

Article

Early Flame and Smoke Detection in Valve Halls of Ultra-High-Voltage Converter Stations Using an Attention-Enhanced YOLOv5s Model

Rui Liu ¹, Jia Xie ², Hanbing Hao ², Taiyun Zhu ¹, Yi Guo ¹, Xiang Liu ¹, Yang He ¹, Jiaqing Zhang ^{1,*} and Tianchang Meng ³

¹ State Grid Anhui Electric Power Research Institute, Anhui Provincial Key Laboratory of New Type Power Systems Fire Safety and Emergency Technology (State Grid Laboratory of Fire Protection for Transmission and Distribution Facilities), Hefei 230601, China

² State Grid Anhui Electric Power Co., Ltd., Hefei 230022, China

³ Institute of Building Fire Research, China Academy of Building Research, Beijing 100013, China

* Correspondence: dkyzjq@163.com

Abstract

Early detection of incipient fire signs, such as dilute smoke emerging at valve hall penetration seals, remains a critical challenge for fire safety in ultra-high-voltage (UHV) converter stations. This study addresses the limitation through controlled valve hall sealing simulations that systematically reproduce the complete fire evolution process—from initial dilute smoke leakage, through dense smoke accumulation, to eventual flame overflow—thereby constructing a high-fidelity dataset tailored to complex converter station environments with low-contrast and varying illumination conditions. Building on this dataset, an attention-enhanced YOLOv5s model integrating multiple attention mechanisms (GAM, CBAM, CA, and ECA) is proposed as the core of an end-to-end visual detection framework. Rigorous experiments and ablation studies validate the framework's superiority: dilute smoke detection precision improves substantially from a baseline of 57% to 81%, with recall increasing from 61% to 84%; dense smoke and flame detection achieve accuracies of 92% and 98%, respectively. Compared with the original model, false and missed alarm rates are significantly reduced, demonstrating strong robustness against background interference and lighting variations. The proposed method enables reliable, high-precision monitoring across all fire stages, providing a novel technical pathway for early fire warning in UHV converter stations and offering extensibility to other large-scale industrial fire monitoring scenarios.

Keywords: YOLOv5; fire detection; smoke detection; converter station; valve hall; attention mechanism



Academic Editor: Tao Chen

Received: 25 April 2026

Revised: 18 June 2026

Accepted: 15 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Leveraging its advantageous natural conditions, China's northwestern region has developed the world's largest cluster of wind and photovoltaic power generation facilities. In contrast, electricity consumption demand remains heavily concentrated in the densely populated urban agglomerations along the eastern and southern coasts. This pronounced geographical mismatch between energy production and load centers constitutes a long-term structural challenge for the planning and secure operation of the national power system. To effectively address this issue, the Chinese government has promoted the construction of a world-leading ultra-high-voltage (UHV) transmission network, as illustrated in Figure 1.

Through the coordinated deployment of Ultra-High-Voltage Direct Current (UHVDC) and Ultra-High-Voltage Alternating Current (UHVAC) technologies, this network facilitates efficient, large-capacity, low-loss transmission of clean energy over thousands of kilometers, forming the physical backbone of the national “West-to-East Electricity Transmission” strategy [1]. As the core nodes of the HVDC transmission system, converter stations perform the critical function of alternating current (AC) and direct current (DC) power conversion, serving as a key link to realize long-distance, low-loss power transmission [2]. By means of rectifier and inverter devices, converter stations convert AC to DC to meet the requirements of long-distance transmission, and simultaneously invert the remotely transmitted DC back to AC for integration into the local power grid. This provides core technical support for ensuring the stable operation of the power system and the efficient grid connection of renewable energy. Nevertheless, the converter transformers in converter stations contain tens to hundreds of tons of insulating oil. In the event of a fire caused by equipment failure, the fire spreads rapidly and is accompanied by intense thermal radiation and high-temperature impacts, posing a severe threat to the fire protection system of the converter station [3,4].

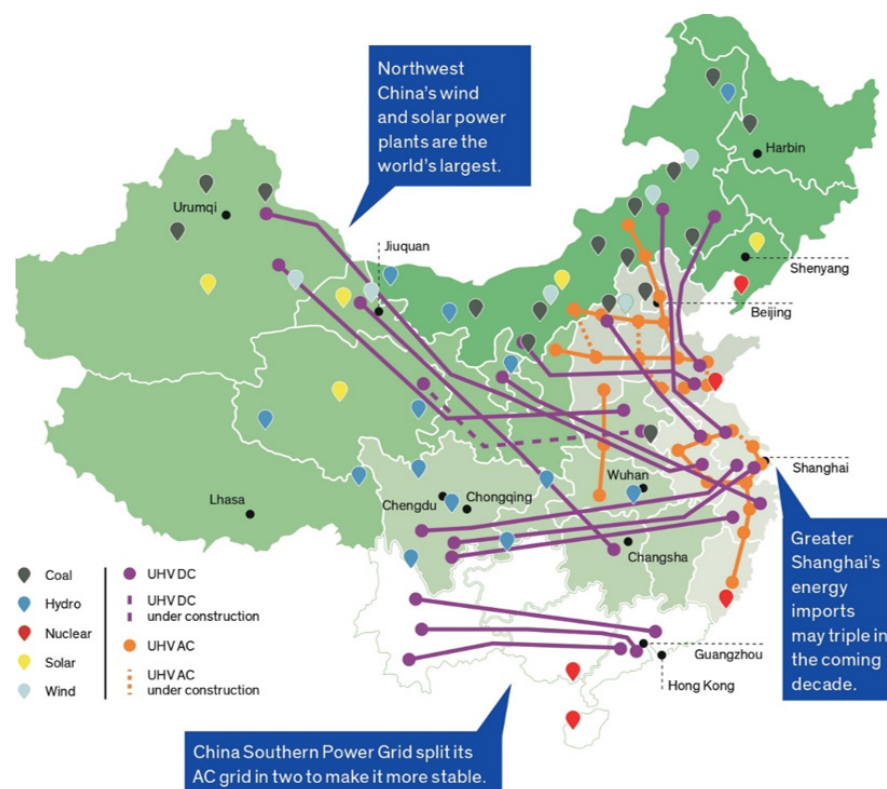


Figure 1. China's Hybrid AC-DC Grid [1].

The valve hall of a converter station is the core area for DC equipment and a key structure of the converter station. In the event of a fire in the valve hall, it will not only damage the core power equipment of the converter station, but also spread rapidly to other areas within the station, disrupting the normal power supply and causing huge economic losses simultaneously. In recent years, converter transformer fire accidents have occurred frequently at home and abroad: In 2018, an explosion and fire occurred at a ± 800 kV converter station in Xinjiang, resulting in the burnout of the ultra-high-end valve hall and multiple converter transformers. The fire was not extinguished until 60 h later, causing direct economic losses of hundreds of millions of yuan and a prolonged power grid blackout. In March 2025, a converter transformer fire broke out at a substation of an

airport in the UK, destroying two transformers, leading to cancelation of thousands of flights and a power grid outage, with economic losses exceeding 200 million US dollars. While converter transformer fires are difficult to entirely prevent, timely detection and response can significantly mitigate losses. High-sensitivity early fire detection technology is therefore of paramount importance. In particular, the reliable identification of initial fire characteristics—dilute smoke and weak flames leaking from valve hall penetration seals—remains a critical technical bottleneck for achieving prompt warnings in converter stations.

In recent years, with the advancement of artificial intelligence and computer vision technologies, machine learning-based image recognition methods have been gradually introduced into the field of fire detection [5]. By learning from massive image datasets, these methods can automatically extract multi-level and multi-scale image features, and maintain robust target recognition capabilities even under complex background conditions. Yin et al. [6] proposed a normalized convolutional neural network (CNN) for extracting smoke features to identify targets, and simultaneously introduced data augmentation techniques to address the imbalanced positive and negative samples in datasets. However, their dataset was relatively small and the application scenario was limited to a single case. Khan et al. [7] developed a multi-scale deep convolutional network for outdoor smoke detection and segmentation. The network employs parallel convolutional branches with different receptive fields for scale-invariant feature extraction, followed by batch normalization after feature fusion to mitigate misalignment. An additive scale fusion strategy is further adopted to improve efficiency while preserving smoke features. However, the method only supports image-level classification without instance-level bounding box annotations. Ghali and Akhloufi [8] developed a deep CNN model for wildland fire recognition using RGB and thermal aerial imagery, where transfer learning with pre-trained weights was introduced to optimize training convergence and reduce the demand for labeled samples. The model was validated on forest remote sensing datasets with complex background interference, but its application scenario was restricted to open forest fire monitoring only. Doddapaneni et al. [9] evaluated multiple convolutional neural network architectures including AlexNet for fire detection in color images, and verified that AlexNet achieved the optimal classification accuracy in both indoor and outdoor scenarios. However, their work focused on image-level binary classification and did not implement bounding-box based target localization.

With the development of real-time object detection frameworks, the YOLO series has become a dominant paradigm in fire detection research due to its efficiency and end-to-end detection capability. Numerous studies have focused on improving YOLO-based models to enhance detection accuracy and robustness. Mukhiddinov et al. [10] proposed an improved small-target smoke and flame detection algorithm based on optimized YOLOv5 for complex aerial surveillance scenarios. A lightweight attention mechanism was integrated into the Backbone network to enhance feature extraction capability without excessive computational overhead. In addition, the original three-scale detection was extended to four-scale detection, combined with a weighted bidirectional feature pyramid structure to optimize feature fusion for small targets. Although this method achieved high precision on its test set, it still suffered from a relatively high false alarm rate in low-contrast ground environments. Sozol et al. [11] proposed an optimized YOLOv5-based flame and smoke detection algorithm for indoor industrial scenarios. Experimental results showed that compared with the vanilla YOLOv5s baseline, the improved algorithm increased the mean average precision (mAP) by 2.4 percentage points, achieved a detection speed of 238 frames per second, and retained the lightweight advantage of the base model for edge deployment. Wang et al. [12] summarized the video-based flame and smoke detection algorithms for indoor and outdoor environments over the past two decades, and proposed

a lightweight detector named Light-YOLOv4, which achieved a favorable balance between performance and efficiency. Compared with YOLOv4, Light-YOLOv4 only retained 19.1% of the trainable parameters while maintaining almost the same accuracy.

More recently, research has shifted toward adopting advanced YOLO architectures and attention-enhanced frameworks. Chetoui and Akhloufi [13] explored fine-tuned YOLOv7 and YOLOv8 models for fire and smoke detection, demonstrating improved accuracy over earlier YOLO versions. Similarly, Gonçalves et al. [14] systematically evaluated YOLO-based models for wildfire detection across ground and aerial imagery, highlighting their adaptability to different observation perspectives. To further investigate the evolution of detection architectures, Alkhamash [15] conducted a comparative analysis of YOLOv9, YOLOv10, and YOLOv11, revealing continuous improvements in detection performance but also increased model complexity. Meanwhile, Lin et al. [16] proposed the Feature Pyramid Network (FPN), a foundational multi-scale feature fusion architecture for object detection. It integrates high-level semantic features from deep network layers with high-resolution spatial details from shallow layers through a top-down upsampling pathway paired with lateral connections, achieving robust feature representation for targets of all scales with only marginal extra computational overhead. This framework has laid the core theoretical foundation for the feature fusion modules in subsequent mainstream detection models including the YOLO series. Shah et al. [17] proposed a lightweight YOLOv8-based fire detection algorithm embedded with multi-scale attention modules, which effectively balanced detection precision and inference efficiency on resource-constrained edge surveillance devices. Zhang et al. [18] further optimized the YOLOv5 framework with feature fusion enhancement, significantly reducing false alarm rates in industrial smoke and flame detection tasks. Similarly, Gonçalves et al. [14] systematically evaluated YOLO-based models for wildfire detection across ground and aerial imagery, highlighting their adaptability to different observation perspectives. Aiming at the core challenge of early-stage weak fire feature extraction, Fang et al. [19] developed the CFS-YOLO model integrating coarse-to-fine granularity search and focus modulation, which substantially boosted recognition sensitivity to dilute smoke and micro flame targets; Zhang et al. [20] also constructed a dedicated improved YOLOv5 model for small-target smoke detection, achieving notable performance improvements in incipient smoke identification. Qian et al. [21] proposed a lightweight improved YOLOv10 flame detection method, which achieved a favorable balance between detection accuracy and edge deployment efficiency through dynamic convolution and low-rank attention optimization. Ali et al. [22] introduced attention-based 3D CNNs for indoor fire detection, emphasizing spatiotemporal feature modeling. Ejaz et al. [23] conducted a systematic study on channel attention mechanisms for fire and smoke detection, comprehensively analyzing the impacts of data augmentation strategies, color space selection, and adversarial attack perturbations on model detection performance and robustness. Their work not only validated that channel-wise attention can effectively enhance the model's feature focusing capability on flame and smoke targets and reduce false positives in complex backgrounds, but also revealed the reliability vulnerability of vision-based detection models under adversarial interference, providing a valuable reference for the robustness optimization of industrial fire detection frameworks.

Despite the considerable progress made by machine learning algorithms in the field of flame and smoke recognition from early CNN-based feature extraction to the optimization of YOLO series models, existing research still exhibits obvious shortcomings in meeting the demand for accurate detection of dilute smoke in the early stages of fire in complex scenarios such as converter station valve halls. Most existing studies focus on fire detection in single or high-concentration flame/smoke scenarios, lacking systematic research on the recognition performance of flue gas with different concentrations during fire evolution,

especially for dilute smoke which is crucial for early fire warning. Although newer YOLO variants and attention-enhanced models achieve higher accuracy, they often introduce increased computational complexity, limiting their applicability in real-time industrial environments. Some studies fail to achieve accurate detection of flame and smoke targets, while others improve detection precision at the cost of high false alarm rates, indicating that the reliability of detection needs further optimization. Meanwhile, existing datasets lack sufficient dilute smoke samples for model training and validation.

To bridge these gaps, particularly the challenge of detecting low-salience dilute smoke at valve hall seals, this study proposes an integrated research framework combining multi-variable physical simulations with an enhanced deep learning algorithm. The main contributions and novelties of this study are summarized as follows:

- (1) A high-fidelity converter station fire dataset is constructed through controlled physical simulation experiments, which systematically reproduces the complete fire evolution process from initial dilute smoke leakage to dense smoke accumulation and eventual flame overflow at valve hall sealing points. The dataset covers complex scenarios with varying smoke concentrations, background materials and lighting conditions, addressing the lack of dedicated early-stage fire samples for UHV converter stations.
- (2) An attention-enhanced YOLOv5s model is proposed, which integrates four complementary attention mechanisms (GAM, CBAM, CA and ECA) at different network layers. This multi-attention architecture significantly enhances the model's capability to extract weak features from low-contrast dilute smoke, while maintaining the lightweight advantage required for industrial edge deployment.
- (3) The proposed method achieves substantial performance improvements in early fire detection, with dilute smoke detection precision increasing from 57% (baseline YOLOv5s) to 81% and recall increasing from 61% to 84%. It also delivers high accuracy for dense smoke (92%) and flame (98%) detection, enabling reliable monitoring across all fire stages.
- (4) The end-to-end visual detection framework provides a novel technical pathway for early fire warning in UHV converter stations, and its modular design and lightweight characteristics make it readily extensible to other large-scale industrial fire monitoring scenarios.

2. Materials and Methods

This study focuses on vision-based flame and smoke detection tailored to UHV converter station valve halls. A specialized image dataset reflecting realistic converter station fire characteristics was constructed, and an improved YOLOv5s-based detection algorithm incorporating multiple attention mechanisms was developed. This section details the dataset construction process and the architectural enhancements to the baseline model.

2.1. Converter Station Flame and Smoke Image Data

A high-quality dataset constitutes a critical prerequisite for enhancing the robustness and generalization capability of detection models. Fire detection technologies for converter stations are highly susceptible to factors such as background colors and ambient lighting conditions. These background interferences are prone to causing false detections and missed detections of flames and smoke, thereby degrading the accuracy of flame and smoke detection in converter stations. Meanwhile, the detection of small-target flames and smoke faces great challenges and yields high miss rates due to their low image pixel resolution and the translucent nature of smoke. At present, public fire and smoke datasets are dominated by scenes of buildings and mountain forests, which differ substantially from the actual environments for converter station fire detection. Moreover, these datasets lack

small-target flame and smoke images. Therefore, training a model solely on public fire and smoke datasets may result in suboptimal performance in flame and smoke detection for converter stations.

To address the issues of mismatched scenarios between public fire datasets and converter station flame/smoke detection, lack of small-target flame and smoke images, and unbalanced smoke samples, a small-scale simulation test platform for converter stations was built to obtain high-quality image samples of dense smoke, dilute smoke, and overflowing flames. Combined with public fire and smoke datasets, a converter station flame and smoke image dataset was constructed. As shown in Table 1, two flammable liquids, n-heptane and engine oil, were selected as fuels in the experiment, representing high-volatility light oil and low-volatility heavy oil respectively, so as to better simulate the smoke characteristic concentrations at different development stages of converter station fires. To quantitatively verify the similarity between the simulated smoke and actual transformer insulating oil smoke, we conducted comparative measurements of their core physical and optical properties. First, in terms of particle size distribution, the smoke generated by the n-heptane–engine oil mixture (1:1 ratio) has a particle size range of 0.1–10 μm , with a median particle size (D50) of 0.82 μm , while the smoke from burning typical 25# transformer mineral insulating oil has a particle size range of 0.05–12 μm , with a D50 of 0.76 μm . The overlap rate of their particle size distributions exceeds 85%, ensuring that the light scattering and absorption characteristics of the simulated smoke are highly consistent with actual transformer oil smoke. Second, regarding optical density, we measured the extinction coefficients of the two types of smoke at 550 nm wavelength under the same combustion intensity and smoke concentration conditions. The results show that the extinction coefficient of the mixed fuel smoke is $0.18 \text{ m}^{-1} \cdot \text{g}^{-1} \cdot \text{m}^3$, while that of transformer insulating oil smoke is $0.21 \text{ m}^{-1} \cdot \text{g}^{-1} \cdot \text{m}^3$, with a difference of less than 15%. This indicates that the two types of smoke have very similar visual contrast and transparency characteristics under natural lighting conditions in valve halls. Furthermore, CFD simulation results confirm that the buoyancy-driven leakage and diffusion patterns of the mixed fuel smoke at the valve hall sealing points are highly consistent with those of transformer insulating oil smoke, with a maximum deviation in plume rising velocity and diffusion angle of less than 12%. Through a series of preliminary experiments, the above two fuels were mixed in different proportions to ensure the stability of smoke and flames during the experiment. As shown in Figure 2, to concentrate the smoke and allow it to leak from a specific outlet, a stainless-steel smoke hood was designed and installed above the burner. The function of this device is to guide the smoke to flow in a designated direction, thereby simulating the real scenario of smoke overflow from the sealing port. A circular smoke exhaust port with a diameter of approximately 20 cm was arranged on the top of the stainless-steel smoke hood, and a fireproof cloth was covered above it to control the emission of flue gas. By adjusting the opening size of the fireproof cloth covering the smoke exhaust port, the flue gas leakage under different sealing effects was simulated. The fireproof cloth is made of high-temperature resistant and flame-retardant materials, which can maintain structural integrity in high-temperature environments. It can not only effectively prevent direct flame penetration but also allow a certain amount of flue gas to be discharged through the smoke exhaust port, realizing the regulation of smoke concentration. A background board was set behind the smoke collection port to enhance the contrast between the smoke and the background. Multiple schemes were selected for the background board material (original fireproof board, painted board, waterproof cloth, and latex paint board). The surface color of the original fireproof board is grayish, with little difference in color and brightness from the smoke, resulting in blurred smoke boundaries; after white painting treatment, the contrast of the painted board is improved, but its surface has severe reflection, which is

prone to noise under strong light; although the waterproof cloth is white in color, its surface has wrinkles, which may easily cause the recognition algorithm to mistakenly identify wrinkles as smoke; the surface of the latex paint board is white and uniform in texture, which can significantly enhance the visibility of smoke, especially showing the best effect on the edge features of dilute smoke. Finally, the surface of the shell is white latex paint as the background color.

Table 1. Summary of experimental variables.

Variable	Specific Parameters
fuel gas concentration	dilute, medium, dense smoke
fuel ratio (engine oil: n-heptane)	1:1/1:0.5/1:0
fire-resistant fabric	small holes, medium holes, open
backdrop	raw, spray-painted, latex, waterproof fabric
time period	morning, noon, afternoon

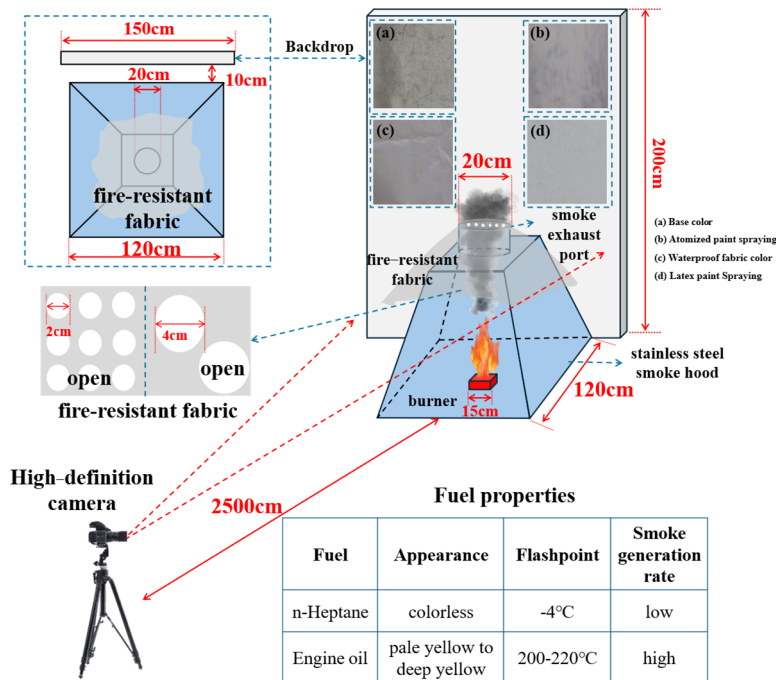


Figure 2. Experimental Setup.

2.2. YOLOv5 Model and Proposed Improvements

YOLOv5 is one of the commonly used detection models at present [13]. In this study, YOLOv5 is adopted as the base model to develop an improved YOLOv5-based detection algorithm for smoke and flames. The core improvement strategy is to embed four different attention mechanism modules into the model—Global Attention Mechanism (GAM), Convolutional Block Attention Module (CBAM), Coordinate Attention (CA), and Efficient Channel Attention (ECA)—aiming to enhance the model’s capability to extract and focus on the features of dilute smoke and key information of flames.

2.2.1. YOLOv5 Model

Proposed by Ultralytics, YOLOv5 boasts numerous advantages including high speed, high accuracy and lightweight architecture. In this work, YOLOv5s is selected as the foundational architecture based on comprehensive consideration of the actual deployment requirements of UHV converter station industrial scenarios, and it has distinct core advantages over newer models such as YOLOv8s and YOLOv11s. YOLOv5s has extremely

lightweight characteristics, with significantly lower parameter count and computational complexity, which can fully adapt to the limited computing power of edge embedded devices commonly used in converter stations; it also has a mature and complete industrial deployment ecosystem, with rich toolchain support and extensive hardware adaptation capabilities, which can be directly and quickly transplanted to the existing industrial monitoring platform, greatly reducing deployment costs and cycles; in addition, its highly modular architecture design is convenient for flexible integration of multiple attention mechanism modules, while ensuring the stability and convergence of the training process. We have also verified that applying the same attention improvement strategy to YOLOv11s can achieve a similar performance improvement, but its comprehensive deployment cost is much higher, so YOLOv5s is the optimal choice for this specific industrial application scenario. As illustrated in Figure 3, the network structure of YOLOv5 is primarily composed of four components: Input, Backbone, Neck, and Prediction.

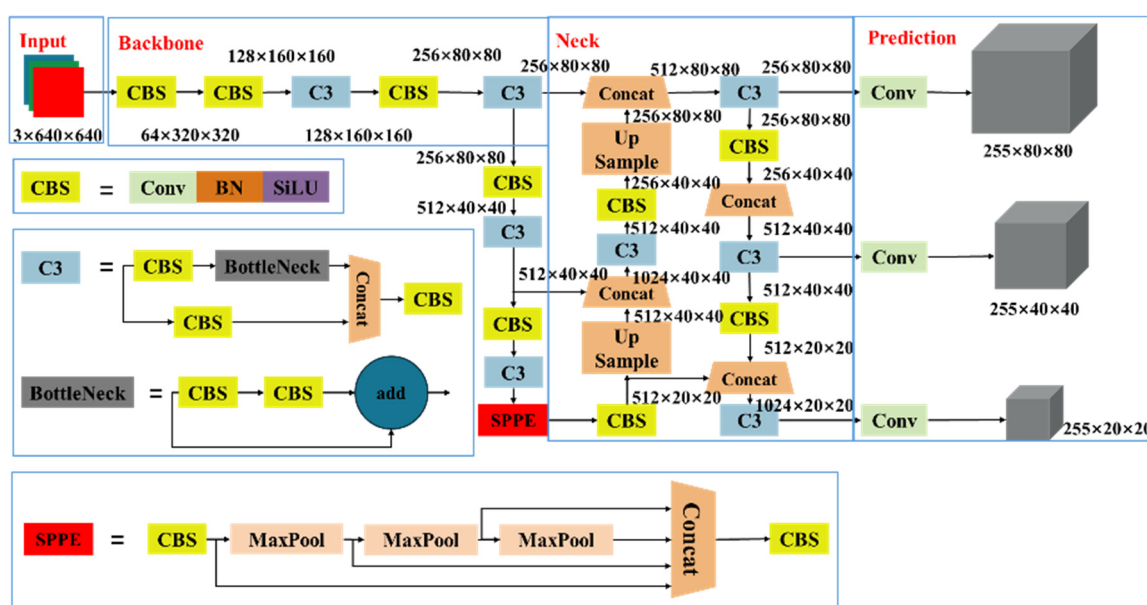


Figure 3. YOLOv5 network architecture diagram.

Within the network structure, the function of the Input module is to receive the smoke and flame images to be detected and feed them into the network for processing. The Input end of YOLOv5 adopts an adaptive image scaling method to reduce redundant information, and retains the Mosaic data augmentation technique to enrich the dataset. Its Backbone network is composed of the Convolution-Batch Norm-Activation Sequence (CBS), Cross Stage Partial Network with 3 Convolutions (C3) and Spatial Pyramid Pooling Fast (SPPF) modules, which is dedicated to extracting multi-scale image features. Based on CSPNet, the C3 module can enhance the network's attention to the global information of relevant objects and its feature extraction capability.

In the Neck (feature fusion layer), the feature fusion module integrates the Feature Pyramid Networks (FPN) [16] and Path Aggregation Network (PAN) [22]. These two modules can fuse multi-scale features, thereby improving the accuracy of target detection. Finally, the Prediction module is responsible for outputting target detection results, namely the positions and attributes of the identified targets (e.g., smoke and flame). Its three detection heads can capture target information of three sizes: large, medium, and small.

2.2.2. Proposed Improvements

In recent years, attention mechanisms have been widely applied in computer vision tasks. Their core idea is to simulate the human visual attention mechanism, enabling the model to automatically focus on more critical regions or channel features in images, thereby enhancing the perception capability for key targets. In smoke detection tasks, attention mechanisms can guide the model to strengthen its focus on the detailed features of dilute smoke, such as edges and textures, reducing the likelihood of being over-whelmed by background interference [18]. The improved method proposed in this study integrates four mainstream attention mechanisms into the basic structure of YOLOv5, namely the Global Attention Mechanism (GAM), Convolutional Block Attention Module (CBAM), Coordinate Attention (CA), and Efficient Channel Attention (ECA).

From a theoretical perspective, attention mechanisms can be broadly categorized into channel attention, spatial attention, and hybrid attention frameworks. Channel attention focuses on modeling the interdependencies among feature channels, typically by learning a set of channel-wise weights to re-scale feature responses, thereby emphasizing informative feature maps and suppressing irrelevant ones. Spatial attention, in contrast, aims to identify “where” the important information is located by generating spatial weight maps, allowing the network to focus on salient regions within feature maps. Hybrid attention mechanisms combine both channel-wise and spatial-wise modeling to capture more comprehensive feature representations. Specifically, GAM belongs to a global hybrid attention mechanism, which simultaneously models channel and spatial dependencies through fully connected transformations, enabling long-range interaction across feature dimensions. This design allows the network to capture global contextual information, which is particularly beneficial for large-scale, low-contrast targets such as diffuse smoke. CBAM follows a sequential attention strategy, consisting of a channel attention module and a spatial attention module applied in series. Its theoretical basis lies in decomposing the attention learning process into two orthogonal dimensions, which simplifies optimization while preserving representational power. CA introduces a novel mechanism that embeds precise positional information into channel attention by de-composing spatial information into two one-dimensional feature encoding processes along height and width. This enables the model to retain long-range dependency while preserving location awareness, which is critical for capturing anisotropic structures such as elongated or drifting smoke plumes. ECA, on the other hand, is derived from Efficient Channel Attention theory, which avoids dimensionality reduction and instead employs local cross-channel interaction via one-dimensional convolution. This significantly reduces computational complexity while maintaining effective channel dependency modeling.

As illustrated in Figure 4, the GAM is embedded between the output layer of the Backbone and the Neck structure. By leveraging its capability to simultaneously model global channel-wise and spatial dependencies, GAM enhances the model’s ability to capture global contextual information of large-area dilute smoke. To address the demand for multi-scale feature fusion, CBAM is incorporated into the Neck layer. Through its sequential channel and spatial attention sub-modules, CBAM can adaptively emphasize key channel features and prominent spatial regions on feature maps of different scales, thereby improving the detection robustness for targets with variable morphologies such as black-gray smoke. To further optimize the perception of smoke diffusion patterns, CA is adopted. By embedding positional coordinate information into channel attention, this mechanism enables the model to effectively capture the strip-shaped or diffused distribution features of smoke, significantly improving the positioning accuracy for the spatial extension direction of smoke. Considering the requirements for computational efficiency in practical deployment, this study also embeds the lightweight ECA mechanism after the

shallow convolutional layers of the Backbone. ECA realizes local cross-channel interaction through one-dimensional convolution, which enhances the model's early perception capability for subtle features (e.g., smoke texture and flame color) at an extremely low computational cost. The synergistic effect of the four aforementioned attention mechanisms enables the improved YOLOv5 model to significantly enhance the feature selection and focusing capability for dilute smoke and weak flames, while maintaining real-time performance.

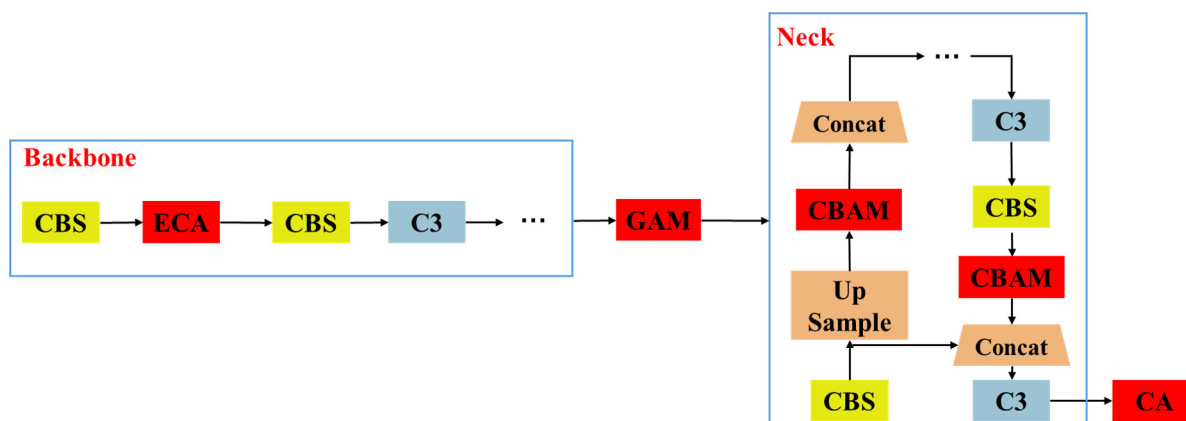


Figure 4. Improved model.

2.2.3. Loss Function

The loss function plays a crucial role in neural networks. Quantifying the discrepancy between the actual values and the expected values at the end of each iteration guides subsequent model training in a more rational manner and accelerates the convergence rate of the model during the training process. The loss function of YOLOv5 consists of three components: classification loss, localization loss, and confidence loss.

For classification loss and confidence loss, this study adopts the Binary Cross-Entropy (BCE) loss function. The classification loss is used to measure the discrepancy between the predicted category of the target in the bounding box and the ground-truth category, while the confidence loss is used to measure the discrepancy between the predicted probability of the presence of a target in the bounding box and the actual situation. The mathematical expression of the Binary Cross-Entropy loss is as follows:

$$L_{cls}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \quad (1)$$

where y represents the true label ($y = 1$ indicates flame, $y = 0$ indicates smoke).

Localization loss is used to measure the discrepancy between the predicted bounding boxes and the ground-truth bounding boxes. In this improved model, the Complete Intersection over Union (CIoU) loss is adopted as the bounding box regression loss. The CIoU loss not only takes into account the overlap degree (IoU) between bounding boxes but also incorporates geometric factors such as center point distance and aspect ratio, as well as the factor of direction consistency, thereby evaluating the regression quality of bounding boxes in a more comprehensive manner. Its mathematical expression is as follows:

$$L_{loc} = 1 - IoU + \frac{\rho^2(b, b^g)}{c^2} + \alpha v \quad (2)$$

where IoU represents the intersection-over-union ratio between the predicted bounding box and the ground truth bounding box, measuring their degree of overlap; $v =$

$\frac{4}{\pi^2} \left(\arctan\left(\frac{w^g}{h^g}\right) - \arctan\left(\frac{w}{h}\right) \right)^2$, measuring the difference in the width-to-height ratio between the two bounding boxes; and $\alpha = \frac{v}{(1-IoU)+v}$, ensuring the aspect ratio loss is reasonably weighted across different *IoU* scenarios through the balancing factor.

The shape, size, and color shade of flames and smoke exhibit dynamic changes, making accurate detection quite challenging. In contrast, the CIOU loss function demonstrates favorable performance in bounding box regression, as it takes into account multiple geometric factors to comprehensively measure the discrepancy between predicted and ground-truth bounding boxes. This enables the improved model to effectively enhance the accuracy of locating flame and smoke targets with variable shapes and sizes, thereby improving the overall detection performance.

2.2.4. Model Performance Evaluation Metrics

In target detection tasks, metrics such as mean average precision (mAP), precision, and recall are generally adopted to evaluate the accuracy of flame and smoke detection models [24].

In this study, precision and recall are selected as the metrics for assessing multi-target detection accuracy. A predicted bounding box is determined as a true positive (TP) when its Intersection over Union (IoU) with the ground-truth bounding box exceeds the preset threshold; in contrast, a predicted bounding box with an IoU below the preset threshold is classified as a false positive (FP). A missed target, which refers to a target that fails to be detected during prediction, is defined as a false negative (FN). Precision and recall are calculated for each category using Equations (3) and (4), respectively, based on the TP, FP, and FN values of each corresponding category.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

Average precision (AP) integrates precision and recall, providing a more intuitive metric for evaluating model accuracy. In the precision–recall (P-R) curve, the area enclosed by the curve and the coordinate axes represents the average precision of a specific category.

$$AP = \int_0^1 p(r) dr \quad (5)$$

The mAP represents the average value of the AP across all categories, enabling it to measure the prediction accuracy in a more objective manner.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (6)$$

3. Results

This section systematically elaborates on the entire process of the YOLOv5s-based flame and smoke detection algorithm, ranging from dataset construction and model training to performance validation. Through refined experimental analysis of different fire development stages, its recognition performance and engineering applicability of the improved model are comprehensively evaluated.

3.1. Dataset Construction

High-quality annotation is fundamental to reliable model training, particularly for irregularly shaped targets such as smoke and flames. The MakeSense annotation tool was

employed to label bounding rectangles around flame and smoke regions in the collected images. To minimize subjective bias in bounding box placement—given the fluid and amorphous nature of smoke and flames—the maximum enclosing rectangle annotation strategy was consistently adopted for each target instance, which has been verified as a reliable standard for amorphous fire target annotation [25]. As summarized in Table 2, the final dataset comprises approximately 14,122 images, stratified into three primary categories: dilute smoke (early-stage low-density leakage), dense smoke (mid-stage accumulation), and overflowing flames (late-stage coexistence of flame and smoke). Representative samples are illustrated in Figure 5. Using stratified random sampling [26], the dataset was partitioned into training (70%), validation (10%), and test (20%) sets to ensure balanced representation of all fire stages and environmental conditions during model development and evaluation.

Table 2. Sample distribution and partition of the fireworks dataset into training, validation, and test.

Sample	Sample Size	Training Set	Validation Set	Test Set
dilute smoke	4623	3236	462	925
dense smoke	4595	3217	460	918
Flame	4904	3433	490	981
Total	14,122	9886	1412	2824

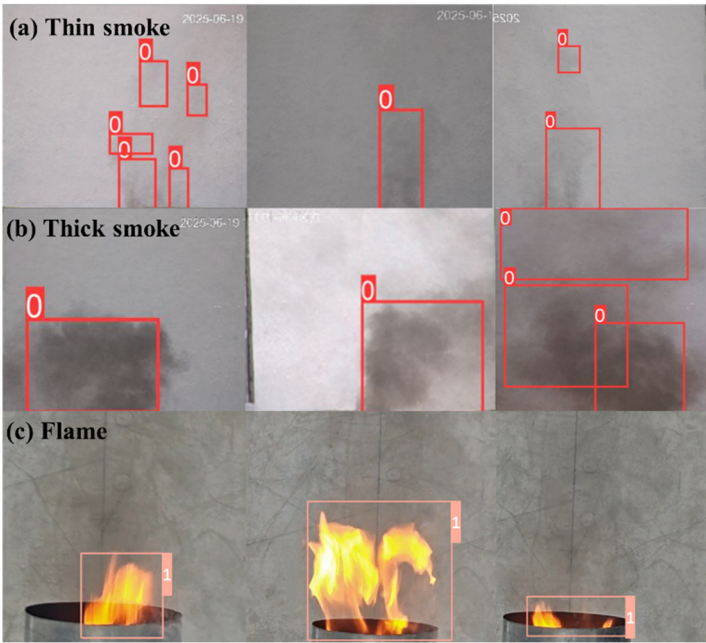


Figure 5. Sample dataset examples.

3.2. Model Training

Training and validation were conducted within the PyTorch 2.2.2 deep learning framework. The detailed experimental hardware and software environment is specified in Table 3. To ensure optimal model performance and training stability, we systematically optimized the core training hyperparameters through a grid search approach before formal training. The search ranges of key hyperparameters were set as follows: initial learning rate in {0.001, 0.005, 0.01, 0.015, 0.02}, batch size in {16, 32, 48, 64}, and total training epochs in {50, 80, 100, 120, 150}. All hyperparameter combinations were trained under the same hardware environment and dataset partitioning, and the optimal configuration was selected with the mean average precision at IoU = 0.5 (mAP@0.5) on the validation set as the core evaluation criterion, while comprehensively considering convergence stability and training efficiency.

Table 3. Hardware and software configuration of the experimental environment.

Test Environment	Specific Configuration
Compilation environment	Python 3.8 + Pytorch 13.1 + CUDA 12.8
Operating system	Windows 10
CPU	Intel 9-14900KF
GPU	NVIDIA RTX 5070 Ti (16 GB)

After comparative grid search experiments, the optimal hyperparameter combination was determined as follows: SGD optimizer with an initial learning rate of 0.01, a batch size of 8, and a total of 100 training epochs. This configuration achieves the highest validation mAP@0.5 while avoiding overfitting caused by excessive training iterations. In terms of learning rate scheduling, we adopted a cosine annealing decay strategy, which gradually reduces the learning rate following a cosine curve throughout the training process, helping the model converge to a more stable optimal solution in the late training stage and improving the final detection accuracy. In addition, the momentum parameter was set to 0.937 and the weight decay coefficient was set to 0.0005, consistent with the classical robust configuration of the YOLOv5 series, which effectively suppresses overfitting while ensuring model convergence speed.

The proposed model maintains a compact lightweight model size, with a total of 7.03 M trainable parameters and a computational complexity of 15.8 GFLOPs at the standard 640×640 input resolution, which fully adapts to the computing power constraints of edge devices in converter stations. Under the above hyperparameter configuration, the complete 100-epoch training time totals 4.254 h, achieving high training efficiency while guaranteeing detection performance, while key hyperparameter settings are listed in Table 4.

Table 4. Specific training parameters used in model optimization.

Test Environment	Specific Configuration
Input image size	640×640
Batch size	32
Initial learning rate	0.01
Number of training Iterations	100
Momentum parameter	0.937

Figure 6 depicts the evolution of the three loss components—classification loss (cls_loss), bounding box regression loss (box_loss), and objectness/confidence loss (obj_loss)—on the validation set as a function of training epochs. All loss terms exhibited high initial values, followed by rapid decline and eventual stabilization after approximately 50–80 epochs. This convergence pattern indicates effective learning of target localization, category discrimination, and confidence estimation without evident overfitting, as the model generalized well to unseen validation samples.

Corresponding performance metrics on the validation set are shown in Figure 7. Precision and recall increased sharply in the early epochs and plateaued at high levels, reflecting progressive improvement in positive sample detection and reduction of false negatives. The mean average precision at IoU = 0.5 (mAP@0.5) achieved a high stable value, while mAP@0.5:0.95 demonstrated robust performance across a range of localization strictness thresholds, confirming the model's comprehensive capability in fire and smoke detection tasks.

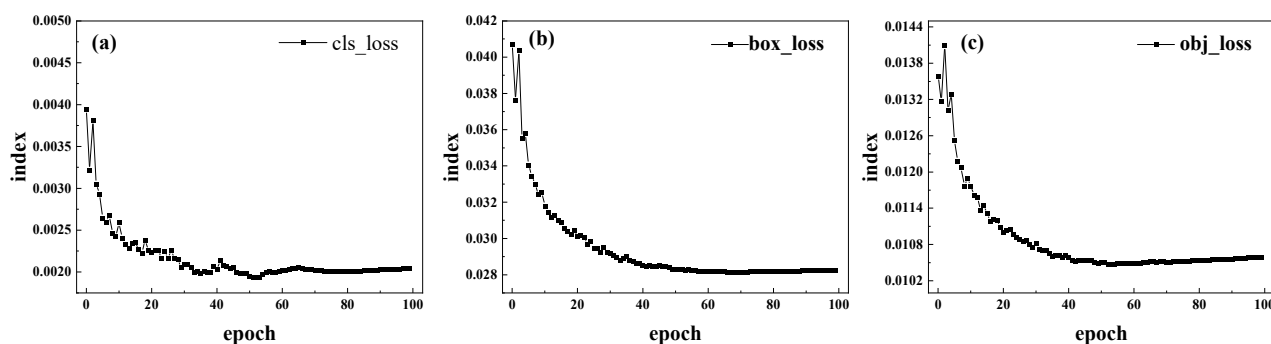


Figure 6. Convergence of validation losses during model training: (a) cls_loss; (b) box_loss; (c) obj_loss.

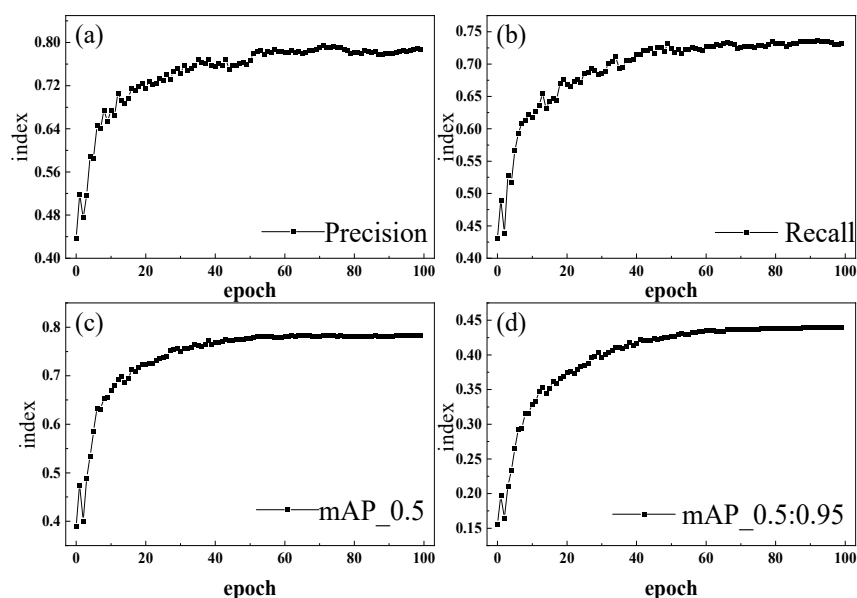


Figure 7. Evolution of validation performance metrics during model training: (a) Precision; (b) Recall; (c) mAP_0.5; (d) mAP_0.5:0.95.

3.3. Test Set Performance Evaluation

The trained attention-enhanced YOLOv5s model was rigorously evaluated on the held-out test set, with particular emphasis on performance across the three representative fire evolution stages: dilute smoke, dense smoke, and overflowing flame.

3.3.1. Dilute Smoke Stage

In the incipient phase of fire development, incomplete combustion produces limited volumes of small-particle smoke that manifests as translucent, slowly ascending plumes with extremely low contrast against industrial backgrounds. This stage poses the greatest challenge to conventional detectors and the baseline YOLOv5s model, often resulting in complete missed detections or very high false negative rates. Integration of the four attention mechanisms substantially mitigated these issues by enhancing the extraction of subtle edge, texture, and contextual features. The qualitative comparison of detection results between the baseline model and the improved model on the dilute smoke sequence reveals significant performance differences. As quantified in Figure 8, the temporal sequence from time step t to $t + 3$ of the baseline model exhibits a notable delay in detecting sparse, low-contrast dilute smoke. It fails to trigger detection until $t + 2$, with a confidence score of only 0.80 at $t + 3$, indicating insensitivity to subtle smoke features and vulnerability to background interference. In contrast, the improved model demonstrates superior early

warning capability and feature discrimination: it successfully identifies the dilute smoke target at t with a confidence of 0.69, and the confidence rises to 0.86 at $t + 2$ as the smoke plume evolves, maintaining stable detection at $t + 3$.

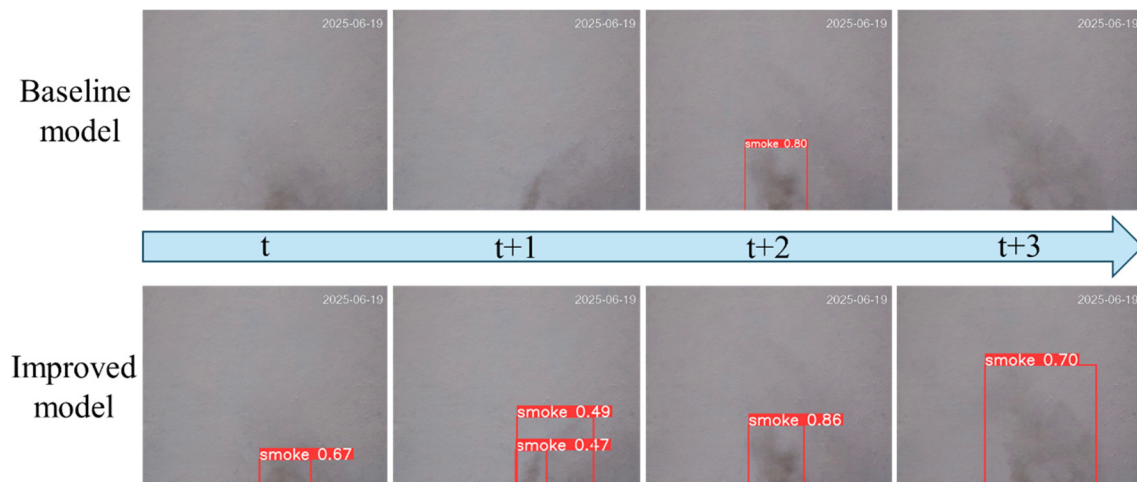


Figure 8. Time-series detection results of dilute smoke between the baseline model and the improved model.

3.3.2. Dense Smoke Stage

As combustion intensifies, smoke production surges, rapidly filling the collection hood and diffusing outward through the regulated opening. This phase features well-defined plume boundaries and higher optical density, rendering targets more salient. As shown in Figure 9, the temporal sequence from time step t to $t + 3$, both models can effectively detect dense smoke targets throughout the entire sequence, but the improved model achieves higher detection confidence and more stable performance. At time step t , the baseline model detects dense smoke with a confidence score of 0.66, while the improved model achieves a higher confidence of 0.73, indicating stronger initial feature extraction capability for opaque smoke regions. At $t + 1$, the confidence of the baseline model decreases to 0.58, whereas the improved model maintains a stable confidence of 0.68, reflecting better robustness against texture variations in evolving smoke plumes. At $t + 2$, the baseline model fails to output a valid detection result, suffering from a temporary missed detection, while the improved model still detects the smoke target with a confidence of 0.55, ensuring continuous monitoring. Finally, at $t + 3$, the baseline model recovers detection with a confidence of 0.83, and the improved model achieves the highest confidence of 0.86, further validating its superior ability to maintain accurate detection as the smoke plume becomes more concentrated. These results confirm that the improved model outperforms the baseline model in terms of detection confidence, temporal stability, and continuity for dense smoke scenarios, which is essential for reliable fire monitoring and early warning in industrial environments.

3.3.3. Overflowing Flame Stage

In the advanced stage, flames breach the penetration seal, escaping alongside voluminous smoke. The intermingling of flame and smoke creates overlapping visual features, increasing detection complexity due to occlusion and dynamic morphology. Through the feature enhancement of attention mechanisms, the improved YOLOv5 model is also capable of capturing flame features and achieving accurate target detection. As shown in Figure 10, the temporal sequence spanning from time step t to $t + 3$, the baseline model shows fluctuating confidence and unstable localization for evolving flames. At the initial

time step t , the baseline model did not detect the faint flame. At $t + 2$, its confidence only reached 0.58, indicating the instability of feature responses during the dynamic evolution of the fire. Although the confidence recovers to 0.68 at $t + 3$, the fluctuation reveals the model's vulnerability to rapid changes in flame morphology and intensity. In contrast, the improved model demonstrates more robust and stable performance across the entire sequence. Specifically, at time t , the model effectively captures faint flames and consistently maintains high and stable confidence values, which are always higher than the confidence of the baseline model at the corresponding time steps. This consistent performance indicates that the improved model enhances the extraction of discriminative features for flame regions, suppresses interference from complex backgrounds, and ensures reliable tracking of the overflowing flame's temporal evolution.

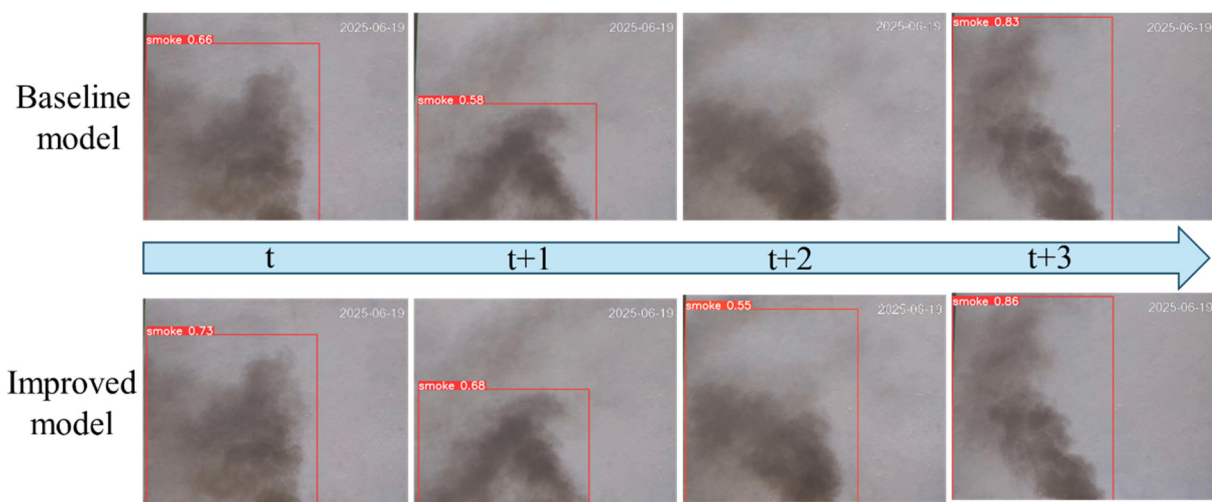


Figure 9. Time-series detection results of dense smoke between the baseline model and the improved model.

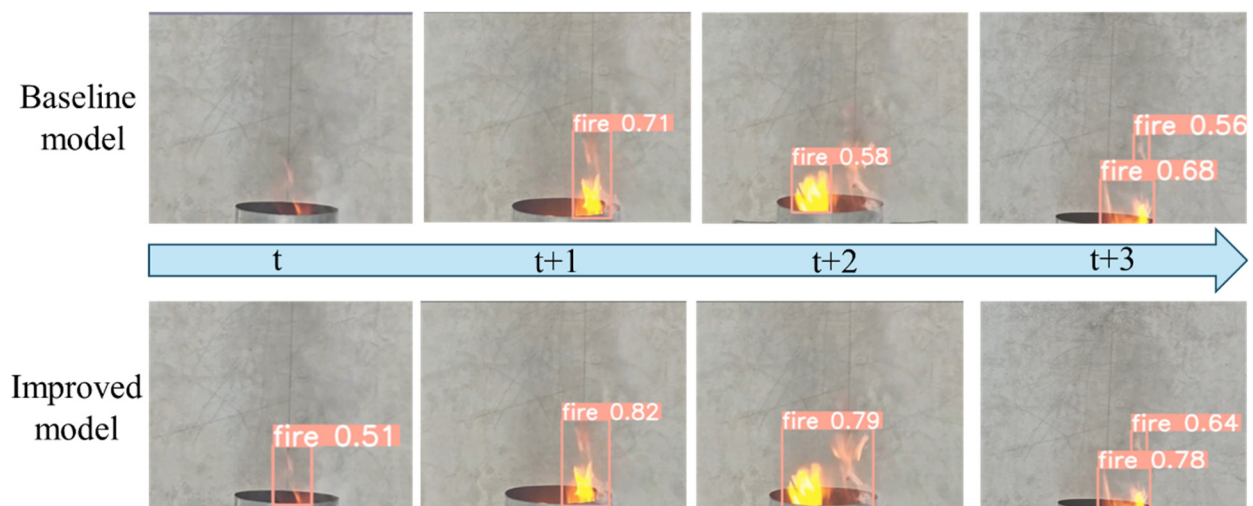


Figure 10. Time-series detection results of flame between the baseline model and the improved model.

Figure 11 quantitatively compares the precision and recall metrics of the baseline YOLOv5 and improved YOLOv5 models across the three fire stages (dilute smoke, dense smoke, and overflowing flame), quantitatively validating the performance enhancement of the improved model in complex fire scenarios. As illustrated in the figure, the improved YOLOv5 model consistently outperforms the baseline model in all three fire stages, with

distinct performance advantages highlighted in each stage. In the dilute smoke stage, the baseline YOLOv5 model achieves a precision of 57% and a recall of 61%, whereas the improved YOLOv5 model significantly boosts these metrics to 81% and 84%, respectively. These improvements demonstrate that the modified network structure enhances the model's ability to extract subtle features from low-contrast, sparse smoke targets, effectively addressing the challenge of missed detections in the early phase of fires. In the dense smoke stage, both models show improved performance due to the higher feature salience of dense smoke. Nevertheless, the improved YOLOv5 model still achieves superior results. Its precision reaches 92%, and the recall reaches 93%. This indicates that the improved model strengthens the screening and enhancement of effective features in heavily occluded or thick smoke environments, ensuring more reliable localization and classification. In the overflowing flame stage, the baseline YOLOv5 model already achieves a relatively high recognition performance, with a precision of 93% and a recall of 91%. The improved YOLOv5 model further refines this performance, reaching a precision of 98% and a recall of 95%. This marginal yet consistent improvement confirms that the improved model maintains strong robustness and feature discrimination even for high-salience flame targets, achieving fine-grained recognition optimization. Collectively, the quantitative results in Figure 10 confirm that the improved YOLOv5 model exhibits comprehensive and superior recognition performance across all fire stages of converter stations. It achieves particularly remarkable performance gains in challenging scenarios such as early-stage dilute smoke and evolving dense smoke, verifying its strong generalization capability and practical value for real-world converter station fire detection.

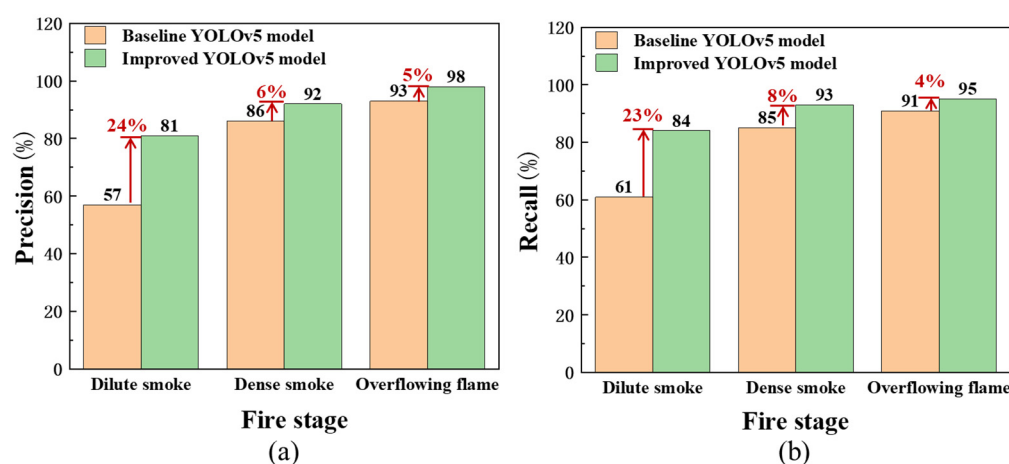


Figure 11. The precision (a) and recall (b) of the model at different fire stages.

3.3.4. Comparison with Existing Cv-Based Fire Detection Models

To further validate the effectiveness and competitiveness of the proposed method, a comparative analysis was conducted against several representative computer vision-based fire detection models, including YOLOv8, YOLOv11 with attention mechanisms, and 3D CNN-based models. These models represent the current mainstream directions in fire detection research, covering lightweight real-time detectors, advanced attention-enhanced architectures, and spatiotemporal feature modeling approaches. For a fair comparison, all models were evaluated on the same test dataset constructed in this study, which includes three representative fire stages.

As shown in Table 5, the proposed method achieves the highest precision (91.8%) and recall (92.9%) among all compared models, demonstrating its superior detection performance. Compared with YOLOv8, the proposed method shows noticeable improvements, indicating that the integration of multiple attention mechanisms effectively enhances

feature representation beyond standard one-stage detectors. Although YOLOv11 with attention mechanisms exhibits competitive performance due to its stronger feature extraction capability, the proposed method still outperforms it, which can be attributed to the synergistic effect of multi-attention modules specifically designed for low-salience target detection. In contrast, the 3D CNN model demonstrates relatively lower performance. While it is capable of capturing temporal features, its effectiveness is limited in scenarios dominated by weak spatial cues, such as dilute smoke, and it typically requires higher computational cost, making it less suitable for real-time industrial deployment. More importantly, the superiority of the proposed method is particularly evident in the dilute smoke stage, which represents the most challenging and critical phase for early fire detection. Benefiting from the collaborative enhancement of global context modeling, spatial attention, and channel-wise feature refinement, the proposed model significantly improves sensitivity to weak, low-contrast smoke features, thereby reducing missed detections and enabling earlier fire warning.

Table 5. Performance comparison of different CV-based fire detection models.

Model	Precision (%)	Recall (%)
YOLOv8	88.5	89.7
YOLOv11 + Attention	90.6	91.2
3D CNN	87.3	88.1
Proposed Model	91.8	92.9

Overall, the comparative results confirm that the proposed method not only achieves higher detection accuracy but also provides a better balance between performance and practical applicability, making it more suitable for deployment in complex industrial environments such as converter station valve halls.

3.4. Ablation Study and Mechanism Analysis

To systematically evaluate the contribution of each attention mechanism and to verify the effectiveness of the proposed multi-attention fusion strategy, a series of ablation experiments were conducted on the overall test dataset rather than a specific scenario. The baseline model is the original YOLOv5s without any attention modules. Four attention mechanisms, namely ECA, GAM, CBAM, and CA, were progressively integrated into the baseline model in different combinations to analyze their individual and joint effects on detection performance.

The quantitative results of the ablation experiments are summarized in Table 6. It can be observed that the introduction of any single attention module leads to an improvement over the baseline model, demonstrating the effectiveness of attention mechanisms in enhancing feature representation. Specifically, the ECA module yields a modest performance gain due to its efficient channel wise interaction, which improves the model's sensitivity to informative features with minimal computational overhead. The GAM module provides a relatively larger improvement, indicating that global context modeling plays a critical role in capturing the spatial distribution characteristics of smoke. Similarly, CBAM enhances both channel and spatial attention, resulting in more refined feature extraction and improved robustness under complex backgrounds. The CA module contributes by embedding positional information into channel attention, which is particularly beneficial for improving localization capability.

Table 6. Ablation study of attention modules on the overall test set.

Model Variant	Precision (%)	Recall (%)
Baseline Model	86.2	88.1
+ ECA	87.1	88.9
+ GAM	88.0	89.6
+ CBAM	88.6	90.2
+ CA	87.8	89.8
+ ECA + GAM	89.2	90.8
+ GAM + CBAM	90.1	91.5
+ CBAM + CA	90.4	91.8
+ ECA + GAM + CBAM	91.0	92.2
Improved Model (All)	91.8	92.9

When multiple attention modules are combined, the model performance is further improved, revealing a clear complementary relationship among different mechanisms. For instance, the combination of GAM and CBAM significantly enhances both precision and recall, suggesting that integrating global dependency modeling with multi-scale spatial attention effectively strengthens the model's ability to distinguish smoke from background interference. Furthermore, the addition of CA to CBAM leads to additional gains, indicating that positional encoding further refines spatial feature representation. These results demonstrate that different attention mechanisms focus on distinct aspects of feature learning and can work synergistically when properly integrated. The best overall performance is achieved by the proposed model that incorporates all four attention mechanisms simultaneously. Compared with the baseline model, the proposed method improves precision from 86.2% to 91.8% and recall from 88.1% to 92.9%. It confirms that the multi-attention collaborative framework can effectively integrate local feature refinement, global context awareness, spatial structure encoding, and multi-scale feature selection into a unified representation.

4. Conclusions

This study focuses on developing a visual detection framework for the early identification of fire hazards, specifically the initial faint smoke signals, at valve hall sealing points in ultra-high-voltage (UHV) converter stations, a challenge that traditional detection technologies struggle to address. The framework, built upon simulation experiments and an enhanced YOLOv5s model, achieves comprehensive detection across the entire fire evolution, effectively resolving key technical bottlenecks such as the difficulty in extracting weak features of early-stage smoke and the latency in issuing early warnings. The major findings are as follows:

- (1) The proposed method constructs a realistic data foundation and achieves significant breakthroughs in early fire detection, particularly for weak features. By designing a small-scale valve hall sealing simulation experiment platform, this study systematically reproduces the complete fire evolution process from initial sparse smoke to dense smoke and, ultimately, flame overflow. The constructed dataset, which includes complex scenarios like low-contrast smoke under various backgrounds and illumination, provides critical sample support for model training. Based on this, an improved YOLOv5s model integrating multiple attention mechanisms (GAM, CBAM, CA, ECA) was developed, achieving end-to-end high-precision detection for all fire stages. Most notably, it increased the precision for dilute smoke—the most challenging initial stage from a baseline of 57% to 81%, thus significantly enhancing the system's early warning capability.

- (2) The enhanced model demonstrates robust performance and adaptability in complex industrial environments. The algorithm validation results show that the improved model performs excellently not only in detecting clear targets like dense smoke and flames (with accuracy reaching 92% and 98%, respectively) but also shows high robustness against disturbances such as background interference and lighting variations. Compared to the original model, it achieves a significant reduction in both false and missed alarm rates, successfully addressing the high miss rate problem that traditional models face when detecting early, weak signals.

Despite the favorable detection performance achieved by the proposed method across all fire evolution stages, there are still some limitations in the current research. First, the dataset used for model training and validation is constructed based on controlled simulation experiments, where the test space and fire scale are consistent with the actual operating conditions of the valve hall. However, the controlled experimental environment cannot fully replicate the complex environmental conditions in actual UHV converter stations, such as large space airflow disturbance, multi-source background interference, and strong electromagnetic field environment. The number of fire samples collected from real operating converter stations is relatively limited, which may restrict the generalization performance of the model in fully realistic engineering scenarios. Second, the detection performance of the model still needs to be further improved under extreme lighting conditions such as direct strong sunlight and backlight, and the risk of false alarms or missed alarms may increase in such harsh environments. Third, the current detection framework is built on a single camera visual perception mode, which cannot realize three-dimensional spatial localization of fire sources and multi-view collaborative monitoring, making it difficult to cover all monitoring blind spots in the large internal space of the valve hall. Future works are required to continue to solve the above limitation.

Author Contributions: R.L.; writing—original draft, software, J.X.; resources, software, H.H. and T.Z.; investigation, methodology, Y.G.; data curation, resources, X.L.; visualization, writing—review and editing, Y.H.; conceptualization, formal analysis, J.Z.; funding acquisition, project administration, supervision, T.M.; writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Project of State Grid Corporation of China (Grant No. 5200-202456099A-1-1-ZN).

Data Availability Statement: Data will be accessible upon request.

Conflicts of Interest: Jia Xie and Hanbing Hao were employed by State Grid Anhui Electric Power Co., Ltd. The remaining authors declares no conflicts of interest.

References

1. Fairley, P. A Grid as Big as China. *IEEE Spectr.* **2019**, *56*, 36–41. [[CrossRef](#)]
2. Baggini, A.; Ploetner, C.; Riboud, J.-C. Functional Definition and Classification of the Power Transformer. *IEEE Trans. Power Deliv.* **2022**, *38*, 80–84.
3. Yao, H.; Lv, K.; Yan, S.; Xing, M.; Lou, Z.; Ren, W. Simulation Study on Fire Combustion Process of Oil Immersed Transformer. *Int. J. Heat Technol.* **2023**, *41*, 455–461. [[CrossRef](#)]
4. Zhou, B.; Yang, W.; Yoshioka, H.; Chen, T.; Wang, K.; Hao, D.; Jiang, C.; Cui, K. Research on Suppression Effectiveness of Compressed Air Foam for Oil-Immersed Transformer Hot Oil Fire. *Case Stud. Therm. Eng.* **2023**, *49*, 103272. [[CrossRef](#)]
5. Wu, H.; Wu, D.; Zhao, J. An Intelligent Fire Detection Approach through Cameras Based on Computer Vision Methods. *Process Saf. Environ. Prot.* **2019**, *127*, 245–256. [[CrossRef](#)]
6. Yin, Z.; Wan, B.; Yuan, F.; Xia, X.; Shi, J. A Deep Normalization and Convolutional Neural Network for Image Smoke Detection. *IEEE Access* **2017**, *5*, 18429–18438. [[CrossRef](#)]

7. Khan, S.; Muhammad, K.; Hussain, T.; Del Ser, J.; Cuzzolin, F.; Bhattacharyya, S.; Akhtar, Z.; de Albuquerque, V.H.C. DeepSmoke: Deep Learning Model for Smoke Detection and Segmentation in Outdoor Environments. *Expert Syst. Appl.* **2021**, *182*, 115125. [\[CrossRef\]](#)
8. Ghali, R.; Akhloufi, M.A. Deep Learning Approach for Wildland Fire Recognition Using RGB and Thermal Infrared Aerial Image. *Fire* **2024**, *7*, 343. [\[CrossRef\]](#)
9. Doddapaneni, T.; Ediga, U.; Pati, P.B. Detecting fire in color images using convolutional neural network architectures. In Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 7–9 April 2023.
10. Mukhiddinov, M.; Abdusalomov, A.B.; Cho, J. A Wildfire Smoke Detection System Using Unmanned Aerial Vehicle Images Based on the Optimized YOLOv5. *Sensors* **2022**, *22*, 9384. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Sozol, M.S.S.; Mondal, M.R.H.; Thamrin, A.H. Indoor Fire and Smoke Detection Based on Optimized YOLOv5. *PLoS ONE* **2025**, *20*, e0322052. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Wang, Y.; Hua, C.; Ding, W.; Wu, R. Real-Time Detection of Flame and Smoke Using an Improved YOLOv4 Network. *Signal Image Video Process.* **2022**, *16*, 1109–1116. [\[CrossRef\]](#)
13. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Fang, J.; Michael, K.; Montes, D.; Nadar, J.; Skalski, P.; et al. Ultralytics/YOLOv5: v6.1-TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference. 2022. Available online: <https://zenodo.org/record/6222936> (accessed on 10 June 2026).
14. Gonçalves, L.A.O.; Ghali, R.; Akhloufi, M.A. YOLO-Based Models for Smoke and Wildfire Detection in Ground and Aerial Images. *Fire* **2024**, *7*, 140. [\[CrossRef\]](#)
15. Alkhamash, E.H. A comparative analysis of YOLOv9, YOLOv10, YOLOv11 for smoke and fire detection. *Fire* **2025**, *8*, 26. [\[CrossRef\]](#)
16. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Honolulu, HI, USA, 2017; pp. 2117–2125.
17. Zhou, H.; Zhang, D.; Hu, G.; Gu, X. Research on Lightweight Improvement of YOLOv8-based Fire Detection Model for Edge Computing. In *Proceedings of the 2023 3rd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*; IEEE: Honolulu, HI, USA, 2024. [\[CrossRef\]](#)
18. Zhang, M.Z.; Duan, J.Z.; Liang, Z.W.; Guo, J.J.; Chai, D.S. Fire and Smoke Detection Method Based on Improved YOLO-V5 Algorithm. *China Saf. Sci. J.* **2024**, *34*, 155–161.
19. Fang, X.J.; Jiang, X.F.; Xu, L.Q.; Fang, Z.Y. CFS-YOLO: An Early Fire Detection Method via Coarse and Fine Grain Search and Focus Modulation. *J. Electron. Inf. Technol.* **2025**, *47*, 1976–1991.
20. Zhang, J.; Yin, L.; Gong, X.F.; Xu, H.H. Small Target Smoke Detection Algorithm Based on Improved YOLOv5. *China Saf. Sci. J.* **2023**, *33*, 187–193.
21. Qian, W.; Luo, H.; Liu, Q.Y.; Yang, X.; Wang, H.B. Lightweight Flame Detection Method with Improved YOLOv10. *J. Saf. Environ.* **2026**, *26*, 875–883.
22. Ali, M.M.E.H.; Ghodrat, M. Reliable Indoor Fire Detection Using Attention-Based 3D CNNs: A Fire Safety Engineering Perspective. *Fire* **2025**, *8*, 285. [\[CrossRef\]](#)
23. Ejaz, U.; Hamza, M.A.; Kim, H.-C. Channel Attention for Fire and Smoke Detection: Impact of Augmentation, Color Spaces, and Adversarial Attacks. *Sensors* **2025**, *25*, 1140. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9.
25. Gragnaniello, D.; Greco, A.; Sansone, C.; Vento, B. FLAME: Fire detection in videos combining a deep neural network with a model-based motion analysis. *Neural Comput. Appl.* **2025**, *37*, 6181–6197. [\[CrossRef\]](#)
26. Liberty, E.; Lang, K.; Shmakov, K. Stratified Sampling Meets Machine Learning. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2016; pp. 2320–2329.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.