fire

MDPI

Article

Real-Time Smoke Detection in Surveillance Videos Using an Enhanced RT-DETR Framework with Triplet Attention and HS-FPN

Lanyan Yang [1], Yuanhang Cheng [1,*], Fang Xu [2,3], Boning Li [2,3,4] and Xiaoxu Li [2,3,4]

[1] School of Information Engineering, Shenyang University, Shenyang 110044, China; yanglanyan@ynu-edu.cn
[2] Shenyang Fire Research Institute of M.E.M., Shenyang 110034, China; xufang@syfri.com.cn (F.X.); liboning@syfri.com.cn (B.L.)
[3] National Engineering Research Center of Fire and Emergency Rescue, Shenyang 110034, China
[4] Key Laboratory of Fire Prevention Technology, Shenyang 110034, China
* Correspondence: chengyuanhang@syu.edu.cn

Abstract: This study addresses the urgent need for an efficient and accurate smoke detection system to enhance safety measures in fire monitoring, industrial safety, and urban surveillance. Given the complexity of detecting smoke in diverse environments and under real-time constraints, our research aims to solve challenges related to low-resolution imagery, limited computational resources, and environmental variability. This study introduces a novel smoke detection system that utilizes the real-time detection Transformer (RT-DETR) architecture to enhance the speed and precision of video analysis. Our system integrates advanced modules, including triplet attention, ADown, and a high-level screening-feature fusion pyramid network (HS-FPN), to address challenges related to low-resolution imagery, real-time processing constraints, and environmental variability. The triplet attention mechanism is essential for detecting subtle smoke features, often overlooked due to their nuanced nature. The ADown module significantly reduces computational complexity, enabling real-time operation on devices with limited resources. Furthermore, the HS-FPN enhances the system's robustness by amalgamating multi-scale features for reliable detection across various smoke types and sizes. Evaluation using a diverse dataset showcased notable improvements in average precision (AP50) and frames per second (FPS) metrics compared to existing state-of-the-art networks. Ablation studies validated the contributions of each component in achieving an optimal balance between accuracy and operational efficiency. The RT-DETR-based smoke detection system not only meets real-time requirements for applications like fire monitoring, industrial safety, and urban surveillance but also establishes a new performance benchmark in this field.

Keywords: smoke detection; RT-DETR; real-time processing; multi-scale feature fusion

check for updates

Citation: Yang, L.; Cheng, Y.; Xu, F.; Li, B.; Li, X. Real-Time Smoke Detection in Surveillance Videos Using an Enhanced RT-DETR Framework with Triplet Attention and HS-FPN. *Fire* **2024**, *7*, 387. https://doi.org/10.3390/fire7110387

Academic Editor: Lizhong Yang

Received: 3 September 2024
Revised: 25 October 2024
Accepted: 26 October 2024
Published: 28 October 2024

1. Introduction

Fires are common and highly destructive disasters that cause irreversible ecological and economic damage and pose significant threats to human lives [1–8]. For example, in China, forest fires frequently occur during the dry season, destroying vast areas of land, resulting in millions of deaths, and causing extensive property losses. Effective fire alarm measures are crucial for reducing casualties. Smoke, as a primary indicator of fire, is often the first sign to appear, making smoke detection an essential component of early fire alarm systems. There are many traditional methods for detecting fires, such as measuring particle size, temperature, and relative humidity, and assessing air clarity. To achieve this, they monitor changes in the surrounding environment [9]. While these methods have been significant in fire detection, enhancing their sensitivity and accuracy is necessary to meet the demands of complex environments and higher early warning requirements. By continually improving fire detection technologies, we can detect fires earlier and more accurately, reducing fire damage and protecting lives and property [10].

Traditional fire detection methods suffer from significant delays and misjudgments, often due to interference from non-fire energy emissions or byproducts from unrelated combustion processes. As an example, picture a real app that checks security videos every ten seconds to see if there is smoke, a model with a missing detection rate (false alarm rate) of 1% has a probability of $1 - 0.99360 \approx 97.32\%$ to generate missing detection (false alarm) within one hour. Additionally, other detection devices, such as infrared and optical sensors, face limitations due to restricted monitoring areas and varying environmental conditions, which fail to meet the requirements for effective safety warning systems [8–10]. To establish a more reliable and effective fire warning system, it is essential to overcome the limitations of current methods and devices and develop detection technologies with high sensitivity and accuracy that can handle complex environmental changes. This will enable earlier detection of fires, reduce casualties and property damage, and improve overall safety levels.

With advancements in visual processing technology, significant breakthroughs have been achieved in fire monitoring and alarm systems through smoke and flame imaging [11–19]. Visual detection of smoke and flames has become mainstream due to its low error rates and high efficiency. Recent developments include several algorithms that have greatly improved the usability and timeliness of smoke detection, such as texture recognition and the extraction of smoke and flame morphological features. Deep convolutional neural networks (CNNs) have also gained popularity in smoke detection due to their low computational cost and superior feature representation capabilities, making smoke detection technologies more efficient and accurate [20].

In the field of smoke detection, numerous methods have been proposed, all aiming to reduce false positives and missed detections, thereby improving the overall accuracy of detection systems [21,22]. These methods range from basic image processing techniques to sophisticated deep learning algorithms. However, achieving optimal detection performance heavily depends on the ability to learn and extract more discriminative smoke features. These features play a crucial role in enabling algorithms to effectively differentiate smoke from other interfering factors, significantly reducing the likelihood of false alarms and missed detections.

Currently, the predominant smoke detection methods can be broadly categorized into two types: spatial feature-based methods and spatiotemporal feature-based methods. Spatial feature-based methods primarily rely on static image characteristics, such as color and texture, to identify the presence of smoke. On the other hand, spatiotemporal feature-based methods combine both temporal and spatial dynamic information by analyzing changes across video frames to capture the movement and morphological variations of smoke. Each of these approaches has its own strengths and limitations, and selecting the most appropriate detection strategy often depends on the specific application scenario to achieve the best detection outcomes.

In the field of smoke detection, spatial feature-based methods aim to accurately identify smoke by thoroughly extracting clues from individual images. These methods primarily rely on the analysis of static images and use advanced techniques to extract and represent smoke features. Some of these methods employ deep learning frameworks, such as dual deep learning frameworks [23], ARGNet [24], and hybrid deep VGG convolutional classifiers [25], which are designed to capture the complex features of smoke in images, thereby improving detection accuracy to some extent. Certain approaches enhance smoke feature representation by combining traditional hand-crafted features with deep learning features. For instance, some techniques increase the amount of supervisory information by integrating saliency maps and deep feature maps [26], or by using waveform deep neural networks to further strengthen smoke representation [27]. These strategies attempt to capture and identify smoke features as accurately as possible by analyzing images from various dimensions.

However, while these spatial feature-based methods are effective in representing high-level semantic information about smoke, they still face challenges due to the lack of spatiotemporal information. The diversity of smoke, coupled with interference from background objects in continuous image frames, can lead to issues such as missed detections and false alarms. This is particularly problematic when dealing with complex dynamic scenes, where the absence of spatiotemporal information may limit the performance of these methods, resulting in higher miss rates.

In practical applications, the losses caused by missed detections are often much greater than those from false alarms. To reduce missed detections, detection systems typically lower the threshold to increase sensitivity. However, this approach often leads to a sharp increase in false alarms, presenting a new set of challenges and trade-offs in real-world operations. Therefore, finding a balance between reducing missed detections and controlling false alarms has become a crucial research focus in the field of smoke detection.

Spatiotemporal feature-based methods identify smoke by analyzing clues in video or image sequences, generally performing better than spatial feature-based methods because they capture temporal information. This capability helps in detecting subtle or small areas of smoke and distinguishing smoke-like objects. Some approaches, such as integrating foreground modules [28] or combining CNN and LSTM networks [29], enhance the learning of smoke's temporal features. However, most methods still extract spatial features directly from image sequences without specifically emphasizing moving targets that could be smoke [23,30,31]. This limitation can lead to confusion between non-moving and moving targets, negatively impacting detection accuracy.

To address this, frame differencing is often used to identify changing regions [28,32], guiding the model to focus on moving targets. However, this approach can also introduce interference from non-smoke object movements or camera shakes, leading to entangled features that interfere with smoke detection, resulting in missed detections and false alarms. Therefore, removing these interference features is crucial for improving the accuracy of smoke detection.

To mitigate the interference caused by similar targets in smoke detection, various attention mechanisms have been extensively studied. These mechanisms enhance the model's discriminative power by selectively amplifying useful features while suppressing irrelevant ones. For example, some approaches combine attention mechanisms with feature-level and decision-level fusion modules [32], implicit deep supervision mechanisms [33], and QuasiVSD technology [28], showing improvements in reducing interference and extracting spatiotemporal features. However, despite these advances, the methods still have limitations. While they successfully reduce some interference through attention mechanisms, they lack sufficient integration of inter-frame relational information, making it difficult to effectively aggregate long-term relationships and short-term dependencies. As a result, these methods remain susceptible to interference from smoke-like objects, impacting detection accuracy.

Interference during the feature fusion process is a critical issue in smoke detection. Although feature fusion aims to enhance detection accuracy by combining different types of high-level semantic features, simple methods such as fixed-weight decision-level fusion [34] or the concatenation of basic and detailed smoke information [17,32] can result in overlooking key discriminative features due to discrepancies between different feature maps. This interference is particularly problematic in complex scenes. Consequently, more advanced fusion strategies have been developed to dynamically adjust fusion weights or adaptively select the most discriminative features, thereby reducing interference and improving the model's detection capability.

This study proposes a smoke detection system based on the real-time detection Transformer (RT-DETR) framework, which significantly enhances the robustness of smoke detection under low resolution, real-time processing constraints, and diverse environmental conditions through the introduction of the triplet attention module, ADown module, and high-level screening-feature fusion pyramid network (HS-FPN). The triplet attention module allows for extracting subtle smoke features, the ADown module reduces computational complexity to achieve real-time performance, and HS-FPN enhances system robustness through multi-scale feature fusion. These innovations help address the many limitations of current smoke detection technologies and provide effective technical support for practical applications such as forest fire monitoring, industrial safety, and urban surveillance.

The structure of this paper is as follows: The first section presents the current state and technical background of smoke detection, the second section details the proposed system architecture and key modules, the third section describes the experimental design and results analysis, and finally, the paper concludes with a summary and an outlook on future work.

2. Related Work

In the field of smoke detection, traditional methods initially relied heavily on static image processing techniques, focusing on features such as color, texture, and shape. These early approaches used basic algorithms to detect smoke based on predefined characteristics in individual images. For instance, some studies utilized color-based segmentation and edge detection to identify potential smoke regions [35,36]. While these methods were straightforward and easy to implement, they often struggled to maintain accuracy in complex environments and were prone to high false alarm rates due to the static nature of the analysis, which failed to account for the dynamic properties of smoke in real-world scenarios [37].

With advancements in computer vision, more sophisticated spatial feature-based smoke detection methods emerged, leveraging deep learning techniques to improve detection accuracy. These methods aimed to extract high-level semantic information from static images using convolutional neural networks (CNNs). For example, dual deep learning frameworks [38], ARGNet [24], and hybrid deep VGG convolutional classifiers [39] were developed to better capture the complex features of smoke in images. By integrating traditional hand-crafted features with deep learning outputs, these approaches enhanced the robustness of detection models. However, despite their effectiveness in static scenes, these methods often encountered challenges in dynamic environments, where the absence of temporal information led to increased rates of false positives and missed detections [40].

To overcome the limitations of spatial-only methods, researchers introduced spatiotemporal feature-based approaches. These methods analyze both spatial and temporal characteristics of smoke by examining changes across video frames, thereby capturing the movement and morphological variations of smoke [41,42]. Techniques combining CNNs with long short-term memory (LSTM) networks [43] and other temporal modeling frameworks have shown promise in improving detection performance in dynamic scenes. For instance, methods like QuasiVSD [28] employed frame differencing and motion analysis to focus on moving targets likely to be smoke. While these spatiotemporal methods generally outperformed their spatial counterparts, they still faced difficulties in accurately distinguishing between smoke and similar-looking objects, particularly in complex and cluttered scenes.

Recent research has explored the use of attention mechanisms to further enhance the discriminative power of smoke detection models. Attention mechanisms enable models to selectively amplify relevant features while suppressing irrelevant ones, thereby improving overall accuracy [44,45]. Several studies have integrated attention modules with feature-level and decision-level fusion strategies [46] or employed technologies like QuasiVSD [28] to reduce interference and improve spatiotemporal feature extraction. These methods demonstrated significant improvements in handling complex scenarios, but challenges remained in effectively integrating long-term relational information across frames, which is crucial for detecting subtle smoke patterns in challenging environments.

A critical issue in smoke detection lies in the process of feature fusion. Although combining different types of high-level semantic features can enhance detection accuracy, simple fusion methods, such as fixed-weight decision-level fusion [47] or basic concatenation [48], can lead to the overlooking of key discriminative features. This is particularly problematic in complex scenes where discrepancies between different feature maps can introduce noise, reducing the effectiveness of the detection system. To address this, more advanced fusion strategies have been developed, dynamically adjusting fusion weights or adaptively selecting the most relevant features to improve detection capability and reduce interference.

Building on the limitations of these previous methods, this study proposes a novel smoke detection system based on the real-time detection Transformer (RT-DETR) architecture. By integrating the triplet attention module, the ADown module, and the high-level screening-feature fusion pyramid network (HS-FPN), the proposed model significantly enhances both the accuracy and efficiency of smoke detection in various environmental conditions. This work addresses the shortcomings of existing approaches by providing a balanced solution that combines the strengths of spatial and spatiotemporal features while also incorporating advanced attention mechanisms and dynamic feature fusion strategies to improve overall detection performance.

3. Method

In recent years, the application of Transformer models in computer vision has developed rapidly, with the DETR method serving as a significant innovation that successfully brings Transformers to object detection tasks. As shown in Figure 1, compared to traditional deep learning methods, DETR makes better use of positional information of objects, demonstrating higher adaptability, especially in handling complex tasks such as smoke detection.

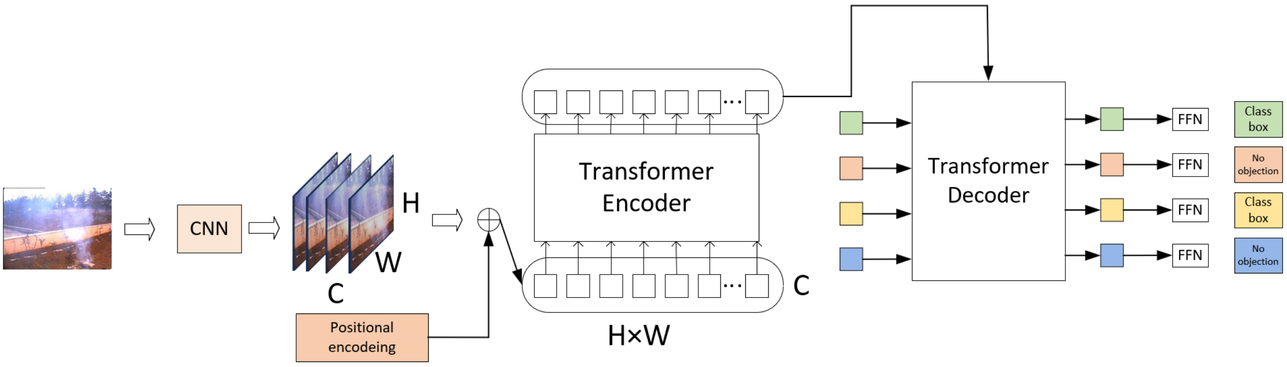


Figure 1. Architecture of the DETR.

This paper proposes a smoke detection system based on deep learning and DETR, aiming to enhance the accuracy of smoke detection in surveillance videos. The system utilizes the advantages of the DETR model, effectively detecting subtle or thin smoke and overcoming the limitations of traditional methods. In the system architecture section, this paper will discuss the network model, detection algorithm, and related design, showcasing its superiority in practical applications. The proposed system architecture is shown in Figure 2.

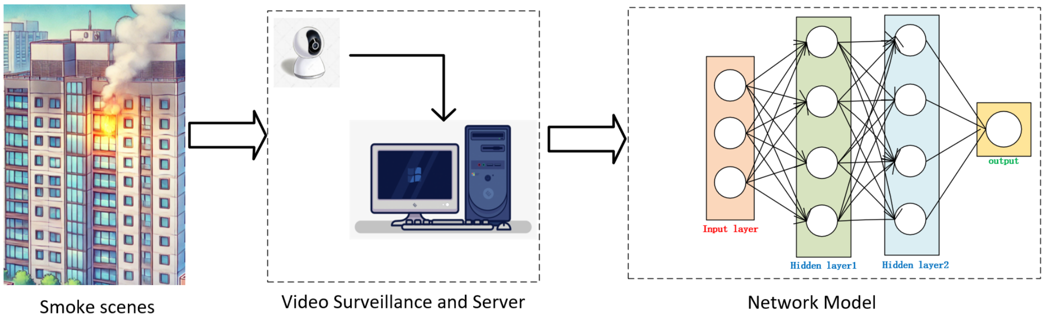


Figure 2. The architecture of the system.

3.1. The Proposed Architecture for Smoke Detection

The proposed smoke detection system is based on the real-time detection Transformer (RT-DETR) framework, integrating the triplet attention module, ADown module, and high-level screening-feature fusion pyramid network (HS-FPN). The triplet attention module enhances sensitivity to subtle smoke features, the ADown module reduces computational complexity for real-time processing, and HS-FPN improves robustness and accuracy through multi-scale feature fusion. This architecture balances precision and efficiency, effectively detecting smoke in diverse and dynamic environments and overcoming the limitations of traditional methods. The structure of the proposed network model is shown in Figure 3.

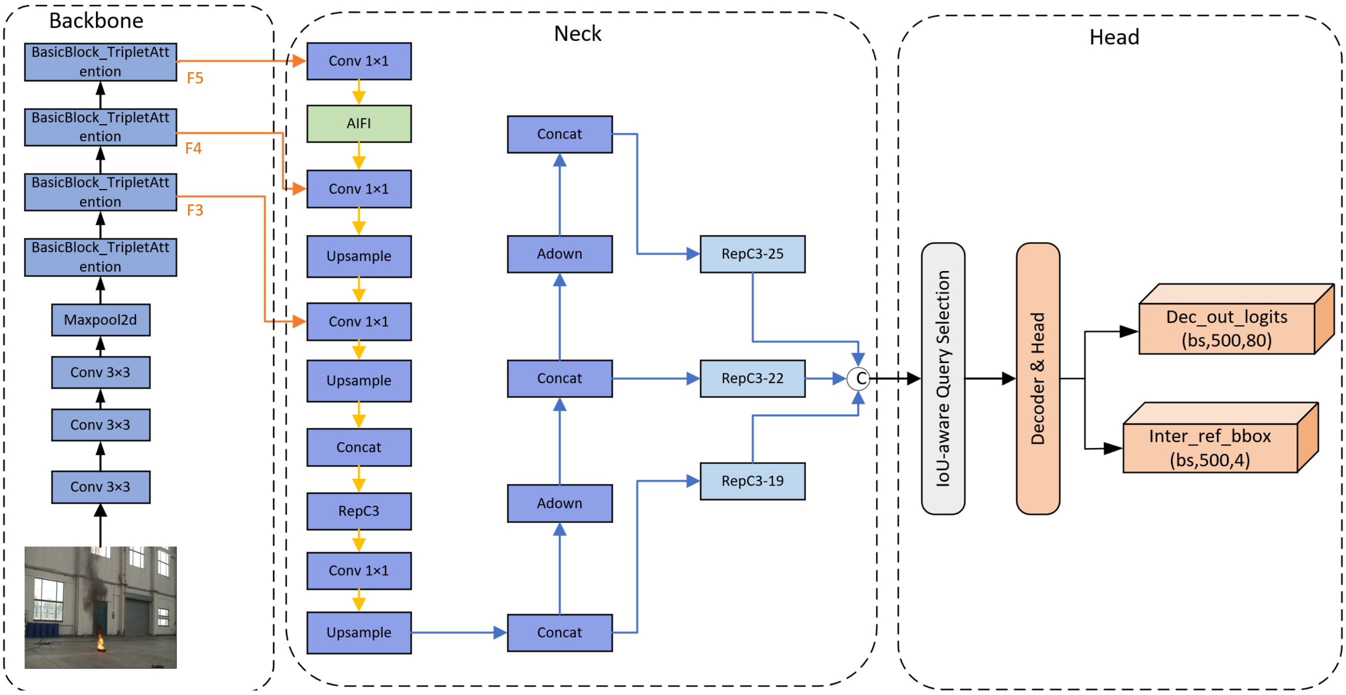


Figure 3. Architecture of the proposed network.

3.2. The Backbone Network of the Proposed Architecture

3.2.1. Triplet Attention

The integration of the triplet attention module into the RT-DETR (Real-Time Detection Transformer) backbone network brings significant improvements to fire smoke detection by enhancing feature extraction, increasing sensitivity to subtle features, and reducing background interference. This module processes attention maps across three spatial dimensions—height, width, and channel—in parallel, capturing correlations between different feature dimensions, which improves feature extraction even in low-pixel conditions. As a result, the model can accurately identify key smoke characteristics in lower-resolution inputs, ensuring high detection accuracy even in low-quality video feeds. Furthermore, the triplet attention module increases the model's sensitivity to subtle features, excelling in detecting thin or faint smoke that often appears diffuse and blurry with unclear boundaries. By focusing on fine-grained variations across each dimension, it captures progressive changes, making it particularly effective at detecting smoke that might be barely noticeable.

Additionally, the module helps the model focus on fire-related smoke features while reducing interference from irrelevant background information such as dust, lighting changes, or other smoke-like elements. This optimized attention distribution decreases the model's sensitivity to background noise and significantly reduces the false detection rate, ensuring reliable performance in complex real-world scenarios. The module is designed to be small and lightweight, enabling real-time processing even on devices with limited computational power. Triplet attention also enhances the model's generalization capability, allowing it to perform effectively in more complex situations. This ensures accurate fire smoke detection across a range of environments, providing more opportunities for experts to refine the algorithm. This feature clearly elevates the performance of RT-DETR in smoke detection.

Triplet attention consists of three parallel branches, each responsible for capturing different interactions. Some branches track the interactions between channels (C) and spatial dimensions (H or W), while the final branch functions similarly to the convolutional block attention module (CBAM) to build spatial attention. The outputs from these three branches are then averaged to produce a combined attention score. The structure of triplet attention is illustrated in Figure 4.

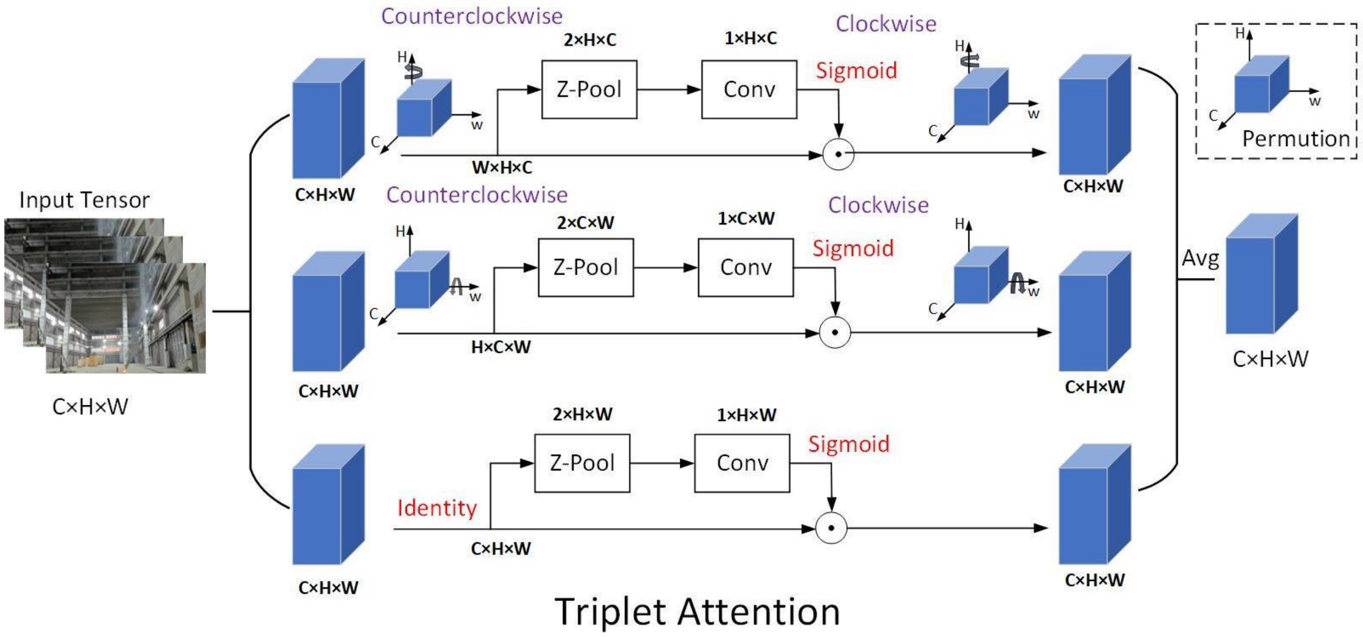


Figure 4. Abstract representation of triplet attention with three branches capturing cross-dimension interaction.

The top branch, the middle branch, and the bottom branch make up the triple attention mechanism. They all do different things. The upper branch is mostly in charge of Z-pooling, which is typically used to reduce data dimensionality or focus on significant features. Then, the input passes through a convolutional layer to extract higher-level features. It is possible to figure out the attention weights for the W space dimension and the C channel dimension. It can change how important the input features are on the fly by looking at how these factors affect each other. This gives a more accurate picture of the important data.

The triple attention mechanism consists of three branches: upper, middle, and lower, each responsible for different functions and operations. The upper branch is primarily responsible for calculating the attention weights for channel dimension C and spatial dimension W. By focusing on the interactions between these dimensions, the model can dynamically adjust the importance of the input features, thereby better representing the critical information in the data. In terms of operations, the input tensor first undergoes Z-pooling, which is typically used to reduce data dimensionality or focus on significant features. Then, the input passes through a convolutional layer to extract higher-level features. The convolution operation captures local spatial information and generates representative feature maps through filters. Finally, attention weights are generated through the Sigmoid function. These weights are applied to the original input features, adjusting the relative importance of each channel and spatial location, thereby enhancing the model's ability to recognize key features.

The main task of the middle branch is to capture the dependencies between the channel dimension C and the spatial dimensions H and W. By modeling the interactions between these dimensions, the model can better understand the relationships between different features, thereby improving overall performance. Similar to the upper branch, the middle branch also performs Z-pooling on the input to reduce dimensionality and highlight the significant parts of the features. Subsequently, the pooled features are passed into a convolutional layer for further feature extraction. This convolution operation combines channel and spatial information, enabling the network to simultaneously consider dependencies between these dimensions. Ultimately, the attention weights generated by the Sigmoid function reflect the mutual influence between the channel and spatial dimensions. These weights are then used to adjust the intensity of the input features, capturing more complex dependencies.

The lower branch is designed to capture the dependencies between the spatial dimensions H and W. Unlike the other branches, this branch focuses on modeling spatial information rather than the interactions between channels. Initially, the input features are preserved through an identity mapping operation, meaning no initial transformation is applied to the input; the original data are directly passed to the subsequent steps. Next, the input undergoes Z-pooling and convolution operations. Although the identity of the input is maintained, the combination of pooling and convolution effectively captures dependencies along the spatial dimensions, allowing the model to better understand the spatial structure of the data. Finally, the weights generated by the Sigmoid function are used to adjust the representation of input features along the spatial dimensions, enabling the model to concentrate more on important spatial regions.

After generating the corresponding attention weights for each of the three branches, the input features undergo a permutation operation to adjust the feature dimensions, making them more compatible with subsequent processing. The permuted features are then passed to the aggregation step, where the outputs of the three branches are averaged to obtain the final triplet attention output. At this stage, the model has effectively integrated information from different dimensions, providing a more comprehensive understanding of the input data.

This design of the triplet attention mechanism, by combining information interactions across channel, spatial, and dimensional aspects, more precisely captures the intrinsic features of the data. Due to its efficient computational approach, this module can be seamlessly integrated into existing network architectures, greatly enhancing the network's ability to process complex data structures. Moreover, this multi-branch attention mechanism can simultaneously consider dependencies across multiple dimensions, making the model more robust and flexible when handling high-dimensional data.

3.2.2. ADown Module

Introducing the ADown module into a smoke detection algorithm can significantly improve computational efficiency, especially in real-time applications or resource-constrained environments like edge devices or surveillance cameras. The ADown module reduces model complexity by lowering the number of parameters, speeding up inference, and enabling faster responses to fire alarms. While it reduces the resolution of feature maps, ADown retains critical details, ensuring that the model can accurately detect smoke even at lower resolutions, thus enhancing detection accuracy.

Additionally, the ADown module is adaptive and capable of automatically adjusting its parameters based on varying environmental conditions such as lighting, background complexity, and smoke patterns. This adaptability helps maintain high detection accuracy across diverse scenarios. Moreover, it helps the model better distinguish smoke from other similar interferences, reducing false alarm rates and enhancing the system's reliability and precision.

The ADown module is a convolutional block used for downsampling operations in object detection tasks. In deep learning models, downsampling is a common technique to reduce the spatial dimensions of feature maps, helping the model capture higher-level image features while reducing computational complexity. The ADown module is designed to perform this operation efficiently, with minimal impact on performance. Through its lightweight design and flexibility, it provides an effective downsampling solution for real-time object detection models. The network structure of the ADown module is illustrated in Figure 5.

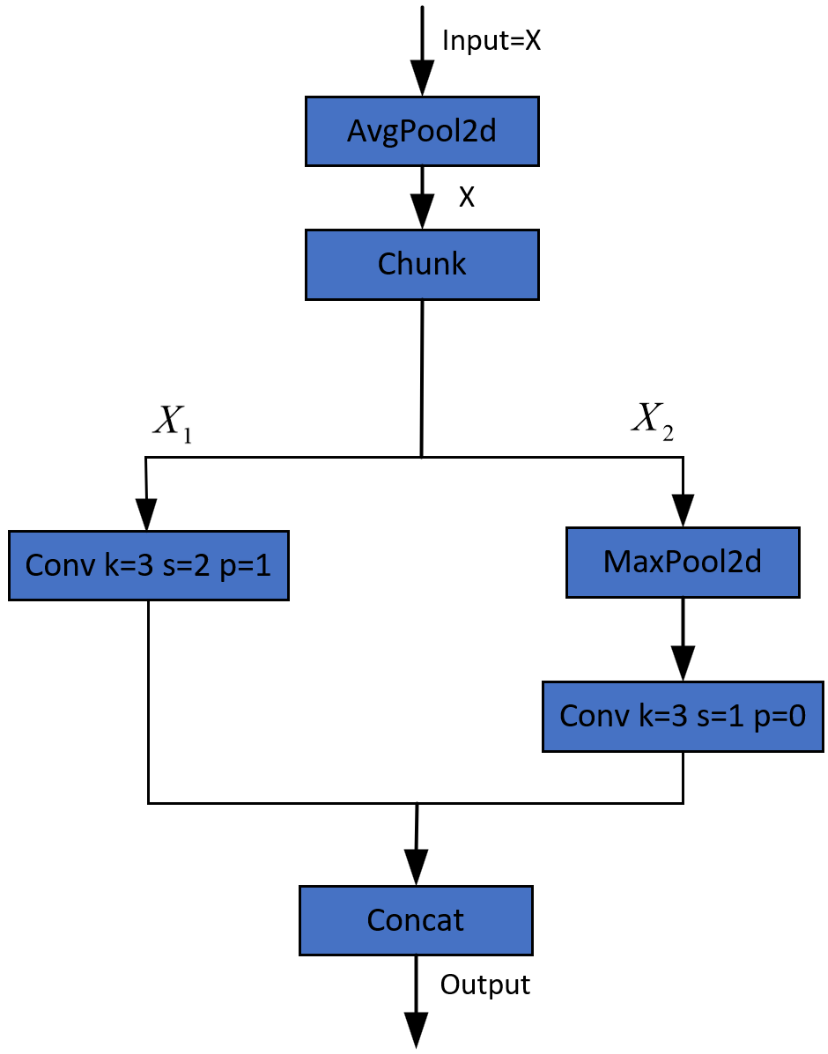


Figure 5. The diagram of the proposed triplet attention.

3.3. *The Head Network of the Proposed Architecture*

The structure of HS-FPN is illustrated in Figure 6. The HS-FPN consists of two primary components: (1) the feature selection module. (2) the feature fusion module. The Feature Selection Module, particularly the channel attention (CA), plays a significant role in smoke detection. Smoke features are often vague and have low contrast in images. The CA mechanism helps the model better recognize smoke-related features by highlighting important channels, reducing background interference, and increasing the model's sensitivity to smoke. Dimension matching (DM) ensures effective alignment of features across different scales through global average pooling and global max pooling. Since smoke can vary greatly in shape and distribution, this matching mechanism ensures that the model can effectively detect smoke across different scales.

In the feature fusion module, selective feature fusion (SFF) effectively combines low-level features (such as edges and textures) with high-level features (such as the shape and spread of smoke), enhancing the model's ability to represent different types of smoke, including white or gray smoke. By adjusting scales through bilinear interpolation or convolution, the model becomes better adapted to various smoke scales, thereby improving detection accuracy.

Multi-scale Feature Extraction and Fusion is another key part of this model. By extracting features at different scales through the Backbone (such as S2, S3, S4, and S5), the model captures rich information ranging from small local features to large-scale global features, as smoke often appears in various forms at different scales (e.g., from small smoke particles to widespread smoke). The top-down feature fusion, achieved through the SFF module, enables the combination of features from different scales, enhancing the model's robustness in detecting smoke targets of various sizes. This fusion process strengthens the detailed features, effectively capturing smoke edges or thin fog areas that are easily overlooked.

The HS-FPN structure, with its carefully designed Feature Selection and Fusion modules, ensures that the model can reliably detect smoke even in complex backgrounds, varying lighting conditions, and different types of smoke. This makes it highly valuable in real-world applications, such as fire warning systems and industrial monitoring.

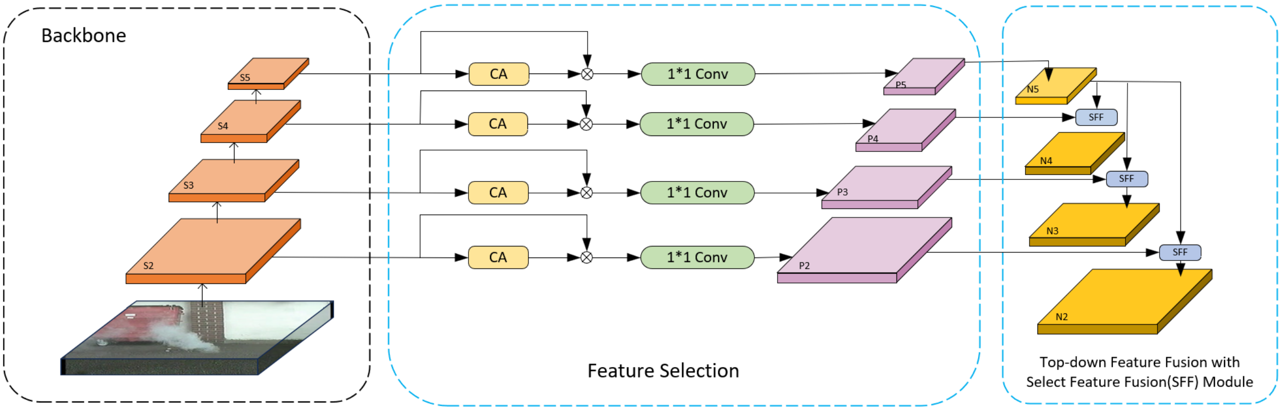


Figure 6. The framework of high-level screening-feature fusion pyramid networks comprises two parts: The feature selection module and the feature fusion module.

3.4. The Prediction Heads and the Loss Function

The final component of the model, the prediction heads, is employed to forecast the object's bounding box and to categorize the decoder's output.

3.4.1. FFN

A network model similar to DETR uses a feedforward neural network (FFN) in its prediction head, treating it as a 1 × 1 convolution, which makes the encoder function like a convolutional neural network (CNN) with enhanced attention mechanisms. The FFN is composed of a three-layer perceptron with ReLU activation functions and d hidden layer nodes, which enhances the model's feature representation capability. During the prediction phase, the FFN is responsible for predicting the bounding box parameters (center position, width, and height) for each object query, while the category prediction is generated through a Softmax activation function. This design enables precise object detection and classification.

3.4.2. The Loss Function

In smoke detection tasks, the issue of imbalanced samples also exists, with samples categorized as difficult or simple based on detection difficulty. Analyzing the scale and form of smoke, typical smoke can be considered simple samples, while subtle or thin smoke, which is harder to accurately detect and localize, is regarded as a difficult sample. In smoke detection tasks primarily involving simple samples, focusing on the bounding box regression of these simple samples can help improve detection performance. Conversely, in tasks where subtle or thin smoke is more prevalent, it is necessary to pay greater attention to the bounding box regression of difficult samples to enhance detection effectiveness. The formula is as follows:

$$IoU^{focaler} = \begin{cases} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \ll IoU \ll u \\ 1, & IoU > u \end{cases} \tag{1}$$

where $IoU^{focaler}$ is the reconstructed Focaler-IoU, IoU is the original IoU value, and [d, u] ϵ [0, 1]. By adjusting the values of d and u, we can make Focaler-IoU focus on different regression samples. Its loss is defined as follows:

$$L_{Focaler-IoU} = 1 - IoU^{focaler} \tag{2}$$

Applying the Focaler-IoU loss to the existing IoU-based bounding box regression loss function, the variants, $L_{Focaler-GIoU}$, $L_{Focaler-DIoU}$, $L_{Focaler-CIoU}$, $L_{Focaler-EIoU}$, and $L_{Focaler-SIoU}$ are as follows:

$$L_{Focaler-GIoU} = L_{GIoU} + IoU - IoU^{Focaler} \tag{3}$$

$$L_{Focaler-DIoU} = L_{DIoU} + IoU - IoU^{Focaler} \tag{4}$$

$$L_{Focaler-CIoU} = L_{CIoU} + IoU - IoU^{Focaler} \tag{5}$$

$$L_{Focaler-EIoU} = L_{EIoU} + IoU - IoU^{Focaler} \tag{6}$$

$$L_{Focaler-SIoU} = L_{SIoU} + IoU - IoU^{Focaler} \tag{7}$$

4. Experimental Results and Discussion

4.1. Dataset Acquisition and Preprocessing

4.1.1. Image Acquisition

Many people use public fire and smoke datasets right now, but they have some issues. For example, they only show one picture at a time, have simple types of smoke, and are low quality. Also, most of the pictures in the files are single frames, and most of them are smoke targets. They do not have any smoke or scenes from real life. Using online tools, smoke tests, and other methods, this study makes a smoke collection with many negative samples from different scenes and types of smoke already available to the public. There are 12,564 pictures in the file. There are 10,154 pictures of smoke and 2500 pictures of other things. It also has 44 videos of different scenes, with 10 videos of other things and 34 videos of smoke. As a negative sample, the dataset mostly gives us darkness, light, white clouds, and dark clouds. Some general cases from the dataset are shown in Figure 7.

4.1.2. Image Preprocessing

Training deep learning models typically requires a large number of data samples to mitigate overfitting and enhance generalization. By collecting and processing diverse data types, the model's recognition performance improves. In the early stages of training, appropriate data augmentation techniques can expand the quantity and diversity of samples. This study employs various data augmentation methods, including morphological operations (such as translation, saturation adjustment, angle adjustment, and image flipping) and the Mosaic augmentation technique, which involves random scaling, cropping, and stitching of four images to create diverse samples. Additionally, unified color space transformations are applied to adjust image hue, exposure, and saturation, further reducing overfitting and improving the model's generalization capability, as illustrated in Figure 8.

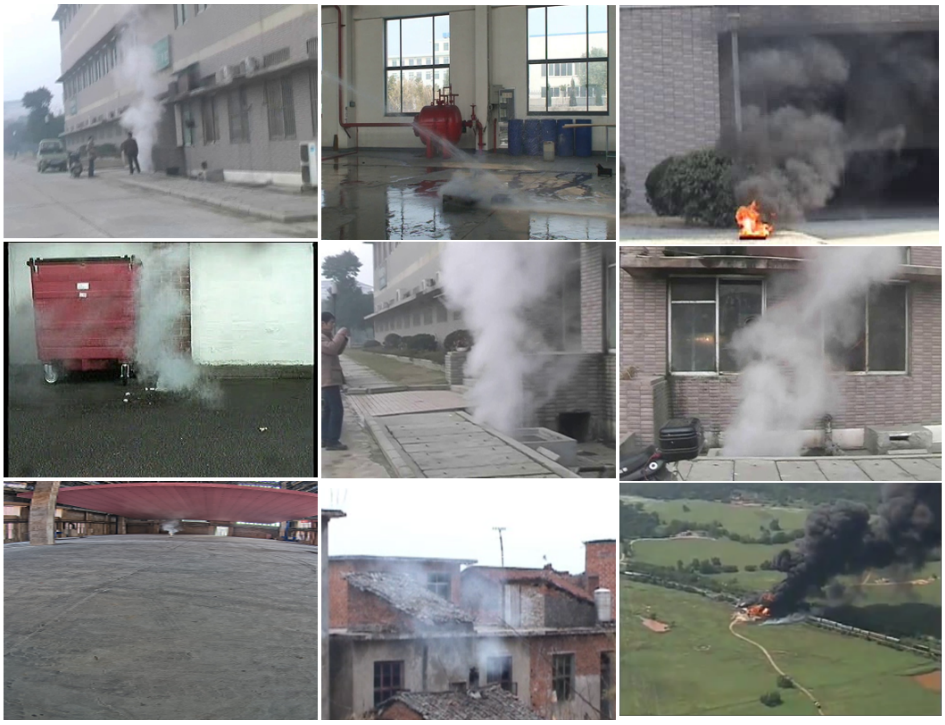


Figure 7. Typical samples of the smoke dataset.

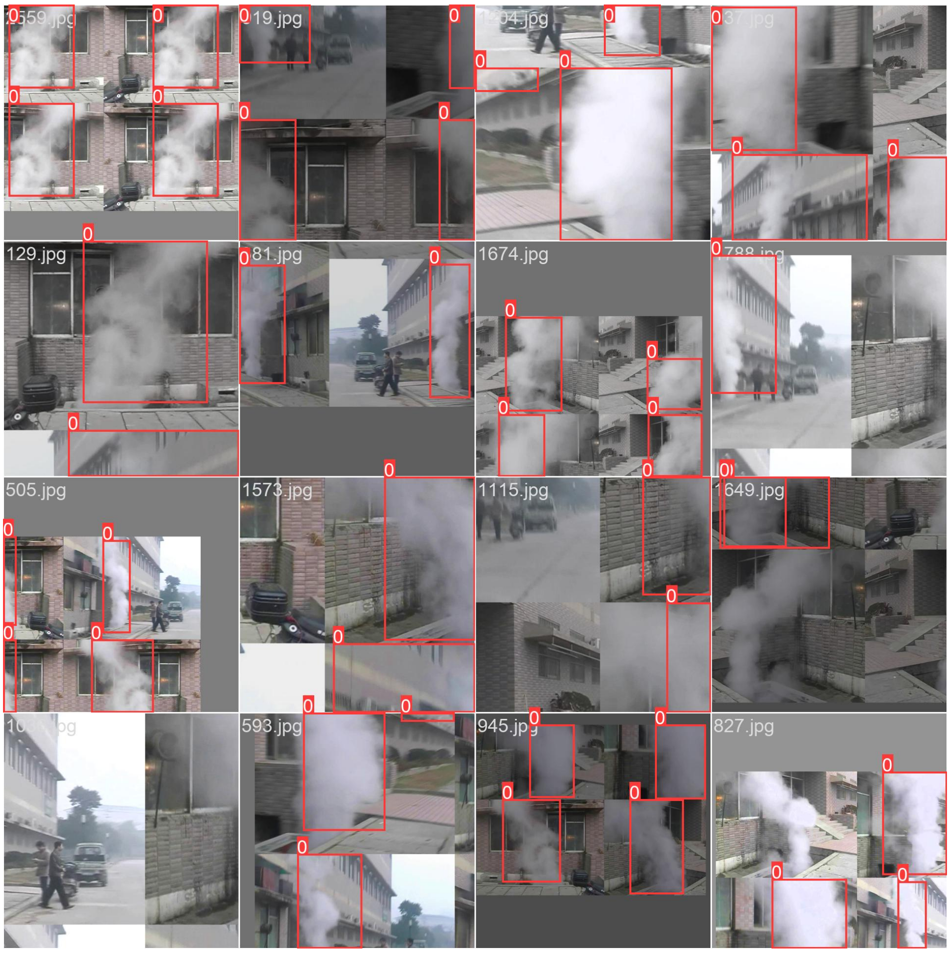


Figure 8. Data enhancement renderings.

4.1.3. Image Database and Label Database

In addition to regularization strategies, this work employed several common techniques during the training phase to prevent model overfitting. Early stopping was utilized to halt training when performance on the validation set stopped improving, thereby avoiding overfitting. Weight regularization added penalty terms to the loss function to limit model complexity.

Batch normalization was applied to normalize inputs at each layer, accelerating training and improving stability. Dropout randomly dropped neuron connections, reducing dependency among neurons and further preventing overfitting. These methods effectively enhanced the model's generalization ability and robustness.

This study utilized LabelImg software to create a dataset in VOC format, with smoke as the labeled category. The dataset was carefully divided into training, validation, and test sets, comprising a total of 10,154 samples. This division ensures that the model encounters a diverse set of samples during training, validation, and testing, thereby enhancing its generalization ability. The training set was used for model learning, the validation set for tuning, and the test set for performance evaluation. The specific distribution of the data is detailed in Table 1, while the distribution of labels is illustrated in Figure 8, ensuring the dataset's balance and effective support for model training.

Figure 9a shows where the marking is placed. When you compare the label center's ordinate to the image's height, you have ordinate y. When you compare the label center's ordinate to the image's width, you have abscissa x. Figure 9a shows that the data are spread out, with most of them in the middle. Figure 9b shows how the width of the label and the width of the picture compare. When you compare the label height to the picture height, you obtain the ordinate height. There are different sizes of data in the file, but the smaller and middle target data are the ones that are used most often.

Table 1. Experimental data.

Dataset	Smoke Image	No-Smoke Image	Total
training set	8123	2000	10,123
validation set	1054	250	1304
test set	977	250	1227
total	10,154	2500	12,564

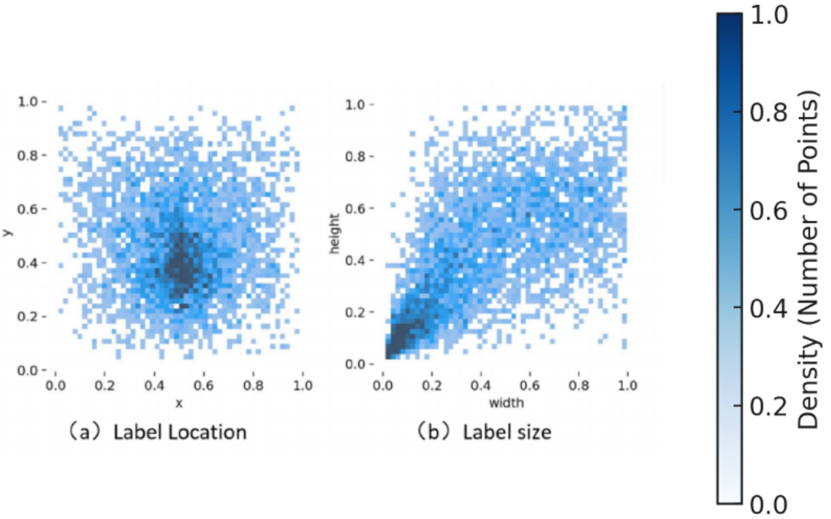


Figure 9. Label distribution.

4.2. Experimental Environment

The experiments detailed in this paper were conducted using a Windows 10 operating system, equipped with dual Intel Xeon 4215R CPUs and four NVIDIA RTX 3080 GPUs, which together deliver 1.84 teraflops of double-precision computing power. The experiments utilized PyTorch version 2.0.1, running on CUDA 11.8, within a Python 3.8 environment.

4.3. Model Evaluation Indicators

To rigorously evaluate the proposed advanced smoke detection algorithms, a comprehensive set of metrics was used. These metrics assess the algorithms' effectiveness across various scenarios, ensuring a thorough evaluation of their accuracy, robustness, and reliability in real-world conditions.

The performance evaluation of the model utilizes several key metrics, including precision, recall, F1 score, and average precision (AP). These metrics are crucial for assessing the effectiveness of the model in target detection algorithms, providing insights into its accuracy and reliability. The specific equations for calculating these metrics are also provided for reference, ensuring clarity and consistency in the evaluation process.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1scores = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{10}$$

$$AP = \int_0^1 Precision \times Recalldx \tag{11}$$

$$MAP = \frac{\sum_1^N \int_0^1 P(R)dR}{N} \tag{12}$$

In Formulas (1)–(5), true positives (TP) are samples correctly identified as particles that are indeed particles. False negatives (FNs) are samples predicted as background but are actually particles. True negatives (TNs) are samples correctly identified as background that are indeed background. False positives (FPs) are samples incorrectly predicted as particles when they are actually background.

Precision is the ratio of correctly identified targets to the total number of identified targets, including false positives. Recall measures the ratio of correctly identified targets to the total actual targets, including those missed. The F1 score combines precision and recall into a weighted average. AP_{50}, or the average precision at a 0.50 IoU threshold, evaluates the model's precision across different recall levels for detections with at least 50% overlap.

Frames per second (FPS) is an important indicator to measure the detection speed, whose formula is defined as follows:

$$FPS = \frac{1}{t} \tag{13}$$

In Equation (13), t represents the time required to process each frame of the image.

4.4. Model Training

During the training process, a 640 × 640-pixel input image size, a batch size of 16, a momentum of 0.937, a weight decay of 0.0005, and a learning rate of 0.01 are employed. To increase the diversity of the training data, several data augmentation techniques are utilized, including rotation, translation, scaling, shearing, mixup, and flipping.

4.5. Method and Effect Analysis

A close look at the chosen loss function, attention mechanism, and non-maximum suppression method showed that the model suggested in this study does work. One model that was used in all of the tests was Deformable-DETR. The important things that were used to judge the performance were FPS, precision (P), recall (R), and average precision (AP). This is because the study's goal is to make smoke recognition faster and more accurate.

Based on Table 2, the proposed model (using ResNet-18 for feature extraction) significantly outperforms Deformable-DETR (which uses ResNeXt50) in both AP50 and FPS metrics. Specifically, the proposed model achieves an AP50 of 0.961 compared to 0.798 for Deformable-DETR, and it reaches 45.4 FPS, whereas Deformable-DETR only achieves 15 FPS. This indicates that the proposed model not only improves detection accuracy but also significantly enhances inference speed, even with a lighter feature extraction network.

Table 2. Results of the ablation experiment on the feature extraction network.

Method	Feature Extraction	AP50	FPS
Deformable-DERT	ResNeXt50	0.798	15
Proposed	**ResNet-18**	**0.961**	**45.4**

In our enhanced model, we use the Focaler-IoU loss function to evaluate the effectiveness of the loss function choice and compare it with other loss functions. This study examined the performance of Focaler-IoU against GIoU, DIoU, CIoU, and SIoU. The results are presented in Table 3.

In smoke detection tasks, the table data indicate that the Focaler-IoU loss function performs better overall. Although SIoU slightly leads in precision and recall, Focaler-IoU achieves an AP50 of 0.945, significantly outperforming the other loss functions. This result suggests that Focaler-IoU has a stronger advantage in improving the accuracy of smoke detection, allowing the model to more effectively identify and locate smoke targets.

Table 3. Performances of different loss functions.

Loss Function	Precision	Recall	AP50
GIoU	0.902	0.884	0.901
DIoU	0.911	0.891	0.910
CIoU	0.934	0.911	0.912
SIoU	0.939	0.917	0.921
Focaler-IoU	**0.938**	**0.916**	**0.945**

Table 3 shows the performance metrics (precision, recall, and AP_{50}) for different loss functions, highlighting that Focaler-IoU achieves the highest. The table showcases the performance of various detection networks on key metrics such as F1 score, precision, recall, AP50, and FPS, highlighting the effectiveness of our improved method in smoke detection. Our method stands out particularly in F1 score (0.916), recall (0.916), and AP50 (0.961), demonstrating a strong balance between precision and recall, along with high accuracy across thresholds. Additionally, the FPS reached 45.4, indicating that this method is both accurate and efficient, making it suitable for real-time detection. Compared to methods like YOLOv5s and YOLOv7, our approach offers superior overall performance and stability, achieving a value of 0.945, which surpasses other loss functions. While SIoU slightly outperforms Focaler-IoU in precision (0.939 vs. 0.938) and recall (0.917 vs. 0.916), Focaler-IoU still delivers the best overall performance due to its significantly higher AP_{50}.

Figure 10 shows that the Focaler-IoU loss function performs the best throughout the entire training process. In the early stages of training, the Focaler-IoU curve descends rapidly and stabilizes around 40 epochs, indicating a faster convergence rate compared to other loss functions. Additionally, when the curves stabilize, the final loss value for Focaler-IoU is significantly lower than that of the other loss functions, which means that the model has the lowest error and higher accuracy at the end of the training. Furthermore, the Focaler-IoU curve exhibits less fluctuation throughout the training process, demonstrating greater stability. Therefore, the Focaler-IoU loss function shows outstanding performance in this experiment and is the best choice for optimizing the model.

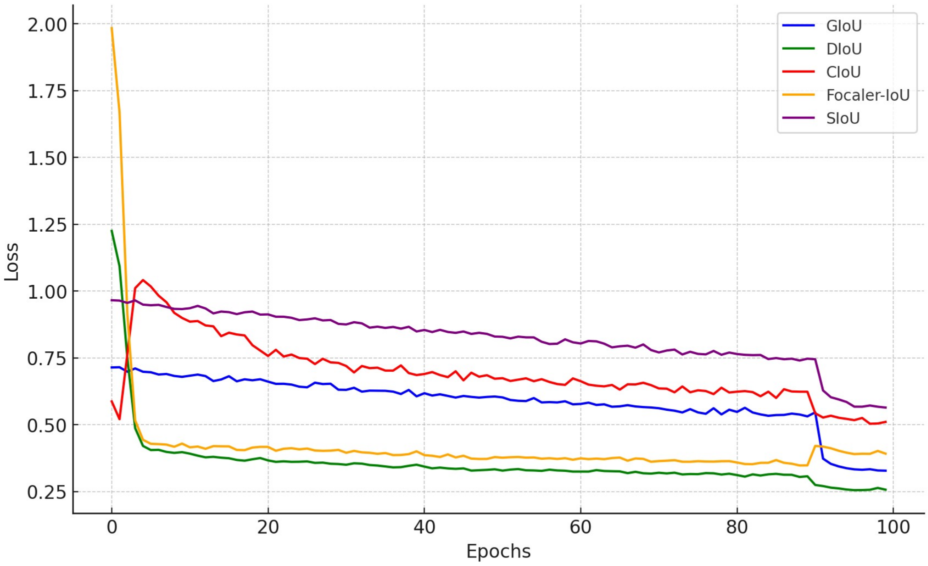
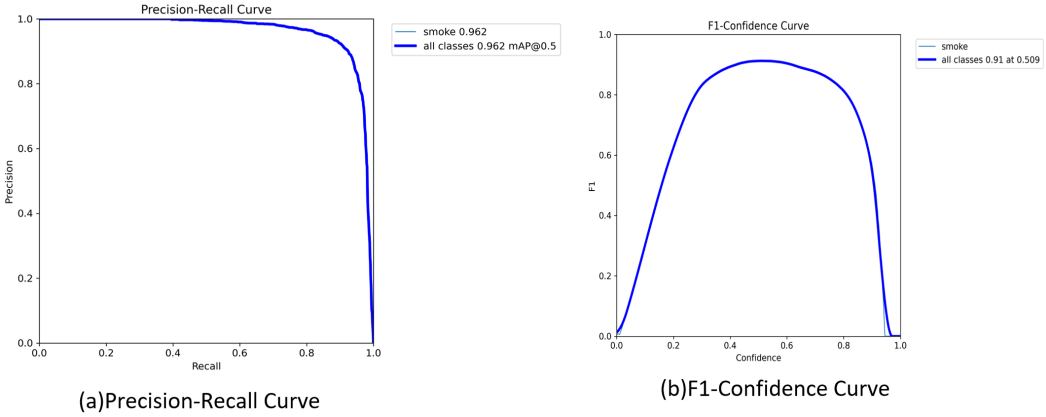


Figure 10. Training curves of different loss functions.

These two charts demonstrate the performance of the model in the task of smoke detection, revealing the relationships between different metrics and the model's stability. The precision–recall curve of the model is shown in Figure 11a, we observe that as the recall increases, precision remains at a high level, particularly when the recall approaches 1.0, at which point the precision begins to significantly decrease. This indicates that in smoke detection, the model can correctly identify smoke in most cases while maintaining a low false positive rate. The mAP@0.5 shown in the figure reaches 0.962, indicating that the model has a very strong detection capability at a threshold of 0.5, effectively distinguishing between smoke and non-smoke targets.



(a)Precision-Recall Curve

(b)F1-Confidence Curve

Figure 11. The model training results of this paper.

The F1 confidence curve in Figure 11b shows how the F1 score varies with confidence. The F1 score peaks at 0.91 when the confidence is around 0.509, meaning that at this confidence level, the model achieves the best balance between precision and recall. This is particularly important in smoke detection because it shows that the model not only can accurately detect smoke but also maintain stable detection performance at different confidence levels. By adjusting the confidence threshold according to specific needs, the model can optimize the balance between precision and recall, thereby enhancing detection accuracy and reliability.

4.6. Comparison of Different Models

We used the same experimental settings and smoke dataset to determine if the improved model proposed in this study outperforms the best existing object recognition networks. It was mostly the F1 score, precision (P), recall (R), average precision (AP), and frames per second (FPS) that were looked at. These showed how correct and useful the model was in every way. The full results of the experiment are shown in Table 4. It shows how well the better model did overall compared to the others in a number of ways.

Table 4. Test results of different detection networks.

Method	Backbone	F1 Score	Precision	Recall	AP50	FPS
SSD	VDD19	0.833	0.835	0.831	0.841	31.2
RetinaNet	ResNeXt101	0.717	0.734	0.701	0.753	-
M2Det	VGG-16	0.758	0.774	0.743	0.751	-
YOLOv4	CSPDarkNet53	0.904	0.907	0.901	0.897	12.1
YOLOv4-tiny	CSPDarkNet53_tiny	0.887	0.899	0.877	0.891	23.5
YOLOv5s	CSPDarkNet53	0.917	0.926	0.910	0.915	67.5
YOLOv7	E-ELAN	0.902	0.904	0.911	0.922	119.7
Faster R-CNN	VGG-16	0.847	0.854	0.841	0.861	-
R-FCN	ResNet-101	0.699	0.712	0.688	0.692	-
FPN	ResNet-101	0.732	0.745	0.721	0.701	-
YOLOv8s	CSPDarkNet53	0.86	0.905	0.832	0.901	44.6
YOLOv10n	CSPDarkNet53	0.89	0.903	0.938	0.89	41.7
Deformable DETR	ResNet50	0.782	0.778	0.802	0.798	15
Proposed	**ResNet-18**	**0.916**	**0.938**	**0.916**	**0.961**	**45.4**

The table showcases the performance of various detection networks on key metrics such as F1 score, precision, recall, AP_{50}, and FPS, highlighting the effectiveness of our improved method in smoke detection. Our method stands out with the F1 score (0.916), recall (0.916), and AP_{50} (0.961), demonstrating a strong balance between precision and recall, along with high accuracy across thresholds. Additionally, the FPS reached 45.4, indicating that this method is both accurate and efficient, making it suitable for real-time detection. Compared to methods like YOLOv5s and YOLOv7, our approach offers superior overall performance and stability.

4.7. Ablation Experiment

To further investigate the impact of different improvements on the performance of the detection algorithm presented in this paper, we conducted experimental validation for each enhancement to the RT-DETR network structure, using RT-DETR as the base algorithm. The ablation experiments, with results presented in Table 5, were performed on the self-built dataset developed specifically for this study.

Table 5. Results of ablation experiments.

Model	Precision	Δ	FPS	Δ	Params	Δ
RT-DETR	0.894	0	119.7	0	36.71	0
+Triple Attention	0.935	+0.001	110.6	−9.1	35.31	−1.4
+ADown	0.908	+0.004	114.4	−5.3	36.72	+0.01
+HS-FPN	0.909	+0.005	119.4	−0.5	36.71	0
+Focaler-IoU	0.907	+0.003	106.2	−12.8	36.71	0
+HS-FPN+ADown+Focaler-IoU	0.915	+0.012	104.9	−14.8	36.41	−1.3
+Triple Attention+ADown+HS-FPN	0.901	+0.017	104.7	−15.0	36.41	−1.3
+Triple Attention+ADown+HS-FPN+Focaler-IoU	0.938	+0.005	108.2	−15.0	36.15	−1.3

Table 5 presents the results of various modifications to the RT-DETR model, focusing on changes in precision, frames per second (FPS), and parameters (Params). The baseline RT-DETR model has a precision of 0.894, an FPS of 119.7, and a parameter count of 36.71 M. After introducing the triple attention module, precision increased to 0.935, with a 9.1 FPS drop. The ADown module raised precision to 0.908, with FPS dropping to 114.4, while the HS-FPN module slightly improved precision to 0.909 with minimal impact on FPS and parameter count. The Focaler-IoU module further increased precision but caused a significant drop in FPS. Ultimately, combining Triple Attention, ADown, HS-FPN, and Focaler-IoU resulted in the highest precision of 0.938, with FPS dropping to 108.2. This significant increase in precision demonstrates the effectiveness of the improvements, enhancing the model's predictive capability.

4.8. Comparison of Visualization Results

4.8.1. Small Target Smoke Detection Results

The experimental detection results are shown in Figure 12. The small target experimental scenarios selected in this paper contain outdoor open scenarios and indoor large-space laboratories. The small targets in the experimental sample pictures contain small target smoke under different light conditions.

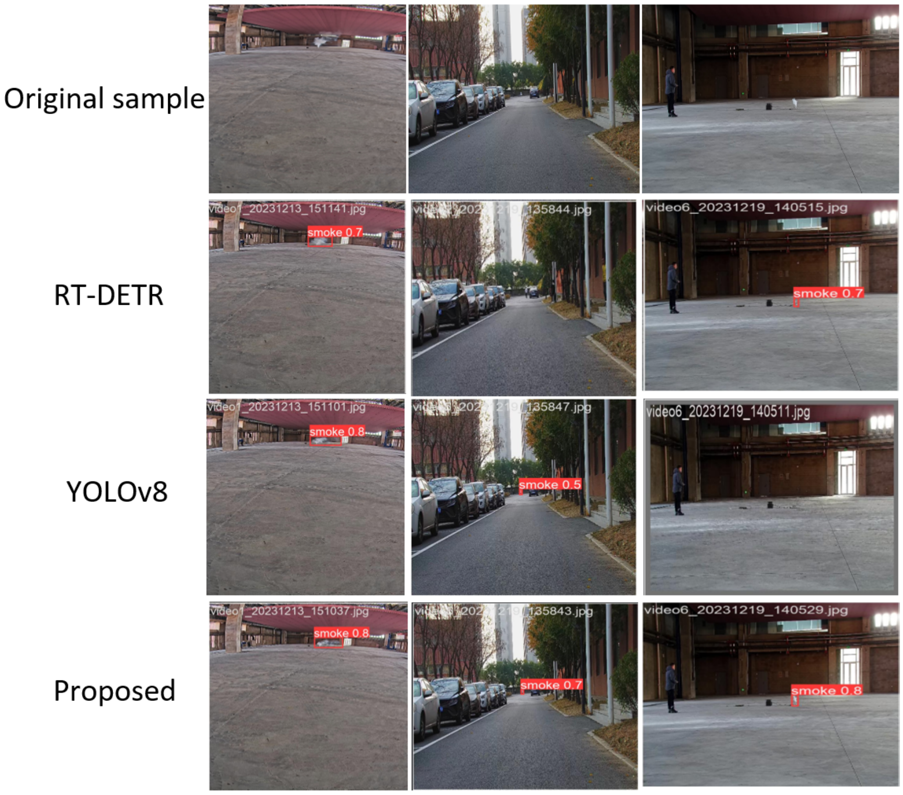

Original sample
RT-DETR
video1_20231213_151141.jpg
smoke 0.7
135844.jpg
video6_20231219_140515.jpg
smoke 0.7
YOLOv8
video1_20231213_151101.jpg
smoke 0.8
135847.jpg
smoke 0.5
video6_20231219_140511.jpg
Proposed
video1_20231213_151037.jpg
smoke 0.8
135843.jpg
smoke 0.7
video6_20231219_140529.jpg
smoke 0.8

Figure 12. Small target detection results.

Figure 12 shows that the proposed method clearly outperforms the existing RT-DETR and YOLOv8 methods. In all the tested scenarios, the proposed method demonstrates a higher confidence level in smoke detection, with more accurately positioned detection boxes. For example, in the outdoor scene, the proposed method not only detects the smoke accurately but also shows a considerably higher confidence score compared to YOLOv8 and greater stability in confidence levels than RT-DETR. In the indoor scene, the proposed method also exhibits superior detection precision, with the labeled smoke areas aligning more closely with the actual conditions.

These results indicate that the proposed method consistently outperforms in various scenarios, both indoors and outdoors. This stable superiority demonstrates that the method can effectively enhance detection accuracy and reliability in smoke detection tasks, providing a significant advantage over the existing RT-DETR and YOLOv8 methods.

4.8.2. Smoke Detection Results in Complex Backgrounds

The detection results are shown in Figure 13. The complex backgrounds selected in this study include building structures in urban environments, diverse lighting conditions, and dense background details, all of which interfere with the accuracy of smoke detection. Additionally, the diversity in smoke patterns and the similarity of background elements increase the difficulty for detection algorithms to distinguish smoke from the background.

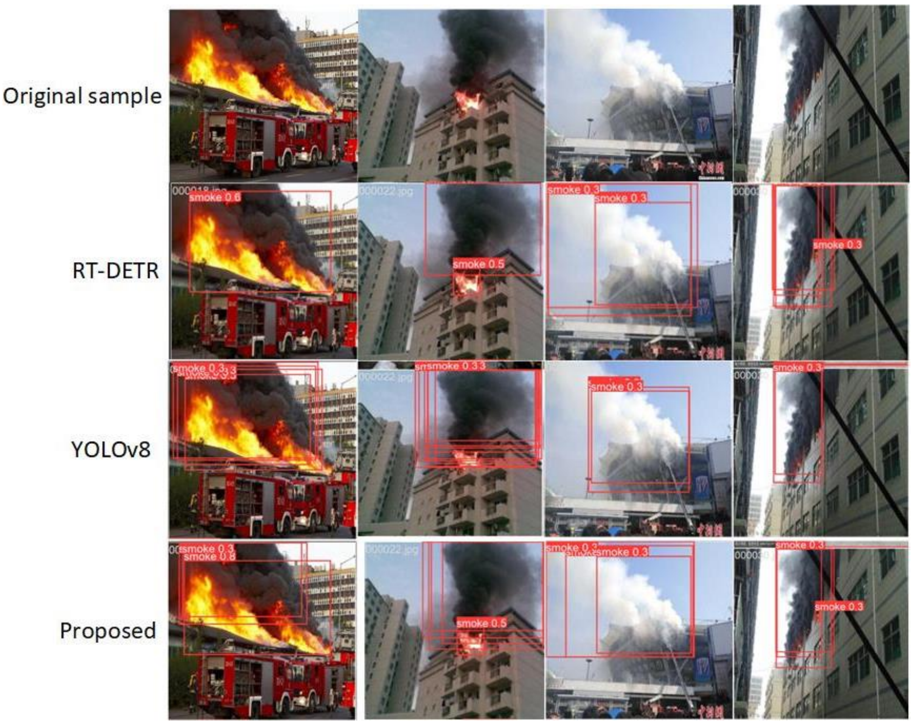


Figure 13. Complex context.

The original samples are located in the top four rows of the image, showing various smoke scenarios without any detection algorithms applied. These images are used for comparison with the detection results of different algorithms. The detection results of the RT-DETR method (in the following four rows) show that this method successfully detects some smoke areas, but the confidence scores are relatively low (around 0.3–0.5), and the number of detection boxes is moderate. This indicates that while the RT-DETR method can detect major smoke areas, it may miss some in more complex scenes. The detection results of the YOLOv8 method, shown in the next four rows, indicate that this method detects a large number of smoke areas, but many detection boxes have low confidence scores (around 0.3). The higher number of detection boxes suggests that YOLOv8 might have an over-detection issue in some scenarios, leading to more false positives. The bottom four rows display the detection results of the improved method (proposed). In comparison, the proposed method shows a moderate number of detection boxes with higher confidence scores (such as 0.5 and 0.8). This suggests that the method is more accurate in identifying smoke areas, reducing false positives, and is more sensitive to small smoke areas. Through these results, it is evident that the proposed method performs better in terms of accuracy and reliability.

The proposed method demonstrates superior performance in smoke detection compared to RT-DETR and YOLOv8, particularly in terms of accuracy and reliability. It achieves better detection accuracy and stability by maintaining a balanced number of detection boxes with higher confidence scores, effectively reducing both false positives and false negatives, even in complex and low-resolution scenarios. This improved performance is achieved through the integration of the triplet attention, ADown modules, and HS-FPN, which collectively enhance the model's ability to handle various smoke detection challenges.

5. Conclusions

Our study introduces a new smoke alarm system based on the real-time detection Transformer (RT-DETR) design, incorporating triplet attention, ADown modules, and the high-level screening-feature fusion pyramid network (HS-FPN). Our method significantly improves the accuracy and speed of smoke detection in security video streams, overcoming challenges associated with low-resolution images, the need for real-time processing, and difficult environmental conditions.

The integration of the triplet attention module enables the model to effectively capture subtle smoke features, particularly in scenarios with thin or dispersed smoke. The ADown module successfully reduces computational complexity, allowing the model to operate efficiently on resource-constrained devices without sacrificing detection performance. Additionally, the HS-FPN module enhances the model's robustness by effectively merging multi-scale features, which is critical for accurate smoke detection across varying sizes and types of smoke.

Our experiments, conducted on a custom-built dataset comprising diverse smoke scenarios, demonstrated that the proposed model outperforms existing state-of-the-art detection networks in key metrics such as average precision (AP_{50}) and frames per second (FPS). The ablation studies further confirmed the importance of each module in achieving a balanced trade-off between accuracy and computational efficiency.

In conclusion, the proposed RT-DETR-based smoke detection system not only achieves superior detection accuracy but also meets the stringent real-time processing demands of practical applications. The system's adaptability and high performance make it an ideal solution for deployment in various environments, including forest fire monitoring, industrial safety, and urban surveillance. Future work may focus on extending the model's capabilities to other related detection tasks and further enhancing its performance in more challenging conditions.

Author Contributions: Conceptualization, L.Y.; methodology, L.Y., Y.C. and B.L.; software, L.Y., Y.C., B.L. and X.L.; validation, B.L. and X.L.; formal analysis, F.X.; investigation, L.Y.; resources, Y.C.; data curation, L.Y.; writing—original draft, L.Y., Y.C. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Liaoning Provincial Science and Technology Plan Project (Project No. 2023JH2/101300205) and the Shenyang Science and Technology Plan Project (Project No. 23-407-3-33).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DETR	detection Transformer
RT-DETR	real-time detection Transformer
YOLO	you only look once
SSD	single-shot multibox detector
M2Det	multi-scale deformable attention network
Faster R-CNN	faster region-based convolutional neural network
R-FCN	region-based fully convolutional network
FPN	feature pyramid network
VDD	video detection and descriptor
VGG	visual geometry group
CSPDarkNet53	cross-stage partial network DarkNet-53
E-ELAN	efficient layer aggregation network
HS-FPN	high-level screening-feature fusion pyramid network

References

1. Appana, D.K.; Islam, R.; Khan, S.A.; Kim, J.M. A video-based smoke detection using smoke flow pattern and spatial-temporal energy analyses for alarm systems. *Inf. Sci.* **2017**, *418*, 91–101. [CrossRef]
2. Muhammad, K.; Khan, S.; Elhoseny, M.; Ahmed, S.H.; Baik, S.W. Efficient fire detection for uncertain surveillance environment. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3113–3122. [CrossRef]
3. Brisley, P.M.; Lu, G.; Yan, Y.; Cornwell, S. Three-dimensional temperature measurement of combustion flames using a single monochromatic CCD camera. *IEEE Trans. Instrum. Meas.* **2005**, *54*, 1417–1421. [CrossRef]
4. Ding, H.; Li, W.; Qiao, J. A self-organizing recurrent fuzzy neural network based on multivariate time series analysis. *Neural Comput. Appl.* **2021**, *33*, 5089–5109. [CrossRef]
5. Qiu, T.; Yan, Y.; Lu, G. An autoadaptive edge-detection algorithm for flame and fire image processing. *IEEE Trans. Instrum. Meas.* **2011**, *61*, 1486–1493. [CrossRef]
6. Aleksic, Z.J. The analysis of the transmission-type optical smoke detector threshold sensitivity to the high rate temperature variations. *IEEE Trans. Instrum. Meas.* **2004**, *53*, 80–85. [CrossRef]
7. Derbel, F. Modeling fire detector signals by means of system identification techniques. *IEEE Trans. Instrum. Meas.* **2001**, *50*, 1815–1821. [CrossRef]
8. Gu, K.; Xia, Z.; Qiao, J. Stacked selective ensemble for PM 2.5 forecast. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 660–671. [CrossRef]
9. Chen, X.; An, Q.; Yu, K.; Ban, Y. A novel fire identification algorithm based on improved color segmentation and enhanced feature data. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–15. [CrossRef]
10. Cao, Y.; Yang, F.; Tang, Q.; Lu, X. An attention enhanced bidirectional LSTM for early forest fire smoke recognition. *IEEE Access* **2019**, *7*, 154732–154742. [CrossRef]
11. Muhammad, K.; Khan, S.; Palade, V.; Mehmood, I.; De Albuquerque, V.H.C. Edge intelligence-assisted smoke detection in foggy surveillance environments. *IEEE Trans. Ind. Inform.* **2019**, *16*, 1067–1075. [CrossRef]
12. Muhammad, K.; Hamza, R.; Ahmad, J.; Lloret, J.; Wang, H.; Baik, S.W. Secure surveillance framework for IoT systems using probabilistic image encryption. *IEEE Trans. Ind. Inform.* **2018**, *14*, 3679–3689. [CrossRef]
13. Filonenko, A.; Hernández, D.C.; Jo, K.H. Fast smoke detection for video surveillance using CUDA. *IEEE Trans. Ind. Inform.* **2017**, *14*, 725–733. [CrossRef]
14. Gotthans, J.; Gotthans, T.; Marsalek, R. Deep convolutional neural network for fire detection. In Proceedings of the 2020 30th International Conference Radioelektronika (RADIOELEKTRONIKA), Bratislava, Slovakia, 15–16 April 2020; pp. 1–6.
15. Muhammad, K.; Ahmad, J.; Baik, S.W. Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing* **2018**, *288*, 30–42. [CrossRef]
16. 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]
17. Gu, K.; Xia, Z.; Qiao, J.; Lin, W. Deep dual-channel neural network for image-based smoke detection. *IEEE Trans. Multimed.* **2019**, *22*, 311–323. [CrossRef]
18. Jiang, L.; Qi, Q.; Zhang, A.; Guo, C.; Cheng, X. Improving the accuracy of image-based forest fire recognition and spatial positioning. *Sci. China Technol. Sci.* **2010**, *53*, 184–190. [CrossRef]
19. Liu, Z.; Du, J.; Wang, M.; Ge, S.S. ADCM: Attention dropout convolutional module. *Neurocomputing* **2020**, *394*, 95–104. [CrossRef]
20. Liu, D.; Kruggel, F.; Sun, L. Elastography mapped by deep convolutional neural networks. *Sci. China Technol. Sci.* **2021**, *64*, 1567–1574. [CrossRef]
21. Peng, Y.; Wang, Y. Real-time forest smoke detection using hand-designed features and deep learning. *Comput. Electron. Agric.* **2019**, *167*, 105029. [CrossRef]
22. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3186–3195.
23. Pundir, A.S.; Raman, B. Dual deep learning model for image based smoke detection. *Fire Technol.* **2019**, *55*, 2419–2442. [CrossRef]

24. Zhan, J.; Hu, Y.; Zhou, G.; Wang, Y.; Cai, W.; Li, L. A high-precision forest fire smoke detection approach based on ARGNet. *Comput. Electron. Agric.* **2022**, *196*, 106874. [CrossRef]
25. Matlani, P.; Shrivastava, M. Hybrid deep VGG-NET convolutional classifier for video smoke detection. *Comput. Model. Eng. Sci.* **2019**, *119*, 427–458. [CrossRef]
26. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, Z.; Jia, Y.; Wang, J. Video smoke detection based on deep saliency network. *Fire Saf. J.* **2019**, *105*, 277–285. [CrossRef]
27. Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A wave-shaped deep neural network for smoke density estimation. *IEEE Trans. Image Process.* **2019**, *29*, 2301–2313. [CrossRef]
28. Cao, Y.; Tang, Q.; Xu, S.; Li, F.; Lu, X. QuasiVSD: Efficient dual-frame smoke detection. *Neural Comput. Appl.* **2022**, *34*, 8539–8550. [CrossRef]
29. Cao, Y.; Lu, X. Learning spatial-temporal representation for smoke vehicle detection. *Multimed. Tools Appl.* **2019**, *78*, 27871–27889. [CrossRef]
30. Valikhujaev, Y.; Abdusalomov, A.; Cho, Y.I. Automatic fire and smoke detection method for surveillance systems based on dilated CNNs. *Atmosphere* **2020**, *11*, 1241. [CrossRef]
31. Muhammad, K.; Ullah, H.; Khan, S.; Hijji, M.; Lloret, J. Efficient fire segmentation for internet-of-things-assisted intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 13141–13150. [CrossRef]
32. Hu, Y.; Lu, X. Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features. *Multimed. Tools Appl.* **2018**, *77*, 29283–29301. [CrossRef]
33. Li, X.; Chen, Z.; Wu, Q.J.; Liu, C. 3D parallel fully convolutional networks for real-time video wildfire smoke detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *30*, 89–103. [CrossRef]
34. He, L.; Gong, X.; Zhang, S.; Wang, L.; Li, F. Efficient attention based deep fusion CNN for smoke detection in fog environment. *Neurocomputing* **2021**, *434*, 224–238. [CrossRef]
35. Chaturvedi, S.; Khanna, P.; Ojha, A. A survey on vision-based outdoor smoke detection techniques for environmental safety. *ISPRS J. Photogramm. Remote Sens.* **2022**, *185*, 158–187. [CrossRef]
36. Chen, X.; Hou, Q.; Fu, Y.; Zhu, Y. A lightweight multiscale smoke segmentation algorithm based on improved DeepLabV3+. *IET Image Process.* **2024**. [CrossRef]
37. Kim, J.M. Reliable Smoke Detection using Static and Dynamic Textures of Smoke Images. *J. Korea Contents Assoc.* **2012**, *12*, 10–18. [CrossRef]
38. Dalal, S.; Lilhore, U.K.; Radulescu, M.; Simaiya, S.; Jaglan, V.; Sharma, A. A hybrid LBP-CNN with YOLO-v5-based fire and smoke detection model in various environmental conditions for environmental sustainability in smart city. *Environ. Sci. Pollut. Res.* **2024**, 1–18. [CrossRef]
39. Yin, H.; Wei, Y.; Liu, H.; Liu, S.; Liu, C.; Gao, Y. Deep convolutional generative adversarial network and convolutional neural network for smoke detection. *Complexity* **2020**, *2020*, 6843869. [CrossRef]
40. Mutar, A.; Dway, H. Smoke detection based on image processing by using grey and transparency features. *J. Theor. Appl. Inf. Technol.* **2018**, *96*, 6995–7005.
41. Zhang, Z.; Jin, Q.; Liu, Z. Video-based fire smoke detection using temporal-spatial saliency features. *Procedia Comput. Sci.* **2022**, *198*, 493–498. [CrossRef]
42. Tao, H.; Lu, X. Smoke vehicle detection based on spatiotemporal bag-of-features and professional convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3301–3316. [CrossRef]
43. Lyu, Z.; Jia, X.; Yang, Y.; Hu, K.; Zhang, F.; Wang, G. A comprehensive investigation of LSTM-CNN deep learning model for fast detection of combustion instability. *Fuel* **2021**, *303*, 121300. [CrossRef]
44. Sun, Y.; Feng, J. Fire and smoke precise detection method based on the attention mechanism and anchor-free mechanism. *Complex Intell. Syst.* **2023**, *9*, 5185–5198. [CrossRef]
45. Li, H.; Ma, Z.; Xiong, S.H.; Sun, Q.; Chen, Z.S. Image-based fire detection using an attention mechanism and pruned dense network transfer learning. *Inf. Sci.* **2024**, *670*, 120633. [CrossRef]
46. Ranipa, K.; Zhu, W.P.; Swamy, M. A novel feature-level fusion scheme with multimodal attention CNN for heart sound classification. *Comput. Methods Programs Biomed.* **2024**, *248*, 108122. [CrossRef]
47. Shao, Y.; Ying, Y.; Chen, X.; Dong, S.; Wei, D. Multi-Scene Smoke Detection Based on Multi-Feature Extraction Method. *J. Shanghai Jiaotong Univ. (Sci.)* **2024**, 1–14. [CrossRef]
48. Li, J.; Xu, R.; Liu, Y. An improved forest fire and smoke detection model based on yolov5. *Forests* **2023**, *14*, 833. [CrossRef]

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.