CSAF-YOLO: An Improved Smoke and Fire Object Detection Algorithm for Complex Scenes Based on YOLOv11n

Yuanpan Zheng*, Chao Wang, Yu Zhang, Xuhang Liu, Wenbin Zhong
Schoolof ComputerScience and Artificial Intelligence, Zhengzhou University of Light Industry, Zhengzhou 450000, China
E-mail: ypzheng@zzuli.edu.cn
*Corresponding author

Keywords: Fire smoke detection, YOLOv11n, multi-scale contextual modelling, bi-branch self-attention feature fusion

Received: March 28, 2026

Early detection of flames and smoke is crucial for fire warning systems, yet blurred smoke boundaries, small-scale targets, and complex background interference still limit existing methods. This paper proposes CSAF-YOLO, an improved detection model based on YOLOv11n. A C3-MCG module is introduced into the backbone to enhance multi-scale contextual feature representation. A DB-SAFM module is designed in the feature fusion stage to improve semantic and spatial feature alignment through a dual-branch self-attention mechanism. AWHIoU is adopted as the loss function, combined with a bi-phase focusing strategy and a scale-adaptive weighting mechanism to achieve adaptive optimisation for sampl0065s with different IoU qualities and targets of different scales. Experimental results show that CSAF-YOLO significantly improves detection performance over mainstream models, achieving 2.2% and 2.3% gains in precision and mAP50 over YOLOv11n, respectively, while maintaining a high FPS of 177.75, demonstrating strong potential for complex monitoring and real-time detection.

Povzetek: Predstavljen je model CSAF-YOLO, kateri izboljša hitro in natančno zaznavanje ognja ter dima v zahtevnih okoljih.

1 Introduction

Fire is a sudden-onset disaster that causes severe casualties and substantial economic losses worldwide, and its destructive impact is particularly prominent among both natural and human-induced hazards. In January 2025, multiple wildfire incidents occurred in Los Angeles County, among which the Palisades and Eaton fires were the two major fire sites, resulting in 31 deaths and the destruction of 16,251 properties [1] In January 2026, multiple wildfires also broke out in south-central Chile, causing 21 deaths, 305 injuries, and the destruction of 2,359 houses, with extremely serious consequences [2]. In the early stage of a fire, smoke usually appears before visible flames. Therefore, accurate smoke detection is of great significance for early fire warning and prevention.

In early fire prevention studies, traditional methods mainly relied on temperature sensors, infrared sensors, and smoke detectors [3]. However, these methods suffer from several limitations, including restricted detection range, delayed response, and susceptibility to false alarms and missed detections under complex environmental interference [4]. Subsequently, flame and smoke detection based on images and videos gradually became an important research direction [5]. Early image-based methods depended on colour, texture, or motion features. In images, smoke often exhibits low contrast, diffuse distribution, and blurred boundaries, making its detection far more challenging than that of flames [6]. In addition, complex environmental factors such as haze, steam, and illumination changes can interfere with judgment, leading to false positives and missed detections [7], which has driven research toward deep learning-based approaches.

The introduction of deep learning has advanced flame and smoke detection into a new stage. Methods represented by convolutional neural networks (CNNs) can better learn smoke textures and flame patterns while reducing interference from complex backgrounds, thus achieving better detection performance across multiple scenarios [8] Kim et al. [9] employed a CNN to construct a smoke detection model and obtained stable results in different environments. Wang et al. [10] proposed DSS-YOLO based on YOLOv8, which improves real-time performance through a lightweight design and feature enhancement mechanism while maintaining satisfactory detection accuracy.

Despite these advances, existing deep learning-based methods still have several limitations. The two-stage detector Faster R-CNN [11] has stronger localisation capability, but its computational cost is high. Single-stage methods such as SSD [12] and RetinaNet [13] offer fast detection speed, but they are prone to missed detections and false detections for small-scale and low-contrast smoke. In comparison, the YOLO series of detection models strikes a better balance between speed and accuracy and provides higher inference efficiency.

Nevertheless, existing YOLO-based models still face challenges in handling blurred smoke boundaries, small-scale flames, and interference from complex backgrounds [14]. To address these issues, this paper proposes CSAF-YOLO (Context and Scale-aware Attention Fusion

YOLO), an enhanced detection framework built upon YOLOv11n. Specifically, the main contributions of this work are summarized as follows.

(1) Considering the limitations of existing public fire and smoke datasets, such as inconsistent annotations, insufficient interference scenarios, and inadequate image quality, a self-built dataset was constructed in this study. The dataset contains complex scenes and multi-scale fire and smoke targets, providing a more challenging benchmark for evaluating model performance in real-world environments.

(2) To better handle the variations in scale and appearance between flames and smoke, a CSP Multi-scale Context Guided Block (C3-MCG) is designed. By combining multi-dilated convolutions with a global channel attention mechanism, C3-MCG enlarges the receptive field, enhances multi-scale contextual representation, and alleviates the interference caused by blurred smoke boundaries and irregular flame textures.

(3) Considering that flames and smoke exhibit distinct visual patterns, and that shallow branches are more sensitive to flame details while deep branches are more effective in representing the global morphology of smoke, a Dual-Branch Self-Attention Fusion Module (DB-SAFM) is proposed. This module explicitly models dependencies among multi-branch features through self-attention, enabling cross-branch information selection and collaborative enhancement, and allowing the network to adaptively emphasize feature representations of flame details and smoke patterns.

(4) An Adaptive Weighted Hybrid IoU (AWHIoU) loss is further introduced by integrating dynamic bi-phase focusing and scale-aware weighting. This design enables adaptive optimisation for samples with different IoU qualities and for small targets from both the quality and scale perspectives.

In summary, the proposed CSAF-YOLO systematically enhances fire and smoke detection from four aspects, including dataset construction, multi-scale contextual modelling, cross-branch attention-based fusion, and adaptive localisation optimisation. Through these improvements, the model achieves more robust and accurate detection of flame and smoke targets in complex scenes while preserving high inference efficiency.

2 Model selection and improvement

2.1 YOLOv11n detection model

YOLOv11n integrates Cross Stage Partial (CSP) connections and multi-scale convolutional operators, thereby significantly enhance feature representation capability [15] Its decoupled head structure further optimises the collaboration between classification and regression, enabling the model to perform better in complex scenarios. However, directly applying YOLOv11n to flame and smoke detection still has certain limitations: it shows limited performance when handling low-contrast smoke with blurred boundaries; its detection capability for small-scale flame targets is insufficient; and the adopted IoU loss function lacks adaptive optimisation for prediction boxes of different qualities, which is unfavourable for detection in complex environments. To address the above deficiencies while achieving a higher frame rate and faster detection speed, this paper adopts structural improvements based on YOLOv11n. The architecture of YOLOv11n is shown in Figure 1.

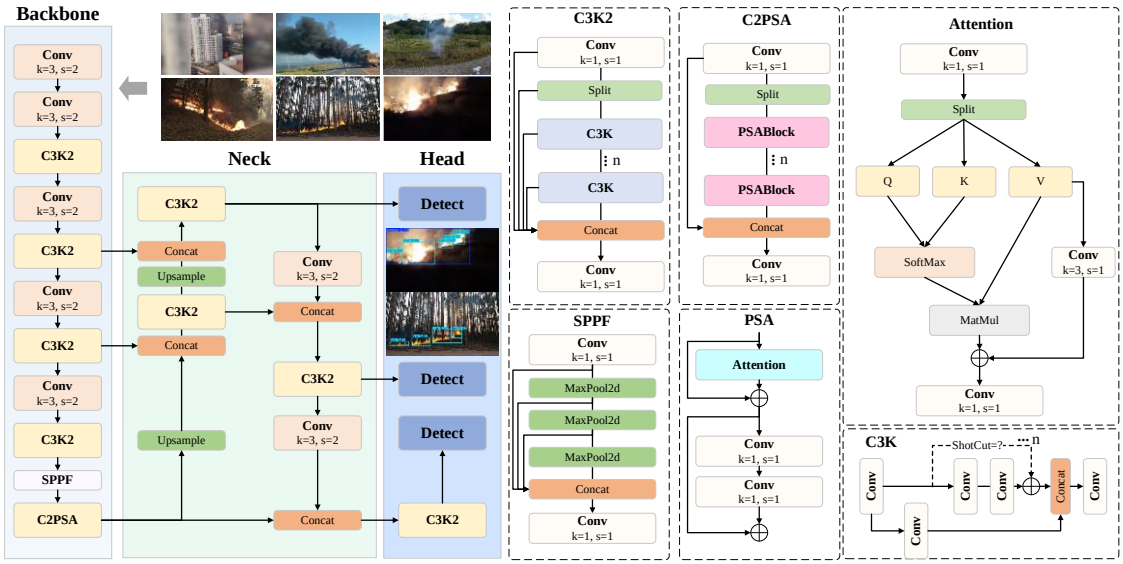

Figure 1: YOLOv11n structure diagram

2.2 CSAF-YOLO model

In this paper, YOLOv11n is improved and renamed CSAF-YOLO, and its overall architecture is shown in Figure 2. The main improvements are as follows: First, in the backbone network, a multi-scale context-guided mechanism is introduced into C3k2 to construct C3-MCG, which alleviates the interference caused by blurred smoke

boundaries and irregular flames while enhancing feature representation capability. Second, in the neck, a dual-branch self-attention fusion module is introduced to replace Concatenate-based feature fusion with DB-SAFM. By jointly modelling channel-wise and spatial attention, this module enhances both global dependency capture and local detail representation, thereby improving the discrimination between smoke and complex backgrounds. Finally, the loss function is redesigned as Adaptive Weighted Hybrid IoU (AWHIoU), which adaptively reweights and optimises conventional IoU-series losses, improves localisation sensitivity to small and elongated targets, accelerates convergence, and enhances boundary fitting accuracy.

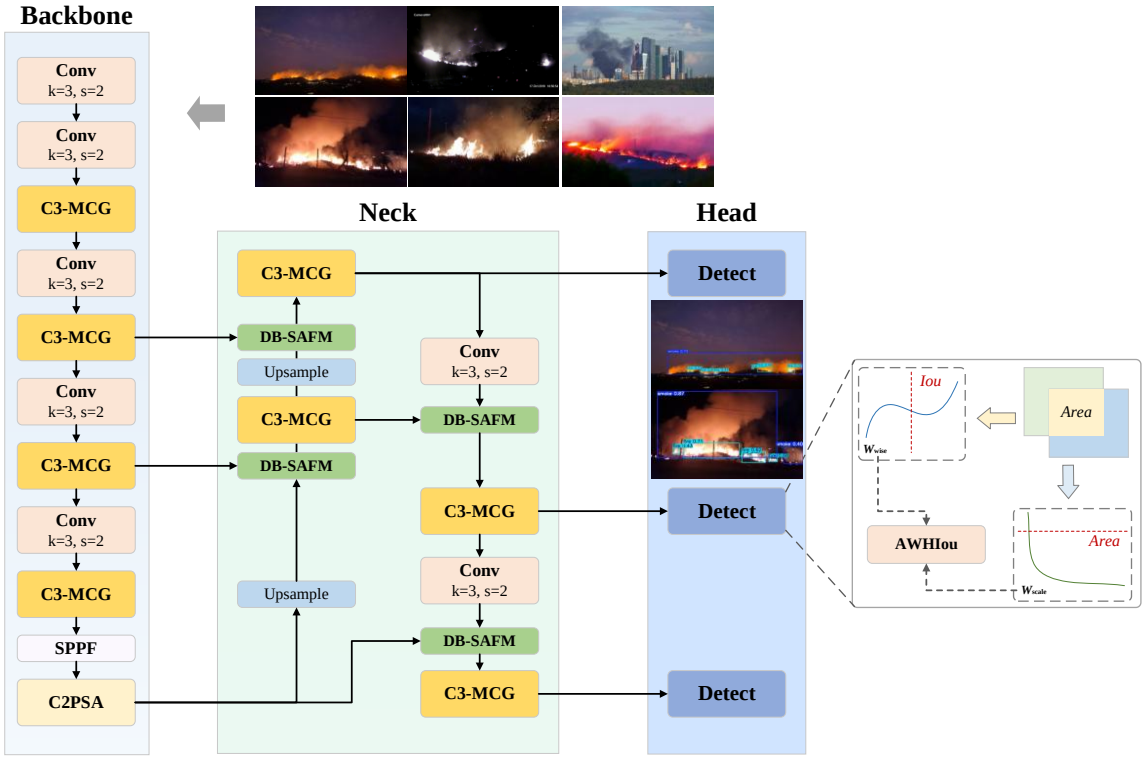

Figure 2 CSAF-YOLO Structure diagram

2.3 C3-MCG Module

Flames and smoke differ substantially in both scale and appearance. Flames are characterized by high brightness and relatively clear local textures, whereas smoke tends to exhibit low contrast, semi-transparency, and unstable diffuse boundaries. Therefore, the model must not only cover the scale range from small flames to large smoke plumes, but also maintain stable responses in weak-texture regions. FPN [16] and PANet [17] achieve cross-layer feature interaction through pyramid structures and play an important role in multi-scale representation; EfficientDet introduces BiFPN [18] to strengthen cross-scale information fusion while maintaining efficiency. Relation Networks proposed by Hu et al. [19] establish a relation modelling mechanism among candidate regions to enhance the utilisation of contextual information. However, such methods usually introduce considerable computational overhead, which is unfavourable for real-time applications.

To improve the backbone network's ability to model long-range dependencies and multi-scale contextual information, this paper proposes a new feature extraction module, C3-MCG. This module incorporates the idea of the Context Guided Block [20] and combines multi-dilated convolutions with a global channel attention mechanism to enlarge the receptive field and alleviate the interference caused by blurred smoke boundaries and irregular flame textures. The design of C3-MCG is shown in Figure 3.

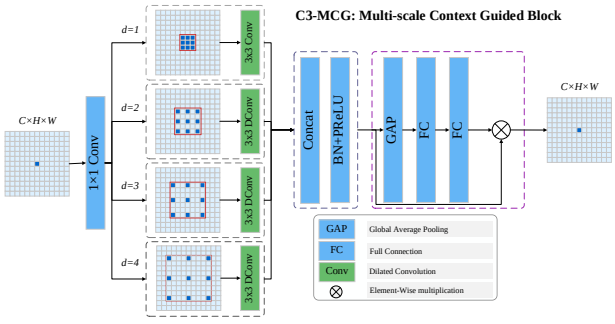

Figure 3: C3-MCG Structure Diagram

The first step of this module is channel compression and alignment. Specifically, given the input feature map $X \in R^{C\times H\times W}$, to reduce computational cost and unify the branch channel dimensions, a 1×1 convolution is applied to perform channel compression and alignment on the input feature, yielding the intermediate feature X':

$$X' = f_{1\times1}(X) \quad (1)$$

Subsequently, to simultaneously capture local details and multi-scale contextual information, four branches are constructed in parallel on X': one local branch employs standard convolution to extract local responses, while three contextual branches use channel-wise dilated convolutions with different dilation rates $d \in \{2,3,4\}$ to enlarge the receptive field.

$$F_1 = f_{3\times3}^{dw}(X') \quad (2)$$

$$F_d = f_{3\times3}^{dw-dil(d)}(X') \quad (3)$$

where, $f_{3\times3}^{dw}(\cdot)$ denotes the depthwise convolution, and $f_{3\times3}^{dw-dil(d)}(\cdot)$ denotes the channel-wise dilated convolution with dilation rate d.

The outputs of the four branches are concatenated along the channel dimension, and BN and PReLU are applied to the concatenated fused features to stabilize the feature distribution after multi-branch fusion and enhance the representation capability:

$$F_m = \phi(BN(Concat(F_1, F_2, F_3, F_4))) \quad (4)$$

where, $Concat(\cdot)$ denotes the concatenation operation along the channel dimension, and ϕ represents the PReLU activation function.

After obtaining the fused feature F_m, global average pooling is first applied, followed by two fully connected layers and nonlinear transformations to generate the channel attention weight vector:

$$s = \sigma(W_2\delta(W_1 GAP(F_m))), F_{ref} = F_m \square\ s \quad (5)$$

where, $GAP(\cdot)$ denotes global average pooling; W_1 and W_2 are learnable weight matrices of the fully connected layers; $\sigma(\cdot)$ denotes the Sigmoid function; s is the channel weight vector; and F_{ref} represents the feature after channel recalibration.

Finally, to preserve the original representation and facilitate stable gradient propagation, a residual connection is adopted to obtain the output feature Y:

$$Y = X + F_{ref} \quad (6)$$

Through this design, C3-MCG is able to simultaneously model local detailed features and multi-scale contextual semantics, while employing a global attention mechanism for channel-wise feature reweighting. Compared with the C3k2 module, C3-MCG significantly enhances the model's sensitivity to irregular textures and diffuse boundaries in fire and smoke.

2.4 DB-SAFM module

In fire and smoke detection, shallow branches are mainly responsible for capturing the fine textures and edge information of flames, whereas deep branches are better at representing the overall morphology and contextual semantics of smoke. Attention mechanisms can enhance feature selectivity to a certain extent. SE-Net [21] was the first architecture to introduce channel attention, while CBAM [22] further incorporated spatial attention on this basis, enabling the model to focus on both important channels and key spatial locations. In smoke detection, Ejaz et al. [23] strengthened the response to smoke regions by using channel attention, and Jin et al. [24] attempted to combine self-attention with multi-scale fusion, achieving improved performance.

Motivated by the above observations and related studies, this paper designs the DB-SAFM module, which explicitly models the dependencies among multi-branch features through a self-attention mechanism, enabling cross-branch information selection and collaborative enhancement, so that the network can adaptively emphasize feature representations of flame details and smoke patterns.

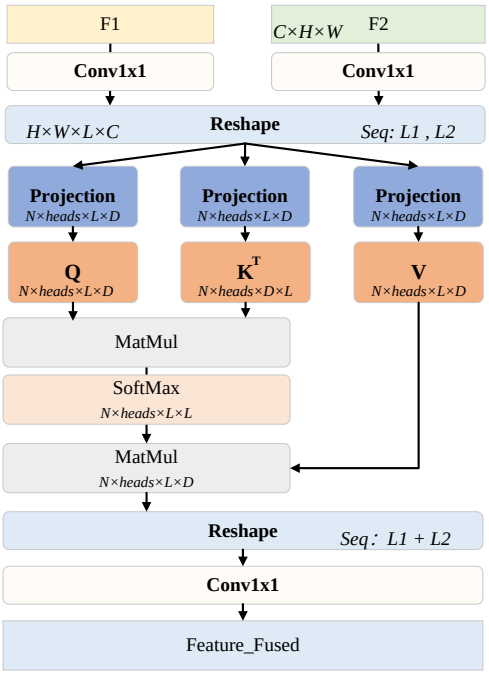

Figure 4 DB-SAFM module structure diagram

As shown in Figure 4, DB-SAFM receives two input feature maps from the backbone, $X_1 \in R^{C\times H\times W}$ and $X_2 \in R^{C\times H\times W}$, and uses 1×1 convolutions to align them to the same embedding dimension C, yielding the aligned features X_1', X_2':

$$X_1' = f_{1\times1}(X_1), X_2' = f_{1\times1}(X_2) \quad (7)$$

The two feature maps are then stacked along the branch dimension, and each spatial position is treated as a sequence of length $L = 2$:

$$S = Reshape(Stack(X_1', X_2')) \in R^{H\times W\times L\times C} \quad (8)$$

where, $Stack(\cdot)$ denotes stacking the two feature maps along the branch dimension, and $Reshape(\cdot)$ denotes flattening the spatial dimensions. With $L = 2$, a sequence input S is thus obtained for each spatial position.

Subsequently, the module employs linear projections to generate the Query, Key, and Value, and performs multi-head self-attention to learn the cross-branch interaction weights between the two branches:

$$Q = SW_Q, K = SW_K, V = SW_V \quad (9)$$

$$Attn(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \quad (10)$$

where, W_Q, W_K, W_V are learnable parameter matrices; d denotes the feature dimension of each attention head, with $d = C/M$; and M is the number of attention heads.

Finally, the attention outputs are fused and then passed through a 1×1 convolution for output mapping to obtain the final output:

$$Y = f_{1\times1}^{\downarrow}(Reshape(\sum_{t=1}^{L} Attn(Q,K,V)_t)) \quad (11)$$

where, Y denotes the output feature map, $f_{1\times1}^{\downarrow}(\cdot)$ represents the 1×1 convolution used for output projection and compression, and t is the sequence position index.

Compared with fixed-weight fusion mechanisms such as SE [21] or MLP-based [25], DB-SAFM can adaptively adjust the weights of flame and smoke features and align them position by position in the spatial dimension, thereby improving detection performance in blurry and complex scenes. Since the sequence length is $L = 2$, the computational complexity is much lower than that of the Transformer architecture, while still preserving the capability of cross-branch modelling.

2.5 AWHIoU loss function

IoU-based loss functions, such as GIoU [26], DIoU and CIoU [27], SIoU[28], and Wise-IoU [29] have been widely applied. However, they tend to treat low-quality and high-quality samples equally, which may lead to insufficient convergence on high-quality samples. Meanwhile, small targets are extremely important in fire and smoke detection, but IoU itself is insensitive to target scale, resulting in suboptimal localisation performance for small objects. Although Focal Loss[13] alleviates the imbalance between positive and negative samples, it mainly operates on the classification task and therefore provides limited improvement in detection performance.

Therefore, the AWHIoU loss function is proposed, which combines Bi-Phase Focusing and Scale-aware Weighting to achieve adaptive optimisation for samples with different IoU qualities and for small targets. The quality- and scale-adaptive weighting strategies are illustrated in Figure 5. In Figure 5(a), along the quality dimension, the sample weight changes in a two-phase manner with IoU, completing a transition from mild to intensified focusing around the threshold, so that high-quality matched samples with high IoU are assigned greater weights. In Figure 5(b), along the scale dimension, the scale weight decreases as the target area increases, and an upper bound (cap) is introduced to prevent excessively large gradients caused by extremely small targets, thereby providing moderate compensation for small targets while suppressing the dominance of large targets in the optimisation process.

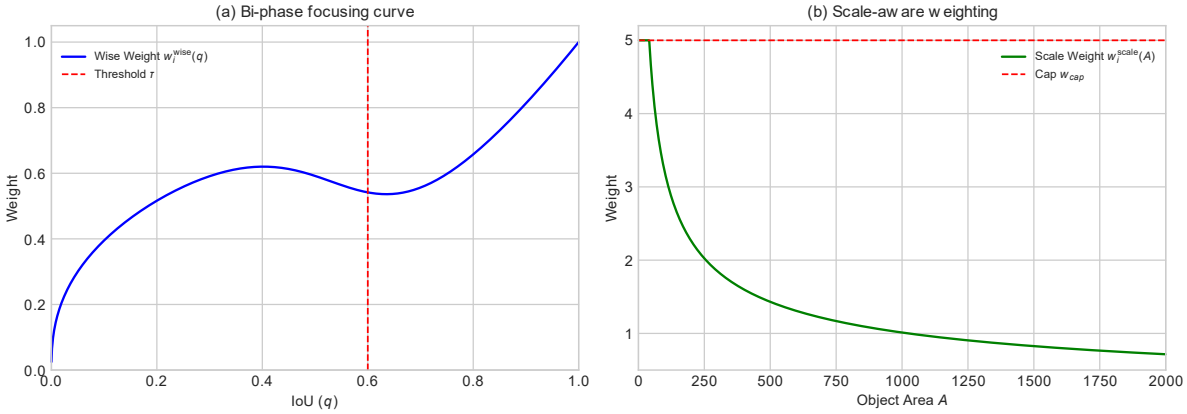

Figure 5: Quality and scale adaptive weighting strategy

In the boundary box regression stage, to incorporate sample quality into weight allocation, $CIoU$ is first used to measure the consistency between the predicted box and the ground-truth box. The sample quality score $q_i \in [0,1]$ is then computed as the basis for subsequent adaptive weighting. Let the (i)-th predicted box and its corresponding ground-truth box be denoted by $b_i = [x_1^b, y_1^b, x_2^b, y_2^b]$ and $g_i = [x_1^g, y_1^g, x_2^g, y_2^g]$, respectively:

$$q_i = clip(CIoU(b_i, g_i), 0, 1) \quad (12)$$

where, $clip(\square, 0, 1)$ denotes the clipping function, which constrains the value to the range from 0 to 1.

To achieve adaptive optimisation for samples of different qualities, a dynamic bi-phase focusing mechanism is adopted to construct a gating coefficient σ_i that varies smoothly with q_i, and a focusing exponent γ_i is continuously interpolated between low-quality and high-quality samples:

$$\sigma_i = Sigmoid(\frac{q_i - \tau}{\kappa}) \quad (13)$$

$$\gamma_i = \gamma_{low}(1-\sigma_i) + \gamma_{high}\sigma_i \quad (14)$$

where, τ controls the boundary between low-quality and high-quality samples, and is set to 0.6 in this paper; κ controls the steepness of the transition; and γ_{low} and γ_{high} correspond to the focusing strengths in the low-quality and high-quality regions, respectively.

On this basis, a quality-aware focusing weight w_i^{wise} is defined to adaptively regulate samples of different qualities, so that the focusing strength can be automatically adjusted according to the variation of q_i during training. When the sample quality exceeds the threshold τ, γ_i tends more toward γ_{high}, thereby increasing the contribution of high-quality positive samples; conversely, when q_i is below τ, γ_i tends more toward γ_{low}, thereby suppressing the gradient oscillation caused by low-quality samples.

$$w_i^{wise} = (q_i)^{\gamma_i} \quad (15)$$

To further enhance sensitivity to small targets, a scale-adaptive weight w_i^{scale} is introduced. The scale weight is determined by the ground-truth box area A_i, amplifies small-area targets, and limits excessive amplification through w_{cap}:

$$A_i = (x_2^g - x_1^g)(y_2^g - y_1^g) \quad (16)$$

$$w_i^{scale} = min((\frac{s_{ref}}{A_i})^{\beta}, w_{cap}) \quad (17)$$

where, s_{ref} denotes the reference scale 32^2, and β is the scale sensitivity factor.

Finally, the base quality score q_i, the quality-aware focusing weight w_i^{wise}, and the scale weight w_i^{scale} are combined to obtain the weighted quality score q_i :

$$q_i = clip(q_i \cdot w_i^{wise} \cdot w_i^{scale}, 0, 1) \quad (18)$$

$$L_{AWHIoU} = \frac{1}{N}\sum_{i=1}^{N}(1-q_i) \quad (19)$$

where, N denotes the number of samples involved in regression (the number of positive bounding boxes).

When $\gamma_{low} = \gamma_{high} = 0$ and $\beta = 0$, AWHIoU degenerates into the standard IoU loss. When only w_i^{wise} is retained, it corresponds to Wise-IoU with dynamic focusing; when only w_i^{scale} is retained, it corresponds to Scale-IoU with scale weighting. AWHIoU can better balance the suppression of low-quality samples and the allocation of weights to small targets in smoke detection tasks, thereby improving the convergence efficiency and detection performance of the model.

3 Experimental results and analysis

3.1 Dataset construction

Existing public fire and smoke datasets still suffer from limitations such as insufficient interference scenarios and inadequate image quality. To verify the applicability of the proposed method in complex environments, this study employs a self-built dataset. Part of the data was derived from the D-Fire dataset [30], while the remaining images were collected from the Internet. After data screening, a total of 13,217 images were retained. The dataset was randomly divided into training, validation, and test sets at a ratio of 7:1:2, as summarized in Table 1.

Table 1: Dataset details

Subset	Number of Images	Percentage
Training set	9250	70
Validation set	1322	10
Test set	2645	20

To provide a more intuitive understanding of the dataset, Figure 6 presents several sample images, covering most typical scenarios, including the coexistence of smoke and flames at multiple scales, the presence of smoke only or flames only, as well as cases under different scene conditions. This requires the model not only to identify the targets accurately, but also to maintain a low false alarm rate in interference-prone environments. Owing to its relatively large scale, this dataset is more suitable for evaluating the stability of the model under complex backgrounds.

Figure 6: Partial data of the dataset

To characterize the target scale distribution in the dataset, the normalized width and height of all annotated bounding boxes were statistically analysed and divided into three categories according to area, namely Small, Medium, and Large. The resulting width–height scatter distribution is shown in Figure 7. Small samples are mainly concentrated in the low-width and low-height region, exhibiting an overall long-tail and multi-scale distribution. In addition, the "×" symbol marks the centroid of each scale category. This distribution indicates that fire and smoke detection involves a large number of small-scale targets with substantial shape variation, further highlighting the challenges of accurate detection in complex scenes.

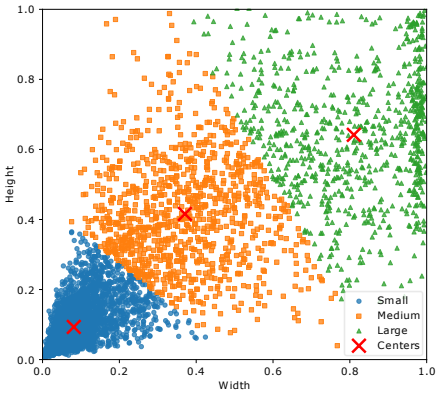

Figure 7: Scatter plot of the width and height distribution of annotation boxes

3.2 Evaluation metrics

In this paper, precision (*P*), mean Average Precision (*mAP*), recall (*R*), and frames per second *(FPS*) are adopted as the evaluation metrics for the model. The formulas for these metrics are given as follows:

$$P = \frac{TP}{TP + FP} \quad (20)$$

$$AP = \int_0^1 P(R)dR \quad (21)$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \quad (22)$$

$$R = \frac{TP}{TP + FN} \quad (23)$$

$$FPS = \frac{N}{T} \tag{24}$$

where, TP denotes the number of correctly detected smoke samples, FP denotes the number of non-smoke samples incorrectly identified as smoke, and FN denotes the number of missed smoke samples. P represents precision, R represents recall, C is the total number of classes, and i is the class index used to traverse all C classes. N denotes the number of inference images, and T denotes the total inference time, measured in milliseconds.

3.3 Experiment setup

All experiments in this paper were conducted under the same conditions. The CPU was an AMD Ryzen 9 9950, the system memory was 128 GB, and the GPU was an NVIDIA RTX 5090D with 32 GB of VRAM. The PyTorch version was 2.10.0, and the CUDA version was 12.6. The experiments were implemented using the Ultralytics YOLO framework with an input resolution of 640 × 640, and all experimental settings were strictly configured according to the parameters listed in Table 2.

Table 2: Training configuration

Parameter	Setup
Epoch	160
Batch size	32
Workers	4
Learning rate	0.01
Optimizer	SGD

3.4 Ablation experiments

Ablation experiments were conducted in this section, and the results are presented in Table 3. YOLOv11n was used as the baseline model, where "√" indicates that the corresponding method was adopted in the model. The training and inference procedures were kept consistent with those of the baseline model. The core evaluation metrics included P, R, mAP50, mAP50(95) and FPS.

Table 3: Ablation test results

Baseline	*C3-MCG*	*DB-SAFM*	*AWHIoU*	*P/%*	*R/%*	*mAP50/%*	*mAP50(95)/%*	*FPS*
YOLOv11n				78.1	72.0	77.8	45.5	181.06
	√			78.8	72.5	79.2	46.3	178.45
		√		78.7	72.4	79.0	47.1	179.63
			√	78.6	72.3	78.5	46.0	181.06
	√	√		80.4	72.4	79.7	46.8	177.75
	√	√	√	80.3	73.3	80.1	47.9	177.75

As shown in Table 3, compared with YOLOv11n, the introduction of C3-MCG increased mAP50 by 1.4%, the introduction of DB-SAFM improved mAP50 by 1.2%, and the introduction of AWHIoU raised mAP50 by 1.6%. When C3-MCG was further incorporated on top of DB-SAFM, mAP50 was further improved by 0.7%. Based on the first two modules, the use of AWHIoU increased mAP50 to 80.1%, while recall improved from 72.4% to 73.3%. After finally integrating DB-SAFM, C3-MCG, and AWHIoU together, mAP50 increased by 2.3%, precision improved by 2.2%, and recall increased by 1.3%, while FPS remained at 177.75, fully satisfying the requirements of real-time detection.

DB-SAFM and C3-MCG mainly enhance feature representation and discriminative capability, whereas the AWHIoU loss function further strengthens localisation supervision and improves recall. As a result, the final CSAF-YOLO model significantly improves detection performance without compromising real-time efficiency, which fully verifies its effectiveness and practical value for fire and smoke detection tasks.

Heatmap-based comparison was adopted in this paper, with particular emphasis on the response patterns and differences at Layers 8 and 21. As shown in Figure 8, five examples are presented to compare the feature heatmaps between the baseline model and CSAF-YOLO, where (a) denotes the original image; (b) and (d) represent the feature heatmaps of Layer 8 for the baseline model and CSAF-YOLO, respectively; and (c) and (e) correspond to the feature heatmaps of Layer 21. The comparison between (b) and (d) demonstrates the effect of replacing C3k2 with the improved C3-MCG, while the comparison between (c) and (e) shows that the improved DB-SAFM exhibits better performance than the baseline model in discriminating confusing targets.

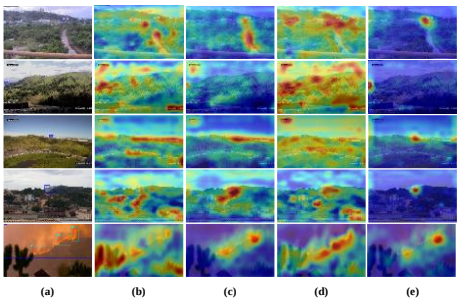

Figure 8: Heatmap comparison

3.5 Comparative experiments

To validate the accuracy and real-time performance of CSAF-YOLO in fire and smoke detection, this paper compares it with various mainstream detection models on the self-built dataset, and uniformly reports evaluation metrics including P, R, mAP50, mAP50–95, as well as FPS.

Table 4: Comparative experiments of different models

Model	*P/%*	*R/%*	*mAP50/%*	*mAP50(95)/%*	*FPS*
SSD	57.9	52.0	51.7	20.5	57.65
Faster R-CNN	64.3	56.3	58.8	24.7	36.76
YOLOv5n	77.6	71.7	77.2	44.6	**182.97**
YOLOv8n	78.1	71.9	77.4	45.4	182.72
YOLOv10n	76.3	71.2	76.7	45.3	180.68
RT-DETR	78.5	72.0	78.1	46.5	146.66
YOLOv11n	78.1	72.0	77.8	45.5	181.06
YOLOv11s	78.5	72.0	78.4	46.3	171.95
CSAF-YOLO	**80.3**	**73.3**	**80.1**	**47.9**	177.75

The specific results are presented in Table 4. The proposed method achieved the best overall performance across multiple evaluation metrics. In terms of precision, CSAF-YOLO improved by approximately 2% over the other models. In terms of recall, it outperformed YOLOv10n by 2.1% and achieved improvements of about 1.5% over the remaining models. For mAP50, CSAF-YOLO improved by 2.3% compared with the baseline model YOLOv11n; except for YOLOv11s, over which the improvement was 1.7%, the gains over all other compared models exceeded 2%. For mAP50–95, the proposed method also achieved substantial improvements over all compared models. Overall, CSAF-YOLO significantly improves recall while maintaining high precision. In terms of real-time detection, it preserves a high level of accuracy without sacrificing too much frame rate.

To provide a more intuitive comparison of the performance of different models, some detection results are visualized in Figure 9, where (a) shows the original images and labels, (b) shows the results of YOLOv5n, (c) shows YOLOv8n, (d) shows YOLOv10n, (e) shows YOLOv11n, (f) shows YOLOv11s, and (g) shows CSAF-YOLO. It can be seen from the figure that, compared with the other models, CSAF-YOLO produces more accurate and comprehensive detection results, without false detections or missed detections.

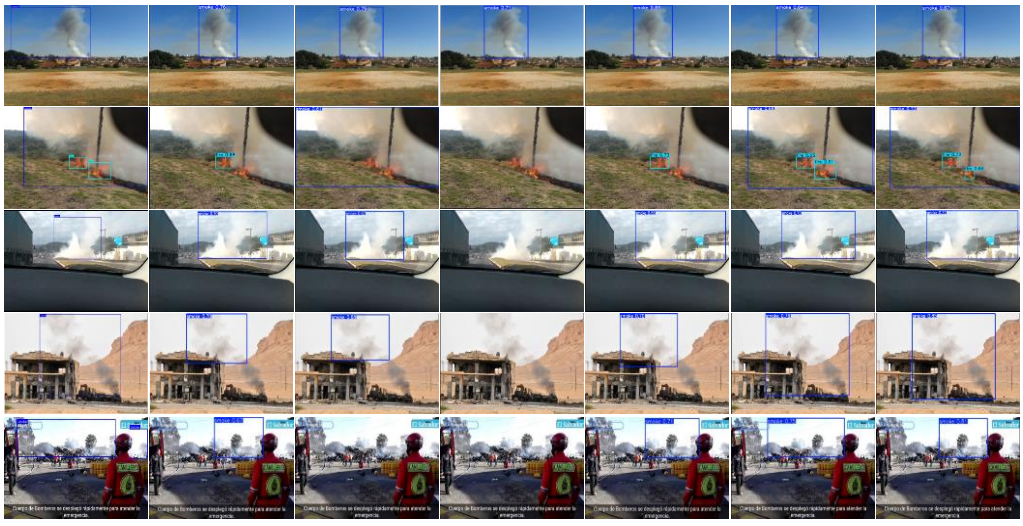

Figure 9: Model detection comparison chart

To further evaluate the detection capability of the model for targets at different scales, the test samples were divided into three groups according to target size, namely Small, Medium, and Large, and the mAP50 of different models was calculated for each group separately. The results are shown in Figure 10.

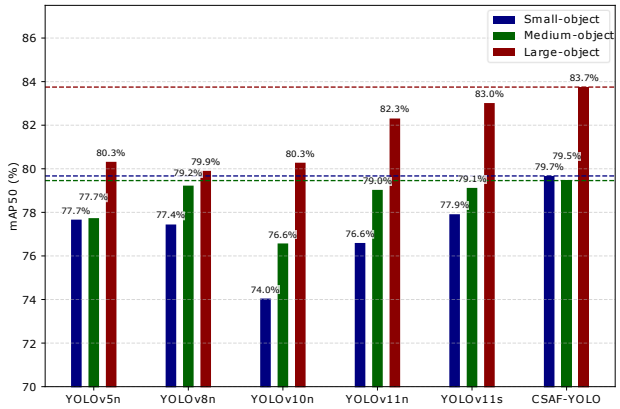

Figure 10: Multi-scale detection performance comparison

CSAF-YOLO achieved the best performance across all three scales. Compared with the Large and Medium groups, the improvement was most pronounced in the Small group, indicating that the C3-MCG and DB-SAFM modules can enhance fine-grained texture representation and cross-layer semantic interaction, thereby significantly improving the recognition of small targets and smoke with blurred boundaries. Meanwhile, AWHIoU further improves the regression quality through adaptive weighting of localisation errors, enabling relatively stable gains for medium- and large-scale targets as well.

CSAF-YOLO exhibits superior detection performance for small-scale smoke targets with blurred boundaries, and the comparative visualization results are presented inFigure 11. In the figure, (a) denotes the original images and corresponding annotations, (b) denotes the detection results of YOLOv11n, and (c) denotes the detection results of CSAF-YOLO. As shown in column (b), YOLOv11n repeatedly suffers from false detections and missed detections when handling small-scale flame and smoke targets. Specifically, small flame targets are missed in the first and fourth rows, whereas small smoke targets are missed in the second and third rows. In contrast, CSAF-YOLO in column (c) does not exhibit either false detections or missed detections for these small targets. These results demonstrate the superiority of the proposed model in small-target detection. In particular, CSAF-YOLO shows stronger discriminative capability for small-scale targets and can more effectively reduce false alarms.

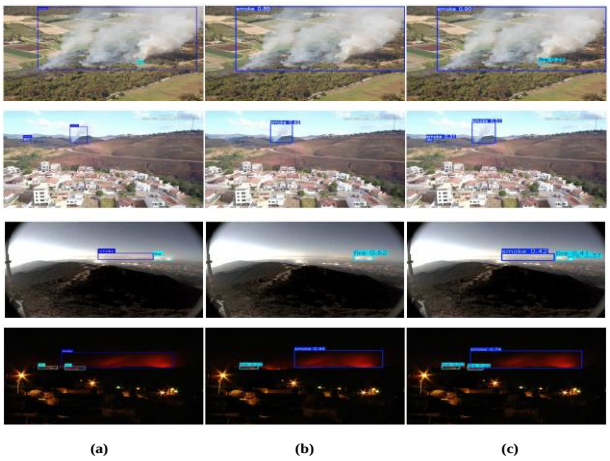

Figure 11: Detection results for some small target smoke

4 Conclusions

To address the challenges of blurred boundaries, small-target recognition, and complex background interference in flame and smoke detection, this paper proposes the CSAF-YOLO model. The model introduces innovations from three aspects: enhancing contextual modelling capability through C3-MCG, strengthening feature interaction and fusion through DB-SAFM, and incorporating AWHIoU together with focusing and weighting mechanisms to improve detection accuracy.

Experimental results show that CSAF-YOLO achieved the best overall performance on the self-built dataset, with a precision of 80.3%, a recall of 73.3%, and an mAP50 of 80.1%. Compared with mainstream detection models such as YOLOv5n, YOLOv8n, YOLOv10n, RT-DETR, YOLOv11n, and YOLOv11s, CSAF-YOLO achieved significant improvements in precision, recall, and mAP. In terms of real-time performance, CSAF-YOLO reached 177.75 FPS, maintaining a high frame rate and satisfying the requirements of real-time monitoring.

Overall, CSAF-YOLO achieves a good balance between detection accuracy and model complexity. While ensuring high precision, it significantly improves recall, enabling efficient flame and smoke detection and demonstrating strong practical application value. Future work will further reduce the computational cost of the model and validate its generalisation ability on larger-scale, multi-scene datasets.

Acknowledgement

This research was supported by the Key Scientific Research Project of Henan Provincial Higher Education Institutions (Grant No.: 25A520033).

Declarations

All authors have been informed and agreed to submit the manuscript, are responsible for its content, and there is no conflict of interest in the submission process.

References

[1] Los Angeles County (2025). *Eaton & Palisades Fires After-Action Reviews*. Available from: https://lacounty.gov/aar/ [Accessed 2026-02-05].

[2] United Nations Chile (2026). *Incendios forestales 2026: Reporte de Situación N°2*. Available from: https://chile.un.org/es/308852-incendios-forestales-2026-reporte-de-situaci%C3%B3n-n%C2%B02 [Accessed 2026-02-05].

[3] Özel, B., M.S. Alam,M.U. Khan (2024). *Review of modern forest fire detection techniques: Innovations in image processing and deep learning*. Information, 15(9), pp. 538. https://doi.org/10.3390/info15090538

[4] Ramadan, M.N.A., T. Basmaji, A. Gad, H. Hamdan, B.T. Akgün, M.A.H. Ali, M. Alkhedher,M. Ghazal (2024). *Towards early forest fire detection and prevention using AI-powered drones and the IoT*. Internet of Things, 27, pp. 101248. https://doi.org/10.1016/j.iot.2024.101248

[5] Gragnaniello, D., A. Greco, C. Sansone,B. Vento (2024). *Fire and smoke detection from videos: A literature review under a novel taxonomy*. Expert Systems with Applications, 255, pp. 124783. https://doi.org/10.1016/j.eswa.2024.124783

[6] Wang, X., M. Li, M. Gao, Q. Liu, Z. Li,L. Kou (2023). *Early smoke and flame detection based on transformer*. Journal of Safety Science and Resilience, 4(3), pp. 294-304. https://doi.org/10.1016/j.jnlssr.2023.06.002

[7] Fernandes, A.M., A.B. Utkin,P. Chaves (2023). *Automatic early detection of wildfire smoke with visible-light cameras and EfficientDet*. Journal of Fire Sciences, 41(4), pp. 122-135. https://doi.org/10.1177/07349041231163451

[8] Chaturvedi, S., C. Shubham Arun, P. Singh Thakur, P. Khanna,A. Ojha (2024). *Ultra-lightweight convolution-transformer network for early fire smoke detection*. Fire Ecology, 20(1), pp. 83. https://doi.org/10.1186/s42408-024-00304-9

[9] Kim, S.-Y.,A. Muminov (2023). *Forest fire smoke detection based on deep learning approaches and unmanned aerial vehicle images*. Sensors, 23(12), pp. 5702. https://doi.org/10.3390/s23125702

[10] Wang, H., X. Fu, Z. Yu,Z. Zeng (2025). *DSS-YOLO: an improved lightweight real-time fire detection model based on YOLOv8*. Scientific Reports, 15(1), pp. 8963. https://doi.org/10.1038/s41598-025-93278-w

[11] Ren, S., K. He, R. Girshick,J. Sun (2017). *Faster R-CNN: Towards real-time object detection with region proposal networks*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp. 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[12] Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu,A.C. Berg (2016). *SSD: Single shot multibox detector*. European conference on computer vision, Springer, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[13] Lin, T.Y., P. Goyal, R. Girshick, K. He,P. Dollár (2020). *Focal Loss for Dense Object Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), pp. 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

[14] Gonçalves, L.A.O., R. Ghali,M.A.J.F. Akhloufi (2024). *YOLO-Based models for smoke and Wildfire Detection in Ground and aerial images*. Fire, 7(4), pp. 140. https://doi.org/10.3390/fire7040140

[15] Khanam, R.,M. Hussain (2024). *Yolov11: An overview of the key architectural enhancements*. arXiv preprint arXiv:2410.17725, pp.

[16] Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan,S. Belongie (2017). *Feature pyramid networks for object detection*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[17] Liu, S., L. Qi, H. Qin, J. Shi,J. Jia (2018). *Path aggregation network for instance segmentation*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[18] Tan, M., R. Pang,Q.V. Le (2020). *EfficientDet: Scalable and efficient object detection*. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781-10790. https://doi.org/10.1109/CVPR42600.2020.01079

[19] Hu, H., J. Gu, Z. Zhang, J. Dai,Y. Wei (2018). *Relation Networks for Object Detection*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3588-3597. https://doi.org/10.1109/CVPR.2018.00378

[20] Wu, T., S. Tang, R. Zhang, J. Cao,Y. Zhang (2021). *CGNet: A Light-Weight Context Guided Network for Semantic Segmentation*. IEEE Transactions on Image Processing, 30, pp. 1169-1179. https://doi.org/10.1109/TIP.2020.3042065

[21] Hu, J., L. Shen,G. Sun (2018). *Squeeze-and-Excitation Networks*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[22] Woo, S., J. Park, J.-Y. Lee,I.S. Kweon (2018). *CBAM: Convolutional Block Attention Module*. Proceedings of the European conference on computer vision (ECCV), pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[23] Ejaz, U., M.A. Hamza,H.-c. Kim (2025). *Channel Attention for Fire and Smoke Detection: Impact of Augmentation, Color Spaces, and Adversarial Attacks*. Sensors, 25(4), pp. 1140. https://doi.org/10.3390/s25041140

[24] Jin, C., A. Zheng, Z. Wu,C. Tong (2023). *Real-time fire smoke detection method combining a self-attention mechanism and radial multi-scale feature connection*. Sensors, 23(6), pp. 3358. https://doi.org/10.3390/s23063358

[25] Tolstikhin, I.O., N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers,J. Uszkoreit (2021). *MLP-Mixer: An all-*

mlp architecture for vision. Advances in neural information processing systems, 34, pp. 24261-24272.

[26] Rezatofighi, H., N. Tsoi, J. Gwak, A. Sadeghian, I. Reid,S. Savarese (2019). *Generalized intersection over union: A metric and a loss for bounding box regression*. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 658-666. https://doi.org/10.1109/CVPR.2019.00075

[27] Zheng, Z., P. Wang, W. Liu, J. Li, R. Ye,D. Ren (2020). *Distance-IoU loss: Faster and better learning for bounding box regression*. Proceedings of the AAAI conference on artificial intelligence, New York, NY, USA, pp. 12993-13000. https://doi.org/10.1609/aaai.v34i07.6999

[28] Song, B., S. Zhao, Z. Wang, J. Chen, W. Liu,X. Liu (2025). *LW-DETR: a lightweight transformer-based object detection algorithm for efficient railway crossing surveillance*. Complex & Intelligent Systems, 11(12), pp. 480. https://doi.org/10.1007/s40747-025-02111-4

[29] Xu, L., Y. Zhao, Y. Zhai, L. Huang,C. Ruan (2024). *Small Object Detection in UAV Images Based on YOLOv8n*. International Journal of Computational Intelligence Systems, 17(1), pp. 223. https://doi.org/10.1007/s44196-024-00632-3

[30] de Venâncio, P.V.A.B., A.C. Lisboa,A.V. Barbosa (2022). *An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices*. Neural Computing and Applications, 34(18), pp. 15349-15368. https://doi.org/10.1007/s00521-022-07467-z