IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2024 Doi Number*

# Improving YOLOv5s Algorithm for Detecting Flame and Smoke

**Li Deng[1,2,3], Jin Zhou[1], and Quanyi Liu[1,2,3]**
[1]College of Civil Aviation Safety Engineering, Civil Aviation Flight University of China, Sichuan, Guanghan, 618307, China
[2]Civil Aircraft Fire Science and Safety Engineering Key Laboratory of Sichuan Province, Guanghan 618307, China
[3]Sichuan Key Technology Engineering Research Center for All-electric Navigable Aircraft, Sichuan Guanghan, 618307, China

Corresponding authors: Li Deng (e-mail: bitdengli@163.com), Quanyi Liu (e-mail: quanyiliu2005@126.com)

Contributing author: Jin Zhou (e-mail: azhou9azhou6@163.com)

This work was supported in part by National Natural Science Foundation of China under Joint Fund of Civil Aviation Research (grant number U2033206); Aviation Science Foundation (grant number ASFC-20200046117001); Key Laboratory Project of Sichuan Province (MZ2022JB01).

**ABSTRACT** Object detection methods can be used to detect flames and smoke from images or videos for the identification and exploration of fire. In this paper, an improved YOLOv5s algorithm, called GAM-ASFF-YOLOv5s is proposed, which introduces an attention mechanism and feature fusion layer. A global attention mechanism (GAM) is introduced into the backbone network of YOLOv5s to focus on the detection information that is conducive to flames and smoke, while suppressing unimportant information. A head network with adaptive spatial feature fusion (ASFF) was designed to extract more complete image features of flame and smoke. In addition, the original object bounding box regression loss function the complete IoU (CIoU) of YOLOv5s was replaced by repulsion loss to enhance the generalization ability of the model and further improve the detection performance of the flame and smoke. The experimental results show that the precision of the GAM-ASFF-YOLOv5s+REP algorithm was 5.7% higher than that of the original YOLOv5s algorithm on the VOC2007 dataset and it also performed well on the flame and smoke dataset, that is, the precision, recall, and mean average precision (mAP) were improved by 0.5 %, 2.1 % and 2.9 % respectively, compared to the original YOLOv5s algorithm.

**INDEX TERMS** adaptive spatial feature fusion, flame and smoke detection, global attention mechanism, repulsion loss, yolov5s

## I. INTRODUCTION

The frequency of fires has increased gradually in recent years. Once a fire occurs, it severe damage society. To avoid the enormous losses caused by fire to people as much as possible, it is of great practical significance to study an accurate and fast detection algorithm for smoke and flame in the early stage of fire.

With the development of deep learning, research on object detection has progress significantly. Researchers have carried out extensive research and obtained many advanced algorithms for object detection. These algorithms widely used in medical imaging [1], vehicles [2], and geological exploration [3]. Owing to the advantages of autonomous learning and fast detection, convolutional neural networks are widely used in object detection and image recognition. Object detection algorithms include one- and two-stage detection algorithms. Compared with the two-stage detection algorithm, which has a slower detection speed, the one-stage detection algorithm has the advantage of fast detection speed, making it more suitable for real-time detection of images or videos. One-stage detection algorithms mainly include single-shot multi-box detector (SSD) algorithms [4], [5], [6], [7], [8] and You Only Look Once (YOLO) algorithms [9], [10], [11], [12], [13], [14]. Currently, CNN [15], Faster R-CNN [16] and the YOLO series of convolutional neural networks are used for fire detection.

In recent years, research on the use of convolutional neural networks for fire detection has gradually increased. Yar et al. [17] developed a fire image detection model with reduced complexity and enhanced accuracy by refining the network architecture of YOLOv5. Uddin et al. [18] evaluated the performance of the YOLOv5 and YOLOv8 models with the aim of developing high-accuracy fire warning models. Dai et al. [19] employed the MobileNet network as a replacement for the backbone network of YOLOv3, yet this modification proved inadequate for achieving the desired level of accuracy in flame detection. Qin et al. [20] integrated deep separable convolution with YOLOv3, thereby enhancing the fire detection capabilities of the model; however, this approach resulted in a detection rate of only 35 fps. Wu et al. [21] used the dual tree-complex wavelet transform as the primary means of image pre-processing to solve the problem of dynamic smoke feature extraction. In addition, this study suggests that deep-learning technology is a worthwhile method for smoke detection. Cao et al. [22] proposed an EFFNet with significant performance advantages for detecting translucent and non-rigid smoke targets. Yang et al. [23] proposed a new network framework, STENet, for smoke detection tasks in videos, which exhibited excellent performance in smoke recognition in a wide range of scenarios. Undoubtedly, EFFNet and STENet achieved excellent smoke detection results. However, flames and smoke are indispensable target categories in fire image detection, and the complexity of fire scenes poses a particular challenge in the research of fire detection algorithms. The dataset used in this study contained complex fire scenarios, such as wildfires, building fires and indoor fires and the detection targets were set in two categories: flames and smoke.

YOLOv5 includes four algorithms, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, which have similar network structures but different network depths and widths. Among them, the YOLOv5s algorithm has fewer network layers and computational complexity, making it a smaller model and faster than the other three algorithms. Therefore, we selected YOLOv5s with a faster detection speed as the baseline algorithm. However, the accuracy of YOLOv5s could be higher when performing detection tasks, and the fire detection algorithm needs to have both high accuracy and speed. Furthermore, because the shapes of the flame and smoke are different, unevenly distributed, and often stacked together, it is not easy to directly apply the YOLOv5s algorithm to flame and smoke detection.

To better cope with these problems, we adopted the following methods enhance the performance of YOLOv5s for flame and smoke detection:

- We added the Global Attention Mechanism (GAM) [24] to the backbone of the YOLOv5s network to focus on the areas that require attention in the image. The attention mechanism is a method to enhance the ability of neural networks to extract important information in the learning process and is widely used in the fields of deep learning and machine training, such as coordinate attention (CA) [25] and efficient channel attention (ECA) [26].
- We introduced adaptive spatial feature fusion (ASFF) [27] to obtain image features of different scales.
- We introduced repulsion loss [28] to improve the regression rate, thereby improving the performance of the model.
- The experimental results demonstrate that the proposed method effectively improves the model performance of YOLOv5 s and shows significant results in flame and smoke detection.

## II. Research Method

This section first introduces the YOLOv5s algorithm and then elaborates on the method of optimizing the YOLOv5s network architecture and the improved YOLOv5s. Finally, the bounding-box regression loss function is referenced in this dissertation.

### A. YOLOv5 Algorithm

The network structure of YOLOv5s is divided into three parts: the backbone network for feature extraction, the neck network for feature fusion, and the output network for object detection. The network structure of YOLOv5s is shown in Figure 1.

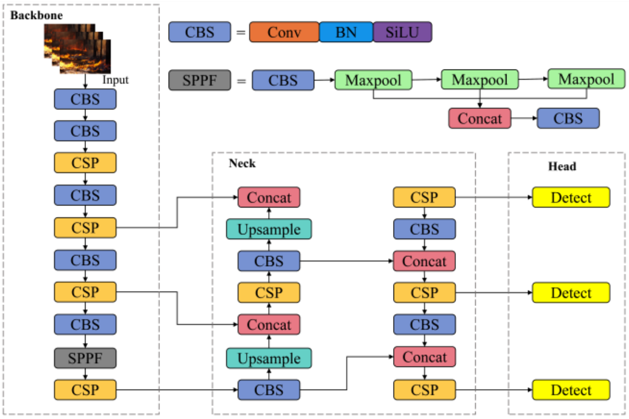


FIGURE 1. Structure of the YOLOv5s network

YOLOv5s have strong feature extraction capability and computational efficiency when using the Darknet53 network with a cross-stage partial network (CSP) [29] structure as its backbone network. The backbone network is responsible for extracting the features by first specifying the size of the input image as 640×640 and then continuing to feed images of this size to the backbone network, which outputs feature maps with sizes of 80×80, 40×40, and 20×20 through three modules.

The neck network uses feature pyramid networks (FPN) [30] and path aggregation networks (PAN) [31]. As shown in Figure 1, the left-hand side of the neck network uses up-sampling to increase the size of the feature maps, which facilitates the fusion of the feature maps from the backbone network. Subsequently, the right-hand side uses down sampling to obtain feature maps of different sizes to combine deep and shallow features to obtain complete features.

The head network detects the category and location information of the feature maps output through the backbone and neck networks. It comprises a detection module responsible for outputting three detection results of different sizes (80×80, 40×40 and 20×20).

### B. Using GAM to obtain more information

GAM is an improvement of the Convolutional Block Attention Module (CBAM) [32] that incorporates channel spatial attention and redesigns the submodule of CBAM. As shown in Figure 2, the GAM includes a channel attention submodule and spatial attention submodule. $F_1 \in R^{C \times H \times W}$ denotes the input feature map, $F_2$ denotes the intermediate state and $F_3$ denotes the output, which can be defined as:

$$F_2 = M_c(F_1) \times F_1 \tag{1}$$

$$F_3 = M_s(F_2) \times F_2 \tag{2}$$

where $M_c$ and $M_s$ represent the channel and spatial attention maps, respectively, × represents the element multiplication operation.

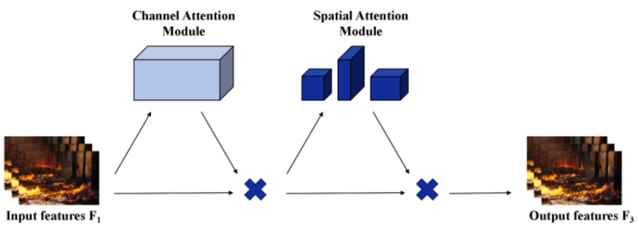


FIGURE 2. Structure of the GAM

Operation process of the channel attention submodule: The channel attention submodule uses a three-dimensional arrangement to preserve the original information. After dimension conversion, the feature map is processed by a Multi-Layer Perceptron (MLP), an encoder-decoder structure used to amplify cross-dimensional channel spatial dependence, which is converted into the original dimension and then output after activation.

The operation process of the spatial attention submodule: Through the convolution operation with a convolution kernel of 7, the number of channels is reduced, and then through the convolution operation with a convolution kernel of 7 again, the number of channels is increased to keep the number of channels consistent, and finally, the output is processed by the activation function. Because the maximum pooling operation reduces the information and produces adverse effects, the pooling operation has been deleted here further to preserve the feature map. However, the spatial attention module occasionally significantly increases the number of parameters. Group convolution with channel shuffle [33] in ResNet50 was adopted to avoid a noticeable parameter increase.

### C. Using ASFF to reduce information dispersion

To address the problem of inconsistency between flame and smoke feature maps of different sizes, the ASFF structure is integrated into the YOLOv5s head network, which can improve the detection ability of the model for flames and smoke of different sizes. Figure 3 describes the specific process of fusing the ASFF structure into the YOLOv5s head network.

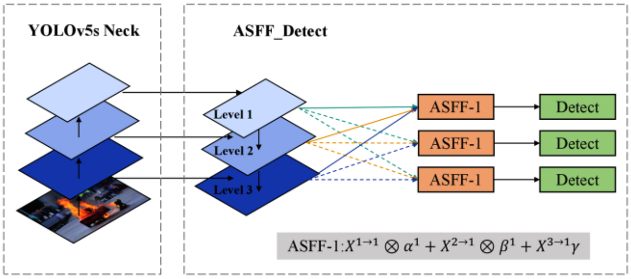

YOLOv5s Neck
ASFF_Detect
Level 1
Level 2
Level 3
ASFF-1
ASFF-1
ASFF-1
Detect
Detect
Detect
ASFF-1:$X^{1\to1} \otimes \alpha^1 + X^{2\to1} \otimes \beta^1 + X^{3\to1}\gamma$

FIGURE 3. Structure of the ASFF_Detect

First, the feature maps of level 1, 2, and 3 were obtained from the neck network structure. They are then fused with the ASFF algorithm to generate three feature maps of corresponding sizes, that is, ASFF-1, ASFF-2, and ASFF-3, and the fusion weights are adaptively adjusted so that the features of different sizes can be fully exploited. Taking ASFF-1 as an example, the feature maps of three different sizes are first re-adjusted to the size of level 1, and then the spatial fusion weights of the feature maps are learned to this size. The fusion calculation of the ASSF was performed according to the following formula:

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot X_{ij}^{1\to l} + \beta_{ij}^{l} \cdot X_{ij}^{2\to l} + \gamma_{ij}^{l} \cdot X_{ij}^{3\to l} \tag{3}$$

where $X_{ij}^{n\to l}$ denotes the feature vector of the feature map at position $(i, j)$ from layers $n$ to $l$. $y_{ij}^{l}$ refers to the new feature map output obtained after using the ASFF module. $\alpha_{ij}^{l}$, $\beta_{ij}^{l}$ and $\gamma_{ij}^{l}$ represent the spatial importance weights of the three feature maps, adaptively learned through the network, satisfying $\alpha_{ij}^{l} + \beta_{ij}^{l} + \gamma_{ij}^{l} = 1$, $\alpha_{ij}^{l}$, $\beta_{ij}^{l}$ and $\gamma_{ij}^{l} \in [0,1]$.

### D. Using Repulsion Loss to improve convergence efficiency

In a fire scene, flames and smoke are often occluded or overlapped, which can easily affect the accuracy of the target location when the model performs a detection task. For example, suppose that a target T overlaps with another target B. In this case, the detector will confuse the target because of the high similarity in the appearance of these two targets, which will cause the prediction box to lock target T to target B, and even target T will become the missing target owing to the influence of non-maximum suppression (NMS). The repulsion loss can suppress the transfer of the prediction box from target T to target B, effectively reducing false and missed detections, as shown in Figure 4, where the attraction mechanism is used to narrow the gap between the proposal and its designated object, which prevents the proposal from moving to the surrounding objects, resulting in more object-robust localization.

The three parts of the Repulsion Loss are defined as:

$$L = L_{Attr} + \alpha \cdot L_{RepGT} + \beta \cdot L_{RepBox} \tag{4}$$

where $L_{Attr}$ is an attraction mechanism to make a predicted box as close as possible to a real target box, whereas $L_{RepGT}$ and $L_{RepBox}$ are repulsion mechanisms to keep a predicted box as far away as possible from other surrounding real objects and other predicted boxes. Coefficients $\alpha$ and $\beta$ are used as weights to balance the auxiliary loss.

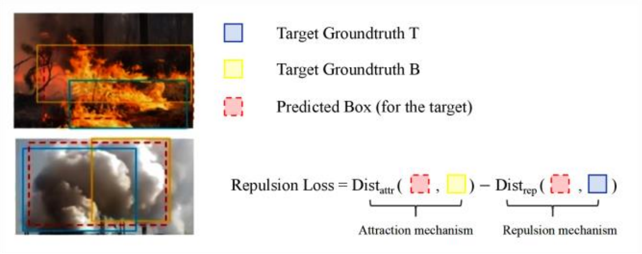

Target Groundtruth T
Target Groundtruth B
Predicted Box (for the target)
Repulsion Loss = Dist_attr ( , ) − Dist_rep ( , )
Attraction mechanism
Repulsion mechanism

**FIGURE 4. Illustration of the repulsion loss. Repulsion loss is driven by two motives: the attraction of the object and the repulsion of other surrounding targets.**

### *E. Improved YOLOv5s*

We improved the network architecture to improve the performance of YOLOv5s in detecting flames and smoke. The improved YOLOv5s network structure is illustrated in Figure 5.

The improved method is divided into:

- Adding GAM to the YOLOv5s backbone network can help focus on flame and smoke information detection while suppressing unimportant information to improve the efficiency of detection tasks.
- The head network integrates the ASFF to reduce the loss of the flame and smoke feature information.
- The boundary box regression loss function is optimized by using the repulsion loss to replace the complete IoU (CIoU) [34] to improve the ability to detect flames and smoke under occlusion and improve the model's regression rate.

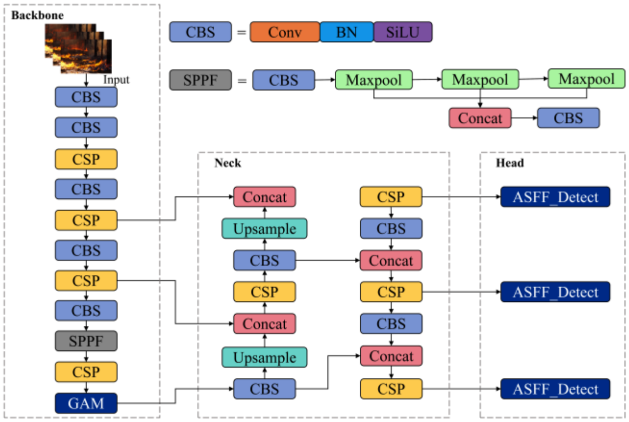


**FIGURE 5. Structure of the improved YOLOv5s network**

## III. Experimental data and processing

### *A. Image processing*

We selected a fire dataset from the public dataset website (https://aistudio.baidu.com/datasetoverview). After manual annotation and division, the fire dataset used in the experiment were formed. In this study, we used the mosaic data enhancement method to increase the generalization ability of the model, and the image enhancement effect is presented in Figure 6. From the visualization results of the data processing, it can be observed that the range of most objects is small, and the object concentration is high in most areas of the image because the image features of the flame and smoke are relatively concentrated. The most significant images in the dataset were wildfires. Images of wildfires typically exhibit a considerable number of long-distance flames and smoke, resulting in smaller flame and smoke feature sizes in the images, which is consistent with the actual situation of fire detection.

It is worth mentioning that the YOLO series algorithm is more suitable for detecting wildfire scenes, particularly for extracting flame and smoke patterns. Many fire detectors with high-speed and high-precision detection performance have appeared in the market and can be used directly in buildings. Therefore, the dataset used in this experiment contained several wildfire images.

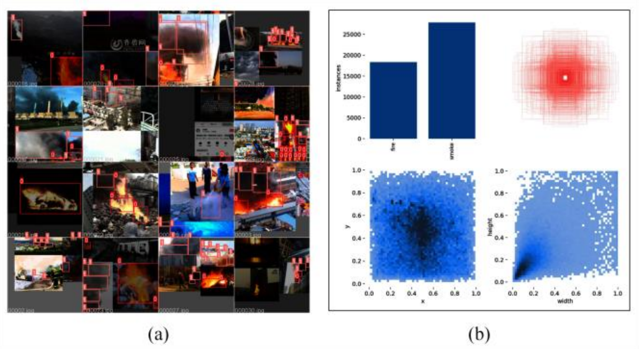


**FIGURE 6. Data processing results: (a) Description of Mosaic image enhancement results and (b) description of visualization results obtained by data processing.**

### *B. Experimental equipment*

The specific configurations of the computer system used in the experiment are listed in Table 1.

TABLE I
COMPUTATION SYSTEM

| Name | Definition |
|---|---|
| Processor | Intel(R) Core (TM)i9-13900CPU@ 3.70GHz |
| Running Memory | 128G |
| Operating System | Windows |
| GPU | NVIDIA GeForce RTX 4090 |
| GPU Memory | 128G |
| Programming Tool | PyCharm |
| Programming Language | Python |
| Deep Learning Framework | PyTorch |

### *C. Evaluating indicator*

In this study, the indicators used to evaluate the model's performance included recall, precision, mean average precision (mAP), frames per second (FPS), parameters and

GFLOPs. The specific meanings of these above indicators are as follows:

- Precision represents the proportion of correct samples in the model prediction results. This is defined by the following equation：

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

where the prediction results include $TP$ and $FP$. $TP$ represents the number of samples with correct predictions, and $FP$ represents the number of samples with incorrect predictions.

- Recall expresses the number of correctly predicted samples in all real object samples and is defined by the following formula:

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

where the number of real object samples is $TP + FN$. $FN$ represents false negatives, that is, the number of negative samples is predicted to be false.

- The two indicators, Precision and Recall, have single-point value limitations and cannot fully evaluate the module performance. Therefore, mAP was introduced to balance the calculation results of Precision and Recall. mAP represents the mean value of all categories of Average Precision (AP) in the entire dataset. The threshold of mAP is generally set to 0.5, that is, the prediction box with IoU greater than 0.5 is valid, the mAP value is calculated every 0.05, and finally the mean value of all mAP is calculated. The AP and mAP are represented as follows:

$$AP = \int_0^1 P(R)dR \quad (7)$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP \quad (8)$$

- Frames Per Second (FPS) represents the number of pictures detected per second and was used to measure the detection speed of the model.

- The parameters represent the variables that the model learns and adjusts during training and are used to measure the complexity of the model and computing resources.

- GFLOPs represents one billion floating point operations per second and are an important metric for evaluating computer performance, particularly in tasks that involve a significant amount of numerical computation, such as the field of deep learning.

## IV. Experimental results and analysis

In this paper, we experimentally compare the GAM-ASFF-YOLOv5s algorithm with several other standard object detection algorithms to verify the effectiveness of the improved YOLOv5s algorithm proposed in this paper. We conducted an ablation experiment to explore the contribution of each module in the algorithm to the entire algorithm. The above experiments were conducted separately on the VOC2007 and fire dataset. During the experiment, the experimental conditions, such as control equipment, training hyperparameters and several iterations, were kept consistent, and the experimental results were analyzed.

### A. Comparison experiments

Here, we compare the results of the proposed method with those of other traditional methods, that is, after training the VOC2007 dataset, two evaluation indexes of the precision and FPS are calculated on the verification set and compared with other object detection algorithms. The detection results are shown in Table 2, where FPS represents the number of images that the target detection algorithm can process per second, which is an important metric used to reflect the real-time detection performance of the network model for images or videos, and the higher the value, the faster the detection speed.

It can be seen from Table 2 that the improved YOLOv5s has better model performance than the other detection algorithms. For example, its precision on the VOC2007 dataset is 3.2 % higher than that of YOLOv4 and 5.7 % higher than that of YOLOv5s, suggesting that the GAM-ASFF-YOLOv5s+REP algorithm has more accurate detection precision. However, because the improved YOLOv5s algorithm has a higher network structure complexity, its effect is lower than that of the YOLOv5s model in FPS.

TABLE II
DETECTION RESULTS OF DIFFERENT NETWORKS

| Dataset | Algorithm | Backbone | Precision | FPS |
|---|---|---|---|---|
| VOC2007 | SSD [35] | VGG-16 | 77.5% | 46 |
| | MDSSD [36] | VGG-16 | 78.6% | 28 |
| | YOLOv3 [37] | Darknet-53 | 74.5% | 36 |
| | YOLOv4 [38] | CSPDarknet53 | 78.1% | 35 |
| | YOLOv5s | CSPDarknet53 | 75.6% | 76 |
| | GAM-ASFF-YOLOv5s+REP | CSPDarknet53 | 81.3% | 57 |

### B. Ablation experiments

We then conducted ablation experiments on the flame and smoke dataset and compared the experimental results of

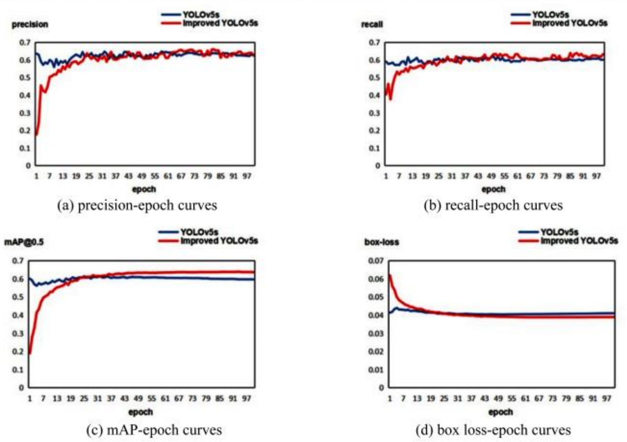
YOLOv5s, GAM-YOLOv5s, ASFF-YOLOv5s, YOLOv5s+REP, GAM-ASFF-YOLOv5s, GAM-YOLOv5s+REP, ASFF-YOLOv5s+REP and GAM-ASFF-YOLOv5s+REP on the dataset. Table 3 lists the data provided by the experimental results, where mAP@0.5 represents the mean average precision when IoU is equal to 0.5. The higher the mAP value, the better the model's performance.

TABLE III
EXPERIMENTAL RESULTS OF DIFFERENT NETWORKS

| Dataset | Algorithm | Precision | Recall | Parameters | GFLOPs | mAP@0.5 | mAP@0.5:0.95 |
|---|---|---|---|---|---|---|---|
| Fire | YOLOv5s | 64.0% | 59.8% | 27.2M | 16.0 | 60.4% | 35.4% |
| | GAM-YOLOv5s | 64.1% | 62.0% | 33.5M | 17.3 | 62.7% | 36.0% |
| | ASFF-YOLOv5s | 63.9% | 60.3% | 47.5M | 24.4 | 61.4% | 35.5% |
| | YOLOv5s+REP | 61.0% | 62.0% | 27.2M | 16.0 | 62.1% | 35.9% |
| | GAM-ASFF-YOLOv5s | 62.5% | 61.2% | 54.2M | 25.7 | 62.4% | 36.1% |
| | GAM-YOLOv5s+REP | 61.9% | 62.8% | 33.5M | 17.3 | 62.7% | 36.2% |
| | ASFF-YOLOv5s+REP | 62.2% | 62.5% | 47.5M | 24.4 | 61.7% | 34.7% |
| | GAM-ASFF-YOLOv5s+REP | 64.5% | 61.9% | 54.2M | 25.7 | 63.3% | 36.5% |

As can be seen from the black underlined values in Table 3, the GAM-ASFF-REP-YOLOv5s has improved precision, recall, and mAP by 0.5%, 2.1% and 2.9%, respectively, compared to the YOLOv5s, which is supported by the curves of experimental results for YOLOv5s and Improved YOLOv5s in Figure 8 and the P-R curves of YOLOv5s and Improved YOLOv5s in Figure 9. The introduction of the GAM module in YOLOv5s obtained not only the information of channel and space but also their interactive information, so that the precision, recall, and mAP increased by 0.1%, 2.2% and 2.3%, respectively, which shows that the global attention mechanism can effectively enhance the model's attention to flame and smoke information and obtain more critical image information. The introduction of ASFF in YOLOv5s strengthens the ability of model feature extraction, which improves recall by 0.5% and mAP by 1% but reduces precision by 0.1%, showing that ASFF-Detect significantly enables higher detection effects of the head network of YOLOv5s. In addition, owing to the large computational requirements of GAM and ASFF, the computation of the GAM-ASFF-YOLOv5s+REP model after the integration of GAM and ASFF increased, for example, the number of parameters increased by 27M and GLFOPs by 9.7. Replacing CIoU with a repulsion loss function improves the model's ability to lock onto flame and smoke objects during detection, improving recall and mAP by 2.2% and 1.7%, respectively, but reducing precision by 3%, which clearly shows that the application of repulsion loss in this network model can effectively improve the object locking ability for flame and smoke in the case of occlusion, thus improving the efficiency of model regression.

Figure 7 shows the experimental results of the YOLOv5s model for precision, recall, and mAP in more detail. The blue curves represent the experimental results of the original YOLOv5s model, whereas the red curves represent the experimental results of the improved YOLOv5s model. The abscissa in plots (a), (b), (c), and (d) represents the epoch value, and the ordinates represent the values of precision, recall, mAP@0.5, and box loss, respectively. From the following four figures, the improved model, that is, YOLOv5s combining the GAM module, ASFF module, and repulsion loss, shows a significant upward trend in precision, recall, and mAP compared to the original YOLOv5s model on the flame and smoke dataset. As can be seen from the introduction of the evaluation indicators, the mAP is a comprehensive metric that can be used to measure the model's overall performance, and a higher mAP value indicates a more robust overall performance.



(a) precision-epoch curves
(b) recall-epoch curves
(c) mAP-epoch curves
(d) box loss-epoch curves

FIGURE 7. Variation curves of experimental results for YOLOv5s and Improved YOLOv5s

As shown in Figure 7, compared with the other graphs, the contrasting effect of the red and blue curves in plot (c) is the most obvious, indicating that the improved method proposed in this study enables the YOLOv5s model to show positive and active feedback in terms of the overall performance enhancement. It can achieve good results for flame and smoke image detection. By observing the box-loss curves of the two models on the dataset, the bounding-box regression loss function value of the improved YOLOv5s remained in a low state during the training process, which was lower than the loss value of YOLOv5s and the convergence effect performed better. The loss curve of the improved YOLOv5s was smoother and more stable. Using repulsion loss as the bounding-box loss was better than using CIoU. It can be inferred that the repulsion loss helps solve the accurate locking of flame and smoke objects in the case of occlusion and improves the flame and smoke detection ability of the model in a complex image background environment, thus contributing to the detection of flame and smoke images with stacked and overlapping shape characteristics.

Subsequently, the detection performance of each object category from the improved YOLOv5s model was evaluated, as shown in Figure 8. Initially, the improved YOLOv5s model demonstrated superior overall flame and smoke detection performance compared to the original YOLOv5s model. Second, the improved YOLOv5s model exhibited a more obvious enhancement in smoke detection compared to flame detection. This may be attributed to the dataset comprising more smoke images than flame images, which may result in the model training converging towards a bias towards smoke.

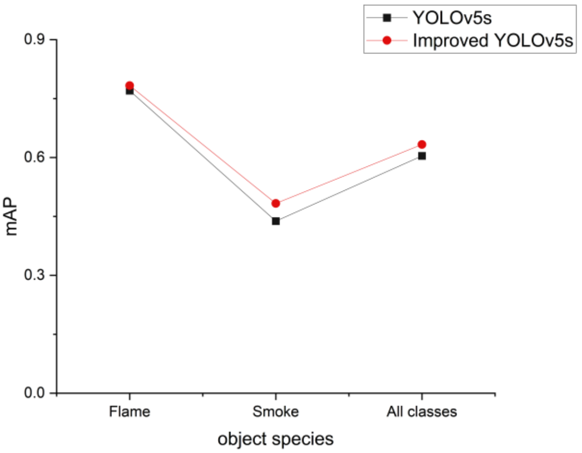


FIGURE 8. **The overall performance of YOLOv5s and Improved YOLOv5s when detecting different object species**

### *C. Comparison of experimental image test results*

Here, the detection results of remote small flame, overlapping and dense flame and smoke, concentrated smoke, mixed flame and smoke are selected for comparison to highlight the advantages of the improved YOLOv5s algorithm, namely, the GAM-ASFF-YOLOv5s + REP algorithm, where the bright blue circle indicates the distinction between the two models in terms of missed detection, as shown in Figure 9.

As shown in Figure 9 that the object detection box confidence of the improved YOLOv5s is significantly higher than that of the original model, and the improved YOLOv5s effectively ameliorates the omitted detection problem of the original model. Specifically, both groups (a) and (b) exhibit object detection leakage (see the area marked by bright blue circles), but the improved YOLOv5s model effectively reduces the omitted detection rate and improves the confidence of many object detection boxes, which suggests that the improved method proposed in this study makes the original YOLOv5s model more robust to flame and smoke detection.

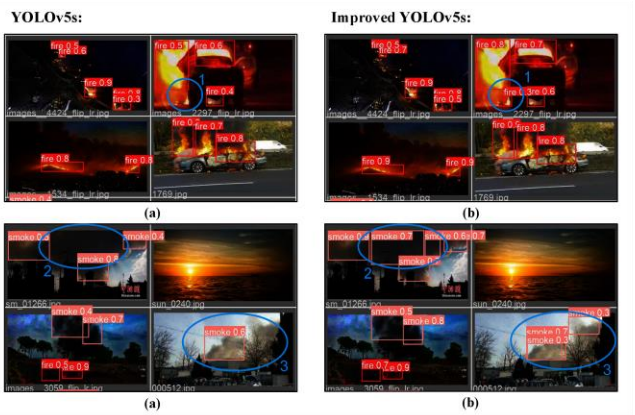


FIGURE 9. **Illustration of detection results for selected partial flame and smoke images, where shows the comparison of the model before and after the improvement on the problem of omitted detection.**

Owing to the misidentification of flames and smoke, it is important to have a fire warning. We also need to discuss cases of misidentification of the detection results. The yellow arrows in Figure 10 indicate the error-detection target box.

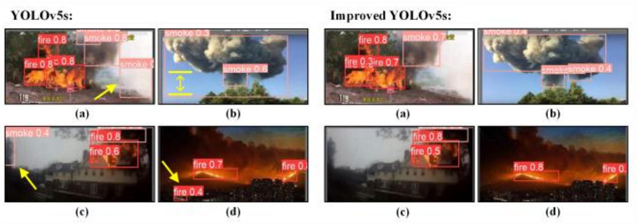


FIGURE 10. **Illustration of detection results for selected partial flame and smoke images, where shows the comparison of the model before and after the improvement on the problem of misidentification.**

As shown in Figure 10, there are many false detections when the original YOLOv5s model detects flames and smoke. For example, in group (a), the model identifies the area without smoke as smoke. In group (b), when the model detected smoke, the target detection frame could not accurately lock the smoke shape, and the target frame gap was too large. In group (c), the model incorrectly identified the branches as smoke because the color and shape of the dense branches in the image were very similar to those of black smoke. In group (d), because the lights in the building resemble flames, particularly at night, the bright lights in the distance are very similar in shape and color to the wildfire. The original YOLOv5s model misinterpreted lights as flames. However, the improved YOLOv5s model takes advantage of the GAM, ASFF, and repulsion loss and improves the model's ability to extract flame and smoke features to effectively distinguish between real flame and smoke features and other objects with similar shapes.

The preceding discussions demonstrate that the improved YOLOv5s model exhibits superior performance to the original YOLOv5s model in the flame and smoke detection domain, which displays good agreement with the analysis of the experimental results in the previous sections.

## V. Conclusions

In this study, the global attention mechanism was introduced into the backbone network of the YOLOv5s original model to reduce the dispersion degree of feature information, and the feature fusion layer was combined with the head network to further enhance the ability of the network to extract flame and smoke feature information. Simultaneously, the CIoU is replaced by the repulsion loss as the object bounding box regression loss function to more accurately detect the flame and smoke images with overlap. Relatively complete experiments were performed on the VOC2007 and the fire dataset to test the performance of the algorithm. The analysis of the experimental results reveals that the GAM-ASFF-YOLOv5s+REP algorithm performed better than the original YOLOv5s algorithm on experimental data such as mAP, and the overall performance, detection accuracy, and detection speed of the model were optimized to a certain extent. The algorithm proposed in this paper enhances the detection performance of the original YOLOv5s for flame and smoke to a certain extent, but the performance of the algorithm has some limitations. For example, it still has a low missed detection rate. The network complexity of the original YOLOv5s model increased because of the introduction of GAM and ASFF. Given this problem, we will continue to optimize and improve the model in the future to achieve a better object detection effect. For example, using the current mainstream lightweight network architecture, MobileNet [39], ShuffleNet [40] or GhostNet [41], to lightweight the network architecture of the model to achieve higher accuracy and lower computational complexity. Moreover, because the change in smoke concentration in the early stage of fire is an important parameter, we will pay more attention to the related research on the change in smoke concentration in the process of fire in a follow-up study and plan to use the smoke sensor to extract the smoke concentration parameters to improve the model's ability to detect smoke.

## REFERENCES

[1] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, and J. J. P. C. Rodrigues, "Identifying pneumonia in chest X-rays: A deep learning approach," *Measurement*, vol. 145, pp. 511-518, 2019.

[2] D. M. Gavrila and V. Philomin, "Real-time object detection for ``smart" vehicles," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999.

[3] L. Wei *et al.*, "SSD: Single Shot MultiBox Detector," *Springer, Cham*, 2016.

[4] S. Chen, J. Hong, T. Zhang, J. Li, and Y. Guan, "Object Detection Using Deep Learning: Single Shot Detector with a Refined Feature-fusion Structure," in *2019 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2019.

[5] Lisha *et al.*, "MDSSD:multi-scale deconvolutional single shot detector for small objects," *Science China(Information Sciences)*, vol. v.63, no. 02, pp. 98-100, 2020.

[6] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional Single Shot Detector," 2017.

[7] P. Shamsolmoali, M. Zareapoor, E. Granger, J. Chanussot, and J. Yang, "Enhanced Single-shot Detector for Small Object Detection in Remote Sensing Images," 2022.

[8] X. Tan, "FASSD: A Feature Fusion and Spatial Attention-Based Single Shot Detector for Small Object Detection," *Electronics*, vol. 9, 2020.

[9] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2017, pp. 6517-6525.

[11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," 2021.

[12] X. Huang, X. Wang, W. Lv, X. Bai, and O. Yoshie, "PP-YOLOv2: A Practical Object Detector," 2021.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Computer Vision & Pattern Recognition*, 2016.

[14] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv e-prints*, 2018.

[15] Y. L. Chung, H. Y. Chung, and C. W. Chou, "Efficient Flame Recognition Method Based on a Deep Convolutional Neural Network and Image Processing," *IEEE*, 2019.

[16] C. Chaoxia, W. Shang, and F. Zhang, "Information-Guided Flame Detection Based on Faster R-CNN," *IEEE Access*, vol. 8, pp. 58923-58932, 2020.

[17] H. Yar, Z. A. Khan, F. U. M. Ullah, W. Ullah, and S. W. Baik, "A modified YOLOv5 architecture for efficient fire detection in smart cities," *Expert Systems with Application*, no. Nov., p. 231, 2023.

[18] M. N. Uddin, M. S. I. Sakib, S. Nawer, and R. T. Mohona, "Improved Fire Detection by YOLOv8 and YOLOv5 to Enhance Fire Safety," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*.

[19] Z. Dai, "Image Flame Detection Method Based on Improved YOLOv3," in *IOP Conference Series: Earth and Environmental Science*, 2021, vol. 693, no. 1: IOP Publishing, p. 012012.

[20] Y. Y. Qin, J. T. Cao, and X. F. Ji, "Fire Detection Method Based on Depthwise Separable Convolution and YOLOv3," *International Journal of Automation and computing*, vol. 18, no. 2, pp. 300-310, 2021.

[21] X. Wu, Y. Cao, X. Lu, and H. Leung, "Patchwise dictionary learning for video forest fire smoke detection in wavelet domain," *Neural Computing and Applications*, vol. 33, no. 13, pp. 7965-7977, 2021.

[22] Y. Cao, Q. Tang, X. Wu, and X. Lu, "EFFNet: Enhanced Feature Foreground Network for Video Smoke Source Prediction and Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1-1, 2021.

[23] F. Yang, Q. Xue, Y. Cao, X. Li, W. Zhang, and G. Li, "Multi-temporal dependency handling in video smoke recognition: A holistic approach spanning spatial, short-term, and long-term perspectives," *Expert Systems with Applications*, vol. 245, p. 123081, 2024.

[24] Y. Liu, Z. Shao, and N. Hoffmann, "Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions," 2021.

[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," 2021.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] S. Liu, D. Huang, and Y. Wang, "Learning Spatial Fusion for Single-Shot Object Detection," 2019.

[28] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion Loss: Detecting Pedestrians in a Crowd," 2017.

[29] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, and I. H. Yeh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[30] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," *IEEE Computer Society*, 2017.

[31] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," *IEEE*, 2018.

[32] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *Springer, Cham,* 2018.

[33] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6848-6856, 2017.

[34] Z. Zheng *et al.*, "Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation," 2020.

[35] W. Liu et al., "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016: Springer, pp. 21-37.

[36] L. Cui, R. Ma, P. Lv, X. Jiang, and M. Xu, "MDSSD: multi-scale deconvolutional single shot detector for small objects," *Sciece China. Information Sciences,* vol. 63, no. 2, 2020.

[37] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv e-prints,* 2018.

[38] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020.

[39] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017.

[40] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," 2017.

[41] K. Han, Y. Wang, Q. Tian, J. Guo, and C. Xu, "GhostNet: More Features From Cheap Operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

**FIRST A. AUTHOR, Li Deng.** was born in 1986 in Sichuan Province, China in 1986. He received the the Ph.D degree from Nanjing University of Science and Technology, Jiangsu Province, China. At present, he is an associate professor at Civil Aviation Flight University of China, Guanghan City, Sichuan Province, China.

His research interests include large space fire detection technology, infrared image processing technology, and multi-sensor fusion technology. His research group has long been engaged in civil aircraft fire detection and early warning and fire prevention technology, large space fire detection and early warning technology and other related research work.

Mr. Deng has published more than ten papers such as SCI and EI in his related fields and has been granted more than ten patents. He has guided graduate students to win the second prize in the Sichuan College Students' Challenge Cup Competition, the second prize in the National College Students' Embedded Design Competition, the third prize in the Sichuan College Students' Safety and Emergency Innovation Practice Competition, etc. He has received the School's Educator Award and the Excellent Paper Award at the National Public Security Conference.

**SECOND B. AUTHOR, Jin Zhou.** was born in 1999 in Hechuan District, Chongqing, China. She received the bachelor's degree in communication engineering from Huaiyin Institute of Technology, Jiangsu Province, China, in 2022. She is currently pursuing a master’s degree in Transportation Engineering at Civil Aviation Flight University of China, Guanghan City, Sichuan Province, China.

Her research interest includes the large space fire detection and computer vision. She is currently working on using computers to detect fires.

Ms. Zhou won the Excellent Paper Award of China Public Security Conference in 2023 and the Challenge Cup Silver Award of China Civil Aviation Flight University in 2024. The paper titled "Flame and Smoke Detection Algorithm Based on Improved YOLOv8" submitted by Ms. Zhou to the Journal of Tsinghua University (Natural Science Edition), an EI-indexed journal, has been accepted. She is striving forward in her research field.

**THIRD C. AUTHOR, Quanyi Liu.** was born in 1987 in Henan Province, China. He received the Ph.D degree from Tsinghua University, Beijing Province, China. At present, he is a professor at Civil Aviation Flight University of China, Guanghan City, Sichuan Province, China.

His research interests include the special fires mechanism of aircraft and fire safety assessment, intelligent fire detection and new fire extinguishing technologies, intelligent fire protection and emergency rescue at airports, etc.

Mr. Liu won the First Prize for Technological Progress of the Public Security Science and Technology Award in 2022, the National Civil Aviation Youth Civilization Unit (Unit Leader), the Second Prize for Civil Aviation Teaching Achievements in 2022, the Second Prize for Civil Aviation Science and Technology of the China Air Transport Association in 2020, the honorary title of Outstanding Communist Party Member of the National Civil Aviation in 2021, and the Second Prize for Civil Aviation Science and Technology of the China Air Transport Association in 2019.