

LMFSNet: A Lightweight Multi-Scale Collaborative Network for Fire and Smoke Detection

Yan Li

Southwest Jiaotong University

Peng-Xiao Yang

Southwest Jiaotong University

Ze-Xi Hua

xx_zxhua@swjtu.edu.cn

Southwest Jiaotong University

Hong-Bo Wang

Southwest Jiaotong University

Research Article

Keywords: Fire smoke detection, Lightweight network, Multi-scale, Context Fusion

Posted Date: April 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9305554/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Real-Time Image Processing on August 26th, 2026. See the published version at <https://doi.org/10.1007/s11554-026-01957-2>.

LMFSNet: A Lightweight Multi-Scale Collaborative Network for Fire and Smoke Detection

Yan Li¹ · Peng-xiao Yang¹ · Ze-xi Hua^{1*} · Hong-bo Wang¹

the date of receipt and acceptance should be inserted later

Abstract To address challenges in fire and smoke detection under complex environments—minuscule early-stage targets, large shape variations, weak features, and the need for lightweight real-time deployment—this paper proposes LMFSNet, a lightweight multi-scale collaborative detection network. The network integrates three key components. First, a lightweight multi-scale feature extraction module (MFE-C2f) handles target scale variations with low computational cost. Second, an efficient multi-scale context fusion module (EMCF) uses attention mechanisms to enhance global semantic associations, improving ambiguous feature discrimination while suppressing background interference. Third, a lightweight fusion path based on the Simplified Feature Transfer Module (SFTM) enables efficient multi-scale information exchange under real-time resource constraints. Additionally, combining the SlimHead detection head with a high-resolution P2 detection layer and the multi-scale adaptive loss (MSAW Loss) improves sensitivity to small early-stage targets and regression stability. To support training and evaluation, we construct a high-quality multi-scenario fire-smoke dataset (FSD). Experimental results show that our method significantly boosts detection accuracy while maintaining high inference efficiency, offering an effective solution for real-time fire early warning in complex environments.

Keywords Fire smoke detection · Lightweight network · Multi-scale · Context Fusion

1 Introduction

As one of the most destructive natural disasters, fires pose a severe threat to global public safety due to their sudden onset and rapid spread. According to the United Nations Office for Disaster Risk Reduction (UNDRR), fires cause over \$100 billion in direct economic losses worldwide each year and result in hundreds of thousands of casualties[1]. In complex environments such as forests, chemical industrial parks, and high-rise buildings, early and precise fire monitoring is critical for disaster prevention and mitigation [2].

Traditional fire detection methods focus on extracting visual features such as color[3], shape [4], and texture [5]. For instance, Vipin et al.[6] identified fire pixels by separating luminance and chrominance in the YCbCr color space; Chen et al.[7] combined HSI and RGB color statistics to segment candidate regions, then extracted LBP texture features for SVM classification; Zhang et al.[8] proposed an adaptive threshold fire-smoke color model. However, these methods rely on manually designed shallow features, exhibiting limitations such as insufficient robustness, limited coverage, poor environmental adaptability, and response delays. They struggle to meet the real-time and reliability demands of modern intelligent security systems.

In recent years, deep learning has enabled data-driven object detection methods to offer novel solutions for fire and smoke detection. Two-stage and single-stage detection frameworks like Faster R-CNN [9] and RetinaNet[10], alongside lightweight networks such as MobileNet [11] and EfficientNet [12], have significantly improved detection accuracy through end-to-end training. Concurrently, Transformer-based models [13] such as Swin Transformer[14] and RT-DETR[15] have demonstrated strong performance in smoke detection within

¹ School of Information Science and Technology, Southwest Jiaotong University, Chengdu, 610000, China
E-mail: xx_zxhua@swjtu.edu.cn

* Corresponding author



Fig. 1 Examples of scenes similar to fire and smoke.

complex backgrounds due to their global attention mechanisms.

Among numerous frameworks, the YOLO[16] series models have become widely adopted baselines in fire and smoke detection due to their balance between accuracy and speed. Extensive research has focused on targeted enhancements to the YOLO series: for instance, introducing attention mechanisms in YOLOv5[17] or integrating hybrid attention in YOLOv6[18] to enhance feature representation. Recent improvements targeting YOLOv8 have been particularly active, with strategies such as designing lightweight modules, introducing multi-scale fusion mechanisms, or adding high-resolution detection layers.

However, fires and smoke challenge generic detectors due to their highly dynamic irregular shapes, extreme scale variation (from full-frame smoke to pixel-level fires), and easy confusion with clouds, fog, lights, or red flags [19][20] (Fig. 1). This leads to low sensitivity to tiny initial targets (high false negatives), susceptibility to similar distractors (high false positives), and a trade-off between model complexity and real-time performance—lightweight designs sacrifice features while heavy modules increase compute.

To address these, we propose LMFSNet (YOLOv8 baseline) with five targeted designs: MFE-C2f backbone handles variable scale and diffuse smoke via parallel multi-branch progressive fusion; EMCF tackles smoke edge blurring and background confusion through global semantic understanding; SFTM-based lightweight Neck balances complexity and speed; SlimHead with high-resolution P2 layer enhances sensitivity for small-scale initial fires and sparse smoke; and Multi-Scale Adaptive Weighted Loss (MSAW Loss) optimizes regression across all scales. Together, these achieve high accuracy, speed, and robustness.

- Constructed the FSD dataset with diverse annotated images, providing a benchmark for fire and smoke detection.
- Proposed MFE-C2f backbone that fuses multi-receptive-field features via parallel branches and progressive refinement, reducing parameters and com-

putations while maintaining robust full-scale representation.

- Designed EMCF for global semantics via attention to enhance weak target discrimination, and SFTM to rebuild FPN with lower overhead and real-time fusion.
- Developed SlimHead with P2 layer for small-target sensitivity, and MSAW loss with scale-aware dynamic weighting to boost accuracy and robustness across scales.

2 Related work

2.1 Deep learning-based fire and smoke detection methods

Deep learning enables end-to-end feature extraction for fire smoke detection, boosting accuracy. CNNs are widely used for robust local features. For instance, Shamsoshaara et al. [21] used a U-Net encoder-decoder to segment fire regions; Khan et al. [22] designed a stacked-encoded EfficientNet that reduces false alarms via cost-sensitive learning. To enhance feature representation, multi-scale and attention mechanisms have been incorporated: Feng et al.[23] used dense connections and group-channel attention for cross-layer fusion; Li et al. [24] integrated attention with network pruning in DenseNet for simultaneous smoke and fire detection; Khan et al. [25] proposed Cross-Module Attention Network (CANet) using compression-expansion paths to focus on fire areas.

However, CNN-based methods face limitations. Smoke exhibits diverse morphologies and extreme scale variations, and early-stage minute targets have weak features prone to misclassification. More critically, CNN local receptive fields limit global context and long-range dependency modeling, essential for distinguishing smoke from clouds, haze, or fog.

To overcome these, recent research uses attention mechanisms and Transformers for global perception. Zhang et al. [26] embedded attention gate units in a U-shaped architecture to amplify useful features. Lin et al. [27] adopted Swin Transformer as a backbone, leveraging hierarchical self-attention for global context. Multi-scale perception has been further strengthened: Zhan et al. [28] introduced parallel spatial attention and small-scale transformed feature pyramids in YOLOX-L; Huang et al. [29] combined multi-scale contextual contrast with dense pyramid pooling in deformable DETR; Li et al. [30] designed multi-dimensional window attention and a feature pyramid based on Swin Transformer. Wu et al. [31] explored CNN-Transformer complementarity, using CNNs for shallow details and Transformers for deep global representations. These works laid a foundation

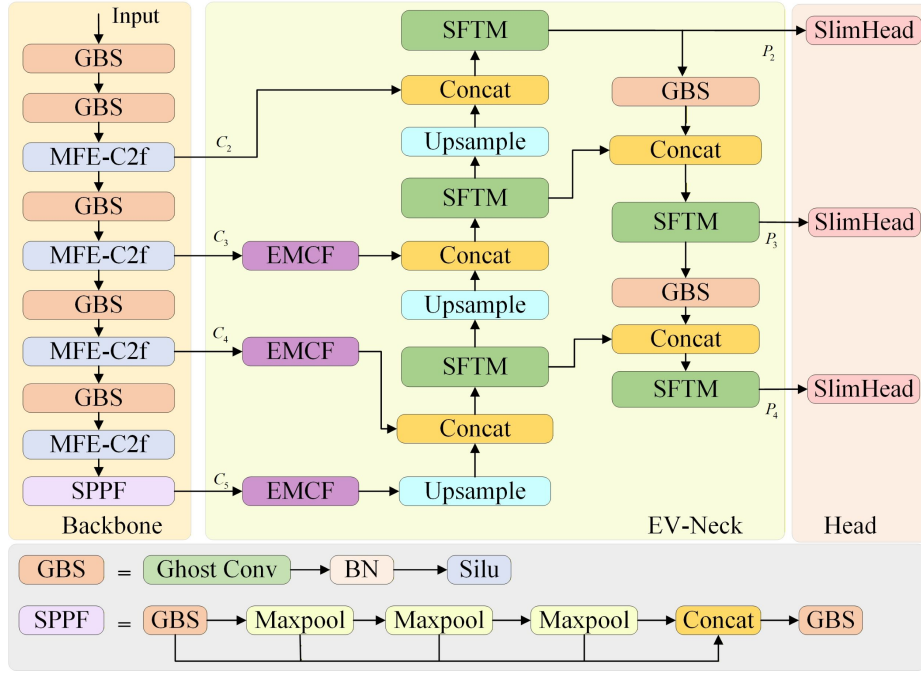


Fig. 2 Overall model structure of LMFSNet.

for precise fire smoke detection in complex dynamic environments.

Deep learning has raised accuracy upper limits, but high computational complexity contradicts real-time, low-power requirements. Thus, embedding high-performance perception into lightweight architectures for resource-constrained deployment remains a critical challenge.

2.2 Lightweight network-based fire and smoke detection methods

Fire smoke detection systems are often deployed on edge devices with limited computational resources. Pursuing lightweight and efficient models while ensuring detection accuracy holds critical practical significance. Existing lightweight research follows two technical paths with distinct design priorities in the accuracy-speed trade-off.

One approach focuses on extreme computational efficiency through structural innovations. For instance, Chen et al. [32] proposed FasterNet with Partial Convolution (PConv) to reduce computational overhead by eliminating redundant operations. Jin et al. [33] developed SWVR, combining reparameterization with parameter-free attention to lower complexity. Cai et al. [34] introduced FalconNet, systematically decomposing lightweight optimization from architecture to operators. While these approaches achieve high inference speeds, they sometimes face challenges in detection accuracy—particularly small object recall—when encountering early-stage fire smoke targets with variable

morphologies and weak features, due to limited feature extraction capability.

Another stream selectively introduces enhancement modules into lightweight frameworks to better balance accuracy and efficiency. For instance, Sung et al. [35] integrated YOLOv8 with iGhostNet and attention modules to enhance accuracy while controlling costs; Chen et al. [36] proposed GS-YOLOv5, using Ghost module to reduce parameters while introducing Coordinate Attention (CA) to strengthen spatial perception; Shen et al. [37] constructed a lightweight backbone combining deformable convolutions with Transformer structures. These approaches trade moderate overhead for higher accuracy, with overall efficiency hinging on optimal co-design between lightweight foundations and enhancement modules.

The core challenge is balancing extreme efficiency with robust representation for diverse scale-varying targets, especially early faint features. Existing approaches trade off: fast inference misses small targets, while accurate designs sacrifice speed. Thus, an integrated architecture fusing efficient computation, robust extraction, and context understanding is needed.

3 Methods

3.1 Architecture overview

To address the diversity in fire smoke target morphology and scale, and real-time requirements, we propose LMF-

SNet, a lightweight detection model (Figure 2). Its core is a multi-scale feature extraction backbone that uses an MFE-C2f module for progressive multi-scale perception, providing feature primitives rich in detail and semantics. The Efficient Multi-Scale Context Fusion (EMCF) module endows features with global understanding, enabling clear distinction of fire/smoke from complex backgrounds. To efficiently integrate multi-scale information, a lightweight SFTM module reconstructs the feature pyramid network (Neck), reducing computational overhead while preserving feature integrity. The SlimHead detection head enhances sensitivity and localization for small targets (e.g., early flames, faint smoke) by introducing a high-resolution P2 layer and a lightweight attention mechanism.

3.2 Lightweight backbone network

To handle scale variation, ambiguity, and real-time constraints, we replace YOLOv8’s deep convolutions with MFE-C2f (Figure 3), enabling low-cost progressive multi-scale extraction from minute fire to diffuse smoke. Stacking it yields an efficient, accurate lightweight backbone.

The module uses Depthwise Convolution (DW-Conv) as its core operator, reducing complexity from $\mathcal{O}(HWCK^2C)$ to $\mathcal{O}(HWC(K^2 + C))$ (H, W : spatial dimensions, C : channels, K : kernel size). We select $(3 \times 3, 5 \times 5, 7 \times 7)$ kernels for parallel branches, forming a progressive receptive field from local details to mid-to-long-range context. The 3×3 kernel is the base; 5×5 and 7×7 expand perception. This balances efficiency: a single moderately enlarged kernel reduces inter-layer computation compared to stacking two 3×3 kernels. Larger kernels (e.g., 9×9) are avoided because their $\mathcal{O}(K^2)$ cost rises faster than the marginal benefit.

The module achieves multi-scale extraction via structured information flow. Input $F_{in} \in \mathbb{R}^{H \times W \times C}$ is split into three branches:

$$F_1, F_2, F_3 = \text{Chunk}(F_{in}) \quad (1)$$

where $F_1, F_2, F_3 \in \mathbb{R}^{H \times W \times C/3}$.

The finest-scale branch F_1 passes through a 3×3 depthwise separable convolution to obtain F'_1 , then a parameter-free parallel pooling attention (GAP+GMP summed and Sigmoid-activated) calibrates features into F_{11} , focusing on local cues like initial fires or smoke textures:

$$F'_1 = \text{DWConv}_{3 \times 3}(F_1) \quad (2)$$

$$F_{11} = F'_1 \otimes \delta(\text{GAP}(F'_1) + \text{GMP}(F'_1)) \quad (3)$$

where δ is sigmoid, GAP and GMP denote global average/max pooling.

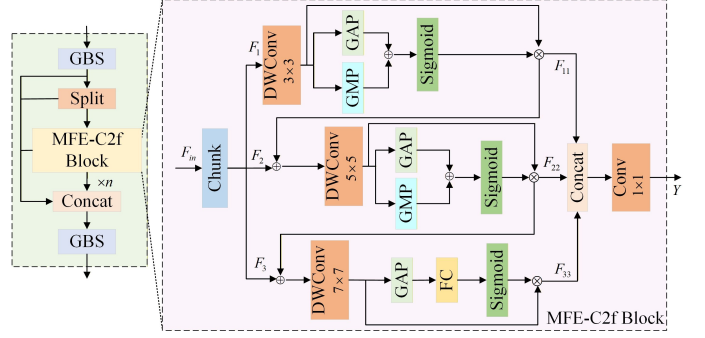


Fig. 3 Schematic diagram of the MFE-C2f model structure.

The medium-scale branch F_2 is combined with F_{11} for cross-scale fusion, then passed through a 5×5 depthwise separable convolution to expand the receptive field, yielding F'_2 which captures broader context. The same attention produces F_{22} :

$$F'_2 = \text{DWConv}_{5 \times 5}(F_2 \oplus F_{11}) \quad (4)$$

$$F_{22} = F'_2 \otimes \delta(\text{GAP}(F'_2) + \text{GMP}(F'_2)) \quad (5)$$

The maximum-scale branch F_3 is combined with F_{22} , and a 7×7 separable convolution extracts global features. A lightweight channel attention module (GAP + simplified FC) generates global guidance weights, yielding F_{33} :

$$F'_3 = \text{DWConv}_{7 \times 7}(F_3 \oplus F_{22}) \quad (6)$$

$$F_{33} = F'_3 \otimes \delta(\text{FC}(\text{GAP}(F'_3))) \quad (7)$$

Finally, all refined features are aggregated: F_{11}, F_{22}, F_{33} are concatenated along channels. A 1×1 convolution fuses and unifies dimensions, producing multi-scale representation Y :

$$Y = \text{Conv}_{1 \times 1}(\text{Concat}(F_{11}, F_{22}, F_{33})) \quad (8)$$

This module parallel multi-branch architecture combines feature aggregation with attention, allowing larger-scale features to absorb smaller-scale information under globally guided weights, enhancing representation for ambiguous multi-scale targets like fires and smoke.

3.3 Efficient multi-scale context fusion module

Flame and smoke detection faces morphological variations and background clutter, requiring global context. The Neck’s multi-scale pyramid depends on backbone features; lacking global semantics and long-range dependencies degrades fusion and recognition.

To overcome this, we propose the Efficient Multi-scale Context Fusion Module (EMCF, Figure 4). EMCF

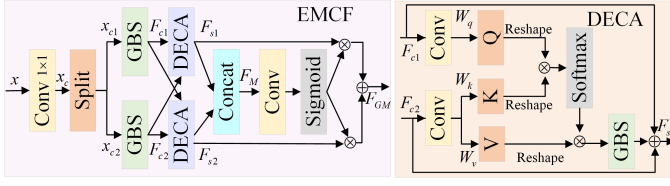


Fig. 4 Structure diagram of the EMCF model.

operates concurrently on multi-scale feature maps (C3, C4, C5) from the backbone. At each scale, self-attention enables each pixel to interact with all others at the same scale, enhancing contextual awareness in a scale-adaptive manner.

For high-resolution C3 features (fine-grained small-target details), EMCF emphasizes local to medium-range context to relate subtle smoke textures to surroundings and suppress noise. For mid-resolution C4 features (balancing detail and semantics), it integrates local and global cues for robust multi-scale recognition. For low-resolution C5 features (rich semantics for large objects), EMCF strengthens global context to separate large fires from confusing scenes like red buildings, clouds, or haze.

A 1×1 convolution reduces input x to x_c , then an ESA unit [38] performs context modeling. ESA lowers cost via dimensionality reduction before generating query/key representations.

In ESA, attention maps value vectors to output based on query-key similarity. Query (Q), key (K), and value (V) are derived from the same input, each transformed by separate linear projections:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V \quad (9)$$

For the dual-branch architecture, x_c is split along channels into x_{c1} and x_{c2} to minimize inter-channel redundancy. The GBS module processes each branch independently with lightweight skip connections, enriching features while balancing efficiency and performance. The outputs F_{c1}, F_{c2} are passed to the next fusion stage.

We introduce the Dual-branch Efficient Cross-Attention Module (DECA, Figure 4). DECA takes F_{c1}, F_{c2} and performs bidirectional guided fusion via cross-attention, avoiding simple concatenation. F_{c1} generates Q_1, K_1, V_1 , F_{c2} yields Q_2, K_2, V_2 ; Q_1 interacts with K_2, V_2 to capture cross-scale long-range dependencies. This integrates information from various image regions [39], leveraging multi-scale features F_s :

$$F_s = f_{GBS}(Attention(Q_1, K_2, V_2)) + F_{c1} + F_{c2} \quad (10)$$

To fuse across channels, each channel serves as a guide query for the other, and results are concatenated

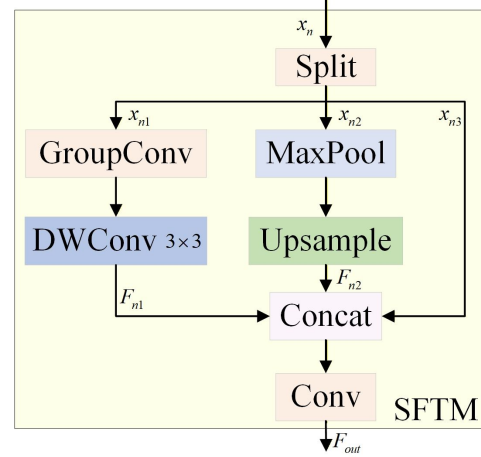


Fig. 5 SFTM model architecture diagram.

to obtain multi-scale features F_M :

$$F_M = Concat(F_{s1}(F_{c1}, F_{c2}), F_{s2}(F_{c2}, F_{c1})) \quad (11)$$

where F_{s1}, F_{s2} are fused features from DECA, which are refined by a lightweight gated prediction network (convolution + Sigmoid) and combined with the original input via a residual connection to produce globally enhanced features F_{GM} :

$$F_{GM} = [\delta(Conv(F_M))] \otimes F_{s1} + [\delta(Conv(F_M))] \otimes F_{s2} \quad (12)$$

DECA integrates dual-branch features via coarse-to-fine attention, capturing global context with low overhead. EMCF preserves resolution and semantics while enriching context, producing robust features for multi-scale detection.

3.4 Lightweight neck

The Neck network balances efficiency and feature retention. YOLOv8's C2f ensures gradient flow via extensive cross-layer connections, but its complex multi-branch structure incurs high overhead and spatial detail loss critical for small targets. Channel redundancy in Bottleneck blocks further increases cost. We thus propose SFTM to replace C2f (Figure 5). SFTM uses a three-path architecture that reduces complexity while preserving feature quality.

Input x_n is split into three channel-equal parts. Path1: lightweight processing on x_{n1} via group ($g=4$) and depthwise separable convolutions. Path2: spatial enhancement on x_{n2} via max pooling and bilinear interpolation (no convolution). Path3: identity mapping on

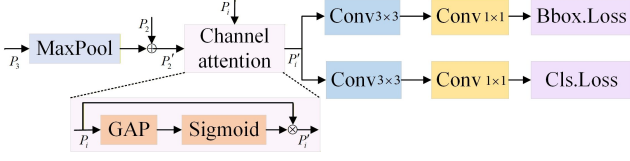


Fig. 6 SlimHead model diagram

x_{n3} (zero cost). Outputs are concatenated and fused by 1×1 convolution:

$$F_{n1} = \text{DWConv}_{3 \times 3}(\text{GroupConv}_{1 \times 1}(x_{n1}, g = 4)) \quad (13)$$

$$F_{n2} = \text{Upsample}(\text{MaxPool}(x_{n2})) \quad (14)$$

$$F_{out} = \text{Conv}_{1 \times 1}(\text{Concat}(F_{n1}, F_{n2}, x_{n3})) \quad (15)$$

where F_{n1} is features from two-stage lightweight convolutions, F_{n2} spatially enhanced features, and F_{out} the final output.

This design provides a lightweight solution through three collaborative strategies: lightweight processing, spatial enhancement, and feature retention.

3.5 Lightweight detection head based on P2

To enable efficient deployment in resource-constrained environments while maintaining accuracy for small-scale targets, we propose SlimHead, a lightweight detection head based on the P2 layer. Building on the traditional three-layer architecture (P3–P5), SlimHead adds a high-resolution P2 layer, forming a four-layer multi-scale system that enhances perception of small targets like fine smoke textures and nascent fires.

SlimHead decouples classification and regression branches (Figure 6), enabling independent parallel operation, preventing cross-task interference, and improving precision. Each branch uses two layers: 3×3 for spatial features and 1×1 for channel adjustment and prediction, simplifying heads and reducing parameters/overhead.

To address the high resolution but relatively weak semantics of P2 features, SlimHead uses a concise cross-layer fusion strategy: adding downsampled P3 features to P2 features enhances semantic expressiveness, improving small-object detection:

$$P_{2'} = P_2 + \text{MaxPool}_{2 \times 2}(P_3) \quad (16)$$

where $\text{MaxPool}_{2 \times 2}(\cdot)$ denotes 2×2 max pooling with stride 2, incurring minimal cost by simply downsampling to align spatial resolution.

To further optimize feature representation, SlimHead introduces a lightweight channel attention mechanism. It acquires channel-wise statistics via GAP and uses a

Sigmoid to generate channel weights, modulating the feature map:

$$P'_i = P_i \otimes \delta(\text{GAP}(P_i)) \quad (17)$$

where P'_i and P_i are the enhanced and original features for $i = 2', 3, 4$, respectively.

SlimHead achieves synergistic optimization of accuracy and efficiency by introducing the P2 detection layer, adopting a streamlined two-layer convolution architecture, designing cross-layer fusion, and integrating a lightweight channel attention mechanism.

3.6 Loss Function

The loss comprises classification (binary cross-entropy) and regression. For regression, we propose MSAW Loss based on CIoU [40], which uses scale-aware dynamic weighting to adapt regression emphasis across object scales, improving multi-scale detection accuracy.

MSAW retains CIoU's three geometric constraints: area overlap, center point distance, and aspect ratio consistency. For predicted and ground-truth boxes:

$$L_{MSAW} = 1 - IoU + \frac{\rho^2(b_p, b_{gt})}{c^2} + \gamma(S) \cdot \nu \quad (18)$$

where b_p and b_{gt} are center coordinates, ρ^2 is Euclidean distance, c the diagonal of the minimum bounding rectangle, and ν the aspect ratio consistency metric: $\nu = \frac{4}{\pi^2} (\arctan \frac{w_{gt}}{h_{gt}} - \arctan \frac{w_p}{h_p})^2$.

Inspired by SDL [41] (Scale-based Dynamic Loss), MSAW introduces a scale-adaptive weighting coefficient $\gamma(S)$ that dynamically adjusts based on the bounding box's absolute scale. $\gamma(S)$ is defined as the Sigmoid function of target area S :

$$\gamma(S) = \frac{\alpha}{1 + \exp(-\lambda(S - S_0))} \quad (19)$$

where $S = w_{gt} \times h_{gt}$ is the area of the ground-truth box. S_0 is the scale threshold, set to the median target area in the dataset. λ is the smoothing factor (set to 0.01), controlling transition smoothness from 0 to α . α is the maximum weight coefficient (set to 0.05), determining the maximum contribution of the aspect ratio loss term.

$\gamma(S) \rightarrow \alpha$ for large targets ($S \gg S_0$), prioritizing shape matching, and $\rightarrow 0$ for small targets ($S \ll S_0$), focusing on center alignment and overlap. λ ensures smooth transitions, enabling scale-adaptive optimization for multi-scale detection.

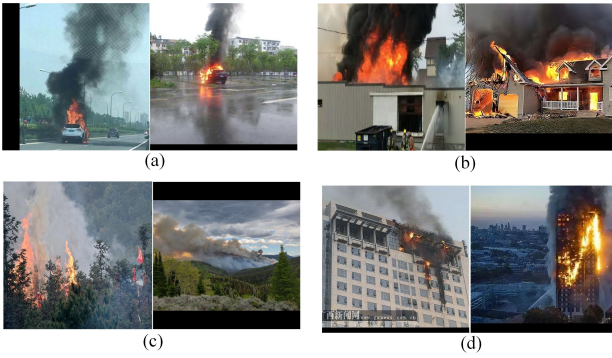


Fig. 7 Sample images captured in various natural environments within the FSD dataset.

4 Experimental validation

4.1 Experimental Settings

Datasets. To evaluate performance and generalization, we trained on our self-built Scene-FSD dataset and used the public DFS dataset [42] as an external benchmark.

Self-built Dataset (Scene-FSD): To address the lack of real surveillance scenarios, we constructed Scene-FSD, a large-scale, scene-oriented fire and smoke dataset. As shown in Figure 7, it covers four key early-warning scenarios: (a) Transportation (3,150 images): vehicle fires/smoke on roadways and tunnel entrances; (b) Open Industrial Site (3,205 images): fires in storage and open-air facilities; (c) Forest monitoring (2,890 images): distant wildfires and forest smoke; (d) High-rise building monitoring (2,780 images): fires on exterior walls and windows. All images are from real surveillance systems.

The dataset has 12,025 high-resolution images. Using LabelImg, we manually annotated each fire/smoke instance, yielding 31,842 bounding boxes. To handle blurred shapes and indistinct boundaries, we defined annotation guidelines (e.g., for semi-transparent smoke edges) and ensured consistency via manual review. Raw images were cropped to 1024×1024 and split 8:1:1 into training (9,621), validation (1,202), and testing (1,202) sets; during training, inputs were dynamically scaled to 512×512 .

Public Dataset (DFS): DFS is widely recognized, with images sourced from multiple public internet sources and annotated to highlight distractors easily confused with fires. It contains 9,462 high-quality images covering diverse real-world fire scenarios (indoor, outdoor, industrial). We directly used its official data partitioning for cross-dataset evaluation.

Implementation Details. All experiments use PyTorch on a single NVIDIA RTX 4090 GPU (24 GB). We initialize the model with YOLOv8 pre-trained weights. Initial learning rate is $1e-2$ with a cosine annealing sched-

uler ($\eta_{\min} = 0.00001$, $T_{\max} = 50$). Training runs for 100 epochs with batch size 16.

Evaluation Metrics. We evaluate using Precision, Recall, Average Precision (AP), mean Average Precision (mAP), parameter count (Params, in millions), GFLOPs, and FPS. mAP@0.5 (IoU threshold 0.5) indicates overall detection accuracy. Lower Params and GFLOPs imply lighter models and lower computational cost, suitable for resource-constrained deployment. FPS measures real-time inference speed, critical for fire/smoke detection.

4.2 Comparative Experiments

4.2.1 Selection of baseline models

We selected YOLOv8 [43] as the baseline for its balanced speed, accuracy, and architectural flexibility. To determine the best variant for our task, we compared YOLOv8n, YOLOv8s, and YOLOv8m on Scene-FSD (Table 1). YOLOv8n is fast but less accurate; YOLOv8m is accurate but heavy; YOLOv8s offers the best overall balance—high inference speed, significantly better accuracy than YOLOv8n (especially recall), and moderate resource demands, making it suitable for edge deployment.

Table 1 Comparison of YOLOv8-Related models on the Scene-FSD dataset.

Model	Precision	Recall	mAP@50	FPS	Params (m)	GFLOPs
YOLOv8n	75.64	70.86	74.12	88	6.86	14.29
YOLOv8s	77.66	76.17	78.24	73	11.02	27.25
YOLOv8m	80.43	79.36	80.35	62	26.98	75.46

Thus, we choose YOLOv8s as the baseline. Our core innovation is a universal lightweight multi-scale enhancement module. YOLOv8s serves as a clear, representative baseline for systematic analysis, but the proposed method is not limited to it. As shown in Tables 3 and 4, our LMFSNet ultimately outperforms even newer models (YOLOv11, YOLOv12) in mAP@50, recall, and efficiency (params, GFLOPs, FPS), demonstrating substantial, generalizable advancements beyond baseline iterations.

4.2.2 Comparison of loss functions

To validate MSAW Loss, we compared it with CIOU[41], GIoU[44], and WIoU[45] on Scene-FSD under identical conditions. Results are in Table 2.

MSAW Loss achieves the best performance across all metrics. It reaches 78.80% mAP@0.5, surpassing CIOU, GIoU, and WIoU by 0.56, 0.62, and 0.48 points, respectively. It also attains the highest recall (76.35%)

and precision (78.02%). The gains stem from its scale-aware dynamic weighting: for small targets, it relaxes aspect ratio constraints to focus on center alignment and overlap, reducing false negatives while maintaining low false positives.

Thus, MSAW Loss effectively addresses the wide scale distribution of fire/smoke targets, improving detection robustness across all object sizes.

Table 2 Comparison of relevant loss functions on the Scene-FSD dataset.

Model	Precision	Recall	mAP@50
CIoU	77.66	76.17	78.24
GIoU	77.45	75.98	78.18
WIoU	77.88	75.98	78.32
MSAW Loss	78.02	76.35	78.80

4.2.3 Comparative experiments with state-of-the-art models

(1) Comparison of YOLO series algorithms

To evaluate LMFSNet, we compared it with mainstream YOLO series (YOLOv5s[46], YOLOv6s[47], YOLOv8s[43], YOLOv9s[48], YOLOv10s[49], YOLOv11s[50], YOLOv12s[51]) on Scene-FSD and DFS (Tables 3 and 4).

Performance on Scene-FSD (Table 3): LMFSNet achieves the highest recall (79.21%), surpassing YOLOv11s by 1.53 pp, validating its capability for small initial fires and sparse smoke. This gain stems from EMCf and SlimHead’s P2 layer. It also leads in precision (79.85%) and mAP@50 (82.11%), ranking first. Training loss declines smoothly and mAP steadily increases before stabilizing (Figure 8), confirming the robust optimization of our loss function and architecture.

Model efficiency: LMFSNet has 6.53M parameters and 13.21 GFLOPs—the lowest among all models, representing reductions of 33.8% and 36.3% compared to YOLOv12s. This efficiency comes from the MFE-C2f backbone and SFTM lightweight Neck. Despite extreme lightweighting, LMFSNet achieves the fastest inference speed of 92 FPS, demonstrating an exceptional accuracy-speed balance.

Generalization on DFS (Table 4): LMFSNet achieves 85.24% mAP@50, outperforming YOLOv11s by 1.28 pp. Its recall (83.4%) and precision (87.2%) are also highest, indicating no overfitting to specific datasets. The modular design adapts well to diverse data distributions. Efficiency remains superior: 4.56M parameters, 5.79 GFLOPs, and 91 FPS, enabling stable, accurate, and rapid detection across real-world scenarios with minimal resources.

Table 3 Comparison of relevant YOLO series models on the Scene-FSD dataset.

Model	Precision	Recall	mAP@50	FPS	Params (m)	GFLOPs
YOLOv5s	72.43	71.26	73.58	65	10.08	26.14
YOLOv6s	75.28	72.51	76.49	68	15.46	29.65
YOLOv8s	77.66	76.17	78.24	73	11.02	27.25
YOLOv9s	78.51	75.27	79.86	69	12.26	26.58
YOLOv10s	79.08	74.60	79.78	68	13.33	27.15
YOLOv11s	79.24	77.68	81.05	80	11.27	24.68
YOLOv12s	79.53	76.28	80.20	83	9.86	20.74
LMFSNet(ours)	79.85	79.21	82.11	92	6.53	13.21

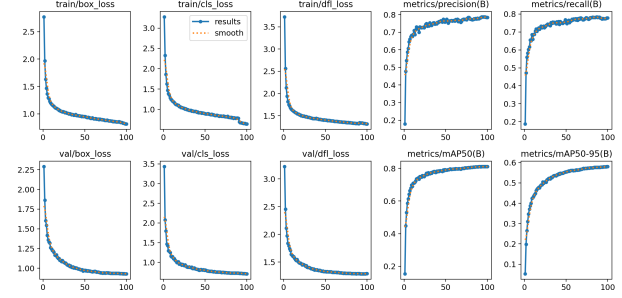


Fig. 8 Training and performance curves of LMFSNet on the Scene-FSD dataset

Table 4 Comparison of relevant YOLO series models on the DFS dataset

Model	Precision	Recall	mAP@50	FPS	Params (m)	GFLOPs
YOLOv5s	79.2	72.1	78.26	57	12.19	15.12
YOLOv6s	81.6	74.8	80.79	61	11.54	14.46
YOLOv8s	83.5	77.2	82.10	68	11.17	13.78
YOLOv9s	84.3	78.1	82.78	71	10.74	12.88
YOLOv10s	82.9	76.4	81.57	74	9.53	11.82
YOLOv11s	85.1	79.3	83.96	86	7.23	9.17
YOLOv12s	84.5	78.6	83.10	85	7.45	9.89
LMFSNet(ours)	87.2	83.4	85.24	91	4.56	5.79

(2) Comparison with mainstream framework models

We further compared LMFSNet with Faster R-CNN[9], RetinaNet[10], MobileNet[11], EfficientNet[12], Swin Transformer[14], and RT-DETR[15] on Scene-FSD and DFS (Tables 5 and 6).

LMFSNet achieves superior accuracy and efficiency. It attains the highest mAP@50 and recall on both datasets, demonstrating strong detection of small and weak-featured smoke, reducing missed detections, while maintaining leading precision. Efficiency gains are more pronounced: LMFSNet uses about one-tenth the parameters and computation of large models (e.g., Faster R-CNN, Swin Transformer) and runs fastest. Compared to lightweight models (MobileNet, EfficientNet), LMFSNet delivers significant accuracy gains with fewer parameters, lower FLOPs, and faster speed. For example, on Scene-FSD, LMFSNet’s mAP@50 outperforms MobileNet by 12 pp, with 31% fewer parameters and 6% lower computation.

Overall, LMFSNet synergistically achieves high accuracy and efficiency through integrated design, sur-

passing general-purpose frameworks and validating its lightweight multi-scale architecture for fire/smoke detection.

Table 5 Comparison of relevant state-of-the-art models on the Scene-FSD dataset.

Model	Precision	Recall	mAP@50	FPS	Params (m)	GFLOPs
Faster R-CNN	66.58	60.52	64.56	23	60.17	150.89
RetinaNet	70.20	64.12	68.18	52	38.10	80.28
MobileNet	72.28	66.22	70.26	78	9.46	14.02
EfficientNet	71.56	65.50	69.54	76	11.12	15.74
Swin Transformer	78.45	76.39	80.43	43	60.25	140.88
RT-DETR	81.14	76.08	80.12	42	63.01	142.67
LMFSNet(ours)	79.85	79.21	82.11	92	6.53	13.21

Table 6 Comparison of relevant state-of-the-art models on the DFS dataset.

Model	Precision	Recall	mAP@50	FPS	Params (m)	GFLOPs
Faster R-CNN	69.42	64.62	67.42	27	50.28	146.67
RetinaNet	72.21	67.41	70.21	55	20.06	61.78
MobileNet	71.56	66.76	69.56	79	6.53	10.65
EfficientNet	73.40	68.60	71.40	78	9.36	12.59
Swin Transformer	83.31	78.51	81.31	46	45.34	130.46
RT-DETR	84.25	79.45	82.25	44	48.65	135.29
LMFSNet(ours)	87.2	83.4	85.24	91	4.56	5.79

(3) Small Object Detection Performance Analysis

To validate LMFSNet’s detection of small targets (e.g., early-stage fire, faint smoke), we defined small objects as those with bounding box area $\leq 0.5\%$ of image area. From Scene-FSD, we selected 1,243 images containing such targets (3,892 instances) as a small-object subset. Table 7 compares LMFSNet with YOLOv8s and other lightweight models on this subset.

LMFSNet achieves the highest mAP@50 (78.3%), surpassing YOLOv8s (70.1%) by 7.2 pp, and recall of 75.6% vs. 66.8% (+8.8 pp). These gains stem from the synergistic effect of MFE-C2f’s progressive multi-scale perception, EMCF’s global context enhancement, and SlimHead’s P2 layer, which collectively improve feature extraction and localization for small objects.

Table 7 Small object detection performance comparison on the Scene-FSD small object test subset.

Model	Precision	Recall	mAP@50
YOLOv8s	73.5	66.8	70.1
YOLOv9s	75.2	68.9	72.3
YOLOv12s	77.6	71.5	74.8
LMFSNet(ours)	80.4	75.6	77.3

4.3 Ablation Study

To validate each proposed module, we conducted ablation studies on Scene-FSD using YOLOv8s as baseline.



Fig. 9 Comparison of detection effects similar to those of fire and smoke.

We sequentially added MSAW Loss, MFE-C2f, EMCF, SFTM, and SlimHead. Results are in Table 8.

Lightweight Backbone MFE-C2f (Model c): MFE-C2f reduces parameters by 21.8% and FLOPs by 40.7%, increases recall by 1.55 pp (to 78.40%), mAP by 1.75 pp (to 80.55%), and FPS to 81. Its parallel multi-branch and progressive fusion reduce redundancy while enhancing multi-scale feature representation.

Context Fusion Module EMCF (Model d): EMCF adds global context, raising mAP by 0.65 pp (to 81.20%) and recall by 0.50 pp (to 78.90%). Its cross-attention models long-range dependencies across scales, improving discrimination of ambiguous targets (blurred edges, semi-transparent smoke, background-like smoke), reducing missed detections and false positives.

Lightweight Neck SFTM (Model e): SFTM optimizes feature fusion: parameters drop by 18.4%, FPS increases by 10.1% to 87, while mAP@50 slightly rises to 81.75%. Its three-path design reduces computational and memory overhead while ensuring effective multi-scale information flow.

SlimHead (Model f): Adding the P2-based SlimHead yields the final LMFSNet: 6.53M parameters, 13.21 GFLOPs, 92 FPS, mAP@50 82.11%, recall 79.21%. The P2 layer leverages high-resolution details, synergizing with multi-scale features and global context to achieve optimal accuracy-efficiency balance.

5 Analysis and Discussion

5.1 Visualization of detection effect on complex backgrounds similar to fire and smoke

We constructed a test set with challenging negative samples: variable clouds, haze, sun glare, artificial lights, and warm-toned buildings. As shown in Figure 9, baseline and competing models often produce high-confidence false detections in such confusing scenarios, over-relying on local superficial features.

In contrast, LMFSNet maintains high classification accuracy without false alarms across all test cases. This

Table 8 Ablation experiment results.

Model	MSAW	MFE-C2f	EMCF	SFTM	SlimHead	Precision	Recall	mAP@50	FPS	Params (m)	GFLOPs
(a) YOLOv8s	—	—	—	—	—	77.66	76.17	78.24	73	11.02	27.25
(b)	✓	—	—	—	—	78.02	76.85	78.80	73	11.02	27.25
(c)	✓	✓	—	—	—	78.50	78.40	80.55	81	8.62	16.16
(d)	✓	✓	✓	—	—	79.10	78.90	81.20	79	8.98	16.32
(e)	✓	✓	✓	✓	—	79.55	79.05	81.75	87	7.33	15.08
(f) LMFSNet(ours)	✓	✓	✓	✓	✓	79.85	79.21	82.11	92	6.53	13.21

**Fig. 10** Comparison chart of experimental results.**Fig. 11** Visualization results using Grad-CAM for the entire image.**Fig. 12** Visualization results using Grad-CAM for fire and smoke, respectively.

robustness comes from the EMCF module's global semantic understanding, which transcends local texture and color similarities by analyzing scene layout and semantic relationships, effectively suppressing false alarms

and enhancing reliability in open, complex environments.

5.2 Visual analysis of fire and smoke detection performance

We compared models on positive samples containing small incipient fires, sparse smoke, lighting variations, shadows, and cluttered backgrounds (Figure 10). Baselines show significant false negatives for faint fires and blurred smoke. Later models partially improve recall but lack localization accuracy and stability.

LMFSNet achieves comprehensive gains: (1) SlimHead with high-resolution P2 layer precisely localizes early-stage minute fires, significantly improving recall. (2) MFE-C2f multi-scale extraction and EMCF context enhancement capture diffuse, semi-transparent smoke contours more completely, producing tighter bounding boxes. Under dynamic lighting and complex backgrounds, LMFSNet detection results show excellent robustness and consistency.

5.3 Feature visualization analysis

We used Grad-CAM to generate feature activation heatmaps, revealing LMFSNet's perceptual characteristics at two levels.

First (Figure 11), LMFSNet attention distributions align well with target semantics: fire heatmaps concentrate on high-temperature regions, smoke heatmaps cover diffuse, semi-transparent contours. In contrast, YOLOv8s activations are more dispersed with false activations on background distractors sharing similar colors/textures, showing LMFSNet's features are more discriminative. Second (Figure 12), Grad-CAM heatmaps for independent fire and smoke channels show LMFSNet's finer perception: on the fire channel, it responds to early, small, isolated fires; on the smoke channel, activation covers smoke body with uniform, continuous contours matching delicate edges. Baseline activations are coarse and confused, with weak small-target responses and insufficient smoke edge capture. This scale-sensitive, morphologically adaptive pattern enables LMFSNet to precisely distinguish fire from smoke.

6 Conclusion

This paper proposes LMFSNet, a lightweight multi-scale detection network for fire and smoke, addressing extreme scale variations, ambiguous shapes, and real-time requirements. The model employs MFE-C2f, EMCF, SFTM, and SlimHead with P2 layer and MSAW loss to improve efficiency, background discrimination, and small-target sensitivity. Experiments on Scene-FSD and DFS show LMFSNet outperforms mainstream models in accuracy and efficiency, achieving an excellent balance for fire early warning.

Training was done on GPUs. While LMFSNet is suitable for edge deployment, validation on edge devices remains unfinished—the main limitation. Future work includes adapting to edge hardware (e.g., Jetson series), evaluating latency, power, and stability, exploring compression and hardware co-optimization, as well as continuous learning and algorithmic interpretability to enhance system credibility.

Acknowledgments

The work was sponsored by National Key Research and Development Program of China (Program No. 2020YFB1711902).

Data availability

Data will be made available on request.

References

1. Dener Silva, Tiago Miguel Ferreira, and Hugo Rodrigues. Wildland-urban interface fire exposure of rural settlements: The case of montesinho natural park. *International Journal of Disaster Risk Reduction*, 112:104790, 2024.
2. Yaowen Hu, Jialei Zhan, Guoxiong Zhou, Aibin Chen, Weiwei Cai, Kun Guo, Yahui Hu, and Liujun Li. Fast forest fire smoke detection using mvnnet. *Knowledge-Based Systems*, 241:108219, 2022.
3. Seong G Kong, Donglin Jin, Shengzhe Li, and Hakil Kim. Fast fire flame detection in surveillance video using logistic regression and temporal smoothing. *Fire Safety Journal*, 79:37–43, 2016.
4. Pengkun Liu, Shuna Ni, Stoliarov Stanislav, Pingbo Tang, et al. Automated image-based identification and consistent classification of fire patterns with quantitative shape analysis and spatial location identification. *Developments in the Built Environment*, 21:100612, 2025.
5. Kosmas Dimitropoulos, Panagiotis Barmoutis, and Nikos Grammalidis. Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. *IEEE transactions on circuits and systems for video technology*, 25(2):339–351, 2014.
6. V Vipin. Image processing based forest fire detection. *International Journal of Emerging Technology and Advanced Engineering*, 2(2):87–95, 2012.
7. Xiao-Han Chen, Xue-Yin Zhang, and Qian-Xi Zhang. Fire alarm using multi-rules detection and texture features classification in video surveillance. In *2014 7th International Conference on Intelligent Computation Technology and Automation*, pages 264–267, 2014.
8. Zhang Qian, Liu Xiao-Jun, and Huang Lei. Video image fire recognition based on color space and moving object detection. In *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pages 367–371, 2020.
9. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
10. T-YLPG Ross, GKHP Dollár, et al. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.
11. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
12. Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114, 2019.
13. Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
14. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
15. Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024.
16. Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine learning and knowledge extraction*, 5(4):1680–1716, 2023.
17. Bensheng Yun, Yanan Zheng, Zhenyu Lin, and Tao Li. Ffyolo: a lightweight forest fire detection model based on yolov8. *Fire*, 7(3):93, 2024.
18. Linsong Xiao, Wenzao Li, Xiaoqiang Zhang, Hong Jiang, Bing Wan, and Dehao Ren. Emg-yolo: An efficient fire detection model for embedded devices. *Digital Signal Processing*, 156:104824, 2025.
19. Xinzhi Wang, Mengyue Li, Mingke Gao, Quanyi Liu, Zhennan Li, and Luyao Kou. Early smoke and flame detection based on transformer. *Journal of Safety Science and Resilience*, 4(3):294–304, 2024.
20. Praveen Kumar Sekharamanthy, Farid Melgani, Jonni Malacarne, Riccardo Ricci, Rodrigo de Almeida Silva, and Jose Marcato Junior. A seamless deep learning approach for apple detection, depth estimation, and tracking using yolo models enhanced by multi-head attention mechanism. *Computers*, 13(3):83, 2024.
21. Alireza Shamsoshoara, Fatemeh Afghah, Abolfazl Razi, Liming Zheng, Peter Z Fulé, and Erik Blasch. Aerial imagery pile burn detection using deep learning: The flame dataset. *Computer Networks*, 193:108001, 2021.

22. Zulfiqar Ahmad Khan, Tanveer Hussain, Fath U Min Ullah, Suneet Kumar Gupta, Mi Young Lee, and Sung Wook Baik. Randomly initialized cnn with densely connected stacked autoencoder for efficient fire detection. *Engineering Applications of Artificial Intelligence*, 116:105403, 2022.
23. Jian Feng and Yu Sun. Multiscale network based on feature fusion for fire disaster detection in complex scenes. *Expert Systems with Applications*, 240:122494, 2024.
24. Hai Li, Zheng Ma, Sheng-Hua Xiong, Qiang Sun, and Zhen-Song Chen. Image-based fire detection using an attention mechanism and pruned dense network transfer learning. *Information Sciences*, 670:120633, 2024.
25. Zulfiqar Ahmad Khan, Fath U Min Ullah, Hikmat Yar, Waseem Ullah, Noman Khan, Min Je Kim, and Sung Wook Baik. Optimized cross-module attention network and medium-scale dataset for effective fire detection. *Pattern Recognition*, 161:111273, 2025.
26. Jianmei Zhang, Hongqing Zhu, Pengyu Wang, and Xiao-feng Ling. Att squeeze u-net: A lightweight network for forest fire detection and recognition. *Ieee Access*, 9:10858–10870, 2021.
27. Ji Lin, Haifeng Lin, and Fang Wang. Stpm_sahi: A small-target forest fire detection model based on swin transformer and slicing aided hyper inference. *Forests*, 13(10):1603, 2022.
28. Jialei Zhan, Yaowen Hu, Weiwei Cai, Guoxiong Zhou, and LiuJun Li. Pdam-stpnnet: A small target detection approach for wildland fire smoke through remote sensing images. *Symmetry*, 13(12):2260, 2021.
29. Jingwen Huang, Jiashun Zhou, Huizhou Yang, Yunfei Liu, and Han Liu. A small-target forest fire smoke detection model based on deformable transformer for end-to-end object detection. *Forests*, 14(1):162, 2023.
30. Rui Li, Yaowen Hu, Lin Li, Renxiang Guan, Ruoli Yang, Jialei Zhan, Weiwei Cai, Yanfeng Wang, Haiwen Xu, and LiuJun Li. Smwe-gfpnnnet: A high-precision and robust method for forest fire smoke detection. *Knowledge-Based Systems*, 289:111528, 2024.
31. Shengbao Wu, Buyun Sheng, Gaocai Fu, Daode Zhang, and Yuchao Jian. Multiscale fire image detection method based on cnn and transformer. *Multimedia Tools and Applications*, 83(16):49787–49811, 2024.
32. Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12021–12031, 2023.
33. Li Jin, Yanqi Yu, Jianing Zhou, Di Bai, Haifeng Lin, and Hongping Zhou. Swvr: A lightweight deep learning algorithm for forest fire detection and recognition. *Forests*, 15(1):204, 2024.
34. Zhicheng Cai and Qiu Shen. Falconnet: Factorization for the light-weight convnets. In *International Conference on Neural Information Processing*, pages 368–380, 2023.
35. Wen-Tsai Sung, Indra Griha Tofik Isa, Ching-Hsiang Tseng, and Sung-Jung Hsiao. Constructing a lightweight fire and smoke detection through the improved ghostnet architecture and attention module mechanism. *IEEE Access*, 2025.
36. Yinsheng Chen, Jiahao Li, Kun Sun, and Ying Zhang. A lightweight early forest fire and smoke detection method. *Journal of supercomputing*, 80(7), 2024.
37. Pengfei Shen, Ning Sun, Kai Hu, Xiaoling Ye, Pingping Wang, Qingfeng Xia, and Chen Wei. Firevit: An adaptive lightweight backbone network for fire detection. *Forests*, 14(11):2158, 2023.
38. Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12021–12031, 2023.
39. Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.
40. Hulin Li, Jun Li, Hanbing Wei, Zheng Liu, Zhenfei Zhan, and Qiliang Ren. Slim-neck by gsconv: A lightweight-design for real-time detector architectures. *Journal of Real-Time Image Processing*, 21(3):62, 2024.
41. Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rong-guang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics*, 52(8):8574–8586, 2021.
42. Siyuan Wu, Xinrong Zhang, Ruqi Liu, and Binhai Li. A dataset for fire and smoke object detection. *Multimedia Tools and Applications*, 82(5):6707–6726, 2023.
43. Mupparaju Sohan, Thotakura Sai Ram, and Ch Venkata Rami Reddy. A review on yolov8 and its advancements. pages 529–545, 2024.
44. Hamid Reza Tofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. pages 658–666, 2019.
45. Zanjia Tong, Yuhang Chen, Zewei Xu, and Rong Yu. Wise-iou: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*, 2023.
46. Qifeng Yu, Yudong Han, Wuguang Lin, and Xinjia Gao. Detection and analysis of corrosion on coated metal surfaces using enhanced yolo v5 algorithm for anti-corrosion performance evaluation. *Journal of Marine Science and Engineering*, 12(7):1090, 2024.
47. Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.
48. Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. pages 1–21, 2024.
49. Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.
50. Priyanto Hidayatullah, Nurjannah Syakrani, Muhammad Rizqi Sholahuddin, Trisna Gelar, and Refdinal Tubagus. Yolov8 to yolo11: A comprehensive architecture in-depth comparative review. *arXiv preprint arXiv:2501.13400*, 2025.
51. Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.