Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatiotemporal Prior-Enhanced Fire Perception Network for Edge Alarms in Industrial LANs

Bin Hu
Tianjin University of Technology,Technical College for the Deaf
Jun Li
604372036@qq.com

Shigatse Everest Urban Investment and Development Group Co., Ltd.
Xufeng Lou
Sinohydro Foundation Engineering Co.
Yuqiao Li
Sinohydro Foundation Engineering Co.
Qianhong Wan
Tianjin University,School of Microelectronics
Jiaji He
Tianjin University,School of Microelectronics
Ju Zhang
Tianjin University of Technology,Technical College for the Deaf

Research Article

Keywords: Industrial LAN, Edge Intelligence, Fire Detection, Event-driven Alarm, YOLO, Communication Payload Optimization

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11171049/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spatiotemporal Prior-Enhanced Fire Perception Network for Edge Alarms in Industrial LANs

Bin Hu[1], Jun Li[2*], Xufeng Lou[3], Yuqiao Li[3], Qianhong Wan[4], Jiaji He[4], Ju Zhang[1]

[1]Tianjin University of Technology,Technical College for the Deaf, Tianjin, China.
[2*]Shigatse Everest Urban Investment and Development Group Co., Ltd., Linzhi, China.
[3]Sinohydro Foundation Engineering Co., Tianjin, China.
[4]Tianjin University,School of Microelectronics, Tianjin, China.

*Corresponding author(s). E-mail(s): 604372036@qq.com;

Abstract

Multi-camera fire monitoring in industrial local area networks is a challenging research problem requiring low-latency alarms under stringent uplink bandwidth constraints. Due to intense lighting and specular reflections, conventional edge vision models struggle with false positives, triggering unnecessary uploads of invalid region-of-interest ROI evidence. To address this, this empirical study improves the YOLO baseline by introducing a Prior-Aware Gating Block, PAGB, into the Backbone, which suppresses background noise and generates minimal semantic ROI evidence. Additionally, to mitigate strong-light false alarms without misfiltering steady-state flames, a Prior-guided Kinematic Attention Block, PKAB, is embedded into the Neck to capture short-term flicker and non-rigid motion cues via a negative-bias dual-branch fusion mechanism. Compared with the conventional YOLO baseline, the proposed PGENet achieves a Fire single-class **mAP@0.5** of 0.963 and, on the dedicated high-exposure interference subset, reduces the false-positive event rate from 36.2% to 2.1% at 32 FPS on the Huawei Atlas 200 DK. Under a detection-triggered edge architecture, PGENet reduces the measured single-alarm payload from 152 KB to 113 KB. Edge-device measurements show improvements in both detection accuracy and communication payload.

Keywords: Industrial LAN, Edge Intelligence, Fire Detection, Event-driven Alarm, YOLO, Communication Payload Optimization

1 Introduction

Industrial safety monitoring is transitioning toward intelligent systems powered by edge computing to alleviate network transmission burdens and minimise response latency. Across diverse environments like expansive forests [1] and specialised mining operations [2], hundreds of cameras routinely connect to industrial local area networks. During high-risk incidents like fires, continuous high-definition video streaming heavily burdens uplink capacity. Reliable edge perception must therefore filter non-event frames at the source and upload only event-level alarms and region-of-interest evidence. However, high-intensity interference prevalent in complex settings such as indoor facilities [3] and aviation hangars [4] hampers this closed-loop efficiency. Strobe searchlights and specular metallic reflections create brightness patterns that closely mimic flames, making conventional spatial detectors prone to false positives. Consequently, every visual misdetection triggers an invalid alarm and a redundant image upload, inadvertently turning visual errors into communication bottlenecks.

To improve visual fire detection, early research primarily focused on refining general YOLO architectures. Comparative studies of YOLOv9 through YOLOv11 confirm that newer single-stage detectors offer a strong baseline for accuracy and speed in fire and smoke recognition [5]. Building upon this, Bayesian hyperparameter optimisation [6] and multi-scale attention mechanisms [7] have been utilised to enhance feature extraction. To address the diverse environmental challenges mentioned above, recent research has introduced numerous structural innovations. Methods like YOLOGX, DSS-YOLO, and FCMI-YOLO optimise feature fusion and real-time inference [8–10], while YOLO-ULNet targets ultra-lightweight embedded operation [11]. Other advanced designs explore longer-range spatial state modelling [12], lightweight wildfire classification [13], and end-to-end feature constraints embedded directly within the Backbone and Neck [14].

Meanwhile, embedding visual models in edge computing frameworks is gaining traction in applications such as vehicular radar video fusion [15], UAV target localisation [16], drone swarm firefighting [17], and federated learning-based forest monitoring [18]. However, existing literature primarily evaluates these models using single-frame spatial visual metrics, implicitly assuming flawless front-end perception. This overlooks the system-level consequences of harsh-lighting false positives, where each misdetection inflates the semantic evidence payload and exacerbates network congestion. To address this in resource-constrained industrial local area networks, edge detectors must use dynamic priors to suppress strong light artefacts and compress the uploaded region-of-interest evidence. Consequently, this paper proposes PGENet, a prior-guided edge network. Evaluated under a unified edge-triggered architecture that uploads evidence slices strictly upon positive detection, the main contributions of this study are threefold:

- A lightweight spatiotemporal prior-enhanced architecture is proposed to augment the YOLO framework with a spatial gating mechanism in the backbone and a short-term dynamic attention mechanism in the neck. Using flicker and non-rigid motion cues, PGENet filters out static highlight artefacts and achieves reliable detection in

high-exposure industrial environments without changing the native detection head or post-processing pipeline.

- A perception-driven compression strategy is proposed to minimise the uploaded semantic region-of-interest evidence: rather than lowering image quality, the spatial gating mask zeros out low-response background pixels inside the detection box, compacting the payload at the perception source while preserving core flame textures for secondary manual review.
- An edge-device-in-the-loop payload measurement framework is proposed to decouple measured single-link payloads from scenario-level open-loop estimates, thereby quantifying the reduction in source-side invalid offered load achieved by jointly lowering the false-positive frequency and the region-of-interest evidence volume.

2 Related Work

2.1 Fire Detection in Complex Environments

Early fire detection methods relied on static colour features, such as RGB and HSV colour spaces, or dynamic textural properties for classification using support vector machines [19, 20]. The advent of deep convolutional neural networks has improved flame and smoke representation [21]. YOLO-based single-stage detectors have become a standard choice for visual monitoring due to their effective balance of accuracy and speed [22]. Comparative analyses of recent miniaturised models show competitive baseline performance on public fire datasets [5]. From an optimisation perspective, Bayesian hyperparameter tuning accelerates convergence [6], while the integration of multi-scale and spatial attention mechanisms improves localisation performance [7]. A major research direction involves model lightweighting for edge deployment in resource-constrained environments. Architectures such as FFYOLO and YOLOv8n-SMMP enhance forest fire detection efficiency by restructuring lightweight convolutions and feature fusion pathways [23, 24]. Similarly, YOLO-LFD and YOLO-MP seek to maintain deep semantic extraction while reducing model size [25, 26]. LEF-YOLO incorporates a streamlined backbone specifically designed to distinguish between varying wildfire intensities [13]. FCMI-YOLO addresses the trade-off between accuracy and inference speed to enable real-time detection on constrained edge devices [10].

Recent research has increasingly focused on adapting lightweight detection models to suppress scene-specific false positives in complex environments. In indoor scenarios, models such as YOLO-HF are designed to detect early-stage small fire sources, thereby reducing missed detections and false alarms that frequently occur in cluttered home environments [3]. Optimization studies of YOLOv5 have shown that dynamic indoor lighting conditions can lead to delayed responses and false alarms in conventional visual algorithms [27]. For specialized confined spaces, such as aircraft cargo holds and hangars, detection models are calibrated to minimize background noise while ensuring rapid localization of fire sources [4]. Research has also addressed challenging outdoor conditions, including the detection of small-scale smoke targets and the management of extreme scale variations in forest fires [1]. YOLO-SIFD utilizes slice inference and fractal dimension analysis to improve recognition accuracy [28].

While these approaches perform well and remain lightweight, most rely primarily on single-frame spatial features. This reliance increases susceptibility to unique temporal patterns caused by industrial strobe searchlights or sensor overexposure, highlighting the need for models that incorporate short-term kinematic priors.

2.2 Edge Computing and Communications for AI

In the context of industrial local area networks and edge-assisted Internet of Things, communications for artificial intelligence and semantic communication have shifted from transmitting raw pixel data toward conveying task-relevant semantic information, so that field devices transmit compact semantic features or event-triggered region-of-interest slices. Representative applications include vehicular edge networks that fuse radar point clouds with video streams for three-dimensional perception [15], low-altitude unmanned aerial vehicles that use enhanced YOLOv11 models to localise targets in complex terrain [16], drone swarms that coordinate fire source detection with multi-sensor path planning [17], and distributed systems such as ForestGuard that combine YOLO detection with federated learning to reduce centralised data aggregation [18]. These paradigms, however, largely assume that front-end visual perception operates without error and rarely account for false positives under challenging industrial lighting, where each misdetection triggers unnecessary event signalling, region-of-interest image encoding, and uplink transmission. To close this gap, PGENet incorporates spatiotemporal visual weak priors to suppress environmental artefacts within the feature pipeline, reducing both the frequency of false alarms and the volume of semantic evidence transmitted.

3 System and Communication Model

3.1 Edge Alarm Architecture

This paper considers a monitoring LAN composed of cameras, edge inference boxes, such as embedded neural processing units (NPUs), wired switches or access points, and local central control platforms. Edge nodes execute local inference on short video clips $\mathcal{V} = \{I_1, I_2, \ldots, I_T\}$; only when the model determines the presence of a fire target does it upload event semantics and ROI evidence images to the central platform. A single edge alarm can be denoted as $\mathcal{S} = \{y, p, \mathcal{B}, \tau, \mathcal{E}, d\}$, where y is the event type, p is the confidence, $\mathcal{B}$ is the bounding box, τ is the timestamp, $\mathcal{E}$ is the ROI evidence image, and d is the device ID. The goal of this design is not to backhaul continuous video, but to make the uploaded content as close to the "minimum effective flame ROI" as possible while ensuring the visual evidence required for manual review. Symbols in subsequent formulas are explained locally upon first appearance, and conventional object detection notations follow common definitions in YOLO series literature.

3.2 Communication Payload Model

This section establishes an application-layer communication payload model to quantify how visual false positives burden the LAN uplink. Under the proposed detection-triggered architecture, a fire alarm packet comprises semantic metadata and an ROI

evidence image. Let s_m denote the size of the metadata, such as category, confidence, coordinates, timestamp, and device ID and s_r denote the size of the JPEG-encoded ROI image. The single-alarm payload s_e is expressed as:

$$s_e = s_m + s_r. \tag{1}$$

Because s_m is typically under 1 KB, the communication overhead is dominated by s_r. While standard YOLO baselines upload the entire rectangular bounding box, PGENet utilises the binary mask generated by the PAGB to zero out low-response background pixels. This operation preserves the detection trigger rules but reduces the JPEG-encoded s_r. To quantify this compression, we define r_u as the relative upload volume ratio of masked ROI bytes to unmasked bounding box bytes.

In industrial scenarios devoid of actual fires but prone to strong light interference, invalid ROI uploads constitute the primary uplink load. Let P_{fp} denote the false-positive event rate, defined as the fraction of high-exposure interference events that produce at least one false positive under the fixed confidence threshold τ_p. Each interference event is a short clip of consecutive frames induced by one strobe or specular-reflection episode; an event is counted as a false positive if at least one of its frames yields a detection box. Assuming a facility operates N_c cameras, each encountering an average of λ_h intense lighting events per hour, the expected invalid-upload frequency in events per hour is

$$\nu_f = N_c \, \lambda_h \, P_{\text{fp}}. \tag{2}$$

To avoid redundant uploads for events with multiple false boxes, the edge device transmits only the highest-confidence ROI per event. Consequently, the hourly invalid offered load L_f injected from the perception source is calculated as:

$$L_f = \nu_f s_e. \tag{3}$$

Therefore, PGENet reduces communication load through two orthogonal pathways: decreasing the invalid upload frequency ν_f by reducing P_{fp} via the PKAB module, and compressing the single-upload volume s_e via the PAGB module. Because L_f is linear in $N_c\lambda_h$, the overall invalid-load reduction rate relative to a standard edge baseline, defined as $\eta = 1 - L_f^p/L_f^b$, is independent of the camera/event scale.

4 Proposed Method

4.1 Overall Architecture

To mitigate false positives and redundant uplink loads caused by high exposure artefacts, this study introduces PGENet as illustrated in Figure 1. Rather than appending external filters, PGENet applies lightweight enhancements directly within the YOLO feature pipeline. A Prior Aware Gating Block, PAGB, integrates into multi-scale backbone outputs to generate spatial masks, suppress background noise, and produce ROI

evidence. Concurrently, a Prior-guided Kinematic Attention Block, PKAB, is embedded at the neck cross-scale fusion nodes. By maintaining a historical feature buffer of $T = 4$, PKAB utilizes short-term dynamic attention to differentiate genuine flame kinematics from static highlight artefacts.

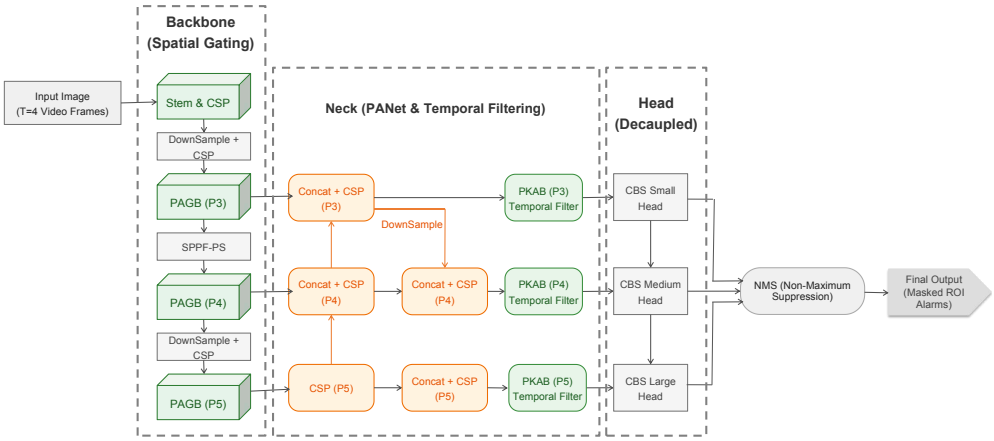


Fig. 1: PGENet Network Architecture.

The prior modelling in PGENet does not rely on physics-informed neural networks to solve thermodynamic equations. Instead, it translates observable visual weak priors, such as local sparsity, edge flicker, and forward expansion, into lightweight tensor transformations suitable for real-time edge inference. Formally, given a sliding analysis window of the input video stream $\mathcal{V}_T = \{I_{t-T+1}, \ldots, I_t\}$, the Backbone extracts single-frame spatial features $F^b_{l,\tau} \in \mathbb{R}^{C_l \times H_l \times W_l}$ across the multi-scale set $\mathcal{L}_b$. Following spatial modulation by the PAGB, these features are pushed into an on-device circular queue Q_g. The Neck stage then retrieves T historical frames from Q_g, where the PKAB injects dynamic kinematic priors. Finally, the decoupled detection head processes the fused spatiotemporal features to predict the event alarm tuple $\{y, p, \mathcal{B}\}$ for the current frame I_t. Here, the term *prior* denotes a learned, task-specific inductive bias expressed as a gating or attention tensor, not a Bayesian or physical prior.

4.2 PAGB Module

The industrial monitoring field of view often contains metal components, floors, and factory backgrounds, which typically contribute minimally to open flame recognition. PAGB is embedded at multi-scale nodes $\mathcal{L}_b$ to adaptively establish single-channel spatial saliency topologies:

$$M_{l,\tau} = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{SiLU}\left(\mathrm{BN}\left(F^b_{l,\tau}\right)\right)\right)\right). \tag{4}$$

where $M_{l,\tau} \in [0,1]^{1\times H_l \times W_l}$ is broadcast to C_l channels along the channel dimension during feature modulation. To prevent drastic gating contraction in early training from

causing gradient vanishing or feature representation collapse, the feature modulation adopts a residual gating with an elastic lower bound ϵ:

$$\widetilde{F}_{l,\tau} = F^{b}_{l,\tau} \odot (\epsilon + (1-\epsilon) M_{l,\tau}), \quad 0 < \epsilon < 1, \tag{5}$$

where $\odot$ represents the Hadamard product broadcast along the channel dimension; $\mathrm{Conv}_{1\times1}$ indicates a lightweight convolution compressing the C_l channel features of the l-th scale into a single-channel spatial gating map without changing the spatial resolution. In every scale, the gating convolution maps $\mathbb{R}^{C_l \times H_l \times W_l}$ to $\mathbb{R}^{1\times H_l \times W_l}$, and the elastic lower bound ϵ is frozen on the validation set. The intermediate layer spatial active ratio ρ_a is defined to evaluate spatial focus:

$$\rho_a = \frac{1}{|\mathcal{L}_b| T} \sum_{l\in\mathcal{L}_b} \sum_{\tau=t-T+1}^{t} \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \mathbb{I}\left(M_{l,\tau,h,w} > \tau_m\right).$$

where τ_m is the binarisation decision threshold. In experiments, τ_m is frozen at 0.5 on the validation set; this metric is solely used to analyse intermediate feature spatial focus and is not directly equivalent to the actual uploaded byte ratio r_u after JPEG encoding. It should be noted that PAGB participates in feature modulation internally via soft gating $M_{l,\tau}$; in the ROI evidence encoding stage, to achieve stable background zeroing, this paper applies the highest-resolution PAGB mask, binarised by τ_m, to the RGB ROI.

In the evidence mask export stage, a pixel-level RGB ROI cropping mask must be generated based on the feature maps. Since high-level semantic features are discriminative but suffer considerable spatial downsampling loss, this paper therefore locks the highest spatial resolution scale $l^\star = \arg\max_{l\in\mathcal{L}_b} H_l W_l$, that is, the P3, stride-8 node, for evidence mask export. This scale mask is restored to the original image coordinates via bilinear interpolation and intersected with the target box $\mathcal{B}$, balancing boundary alignment accuracy and computational overhead. The final ROI evidence is $\mathcal{E} = \mathcal{B} \cap \{M_{l^\star} \geq \tau_m\}$, i.e., the intersection of the detected box and the binarised high-resolution mask.

4.3 PKAB Module

Constant or slowly drifting high-light artefacts, including direct searchlights and vehicle light reflections, easily induce single-frame spatial misjudgments. To address this issue, PKAB is located at the Neck cross-scale fusion nodes, stacking the cached gated features into a time-tube tensor $\mathbf{F}_l$. The module first maintains spatial alignment and applies a lightweight depthwise separable temporal convolution along the time dimension, forming a bandpass filtering response to capture combustion pulsations:

$$M^{f}_{l} = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Pool}_{T}\left(\mathrm{Conv}_{1\mathrm{D}}\left(\mathbf{F}_l\right)\right)\right) + b_0\right). \tag{6}$$

The temporal pooling operation compresses the time dimension into a two-dimensional attention map, while a subsequent point-wise convolution aligns the channel dimension. Specifically, $\mathbf{F}_l \in \mathbb{R}^{B\times T\times C_l\times H_l\times W_l}$, the depthwise temporal convolution uses a

small kernel with causal padding to preserve temporal alignment, and Pool_T averages over the T dimension. A negative bias initial value is specifically set to suppress static zero responses.

Simultaneously, rather than relying on computationally exorbitant optical flow methods, PKAB utilises adjacent frame differences to represent bright area expansion. Real flames exhibit both intensification and attenuation. However, automatic exposure fallback and sensor saturation recovery commonly encountered in industrial environments produce large areas of negative differences. Using absolute differences would inherently amplify these camera imaging artefacts. Therefore, a residual-preserving forward difference operator $D_{l,\tau}$ is designed to primarily respond to feature enhancement and local expansion, leaving negative changes to be handled jointly by the flicker branch and the main detection loss:

$$D_{l,\tau} = \max\left(0,\ \widetilde{F}_{l,\tau} - \widetilde{F}_{l,\tau-1}\right), \tag{7}$$

where $\widetilde{F}_{l,\tau}$ denotes the PAGB-modulated backbone feature at scale l and time τ.

$$M_l^k = \sigma\left(\mathrm{Conv}_{1\times1}\left(\frac{1}{T-1}\sum_{\tau=t-T+2}^{t} D_{l,\tau}\right) + b_0\right). \tag{8}$$

Fusing these two branches via simple element-wise multiplication would suppress the entire attention pathway when a real fire enters a stable combustion phase with minimal area expansion. To prevent this, a union-style complementary fusion is adopted:

$$A_l = M_l^f + M_l^k - M_l^f \odot M_l^k. \tag{9}$$

This formulation supports dual triggering. The early spread phase remains sensitive primarily through the motion branch, while the steady combustion phase remains activated through high-frequency flicker. This approach avoids the need for additional fusion weights, retains continuous gain when both branches respond, and prevents the complete suppression of steady flame responses. Since $M_l^f, M_l^k \in [0,1]$, Eq. 9 equals $1-(1-M_l^f)(1-M_l^k)$ and is monotonically non-decreasing in both branches, which guarantees that neither a static flicker response nor a weak motion response can zero the fused attention.

A standard Sigmoid activation outputs 0.5 under zero input, leaving static reflective areas that exhibit neither flicker nor expansion with a high activation response. For this reason, a negative bias b_0 is injected into both branches so that $\sigma(0+b_0) < 0.5$. Although the fused response in Eq. 9 remains strictly positive, min-max normalisation maps static backgrounds to the low end of the attention range:

$$\overline{A}_l = \frac{A_l - \min_{h,w} A_l}{\max_{h,w} A_l - \min_{h,w} A_l + \delta_A}\, \mathbb{I}\left(\max_{h,w} A_l - \min_{h,w} A_l \geq \epsilon_A\right), \tag{10}$$

where $\delta_A > 0$ is a small constant that prevents division by zero, and ϵ_A is a dynamic-range noise threshold, empirically set to 10^{-3}. When the range of A_l is below ϵ_A, the

scene is treated as static, $\overline{A}_l$ is set to zero, and the attention branch does not amplify numerical noise. The normalised attention is then applied with a residual lower bound ϵ_k for gradient stability; the final suppression of static regions is enforced jointly by the detection loss and, at inference, by the binarised PAGB mask used for ROI encoding:

$$\hat{F}_{l,t} = F^n_{l,t} \odot \left(\epsilon_k + (1-\epsilon_k)\overline{A}_l\right), \quad 0 < \epsilon_k < 1. \tag{11}$$

The bracketed term expands along the channel dimension to match the input feature channels.

4.4 Training Objectives and Fallback Strategy

To achieve unified optimisation of visual perception accuracy and communication payload compression, the network undergoes end-to-end training driven by a composite loss function. Relying solely on the native detection loss $\mathcal{L}_{\text{det}}$ optimises bounding box regression and classification but does not constrain the spatial activation of background regions. To bridge this gap, a dedicated mask sparsity term is introduced to penalise redundant feature activations outside the primary flame areas. By combining these objectives, the model learns to localise targets and suppress background noise. The composite total loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_s \mathcal{L}_{\text{m}}. \tag{12}$$

The sparsity term $\mathcal{L}_{\text{m}}$ penalises soft gating responses outside the ground-truth fire regions and is defined as

$$\mathcal{L}_{\text{m}} = \frac{1}{|\mathcal{L}_b|\,T} \sum_{l\in\mathcal{L}_b} \sum_{\tau=t-T+1}^{t} \frac{1}{H_l W_l} \left\|(1-\mathcal{G}_{l,\tau}) \odot M_{l,\tau}\right\|_1, \tag{13}$$

where $\|\cdot\|_1$ denotes the element-wise ℓ_1 norm of the flattened spatial map, and $\mathcal{G}_{l,\tau}$ is the binary fire mask obtained by downsampling the ground-truth box to scale l. The penalty therefore acts only outside the fire region and does not suppress weak flames inside the box, while $\mathcal{L}_{\text{det}}$ preserves flame-region responses; the balance is controlled by λ_s, whose value is frozen on the validation set. During edge deployment, this learned sparsity lets the spatial mask zero out low-response backgrounds prior to image encoding, minimising the uploaded payload at the perception source. Additionally, evaluating dynamic temporal models on static images inherently yields zero adjacent frame differences, which would falsely trigger temporal suppression. The network resolves this by implementing a static fallback protocol that preserves the backbone gating weights while safely bypassing the temporal attention branch.

5 Experiments and Discussion

5.1 Experimental Setup

The experiments evaluate model performance using a self-collected industrial fire monitoring dataset comprising 120 continuous video sequences. These sequences capture brazier fires and background interference under varying exposure conditions. Annotations target only visible open flames and exclude smoke or formless, high-brightness regions. To prevent temporal data leakage, we partition the training, validation, and testing subsets mutually exclusively by video sequence, as detailed in Table 1. Alongside the fire samples, we independently collected a dedicated high-exposure optical interference subset comprising 5200 non-fire frames. This subset serves exclusively as a stress-test benchmark for false-positive suppression and communication-payload estimation. An interference event represents a short clip of consecutive frames capturing a single strobe or specular reflection episode, where the event-level false-positive rate P_{fp} is computed by checking whether at least one false box appears within the pooled frames of the event.

Table 1: Self-collected dataset statistics.

Subset	Purpose	Frames	Labeled	Fire boxes
HEFD-Train	Model training	7,411	7,392	7,923
HEFD-Val	Hyperparam & threshold freezing	958	956	1,002
HEFD-Test	General metric evaluation	498	494	523
HEOI-Blind	Non-fire high-exposure stress	5,200	0	0

All models are trained using PyTorch 2.0 with a 640 by 640 input resolution, a batch size of eight, and the stochastic gradient descent optimizer over 100 epochs. For the proposed architecture, we freeze the spatial gating binarization threshold and temporal negative bias after validation set convergence. We measure deployment performance on the Huawei Atlas 200 DK embedded edge device in half-precision floating-point format. The reported time statistics include image preprocessing and non-maximum suppression overhead but exclude cold-start buffer filling at a single batch size.

5.2 Evaluation Metrics

To align with general evaluation standards in object detection research, this study concurrently reports four types of metrics: classification, localisation quality, edge deployment, and communication payload. True positives, false positives, and false negatives are represented by TP, FP, and FN respectively. For the high-exposure interference benchmark, the false-positive event rate is $P_{\text{fp}} = N_{\text{fp}}/N_{\text{int}}$, where N_{int} is the number of interference events and N_{fp} is the number of events containing at least one false-positive box; the invalid-upload frequency ν_f follows from Eq. 2.

Precision, denoted P, measures the accuracy of positive predictions:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{14}$$

Recall, denoted R, measures the fraction of actual fire samples that are detected:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{15}$$

The F1 score balances precision and recall:

$$\mathrm{F1} = \frac{2PR}{P + R}. \tag{16}$$

For localisation, average precision integrates the precision-recall curve $P(R)$:

$$\mathrm{mAP} = \int_0^1 P(R)\, dR. \tag{17}$$

The metric mAP@0.5 is evaluated at a fixed intersection-over-union threshold, while mAP@0.5:0.95 averages over thresholds from 0.50 to 0.95 following standard evaluation protocols. Edge deployment efficiency is represented by Atlas FPS, and computational complexity uses end-to-end GFLOPs covering the complete forward path.

Communication payload metrics follow the definitions established previously. The upload volume ratio is $r_u = s_r^m / s_r^u$, where s_r^m and s_r^u denote the JPEG-encoded ROI sizes with and without mask zeroing. The single-alarm payload is s_e, the invalid-upload frequency is ν_f, and the hourly invalid offered load is L_f, with superscripts p and b denoting the proposed and baseline systems. To quantify the communication efficiency gains, the load reduction rate relative to the standard edge baseline is formulated as:

$$\eta = 1 - \frac{L_f^p}{L_f^b}. \tag{18}$$

5.3 Comparative Experiments

To verify PGENet's performance in industrial scenarios, Figure 2 first visualises the detection results of the compared detectors under high exposure. Table 2 then compares PGENet with mainstream lightweight general detectors and representative models optimised for fire detection [8, 14], and also includes the heavyweight vision transformer RT-DETR-L [29] as a cloud-level reference. RT-DETR-L is also profiled on the Atlas 200 DK, where its Transformer backbone reaches only 11 FPS, reinforcing why lightweight detectors are preferred at the edge. We evaluate all lightweight baselines under identical training sequences, input dimensions, and data augmentation strategies to ensure a fair comparison. To validate the need for structural temporal modelling, we introduce an early-fusion control group by concatenating four frames at the input channel level of a standard YOLOv11n network.

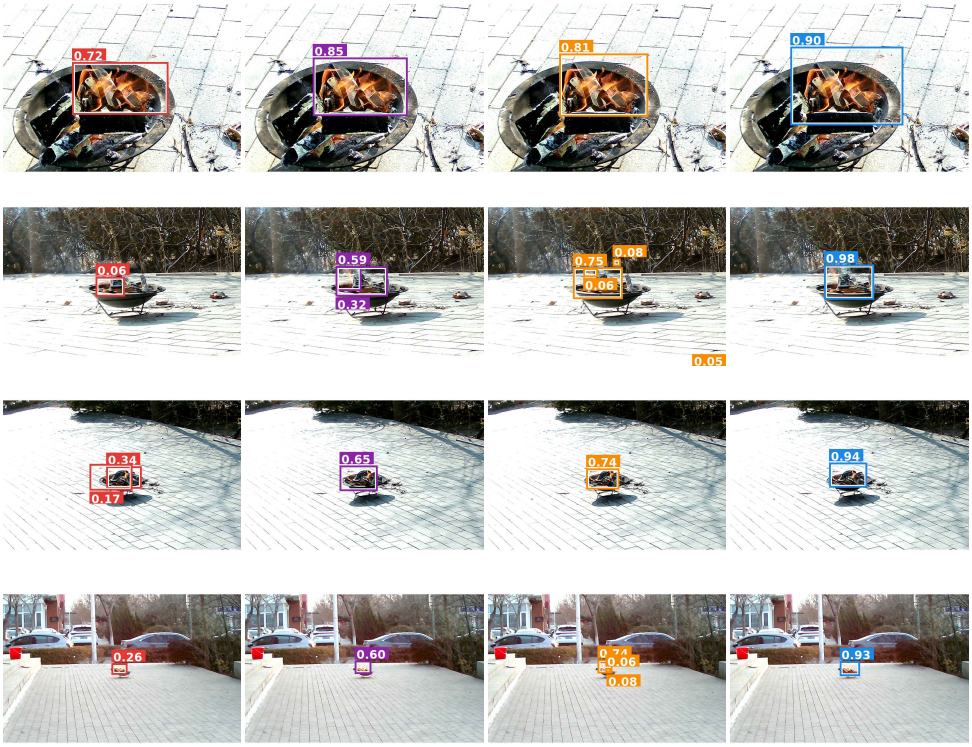


Fig. 2: Detection results of the compared edge detectors on high-exposure HEFD scenes. Columns from left to right: YOLO-FireAD, YOLOv11n, RT-DETR-L, PGENet.

As shown in Figure 2, for qualitative visualization, we also include the heavyweight vision transformer RT-DETR-L [29] as a cloud-level reference. Under high-exposure backgrounds, both the lightweight edge baselines and the heavyweight RT-DETR-L are easily pulled toward ground reflections and saturated textures, producing loose or repeated boxes that enclose large background regions. Notably, RT-DETR-L emits up to five candidate boxes, most of them false, on the mid-range scene, while the lightweight baselines shift markedly from the flame center on distant targets. In contrast, the spatial gating in PGENet constrains the response to the combustion region, yielding a single tight box that stays close to the ground truth across all scenes.

With spatiotemporal multi-frame prior enhancement, PGENet outperforms all baselines across all accuracy metrics. The early-fusion control group yields only marginal improvements over the single-frame baseline, confirming that networks lacking explicit prior guidance cannot distinguish short-term flickers from static highlight artefacts. PGENet's performance gain arises directly from its structured prior design, which integrates historical feature caching to isolate transient flame dynamics from environmental noise while maintaining real-time processing on the edge device. In

Table 2: Performance comparison of different detection models.

Method	F1	mAP50	mAP50:95	Atlas FPS
YOLOv8n [24]	0.81	0.824	0.512	58
YOLOv10-N [5]	0.82	0.831	0.526	52
YOLO-FireAD [14]	0.86	0.875	0.568	38
YOLOv11n [7]	0.85	0.868	0.554	45
YOLOGX [8]	0.85	0.862	0.549	36
RT-DETR-L [29]	0.88	0.891	0.598	11
YOLOv11n T=4 Early-Fusion (control)	0.86	0.879	0.562	41
PGENet Proposed T=4	**0.96**	**0.963**	**0.685**	32

T denotes the temporal window length; Early-Fusion is a temporal-modelling control group, and all other baselines process a single frame.

contrast, PGENet maintains high recall at standard thresholds while reducing the false-positive event rate. This demonstrates that the capability to resist strong light artefacts stems primarily from the structural decoupling and suppression of static bright area features within the network architecture, a robustness that threshold tuning alone cannot easily achieve.

5.4 Ablation and Feature Activation Analysis

To explore the orthogonal suppression effects of the proposed modules on high-exposure artefacts, a hard subset is selected from the 498 test frames using a deterministic criterion: a frame is retained if it contains intense reflections, strong backlighting, or local saturation, identified by the baseline YOLOv11n producing a false-positive box or a true-positive box with an IoU below 0.5. The F1 score is evaluated on this hard subset while the false-positive event rate is calculated on the non-fire interference subset. The corresponding results are presented in Table 3.

Table 3: PGENet module ablation study.

Model Config	F1 Hard	FPR on HEOI (%)	GFLOPs
YOLOv11n Baseline	0.55	36.2	6.5
YOLOv11n Early Fusion	0.58	31.5	6.8
YOLOv11n plus PAGB	0.82	32.1	6.9
YOLOv11n plus PKAB	0.89	6.2	7.3
PGENet Proposed	**0.95**	**2.1**	**7.7**

Table 3 reports the module contributions across two dimensions. Regarding upload frequency, the temporal attention block independently lowers the false-positive event rate on the HEOI benchmark from 36.2% to 6.2%, whereas the early-fusion baseline remains prone at 31.5%, confirming that modelling short-term flicker and non-rigid motion is essential to distinguish genuine flames from static highlight artefacts.

Regarding feature sparsity, the spatial gating block compresses the feature-layer spatial active-area ratio to 23.5 percent at a marginal computational cost, and PKAB leaves this ratio unchanged because it operates only along the temporal dimension, providing the basis for subsequently compressing the single-alarm evidence volume. As a result, the standard baseline maintains a high F1 score on the overall test set but collapses to 0.55 on the hard samples, whereas PGENet sustains 0.95 on the same hard batch, demonstrating the benefits of spatiotemporal weak priors under industrial lighting interference.



Fig. 3: Prediction-ROI-conditioned feature activation visualizations for scene 238 under increasing exposure. Columns from left to right: YOLO-FireAD, YOLOv11n, RT-DETR-L, PGENet.

Figure 3 visualizes the intermediate feature activations conditioned on each model's predicted ROI under three increasing exposure levels. As exposure rises, the single-frame baselines rely purely on spatial appearance, so saturated highlights and specular reflections—whose bright, flame-like textures are spatially indistinguishable from genuine fire—progressively capture their attention and scatter activation toward the background. PGENet, by contrast, keeps its response anchored on the combustion region through two complementary priors: the PAGB gate first enforces spatial sparsity by suppressing low-response background, concentrating the representation on the salient foreground; the PKAB branch then injects short-term kinematic priors such as flicker and non-rigid expansion, so attention is reinforced only where the evidence behaves like a genuine flame over time, while static glare is penalised by the negative bias. As a result, the baselines' activation becomes increasingly diffuse as exposure rises. In contrast, PGENet retains a compact, task-aligned activation on the actual flame, reflecting the benefit of reasoning jointly over spatial gating and temporal kinematics rather than appearance alone.

Table 4: Ablation study on temporal window lengths.

Window	F1 Hard	FPR on HEOI (%)	Edge FPS	GFLOPs
1	0.82	32.1	39	6.9
2	0.90	14.3	35	7.2
4	**0.95**	**2.1**	**32**	**7.7**
8	0.96	1.8	27	8.8

As shown in Table 4, extending the temporal window from one to two frames reduces the false-positive event rate. Still, adjacent-frame differences remain insufficient to capture the full cycle of typical industrial searchlight flickers. Increasing the window to eight frames further decreases the false-positive event rate, but the edge device must maintain a larger historical feature cache. This increased cache dependency raises cold-start buffer latency, given by $(T-1)/\text{fps}$, and peak memory usage. Weighing detection accuracy against anti-interference capability and memory consumption, this study selects a four-frame window as the preferred configuration.

5.5 Communication Payload Analysis and Practical Limitations

Traditional communication payload reduction typically relies on lowering overall image resolution or image quality, which unavoidably degrades fire-event texture and compromises secondary visual review by central control personnel. The proposed spatial gating mechanism instead performs task-semantics-preserving lossy compression: by zeroing out static backgrounds while preserving core flame morphology, the discrete cosine transform distribution during JPEG encoding shifts towards zero-valued coefficients, reducing the encoded byte count. With the JPEG quality factor fixed at 90, Table 5 reports the measured application-layer payloads.

Table 5: Atlas-to-Server single link payload records.

Model Config	Type	ROI Processing	Upload Ratio	Payload KB
YOLOv11n Baseline [7]	Measured	Unmasked BBox	1.00	152
PGENet Control	Control Var	Unmasked BBox	1.00	152
PGENet Proposed	**Measured**	**Mask Zeroing**	**0.74**	**113**

Table 5 shows that spatial mask zeroing reduces the single-upload volume from 152 to 113 KB at the perception source at a JPEG quality factor of 90. The *PGENet Control* row keeps the ROI crop identical to the baseline and turns off only the mask-zeroing step, isolating the mask effect from any box-size change. To evaluate the systemic impact of this compression alongside false-positive suppression, we compute an hourly invalid offered-load estimate for a facility with ten cameras, each encountering ten strong-light interference events per hour (i.e., $N_c\lambda_h = 100$ events per hour); we report the resulting metrics in Table 6.

Table 6: Estimated invalid uplink offered load on an open-loop link.

Config	Events/h	Uploads/h	MB/h	Reduction
YOLOv11n Baseline	100	36.2	5.50	Baseline
PGENet Unmasked	100	2.1	0.32	94.2%
PGENet Mask Zeroing	**100**	**2.1**	**0.24**	**95.7%**

Loads use 1 MB = 1000 KB and exclude MAC-layer queuing and packet-drop overhead.

As shown in Table 6, the communication benefits stem from two orthogonal sources: the temporal kinematic attention lowers the false-positive event rate. At the same time, the spatial gating mask compresses the single-upload payload. For the assumed $N_c\lambda_h = 100$ interference events per hour, the false-positive event rate in Table 3 matches the upload count per hour, and the combined effect reduces the source-side invalid offered load by 95.7%, matching Eq. 18. This estimate is deliberately open-loop: it isolates application-layer payload gains and excludes MAC-layer queuing latency, contention, and retransmission overhead. Despite these reductions, practical limitations remain for real-world deployment. The volume compression depends on background visual distinctiveness: dark environments offer less compression margin because the raw slice has low information entropy. Extracting kinematic priors via a four-frame temporal window introduces a cold-start buffer latency of approximately 94 milliseconds at the measured 32 FPS, which instantaneous response scenarios must budget for. Because these priors originate from visible-light phenomena, tasks requiring temperature profiling or chemical diffusion analysis will require multimodal infrared sensing in future network testbeds.

6 Conclusion

To enable communication-efficient fire perception in industrial local area networks, this study proposes PGENet. By embedding the spatial gating block and the kinematic attention block directly into the native feature flow, the architecture uses spatiotemporal priors to suppress static background noise and dynamic strong-light artefacts without requiring external filtering networks. On the edge device, the false-positive event rate drops from 36.2% to 2.1% under high-exposure interference. Spatial mask zeroing reduces the single-alarm evidence payload from 152 to 113 kilobytes. Combining these two effects yields an estimated 95.7% reduction in the invalid offered load injected from the perception source. Future work will transition from this open loop application layer estimation to closed loop medium access control evaluations measuring queuing latency in physical multi-node environments.

Acknowledgements

Not applicable.

Author contributions

Bin Hu: Conceptualization, Methodology, Writing. Jun Li: Data Curation, Validation. Xufeng Lou: Investigation. Yuqiao Li: Supervision, Funding acquisition. Jiaji He: Review & Editing. Qianhong Wan: Visualization, Project administration. Ju Zhang: Software. All authors read and approved the final manuscript.

Funding

This work was supported by the Key Technologies and Demonstration Applications for Intelligent Safety Production Management throughout the Lifecycle of High-Altitude Water Conservancy Projects Based on Multimodal Large Models under Grant QYXTZX-RKZ2026-05-1, and by the Industrial Multimodal Safety Perception and Emergency Early Warning Equipment Verification and Transformation Service Platform Capability Enhancement Project under Grant 26ZYCGCG01180.

Data availability

Not applicable.

Materials availability

Not applicable.

Code availability

The training and inference code is available from the corresponding author on reasonable request, subject to industrial confidentiality review.

Declarations

Ethics approval and consent to participate

Ethics approved. Consent obtained.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

[1] Zhang, P., Zhao, X., Yang, X., Zhang, Z., Bi, C., Zhang, L.: F3-YOLO: A robust and fast forest fire detection model. Forests **16**(9), 1368 (2025) https://doi.org/10.3390/f16091368

[2] Li, X., Fan, Y., Liu, Y., Li, X., Liu, Z.: Industrial and mining fire detection algorithm based on improved YOLO. Fire Technology **61**(2), 709–728 (2025) https://doi.org/10.1007/s10694-024-01635-4

[3] Peng, B., Kim, T.-K.: YOLO-HF: Early detection of home fires using YOLO. IEEE Access **13**, 79451–79466 (2025) https://doi.org/10.1109/ACCESS.2025.3566907

[4] Wang, K., Zhang, W., Song, X.: A fire detection method for aircraft cargo compartments utilizing radio frequency identification technology and an improved YOLO model. Electronics **14**(1), 106 (2025) https://doi.org/10.3390/electronics14010106

[5] Alkhammash, E.H.: A comparative analysis of YOLOv9, YOLOv10, YOLOv11 for smoke and fire detection. Fire **8**(1), 26 (2025) https://doi.org/10.3390/fire8010026

[6] Hoang, V.-H., Lee, J.W., Park, C.-S.: Enhancing fire detection with YOLO models: A bayesian hyperparameter tuning approach. CMC-Computers Materials & Continua **83**(3), 4097–4116 (2025) https://doi.org/10.32604/cmc.2025.063468

[7] Li, Y., Nie, L., Zhou, F., Liu, Y., Fu, H., Chen, N., Dai, Q., Wang, L.: Improving fire and smoke detection with YOLO11 and multi-scale convolutional attention. Fire **8**(5), 165 (2025) https://doi.org/10.3390/fire8050165

[8] Li, C., Du, Y., Zhang, X., Wu, P.: YOLOGX: an improved forest fire detection algorithm based on YOLOv8. Frontiers in Environmental Science **12**, 1486212 (2025) https://doi.org/10.3389/fenvs.2024.1486212

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

[10] Lu, J., Zheng, Y., Guan, L., Lin, B., Shi, W., Zhang, J., Wu, Y.: FCMI-YOLO: An efficient deep learning-based algorithm for real-time fire detection on edge devices. PLOS ONE **20**(8), 0329555 (2025) https://doi.org/10.1371/journal.pone.0329555

[11] Huang, L., Ding, Z., Zhang, C., Ye, R., Yan, B., Zhou, X., Xu, W., Guo, J.: YOLO-ULNet: Ultralightweight network for real-time detection of forest fire on embedded sensing devices. IEEE Sensors Journal **24**(15), 25175–25185 (2024) https://doi.org/10.1109/JSEN.2024.3416548

[12] Li, Y., Wang, Y., Shao, X., Zheng, A.: An efficient fire detection algorithm based on mamba space state linear attention. Scientific Reports **15**(1), 11289 (2025) https://doi.org/10.1038/s41598-025-95162-z

[13] Li, J., Tang, H., Li, X., Dou, H., Li, R.: LEF-YOLO: a lightweight method for intelligent detection of four extreme wildfires based on the YOLO framework. International Journal of Wildland Fire **33**(1), 23044 (2024) https://doi.org/10.1071/WF23044

[14] Pan, W., Xu, B., Wang, X., Lv, C., Wang, S., Duan, Z.: YOLO-FireAD: Efficient fire detection via attention-guided inverted residual learning and dual-pooling feature preservation. In: Proceedings of the 21st International Conference on Intelligent Computing (ICIC 2025), Ningbo, China, pp. 438–452 (2025). Poster Volume I

[15] Qi, Y., Sun, J., Liu, C., Zheng, Z., Liu, Y., Wu, D., Zhao, J.: A radar-video fusion method for 3D object detection based on on-board edge computing in intelligent connected environments. EURASIP Journal on Wireless Communications and Networking **2025**(1) (2025) https://doi.org/10.1186/s13638-025-02526-0

[16] Jin, H., Ouyang, A., Wang, Q., Yang, D., Gan, X., Yue, X.: Helipad target detection method for low-altitude rotor UAVs based on improved YOLOv11 network model. EURASIP Journal on Wireless Communications and Networking **2025**(1) (2025) https://doi.org/10.1186/s13638-025-02507-3

[17] Yu, B., Yu, S., Zhao, Y., Wang, J., Lai, R., Lv, J., Zhou, B.: Intelligent fire-fighting technology for drone swarms with multi-sensor integrated path planning: YOLOv8 algorithm-driven fire source identification and precision deployment strategy. Drones **9**(5), 348 (2025) https://doi.org/10.3390/drones9050348

[18] Aldughayfiq, B., Alrashdi, I., Tawfeek, M.A., Jamel, L., Shaban, W.M., Talaat, F.M.: Real-time forest fire detection, monitoring, and mitigation to safeguard humanity using YOLO and federated learning. International Journal of Distributed Sensor Networks **2025**(1), 4707734 (2025) https://doi.org/10.1155/dsn/4707734

[19] Çelik, T., Demirel, H.: Fire detection in video sequences using a generic color model. Fire Safety Journal **44**(2), 147–158 (2009)

[20] Çetin, A.E., Dimitropoulos, K., Gouverneur, B., Grammalidis, N., Günay, O., Habiboğlu, Y.H., Töreyin, B.U., Verstockt, S.: Video fire detection – review. Digital Signal Processing **23**(6), 1827–1843 (2013)

[21] Chino, D.Y.T., Avalhais, L.P.S., Rodrigues, J.F., Traina, A.J.M.: BoWFire: Detection of fire in still images by integrating pixel color and texture analysis. In: Proceedings of the 28th SIBGRAPI Conference on Graphics, Patterns and Images, pp. 95–102 (2015)

[22] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)

[23] Yun, B., Zheng, Y., Lin, Z., Li, T.: FFYOLO: A lightweight forest fire detection model based on YOLOv8. Fire **7**(3), 93 (2024) https://doi.org/10.3390/fire7030093

[24] Zhou, N., Gao, D., Zhu, Z.: YOLOv8n-SMMP: A lightweight YOLO forest fire detection model. Fire **8**(5), 183 (2025) https://doi.org/10.3390/fire8050183

[25] Wang, H., Zhang, Y., Zhu, C.: YOLO-LFD: A lightweight and fast model for forest fire detection. CMC-Computers Materials & Continua **82**(2), 3399–3417 (2025) https://doi.org/10.32604/cmc.2024.058932

[26] Zhu, H., Ling, W., Yan, H., Zhong, X., Liao, F.: YOLO-MP: A lightweight forest fire detection model. Ecological Informatics **92**, 103516 (2025) https://doi.org/10.1016/j.ecoinf.2025.103516

[27] Sozol, M.S.S., Mondal, M.R.H., Thamrin, A.H.: Indoor fire and smoke detection based on optimized YOLOv5. PLOS ONE **20**(4), 0322052 (2025) https://doi.org/10.1371/journal.pone.0322052

[28] Ishtiaq, M., Won, J.-U.: YOLO-SIFD: YOLO with sliced inference and fractal dimension analysis for improved fire and smoke detection. CMC-Computers Materials & Continua **82**(3), 5343–5361 (2025) https://doi.org/10.32604/cmc.2025.061466

[29] Zhao, Y., Lv, W., Xu, S., Wei, J., Yang, G., Dang, Q., Liu, Y., Chen, J.: DETRs Beat YOLOs on Real-time Object Detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965–16974 (2024). https://doi.org/10.1109/CVPR52733.2024.01605