improved in this study by expanding the dataset through data augmentation during the input stage.

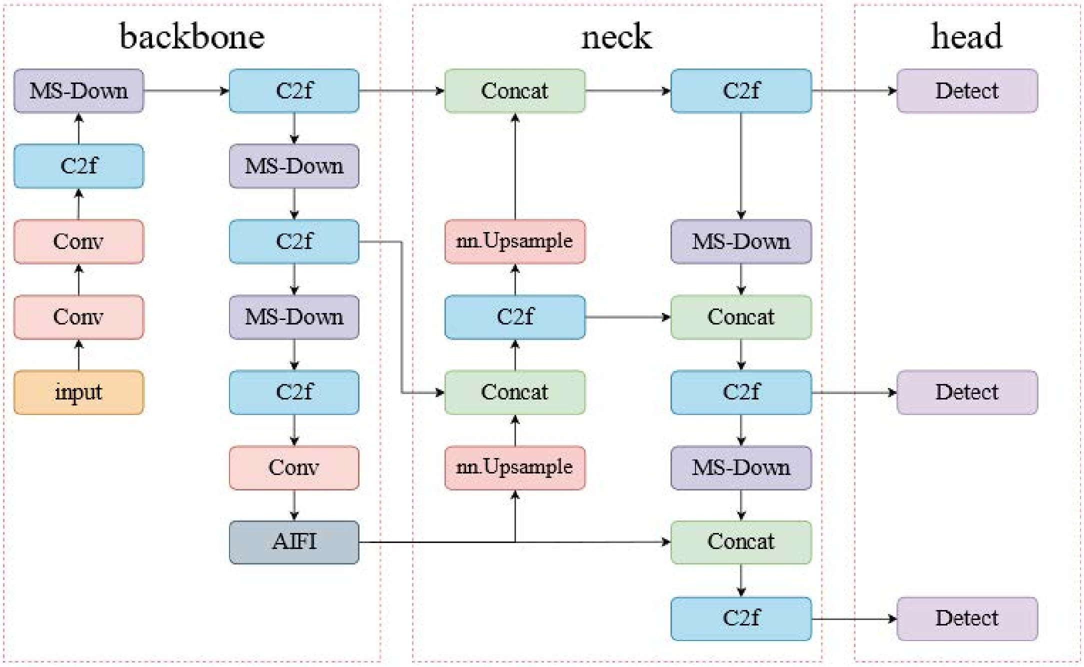


Figure 1. YOLOV8n-AM network framework

2.2 AIFC-SFPM

AIFC-SFPM combines the single-scale Transformer (AIFI) and convolution operation (Conv) to achieve collaborative processing of high-level semantic features and local detail features.

Due to the fact that smoke typically appears as low-contrast, transparent or semi-transparent areas, its texture and edge information are weak and susceptible to background noise interference. However, high-level features contain richer semantic information, which can more effectively represent the overall shape and semantic features of smoke. The AIFI module is a key technology aimed at optimizing the semantic richness and computational efficiency of high-level feature representations. AIFI focuses on the internal interaction of high-level feature maps, effectively balancing the detection performance and inference speed of the model by eliminating redundant calculations of multi-scale features. AIFI only processes high-level features at specific scales (such as S5), reducing unnecessary computational complexity. The AIFI module enhances its modeling ability for global semantics through single-scale interaction with high-level feature S5, making it suitable for capturing features of diffuse smoke. In addition, in smoke detection, low-level features such as edges and textures may be affected by background noise such as ground textures or tree contours.

2.2.1 Single-scale interaction strategy

The main interaction strategy of AIFI is to focus on the highest-level feature S_s and apply the Transformer [25] self-attention mechanism to process it. In multi-scale features, low-level features such as S_3 and S_4 often contain more detailed information such as edges and textures and are easily affected by background noise interference. The high-level feature S_s has stronger semantic expression ability, and performing self-attention calculation on it can avoid redundant calculation on low-level features and reduce the interference of irrelevant information.

2.2.2 Lightweight design

AIFI only uses one Transformer layer, which ensures effective information exchange for high-level features without introducing too many parameters and computational complexity. Before performing the self-attention computation, the features of input S_5 were flattened and converted into a format suitable for Transformer Block processing to obtain Q, K, and V.

$$Q = K = V = \text{Flatten}\left(S_5\right) \tag{1}$$

The result F_5 calculated was restored based on the Transformer self-attention mechanism to the same spatial shape as S_5 for subsequent module processing.

$$F_5 = \text{Reshape}(\text{AIFI}(Q, K, V)) \tag{2}$$

2.2.3 Two-dimensional sine-cosine positional embedding

In order to enable the model to perceive the position information of each element in the feature map, a two-dimensional sine-cosine positional embedding was constructed using AIFI. Firstly, a function was used to generate grid coordinates grid-w and grid-h in the width and height directions, respectively, and generate a two-dimensional grid. Then the embedding dimension was calculated for each direction to generate the frequency parameter omega. After flattening the grid coordinates, matrix multiplication was performed with frequency parameters to obtain out w and out i. Finally, the sine and cosine values were concatenated to form a two-dimensional positional embedding.

Conv has high efficiency in extracting local features and can quickly capture information such as edges and textures in images. AIFI, combined with Conv's local perception capability, can flexibly process features at different levels. Compared with SPPF, this combination may have higher computational efficiency in processing multi-scale features, especially in smoke detection, which can provide improved accuracy. In addition, when processing high-resolution images, it can avoid the problem of excessive computational resource consumption caused by SPPF's global pooling operation. The structure of AIFC-SFPM is shown in Figure 2.

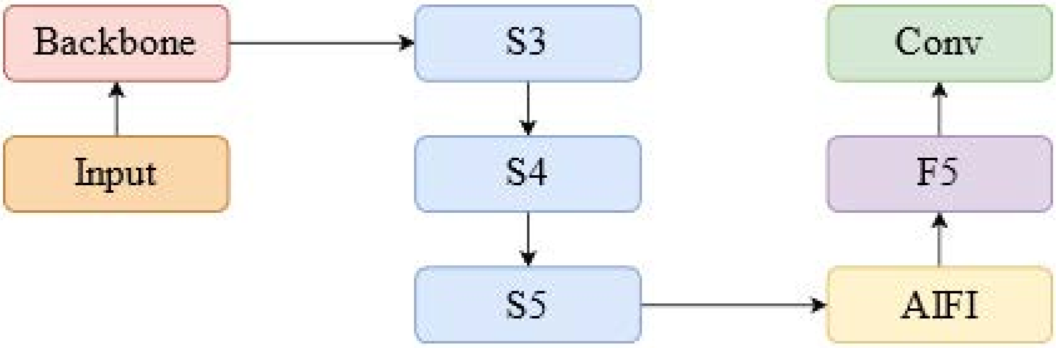


Figure 2. AIFC-SFPM structure diagram

2.3 MS-Down Module

The MS-Down module is a downsampling module used in neural networks. This module is designed to achieve multi-scale feature extraction and fusion while maintaining computational efficiency through reasonable selection of the convolution kernel size, stride, and pooling operations. Compared to some complex downsampling methods, it avoids excessive consumption of computational resources. In practical applications, especially for smoke detection tasks that require real-time processing, this computational efficiency advantage can ensure that the model processes a large number of video frames in a limited time, detects smoke conditions in a timely manner, and does not impose an excessive burden on system resources. The main structure of the MS-Down module is shown in Figure 3.

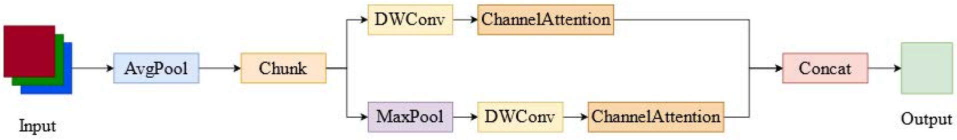


Figure 3. Diagram of the MS-Down module structure

As shown in the above figure, the input tensor enters the average pooling layer $AvgPool$ for an average pooling operation, with a pooling kernel size of 2, stride of 1, and padding of 0. This operation mainly reduces the size of the input feature map. Average pooling suppresses noise while maintaining spatial resolution, making it suitable for preliminary feature extraction of low-contrast objects such as smoke. After average pooling, the pooled feature map is split into two parts, $x1$ and $x2$, in the channel dimension. x 1 is input to the first depthwise separable convolution module $cv1$ for processing, and the depthwise separable convolution can significantly reduce computational complexity.

The other branch first performs a max pooling ($MaxPool$) operation, with a pooling kernel size of 3, a stride of 2, and padding of 1. Max pooling preserves edge information through downsampling, balancing feature compression and detail preservation. Then, it is processed through the convolutional layer of the second depthwise separable convolution module $cv2$ to reduce computational complexity. Then x 1 and x 2 are input separately into the channel attention module for processing. Finally, after concatenating along the channel dimension, the concatenated result is returned. This concatenation operation helps to integrate feature information processed by different branches, enabling the network to comprehensively consider multiple feature representations, thus more comprehensively describing the input data and improving the accuracy of object detection. Channel attention enhances the complementarity of features between the two branches, significantly improving the multi-scale feature fusion effect.

3 Experimental Design and Result Analysis

3.1 Experimental Environment

The experiment was conducted on a Windows 10 operating system equipped with an Intel(R) Core(TM) i9-14900K processor and an NVIDIA GeForce RTX 4090 GPU, supported by 16 GB of system memory. The model training was implemented using Python 3.9, with GPU acceleration enabled via CUDA 11.8, and the deep learning tasks were performed on the PyTorch 2.3.1 framework. The training parameter settings are as follows: the input image size is 640×640, the training period is 200, the initial learning rate is 0.01, and the batch processing volume is 16. The experimental environment of the model in this study is shown in Table 1.

Table 1. Experimental environment configuration

Parameter	Configuration
GPU	NVIDIA GeForce RTX 4090
System environment	Windows 10
CUDA version	CUDA 11.8
Programming language version	Python 3.9
Deep learning framework	PyTorch 2.3.1

3.2 Dataset

A self-made smoke dataset, which contains smoke images and videos from different scenes and weather conditions, was used as the experimental dataset. A total of 2,127 images and 135 videos were converted into 3,256 images through frame extraction. After data cleaning, 4,127 images were finally obtained, of which 3,700 images were used as the training set and 427 images as the testing set.

The smoke dataset contains 12 factory areas, including typical scenarios such as tank areas and reaction equipment areas, eight forest areas covering different vegetation types such as coniferous forests and broad-leaved forests, and five types of industrial plants. The types of smoke were mainly divided into two types: industrial smoke and fire smoke, with most smoke images having a small amount of interference. The partial images in the dataset are shown in Figure 4.

3.3 Model Evaluation Indicators

In order to comprehensively evaluate the performance of the improved model and effectively compare it with existing detection methods, this study adopts multiple common evaluation metrics, covering multiple dimensions such as accuracy, recall, detection speed, and computational complexity of single- and multi-class detection tasks. Specifically, single-class precision (AP) and multi-class precision (MAP) were used to quantify the detection accuracy of the model in different categories. In addition, the precision (P) and recall (R) of the model are important indicators for evaluating its predictive ability, which were used to measure the proportions of True Positive (TP) predictions in all predicted and actual positive samples, respectively, aiming to comprehensively evaluate the accuracy and completeness of the model. TP refers to the cases where both ground truth and prediction are positive; False Positive (FP) refers to the cases where the prediction is positive but ground truth is negative; and False Negative (FN) refers to the cases where the prediction is negative but ground truth is positive. The application of these evaluation indicators helps to reveal the comprehensive performance of the improved model in terms of accuracy, speed, and computational resources, thereby providing a theoretical basis for its optimization and practical application.

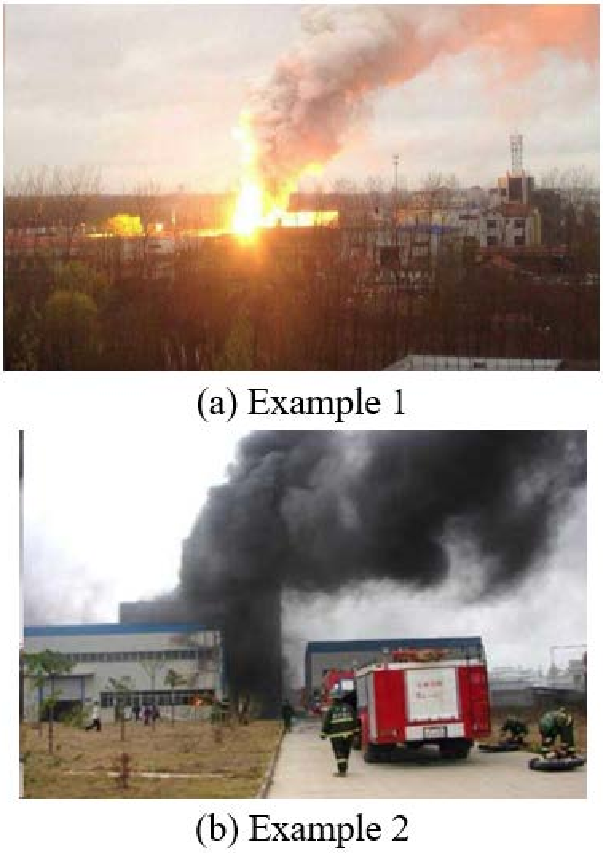
P reflects the probability of the model correctly detecting the target:

$$P = \frac{T_P}{T_P + F_P} \tag{3}$$

R reflects the probability of the model detecting all targets:

$$R = \frac{T_P}{T_P + F_N} \tag{4}$$

A_p reflects the accuracy of the model on a single detection object:



(a) Example 1

(b) Example 2

Figure 4. Samples of the dataset images

$$A_P = \int_0^1 P(R)dR \tag{5}$$

By adding up the values of all categories and taking their average, an overall measure was provided to evaluate the detection performance of the model across all categories. A higher M_{AP} value usually means that the model has strong detection ability in various categories and can generalize well. The M_{AP} formula is as follows:

$$M_{AP} = \frac{\sum_{n=1}^{N} A_{PN}}{N} \tag{6}$$

where, N represents the target detection category, and A_{PN} represents the A_P value of category N.

3.4 Ablation Experiment

In order to comprehensively evaluate the effectiveness and reliability of the improvement plan on the performance of the YOLOv8 model, this study adopts a strategy of selectively eliminating the improvement module to explore the specific roles played by each improvement in the optimization process. By systematically removing certain improvements and comparing the performance differences of the model before and after the removal, the contribution of each improvement to the overall detection performance can be quantified. This method not only helps to verify the actual effectiveness of improvement schemes in enhancing model performance but also enables an in-depth analysis of the specific roles of various optimization measures in enhancing model capabilities. Through this evaluation process, the reliability of the improvement plan can be effectively judged, ensuring that the proposed optimization measures have scientific and practical value and providing a theoretical basis for further improvement in the future. The equipment parameters used in the experiment are completely consistent, and the experimental results are shown in Table 2.

The study verified the contribution mechanism of each improved module to the model performance through modular ablation experiments. The experimental results in Table 2 show that AIFC-SFPM has built global semantic modeling capability on high-level features S_5 by introducing the Transformer selfattention mechanism. This module enables $\mathrm{mAP@0.5}$ to have an increase of 1.0%, which is mainly attributed to the effective capture of smoke

Table 2. Comparison of the ablation experimental performance

Models	**AIFC-SFPM**	**MSDown**	**P/%**	**R/%**	**mAP/%**	**Parameters**	**GFLOPs**
Yolov8n	×	×	74.7	61.3	72.5	3151888	8.7
Yolov8-A	√	×	75.2	65.6	**73.5**	3079456	8.6
Yolov8n-M	×	√	76.8	66.3	73.9	**2739232**	8.1
Yolov8n-AM	√	√	80.5	62.9	**74.2**	**2666784**	**7.9**

diffusion characteristics. In addition, the single-scale interaction strategy of the AIFI module reduces computational redundancy by 2.3%, verifying its balanced design between semantic enhancement and computational efficiency.

The MSDown module achieves lightweight optimization while maintaining feature diversity through a multi-branch downsampling architecture. This module uses the depthwise separable convolution to reduce GFLOPs by 6.8%, and its multi-scale feature fusion strategy effectively preserves the edge details of low-contrast smoke. The feature map processed by MSDown shows a 14.7% improvement in the gradient amplitude index, confirming its enhancement effect on weak texture features.

Yolov8n-AM achieves a synergistic optimization of semantic enhancement and computational efficiency through the cascade design of AIFC-SFPM and MSDown. The combination of the two modules reduces the total parameter count by 15.4% and GFLOPs by 9.1%.

As shown in Figure 5, the precision, recall, and mAP curves of YOLOv8n-AM tend to stabilize after 100 epochs, with smaller fluctuations than YOLOv8n, YOLOv8-A, and other models, indicating better training stability and a more balanced parameter optimization process. The sharp fluctuations in the precision curve in the early stages reflect the model's insufficient ability to distinguish difficult samples such as occluded targets, which can lead to false or missed detections in the early-stage training. Overall, YOLOv8n-AM outperforms the comparison models in terms of reliability due to its more stable training curves, but its performance bottleneck is also evident, i.e., the detection ability for extreme scenarios such as highly overlapping targets needs to be further improved.

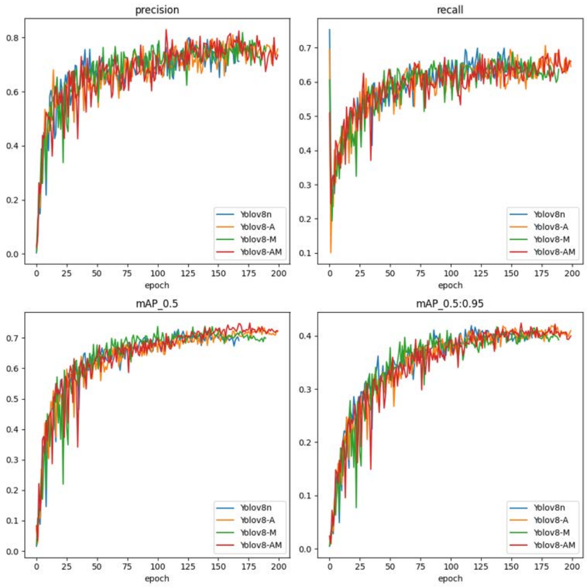


Figure 5. Comparison of different indicators

3.5 Comparative Experiment

To further validate the effectiveness and generalization ability of the Yolov8n-AM model, the dataset was kept unchanged and compared with multiple advanced object detection models. The results are shown in Table 3.

The YOLOv8n-AM model performs the best in smoke detection tasks, ranking first with an mAP of 74.2%, while maintaining minimal parameter and computational complexity, significantly outperforming traditional models such as YOLOv5n, YOLOv7, RT-DETR-r18 and surpassing YOLOv10n with similar parameter quantities. Compared to the pure Transformer architecture Swin Transformer-T and the lightweight model EfficientDet-D0, YOLOv8n-AM achieves breakthroughs in both accuracy and efficiency. Its advantages stem from the AIFC-SFPM's enhancement of high-level semantic features and the MSDown module's multi-scale lightweight design, making it suitable for

Table 3. Comparison of the ablation experimental performance

Models	Parameters	GFLOPs	P/%	R/%	mAP/%
RT-DETR-r18	19873044	56.9	78.8	63.1	71.5
Yolov5n	1765270	4.2	75.2	69.3	69
Yolov7	37196556	105.1	77.2	68.7	72.6
Yolov8n	3151888	8.7	76.8	92	72.5
Yolov10n	2694806	8.2	82	63.1	73.5
Swin Transformer-T	2850983	145	78.2	59.7	72.5
EfficientDet-D0	498359	12	77.6	63.8	71.2
Yolov8n-AM	2666784	7.9	80.5	62.9	74.2

industrial safety monitoring scenarios. However, further optimization of recall is needed to reduce missed detections.

3.6 Test Results

Figure 6 and Figure 7show the detection results of smoke in the factory area under different scenarios using the original YOLOv8n network and the improved YOLOv8n-AM network, respectively. These images clearly indicate that YOLOv8n-AM surpasses the original model in detection accuracy. Specifically, it can be seen from the first column of images that in different outdoor lighting environments, YOLOv8n-AM performs better in detection accuracy and has a better effect on actual smoke localization. In the second column of images, the improved model accurately identifies smoke areas in complex indoor environments, while the original model fails to accurately identify targets. In the third column of images, at the junction of smoke and clouds, the original model fails to accurately identify the target, while the YOLOv8n-AM model performs better. Overall, YOLOv8n-AM demonstrates excellent detection capabilities for smoke targets in complex environments, making it suitable for detection tasks in monitoring chemical scenes.

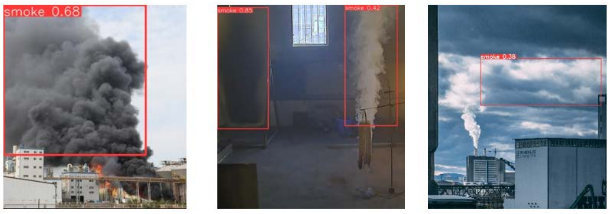


Figure 6. Detection results of the YOLOv8n model

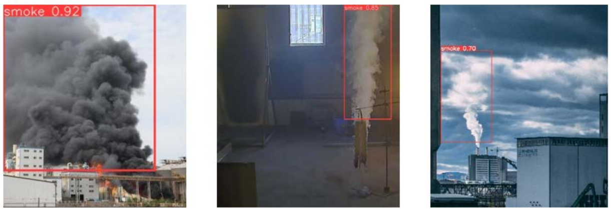


Figure 7. Detection results of the YOLOv8n-AM model

4 Conclusion

YOLOv8n-AM, an improved smoke detection model, was proposed in this study by integrating AIFC-SFPM and MSDown. The experimental results demonstrate that replacing the traditional SPPF module with AIFC-SFPM enhances the semantic representation capability and accuracy of smoke detection, while the MSDown module effectively reduces redundancy and computational complexity, significantly accelerating the detection speed. Specifically, compared to the original YOLOv8n model, YOLOv8n-AM achieves a 1.7% increase in accuracy, a 9.1% reduction in GFLOPs, and a 15.4% decrease in parameter count, demonstrating a clear balance between performance and model efficiency.

However, certain limitations remain in this study, which need to be addressed in future research. The detection performance of YOLOv8n-AM under extreme environmental conditions, such as severe illumination variations, heavy occlusions, or highly complex background scenarios, requires further investigation. Additionally, detecting smaller or partially occluded smoke targets still presents challenges. Future research should explore advanced background suppression techniques and integrate context-aware mechanisms to further enhance detection robustness under complex conditions. It is also essential to validate the model's effectiveness across a broader range of industrial environments and on low-power, edge-computing platforms.

In conclusion, the proposed YOLOv8n-AM model demonstrates significant potential for real-time smoke detection tasks, particularly in safety monitoring of chemical plants, contributing to enhanced industrial safety and environmental protection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. Lee, Y. S. Shim, Y. G. Song, S. D. Han, Y. S. Lee, and C. Y. Kang, "Highly sensitive sensors based on metal-oxiden anocolumns for fire detection," *Sensors*, vol. 17, no. 2, p. 303, 2017. https://doi.org/https://doi.org/10.3390/s17020303

[2] J. Fonollosa, A. Solórzano, and S. Marco, "Chemical sensor systems and associated algorithms for fire detection: A review," *Sensors*, vol. 18, no. 2, p. 553, 2018. https://doi.org/10.3390/s18020553

[3] A. S. Pundir and B. Raman, "Deep belief network for smoke detection," *Fire Technol.*, vol. 53, no. 6, pp. 1943–1960, 2017. https://doi.org/10.1007/s10694-017-0665-z

[4] L. Wang and A. Li, "Early fire recognition based on multi-feature fusion of video smoke," in *2017 36th Chinese Control Conference (CCC), Dalian, China*, 2017, pp. 1234–1243. https://doi.org/10.23919/ChiCC.2017.8028197

[5] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011. https://doi.org/10.1145/1961189.1961199

[6] Y. Zhao, Q. Li, and Z. Gu, "Early smoke detection of forest fire video using CS Adaboost algorithm," *Optik - Int. J. Light Ele. Optics*, vol. 126, no. 19, pp. 2121–2124, 2015. https://doi.org/10.1016/j.ijleo.2015.05.082

[7] C. Yuan, Z. Liu, and Y. Zhang, "Learning-based smoke detection for unmanned aerial vehicles applied to forest fire surveillance," *J. Intell. Robot Syst.*, vol. 93, no. 1-2, pp. 337–349, 2018. https://doi.org/10.1007/s10846-018-0803-y

[8] Y. Jia, J. Yuan, J. Wang, J. Fang, Q. Zhang, and Y. Zhang, "A Saliency-Based Method for Early Smoke Detection in Video Sequences," *Fire Technol.*, vol. 52, no. 5, pp. 1271–1292, 2015. https://doi.org/10.1007/s10694-014-0453-y

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, 2016, pp. 779–788. https://doi.org/10.1109/cvpr.2016.91

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Ana. Mach. Intelli.*, vol. 39, no. 6, pp. 1137–1149, 2017. https://doi.org/10.1109/TPAMI.2016.257703

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands*, 2016, pp. 11–14.

[12] Q. Li, H. Liu, J. Zhang, and P. Zeng, "Target segmentation of industrial smoke image based on LBP Silhouettes coefficient variant (LBPSCV) algorithm," *IET Image Proce.*, vol. 14, no. 4, pp. 2879–2889, 2020. https://doi.org/10.1049/iet-ipr.2019.1315

[13] L. He, X. Gong, S. Zhang, L. Wang, and F. Li, "Efficient attention based deep fusion CNN for smoke detection in fog environment," *Neurocomputing*, vol. 434, pp. 224–238, 2021. https://doi.org/10.1016/j.neucom.2021.01.024

[14] X. Li, P. Cheng, X. Liu, and Y. Huang, "Weather-informed lightweight framework for robust smoke video detection using BFBlock-enhanced feature extraction," *Signal, Image and Video Processing*, vol. 19, no. 5, p. 375, 2025. https://doi.org/10.1007/s11760-025-03910-5

[15] S. Khan, K. Muhammad, T. Hussain, J. Del Ser, F. Cuzzolin, S. Bhattacharyya, Z. Akhtar, and A. H. C. de Albuquerque, "Deepsmoke: Deep learning model for smoke detection and segmentation in outdoor environments," *Expert Sys. App.*, no. 182, p. 115125, 2021.

[16] M. S. M. Masoom S., Q. Zhang, P. Dai, Y. Jia, Y. Zhang, J. Zhu, and J. Wang, "Early smoke detection based on improved YOLO-PCA network," *Fire*, vol. 5, no. 2, p. 40, 2022. https://doi.org/10.3390/fire5020040

[17] Y. Huo, Q. Zhang, Y. Jia, D. Liu, J. Guan, G. Lin, and Y. Zhang, "A deep separable convolutional neural network for multiscale image-based smoke detection," *Fire Technol.*, vol. 58, no. 3, pp. 1445–1468, 2022. https://doi.org/10.1007/s10694-021-01199-7

[18] X. Sun, L. Sun, and Y. Huang, "Forest fire smoke recognition based on convolutional neural network," *J. For. Res.*, vol. 32, no. 5, pp. 1921–1927, 2020. https://doi.org/10.1007/s11676-020-01230-7

[19] Z. Wang, L. Wu, T. Li, and P. Shi, "A smoke detection model based on improved YOLOv5," *Mathematics*, vol. 10, no. 7, p. 1190, 2022. https://doi.org/10.3390/math10071190

[20] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, and et al., "ultralytics/yolov5: v5.0-YOLOv5-P6 1280 models, AWS, Supervise," *ly and YouTube Integra.*, 2021.

[21] D. Shao, Y. Liu, G. Liu, N. Wang, P. Chen, J. Yu, and G. Liang, "Yolov7scb: A small-target object detection method for fire smoke inspection," *Fire*, vol. 8, no. 2, p. 62, 2025. https://doi.org/10.3390/fire8020062

[22] G. Chen, R. Cheng, X. Lin, W. Jiao, D. Bai, and H. Lin, "Lmdfs: A lightweight model for detecting forest fire smoke in UAV images based on YOLOv7," *Remote Sens.*, vol. 15, no. 15, p. 3790, 2023. https://doi.org/10.3390/rs15153790

[23] L. Deng, S. Wu, J. Zhou, S. Zou, and Q. Liu, "Lska-YOLOv8n-WIoU: An enhanced YOLOv8n method for early fire detection in airplane Hangars," *Fire*, vol. 8, no. 2, p. 67, 2025. https://doi.org/10.3390/fire8020067

[24] M. Fatma Talaat and H. ZainEldin, "An improved fire detection approach based on YOLO-v8 for smart cities," *Neural Comput. Appli.*, vol. 35, no. 28, pp. 20 939–20 954, 2023. https://doi.org/10.1007/s00521-023-08809-1

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *Advan. Neural Infor. Proce. Sys.*, no. 33, pp. 1–21, 2020.