remote sensing

MDPI

Article

Real-Time Early Warning of Incipient Fire in Multiple Urban Scenarios: A Deep Learning-Based Monitoring Method

Lingyi Meng [1], Mengquan Wu [1,*], Jinkun Gao [1], Shikuan Wang [2], Xiaodong Song [3], Jie Zhao [4], Hongchun Liu [4], Xindan Cao [1], Longxing Liu [5], Gang Chen [1] and Jinyi Lv [1]

[1] College of Resources and Environmental Engineering, Ludong University, Yantai 264025, China; 2025120750@m.ldu.edu.cn (L.M.); 20212604721@m.ldu.edu.cn (J.G.); 20222604804@m.ldu.edu.cn (X.C.); 20212604715@m.ldu.edu.cn (G.C.); 2023110243@m.ldu.edu.cn (J.L.)
[2] School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China; wangsk@e.gzhu.edu.cn
[3] College of Geomatics & Municipal Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China; xdsong@zjweu.edu.cn
[4] Yantai Geographic Information Center, Yantai 264039, China; zhaojie6711@163.com (J.Z.); 18917332957@163.com (H.L.)
[5] International Institute for Earth System Science, Nanjing University, Nanjing 210023, China; liulx@smail.nju.edu.cn
* Correspondence: ld_wmq@ldu.edu.cn

Highlights

What are the main findings?

- YOLO-Fire achieves a high-precision balance between detection accuracy (75.7% mAP50) and lightweight architecture (10.02 M parameters) for urban environments.
- The novel HFFM, DCD, and GSPPF modules effectively decouple fire–smoke features and suppress complex background clutter in unconstrained scenes.

What are the implications of the main findings?

- The lightweight design facilitates real-time fire monitoring across resource-constrained environments and UAV-based remote sensing platforms.
- The proposed feature decoupling and diffusion strategies provide a robust technological foundation for automated early warning systems in smart city safety management.

Abstract

Urban fire incidents in complex built environments pose severe threats to public safety. However, the unstructured nature of urban scenes presents substantial challenges for existing detection algorithms in reliably identifying incipient flames and diffuse smoke under dynamic visual interference. To address this issue, we propose YOLO-Fire, a lightweight and high-precision detection algorithm based on YOLOv11. Specifically, a Hybrid Feature Fusion Module (HFFM) adopts a parallel dual-stream architecture to structurally decouple high-frequency flame boundaries from low-frequency smoke textures. A Dual-Scale Contextual Diffusion (DCD) mechanism establishes global contextual constraints through an additive diffusion strategy, effectively suppressing fire-like background interference while enhancing semi-transparent smoke features. In addition, a Gaussian Spatial Pyramid Pooling Fast (GSPPF) module further improves multi-scale receptive field aggregation. Evaluated on a self-constructed large-scale urban fire dataset, YOLO-Fire achieves an mAP50 of 75.7%, mAP50-95 of 53.3%, and an F1-score of 73.7%, with only 10.02 M parameters, surpassing the YOLOv11 baseline by 2.4%, 4.5%, and 2.9%, respectively. Ablation studies confirm that each proposed module contributes both independently and synergistically

Check for updates

Academic Editor: Jungho Im

Received: 31 March 2026
Revised: 13 May 2026
Accepted: 20 May 2026
Published: 21 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

to the overall performance gains. Comprehensive comparisons with mainstream detectors and specialized fire detection models further demonstrate that YOLO-Fire achieves superior overall performance, outperforming YOLO-FireAD and FireSmoke-YOLO by 2.7% and 2.4% in mAP50, respectively, while maintaining lower computational complexity. Furthermore, inference evaluation on a single-core CPU achieves 17.28 FPS, validating the practical deployment potential of YOLO-Fire in resource-constrained environments and offering an efficient, lightweight solution for real-time urban fire surveillance and early warning.

Keywords: fire and smoke detection; deep learning; object detection; remote sensing; real-time early warning; multi-scale feature fusion

1. Introduction

Rapid global urbanization and population agglomeration have catalyzed economic prosperity while simultaneously undermining the resilience of urban public safety systems [1]. Urbanization, while driving economic and social progress, concurrently introduces significant challenges, most notably the escalating incidence of urban fires [2]. The intrinsic complexity of urban environments—characterized by dense populations, high-rise building clusters, traffic congestion, and heterogeneous topography—renders fire prevention and emergency response exceptionally difficult [3]. Under such conditions, fires can result in catastrophic casualties and substantial property losses. Research indicates that fires are responsible for approximately 50,000 deaths and 170,000 injuries globally each year, with the cumulative fatalities and economic losses attributable to urban fires frequently exceeding those of forest fires globally [4]. During the period from 2001 to 2020, more than 90 million fire incidents were recorded worldwide [5]. In China alone, a 2018 national fire report documented 237,000 urban fires, resulting in 1407 deaths, 798 injuries, and direct economic losses of 3.67 billion CNY [6]. These statistics highlight the profound threats posed to social stability and public safety, underscoring the urgent need for high-precision, real-time fire detection and perception technologies [7,8].

Addressing this fire safety challenge, traditional technologies such as fire alarm systems and smoke detectors have played a foundational role in urban fire prevention and suppression. However, their inherent technical limitations render them inadequate for meeting current detection and control demands [9,10]. Fire alarm systems are highly susceptible to environmental interference induced by high temperatures, elevated humidity, and equipment aging, causing frequent false alarms and significantly increasing operational costs [11]. Similarly, smoke detectors suffer from limited monitoring coverage and poor environmental adaptability [12]. Their passive triggering mechanisms, which rely on smoke accumulation, often result in delayed detection, and achieving comprehensive coverage necessitates the dense deployment of numerous devices, further escalating system complexity and deployment costs [13].

To overcome the inherent limitations of traditional point-based monitoring technologies, advances in remote sensing satellite technology offer a viable path for wide-area smoke detection [14]. Leveraging its advantages of broad coverage, real-time monitoring, and relatively low operational costs, satellite imagery has demonstrated significant potential in fire smoke monitoring [15,16]. However, elevated false alarm rates arising from complex background environments remain a critical bottleneck limiting its practical efficacy. Smoke manifests highly dynamic characteristics in terms of scale, intensity, and morphology, significantly compounding the difficulty of accurate identification [17]. Moreover, interfering

phenomena such as clouds, fog, and dust exhibit strong similarities to smoke in terms of texture, color, and spectral characteristics, readily causing identification confusion and detection errors [18].

To address the aforementioned technical challenges of accurate fire-related object recognition in complex environments, the rapid advancement of deep learning has provided a robust methodological foundation for fire monitoring algorithms, facilitating the widespread application of object detection methods in fire identification [19–21]. These algorithms, trained on large-scale fire image datasets, automatically identify and classify fire-related features, thereby improving detection accuracy. For instance, Barmpoutis et al. [22] integrated Faster R-CNN with multidimensional texture analysis (Linear Dynamic Systems, LDS) for an enhanced fire detection framework. By analyzing the dynamic texture features of candidate regions, this method effectively distinguishes actual fire from spectrally similar background interferences, significantly reducing false positive rates in complex environments. Similarly, Guan et al. [23] employed an improved Mask R-CNN model to process aerial imagery, reconstructing the mask branch to refine the perception of irregular flame boundaries, achieving pixel-level segmentation of forest fires with high precision and providing reliable data support for fire spread estimation.

Although the aforementioned algorithms excel in detection accuracy, their considerable parameter counts and high computational latency often limit their applicability in real-time scenarios [24,25]. In contrast, one-stage object detection algorithms, owing to their superior inference speed, have gradually become the mainstream choice for real-time early-warning tasks [26,27]. Li and Zhao [28] evaluated multiple architectures, including Faster R-CNN and SSD, confirming the superiority of YOLOv3 in balancing detection speed and accuracy, thus establishing its foundational role in real-time warning tasks. To address the deployment challenges of Transformer-based models on low-power devices, Zheng et al. [29] proposed the FTA-DETR framework. By optimizing encoder efficiency and leveraging TensorRT acceleration, this model achieved an accuracy of 98.32% and an inference speed of 76 FPS, demonstrating that end-to-end detectors can simultaneously satisfy high-precision and real-time requirements in edge computing. Furthermore, P et al. [30] investigated false alarms induced by clouds and sunset afterglow by benchmarking YOLOv8, YOLOv11, and YOLOv12 architectures. Their results indicated that YOLOv12 effectively mitigated complex background interference, achieving an accuracy of 97.17%, whereas YOLOv11 demonstrated clear advantages in resource-constrained field deployments owing to its 15 MB lightweight architecture and millisecond-level inference speed.

Despite these advances in real-time performance and robustness to background interference, existing one-stage detectors still exhibit notable deficiencies in feature extraction for extremely small fire sources and morphologically variable smoke [31]. To address this challenge, Wang et al. [8] proposed incorporating dynamic snake convolutions, FSPPF modules, and dedicated small object detection layers to improve feature representation. Similarly, for small-target fire detection, Wang et al. [32] proposed DCSNet, which employs dynamic contextual aggregation and partial cross-stage feature fusion to enhance boundary perceptibility. However, such strategies yield limited performance gains in scenarios involving extremely weak fire sources or optically thin smoke. Moreover, the stacking of complex modules incurs substantial increases in FLOPs and parameter counts, compromising the lightweight advantage of these models. Consequently, mainstream models continue to face significant limitations in detecting small targets and smoke, underscoring an urgent need to improve their generalization capability.

To overcome the inherent trade-off between computational efficiency and detection accuracy in existing deep learning models, this study proposes YOLO-Fire, a high-precision,

noise-robust framework specifically designed for fire detection in complex, safety-critical environments. Built upon the YOLOv11 architecture, the proposed method effectively addresses the detection bottlenecks associated with identifying incipient fire sources and diffuse smoke against dynamic background interference. Figure 1 illustrates the urban fire detection pipeline proposed in this study, encompassing data processing, model architecture design, and multi-scenario early-warning applications. The main contributions of this study are summarized as follows.

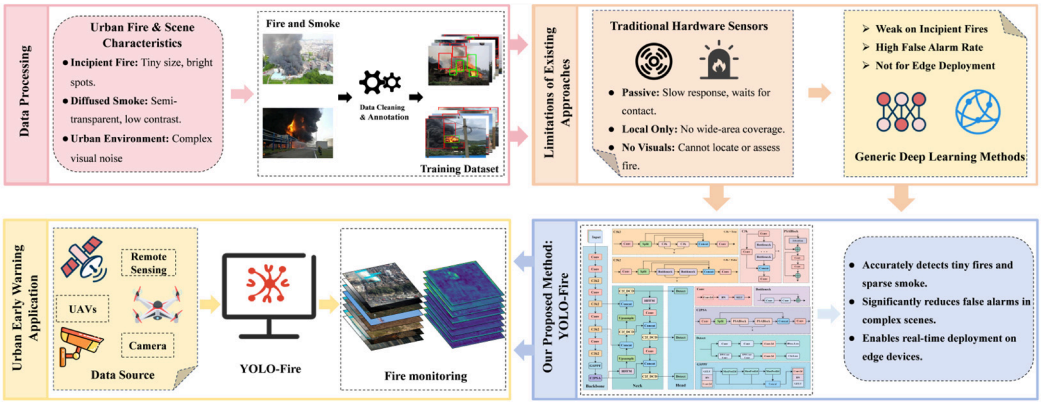

Figure 1. The systematic YOLO-Fire paradigm for urban fire monitoring.

(1) The HFFM is proposed to address the frequency-domain discrepancy between flame and smoke. By integrating a Dual-Attention Grouped Block (DAGB) and a Parallel Depthwise Block (PDB) in parallel, this module implements a dual-branch feature decoupling strategy that separates the high-frequency boundary features of small fire sources from the low-frequency diffusive textures of smoke, significantly improving feature representation at minimal additional parameter cost.

(2) The C2f-DCD module is designed by incorporating a Dual-Scale Contextual Diffusion mechanism. Leveraging the Split-Transform-Fuse strategy, it captures global environmental context via depthwise dilated convolutions and integrates this into local features through an additive fusion scheme. Unlike multiplicative gating, additive injection preserves the intensity of semi-transparent smoke signals while acting as a semantic regularizer to suppress complex background interference, thereby ensuring robust feature discrimination.

(3) The GSPPF module is introduced to handle dynamic morphological variations in fire and smoke targets. By incorporating the GELU activation function for multi-scale feature fusion, this module enlarges the receptive field, enhancing the model's capacity to capture fire targets across multiple scales—from sub-pixel incipient ignition points to large-scale dispersing smoke plumes.

(4) Extensive experiments on a self-constructed cross-scenario fire dataset demonstrate that YOLO-Fire achieves state-of-the-art detection performance. The proposed framework integrates data processing, model architecture, and practical deployment, effectively addressing incipient fire sources and diffuse smoke detection. It provides high-precision technical support for urban safety monitoring and ecological disaster prevention.

2. Materials and Methods

2.1. Dataset

A high-fidelity and comprehensive dataset serves as the cornerstone for training robust data-driven safety systems. To ensure the generalization capability of YOLO-Fire

across complex safety-critical environments, we constructed a large-scale, multi-source fire repository. This process encompasses images captured from diverse platforms including ground-level surveillance cameras, mobile terminals, Unmanned Aerial Vehicles (UAVs), and spaceborne remote sensing satellites, covering a wide spectrum of viewpoints from close-range ground observation to high-altitude remote sensing. Crucially, acknowledging the intensifying risks at the Wildland-Urban Interface (WUI) due to urban expansion, our dataset explicitly incorporates scenarios from these fringe zones. Fire incidents in WUI areas pose direct threats to urban infrastructure and ecological stability. The inclusion of these multi-scale samples ensures that the model possesses reliable detection capabilities not only within concrete building complexes but also in high-risk vegetation areas on the urban periphery, thereby bridging the gap between localized urban safety and broad-scale geospatial monitoring. While UAV and satellite imagery provides invaluable aerial and large-scale perspectives, the scene diversity obtainable solely from these sources is inherently limited; ground-level images captured by surveillance cameras and mobile terminals were therefore also incorporated to ensure comprehensive coverage of diverse urban fire scenarios.

The dataset was constructed through a hybrid collection strategy. A portion of the images was sourced from publicly available fire dataset platforms, including PaddlePaddle AI Studio and similar open-source repositories. However, given the limited scene diversity provided by these public sources, additional urban fire images were further collected from the internet to supplement the dataset, covering a broader range of scenarios including urban building fires, vehicle-induced fires, WUI fires, satellite remote sensing fires, and nighttime low-illumination fires. It should be noted that the UAV and satellite imagery in this dataset was obtained in RGB image format from publicly available platforms and the internet, and therefore does not contain standardized remote sensing metadata such as sensor specifications, spatial resolution, or geographic coordinate information. Although the dataset does not follow a conventional remote sensing data structure with georeferenced imagery, the inclusion of UAV and satellite RGB imagery ensures that the proposed method remains applicable to remote sensing fire monitoring scenarios, extending its applicability from ground-level surveillance to wide-area aerial and satellite observation. To guarantee annotation precision, all images were manually annotated using the Labelme (v5.6.1) tool to ensure consistent annotation quality and labeling standards across the entire dataset. Subsequently, data augmentation techniques were applied during the training phase to expand sample diversity and mitigate overfitting. The final dataset comprises a total of 12,272 finely annotated images, which were randomly partitioned into a training set of 9256 images and a test set of 3016 images. This rigorous dataset partitioning provides a solid foundation for evaluating the model's performance in real-world applications.

As illustrated in Figure 2, this dataset overcomes the limitations of single-scenario datasets by constructing a highly representative multi-source sample repository. In the spatial dimension, the data encompasses a multi-view distribution ranging from ground-level close-range surveillance to high-altitude satellite remote sensing. In the temporal and environmental dimensions, it incorporates all-weather characteristics, spanning diurnal illumination variations and extreme low-light scenarios. This composition, bridging diverse spatiotemporal scales from ground-level surveillance to wide-area satellite remote sensing, establishes the completeness of the dataset and supports the model's applicability across both localized and large-scale fire monitoring scenarios. It effectively simulates complex background interferences found in real-world urban environments, thereby significantly enhancing the model's generalization capability and robustness in practical disaster prevention and mitigation tasks.

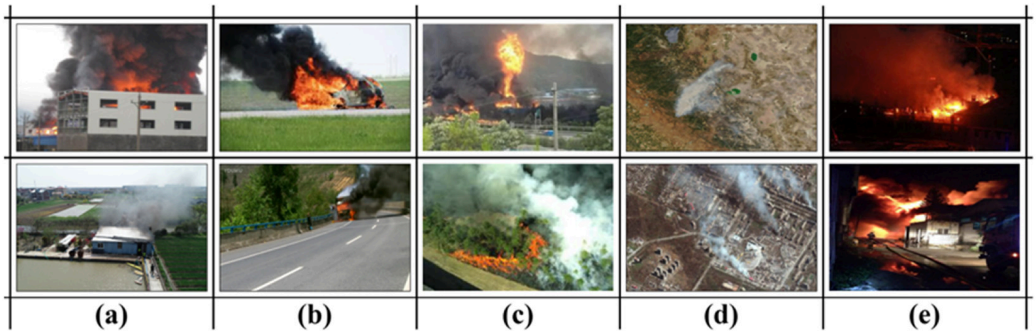

Figure 2. Partial fire datasets. (**a**) Urban building fire. (**b**) Vehicle-induced fire. (**c**) WUI fire. (**d**) Satellite remote sensing fire. (**e**) Nighttime low-illumination fire.

2.2. YOLOv11 Model

Driven by continuous architectural evolution in backbone efficiency and feature integration strategies, the YOLO series has cemented its status as the dominant paradigm in real-time object detection. YOLOv11 [33], the latest generation released by Ultralytics, builds upon the efficient design of YOLOv8 [34], further optimizing parameter efficiency and detection accuracy through advanced feature extraction modules and attention mechanisms. It adheres to the classic Backbone-Neck-Head meta-architecture. The Backbone adopts a Cross Stage Partial (CSP) [35]-based design utilizing C3k2 (Faster Implementation of CSP Bottleneck with 2 convolutions) modules for hierarchical feature extraction. The Spatial Pyramid Pooling-Fast (SPPF) module, distinguished from the traditional Spatial Pyramid Pooling (SPP) [36] design, employs a serial cascade of three max-pooling layers to efficiently expand the spatial receptive field while maintaining fast inference., and a C2PSA (CSP with Position-Sensitive Attention, for enhanced feature extraction and processing) module is introduced at the end of the backbone to capture deep global semantic information. The Neck retains the classic PAN-FPN (Path Aggregation Network with Feature Pyramid Network) [37] architecture, achieving bidirectional aggregation of multi-scale features through top-down upsampling and bottom-up convolutional pathways, with C3k2 modules applied at feature fusion nodes to ensure representational capability and gradient flow stability. The Head adopts the Decoupled Head [38] paradigm with an Anchor-free [39] strategy, directly predicting target centers and boundary distances, jointly optimized using DFL (Distribution Focal Loss) [40] and CIoU Loss [41] for robust and precise bounding box localization.

2.3. Overall Architecture

Although YOLOv11 performs exceptionally well in general object detection tasks, its standard architecture faces challenges in simultaneously addressing the extremely minute scale of early flames and the semi-transparent, blurry characteristics of diffused smoke in early fire warning scenarios. To overcome these challenges, we propose an enhanced detection architecture tailored to the specific characteristics of fire scenarios, as illustrated in Figure 3. While retaining the efficient inference capabilities of YOLOv11, this architecture performs a strategic restructuring of the critical feature extraction and fusion stages.

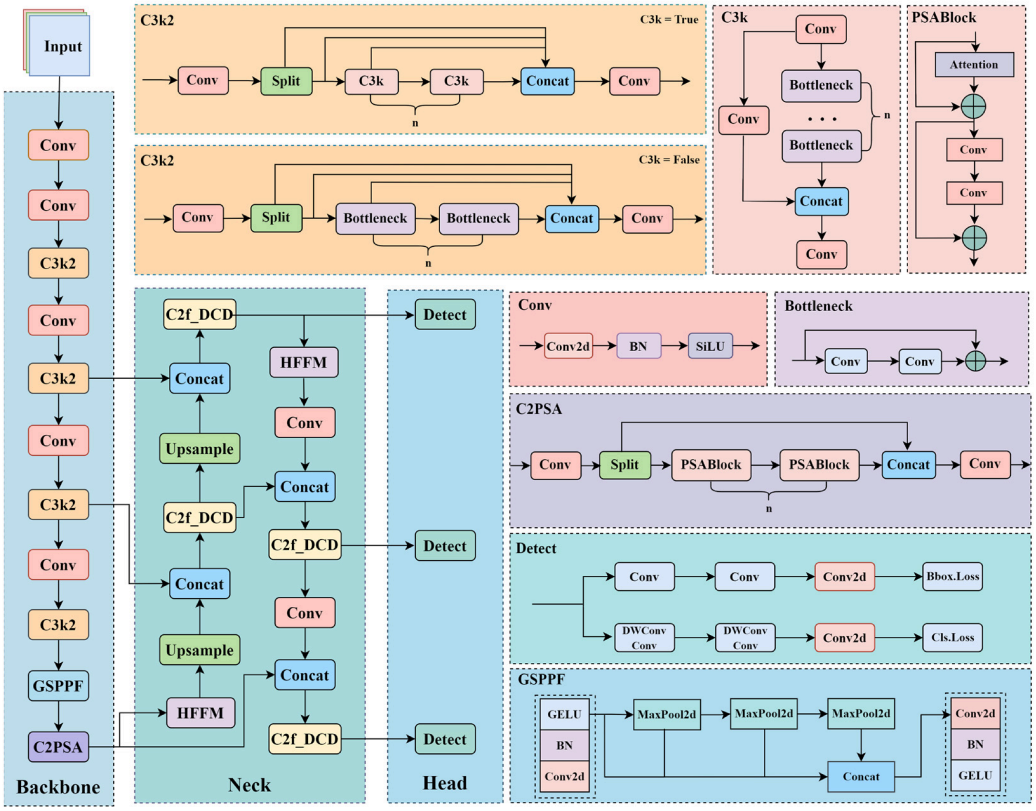

Figure 3. YOLO-Fire Network Architecture. Note: GSPPF replaces the original SPPF in the Backbone; the first HFFM is inserted at the beginning of the Neck before the first upsampling operation; the second HFFM is inserted after the second C2f-DCD and before the first Conv layer in the Neck; all C2f modules in the Neck are replaced by C2f-DCD.

First, to resolve the feature discrepancy inherent in processing two distinct targets—"fire" and "smoke"—using traditional fusion methods, we propose the HFFM in the Neck stage. Moving beyond simple feature concatenation, HFFM adopts a parallel dual-stream paradigm for feature decoupling. Specifically, the PDB branch focuses on preserving high-frequency, point-like features, ensuring that early tiny flames are not smoothed out in deep layers. Conversely, the DAGB branch specializes in enhancing low-frequency, large-scale semi-transparent textures to capture the amorphous morphology of smoke. This hierarchical fusion strategy ensures that the model balances the representation capabilities for both fine-grained fire and diffused smoke during the feature aggregation stage, significantly improving detection robustness in complex fire scenes.

Secondly, addressing the challenge where smoke features are highly coupled with complex backgrounds (e.g., clouds and lighting interference) and tiny fire sources are prone to being overwhelmed by environmental noise (e.g., urban lights and strong reflections) due to pixel scarcity, we introduce the C2f-DCD module into the Neck, replacing the original C2f structure. Through an explicit contextual diffusion mechanism, C2f-DCD constrains local texture extractors to perform feature discrimination conditional on global environmental information. This design not only enhances the network's capability to distinguish blurry smoke boundaries under low-contrast conditions but also provides critical "semantic verification" for tiny objects. By relying on background logic, it effectively suppresses false interferences resembling fire points, thereby significantly improving the recall rate for extremely small fire sources.

Finally, to adapt to the extreme scale variations ranging from local fire points to large-area dense smoke, a GSPPF module is deployed at the end of the backbone. Utilizing an optimized pooling strategy and the GELU [42] activation function, this module aggregates

multi-scale global context with negligible computational overhead, effectively expanding the effective receptive field of deep features.

It is worth noting that the three proposed modules are not only tailored for general urban fire detection but are also specifically designed to address the unique challenges inherent in remote sensing scenarios. In UAV and satellite imagery, fire targets often manifest as sub-pixel ignition points embedded in complex terrain backgrounds, while smoke plumes may span large spatial extents with highly variable morphology. The GSPPF module directly addresses the large-scale scene challenge by aggregating multi-scale receptive fields to capture both localized fire points and wide-area smoke diffusion patterns. The HFFM module tackles the small target detection challenge in remote sensing by preserving high-frequency boundary details of tiny ignition points through the PDB branch, preventing their erosion during deep feature extraction. The C2f-DCD module further enhances robustness in remote sensing scenarios by injecting global terrain context into local feature extraction, effectively suppressing complex background interference such as cloud formations and varied land cover textures that are characteristic of aerial and satellite imagery.

2.3.1. Hybrid Feature Fusion Module

Conventional feature aggregation methods typically couple global and local features within a single processing stream. This design limits the network's ability to recalibrate blurry smoke features in uncontrolled environments and hinders its flexibility in responding to drastic changes in target scale [43]. To break this coupling limitation between global and local representations, we designed the Hybrid Feature Fusion Module (HFFM), as shown in Figure 4. It adopts a parallel dual-stream cooperative framework aimed at decoupling feature recalibration from contextual aggregation.

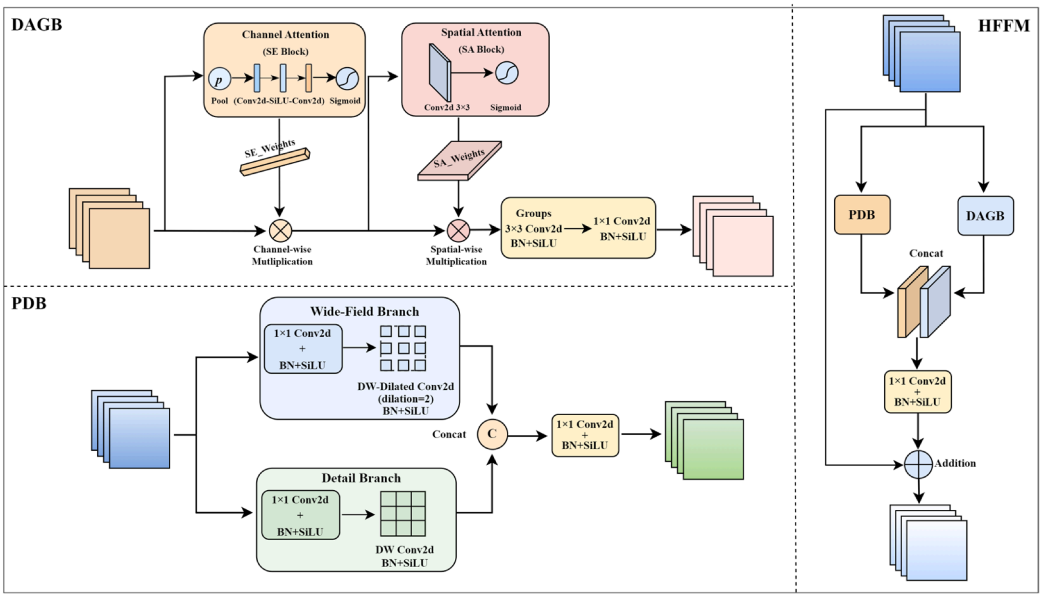

Figure 4. HFFM structure.

Given an input feature tensor $X \in \mathbb{R}^{C \times H \times W}$, HFFM diverts the information flow into two parallel branches: the Dual-Attention Grouped Block (DAGB) for texture refinement and the Parallel Depthwise Block (PDB) for multi-scale perception. These heterogeneous features are subsequently integrated through a fusion layer to produce the refined output $Y \in \mathbb{R}^{C \times H \times W}$. The overall formulation is expressed as:

$$Y = \mathcal{F}_{fuse}(\text{Concat}[\mathcal{F}_{DAGB}(X), \mathcal{F}_{PDB}(X)]) + X \quad (1)$$

where $\mathcal{F}_{fuse}$ denotes a 1×1 convolution followed by Batch Normalization and the SiLU activation function, used to project the concatenated features back to the original channel dimension. $\mathcal{F}_{\mathrm{DAGB}}$ represents the operation function of the DAGB branch, and $\mathcal{F}_{PDB}(\mathrm{X})$ represents the operation function of the PDB branch. X and Y denote the input and output feature tensors, respectively. We employ a residual connection to facilitate gradient flow and prevent network degradation.

Dual-Attention Grouped Block (DAGB). Smoke detection in uncontrolled environments faces a unique challenge: smoke features are typically blurry and lack a rigid geometric structure, resulting in high inter-class similarity with the background and low intra-class variance. To address this, we drew on the design of CBAM [44] and proposed DAGB, DAGB employs a "Coarse-to-Fine" recalibration strategy to progressively purify features before extraction.

The process begins with Channel-Level Recalibration. Since standard convolutions treat all channels equally, they struggle to distinguish informative smoke textures from background noise. We employ the "Squeeze-and-Excitation" (SE) mechanism [45] to explicitly model channel dependencies. By aggregating global spatial information into a channel descriptor vector $\mathrm{z} \in \mathbb{R}^C$ via Global Average Pooling (GAP), we capture the global distribution of features. This descriptor is then mapped to a set of channel weights w_c through a bottleneck Multi-Layer Perceptron (MLP), enabling the network to selectively enhance smoke-relevant channels while suppressing noise:

$$\mathrm{w}_c = \sigma(\mathrm{W}_2\delta(\mathrm{W}_1\mathrm{z})),\ \mathrm{X}' = \mathrm{X} \odot \mathrm{w}_c \tag{2}$$

where δ and σ denote the SiLU and Sigmoid functions, respectively. W_1 and W_2 represent the learnable weight parameters for the dimensionality reduction and expansion layers, respectively; z is the channel descriptor vector obtained via Global Average Pooling (GAP), and w_c is the generated set of channel weights used to selectively enhance smoke-relevant channels. $\odot$ represents element-wise multiplication (Hadamard product), used to reweight the input feature X.

Immediately following, we perform Spatial-Level Recalibration to localize targets. A lightweight 3×3 convolution compresses the channel dimension of the weighted tensor X' to generate a spatial probability map M_s. This map acts as a spatial filter, suppressing background clutter and focusing computational attention on salient regions:

$$\mathrm{M}_s = \sigma\big(\mathrm{Conv}_{3\times3}(\mathrm{X}')\big),\ \mathrm{X}_{att} = \mathrm{X}' \odot \mathrm{M}_s \tag{3}$$

here, M_s is the generated spatial attention mask; X_{att} is the final feature map after dual recalibration, and X' is the feature after channel-level recalibration.

Finally, to efficiently encode these refined features, we utilized Grouped Convolution with the group number set to $g = C/4$. Unlike standard convolution, grouped convolution isolates information flow within channel subsets. This design not only significantly reduces parameter redundancy but also imposes a regularization effect, forcing the network to learn diagonal correlations and preventing the destruction of the carefully recalibrated attention features.

Parallel Depthwise Block (PDB). A critical bottleneck in detecting distant smoke or small fire points is the dilemma of scale variation: standard downsampling operations often erode the details of small targets, while maintaining high resolution limits the perceptual range required to distinguish smoke from background noise. PDB solves this problem by adopting a multi-branch topology that processes local and global information simultaneously.

To ensure computational efficiency, the input is first projected into a low-dimensional space X_{mid} (where the channel count $C_{mid} = C/2$). The feature flow is subsequently split into two parallel branches. The Detail Branch is designed to preserve high-frequency information. It utilizes a standard 3×3 Depthwise Convolution. By performing spatial filtering independently on each channel, this branch excels at maintaining the structural integrity of edges and boundaries, which is crucial for delineating small targets. Simultaneously, the Wide-Field Branch is designed for contextual verification. It utilizes a 3×3 Depthwise Dilated Convolution with a dilation rate of d = 2. This operation inserts "holes" into the convolution kernel, effectively expanding the receptive field to 5×5 without increasing parameters or reducing resolution. This enables the network to integrate surrounding environmental context to verify the presence of smoke or targets:

$$F_{detail} = \text{DWConv}_{3\times3}(X_{mid}) \tag{4}$$

$$F_{wide} = \text{DW Dilated Conv}_{3\times3,d=2}(X_{mid}) \tag{5}$$

In this formula, X_{mid} represents the intermediate feature after dimensionality reduction; F_{detail} and F_{wide} respectively represent the captured local high-frequency detail features and the wide-field large receptive field features.

To synergize these complementary features, the outputs of the two branches are concatenated and fused via a 1×1 convolution. This fusion step dynamically aggregates local details and global context, ensuring that small targets are not lost in the background while possessing a sufficiently broad field of view for correct classification:

$$\mathcal{F}_{PDB}(X) = \text{Conv}_{1\times1}(\text{Concat}[F_{detail}, F_{wide}]) \tag{6}$$

where $\mathcal{F}_{PDB}(X)$ denotes the final output feature of the PDB, and Concat represents the concatenation operation along the channel dimension.

2.3.2. Dual-Scale Contextual Diffusion

Traditional Convolutional Neural Networks (CNNs) often face an inherent contradiction between "Receptive Field Expansion" and "Local Detail Preservation" during feature extraction. Existing multi-scale fusion methods, such as Inception [46] modules or ASPP [47], typically employ parallel branch structures to extract features independently, performing simple aggregation only at the terminal stage. This "Late Fusion" strategy severs the interaction potential between features of different scales during the extraction process, leading to a lack of global semantic guidance for local feature extraction.

To address this issue, we propose the Dual-Scale Contextual Diffusion (DCD) module. Unlike traditional parallel designs, DCD introduces a novel Cascade-style Feature Diffusion Mechanism. By explicitly injecting long-range contextual information into the input of the local branch, we ensure that the extraction of local textures is constrained by the "condition" of global environmental knowledge. This design significantly improves the suppression of fire-like interferences (e.g., urban neon lights, strong reflections) and the disentanglement of semi-transparent smoke from complex backgrounds without adding significant computational burden. We replaced the Bottleneck part in the C2f structure with the DCD module to further enhance model performance.

As shown in Figure 5, the DCD module adheres to the lightweight "Split-Transform-Fuse" paradigm but reconstructs the branch interaction logic. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, it is first projected to a hidden layer dimension via a 1×1 convolution and decoupled in the channel dimension into two independent data streams: the Context Branch (X_{ctx}) and the Local Branch (X_{loc}).

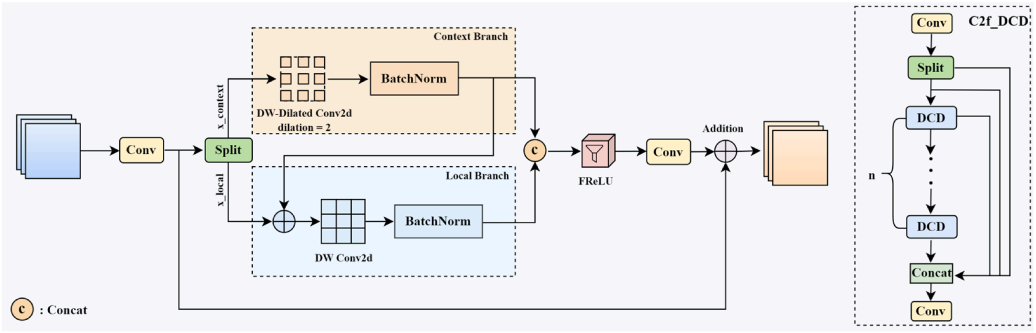

Figure 5. C2f_DCD structure.

Context Perception. To capture context environmental information with low parameter overhead, the Context Branch prioritizes a Depthwise Dilated Convolution with a Dilation Rate of 2. Compared to standard convolution, this design physically expands the receptive field to a 5×5 range without increasing computational cost. The objective of this branch is to extract low-frequency, large-scale semantic features (such as amorphous diffusion trends of smoke or large-scale environmental textures like buildings), denoted as Y_{ctx}:

$$Y_{ctx} = \mathcal{BN}(\text{DW Dilated Conv}_{3\times3,d=2}(X_{ctx})) \tag{7}$$

where Y_{ctx} represents the output feature of the Context Branch. $\mathcal{BN}$ denotes Batch Normalization. DW Dilated $\text{Conv}_{3\times3,d=2}$ refers to the 3×3 depthwise dilated convolution with a dilation rate of 2. X_{ctx} is the input feature for the context perception branch.

Semantic Diffusion Mechanism. This is the core innovation of the DCD module. Distinct from isolated parallel extraction, we constructed a unidirectional information pathway from context to local. We treat the output of the Context Branch, Y_{ctx}, as a Spatial Prior, explicitly "diffusing" and injecting it into the input end of the Local Branch via Element-wise Addition:

$$X'_{loc} = X_{loc} + Y_{ctx} \tag{8}$$

here, X'_{loc} represents the local branch feature after the injection of semantic diffusion, and X_{loc} is the initial input feature of the local branch.

Mathematically, this "diffusion" operation is equivalent to introducing a dynamic Spatial Bias. It enables the local branch to possess the "base color" (global context) of the environment before processing. This design ensures that subsequent local feature extraction is no longer blind but performs a targeted search for object textures that contrast strongly with the background under known environmental context conditions. For tiny fire sources, this implies the network can filter out isolated noise based on environmental logic; for thin smoke, this aids in enhancing foreground feature contrast within turbid atmospheres.

Local Refinement & Reconstruction. The context-modulated local feature X'_{loc} is subsequently processed by a standard Depthwise Convolution, focusing on capturing the high-frequency edge details of early tiny flames, generating Y_{loc}. Finally, to achieve a unified representation of multi-scale information, we perform channel Concatenation on the outputs of both scales. To further enhance the model's spatial modeling capability for irregular objects, we introduced the Funnel ReLU (FReLU) [48] activation function before the final linear projection:

$$Y_{out} = \text{Conv}_{1\times1}(\mathcal{T}(Y_{loc}\ Concat\ Y_{ctx})) + X_{in} \tag{9}$$

where Y_{loc} is the feature output by the local branch. Y_{out} is the final output feature of the DCD module. X_{in} represents the original input of the DCD module, used for the residual

connection. $\mathcal{T}$ represents the FReLU activation function. Unlike the scalar activation of ReLU [49], FReLU integrates a pixel-level Spatial Condition Window, which complements the dual-scale interaction features of DCD, maximizing the discriminability of fire targets.

2.3.3. Gaussian Spatial Pyramid Pooling Fast

To bolster the non-linear representational capacity of the YOLOv11 model during multi-scale feature fusion and to optimize the aggregation of deep features, we propose an improved spatial pyramid pooling module—Gaussian Spatial Pyramid Pooling Fast (GSPPF), as shown in Figure 3. While retaining the efficient receptive field expansion capabilities of the original SPPF, this module replaces the original SiLU activation function with the Gaussian Error Linear Unit (GELU), constructing a composite convolutional unit comprising Convolution, Batch Normalization, and GELU. Compared to ReLU and its variants, GELU introduces the concept of stochastic regularization. By multiplying the input by the Cumulative Distribution Function (CDF) of the standard normal distribution, it achieves activation with smoother gradient propagation characteristics. This property helps mitigate the gradient vanishing problem and accelerates model convergence.

In this study, we adopt the tanh-based approximation for the GELU implementation to balance computational precision and speed. Its mathematical expression is defined as follows:

$$\mathrm{GELU}(x) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right) \tag{10}$$

where x represents the input feature value, and tanh is the hyperbolic tangent function. The constant $\sqrt{2/\pi}$ and the coefficient 0.044715 are fitting parameters introduced to approximate the CDF of the standard normal distribution.

The specific operational workflow of GSPPF begins by passing the input feature map through a 1×1 convolutional layer activated by GELU, which compresses the channel dimension to $C/2$ to reduce parameter redundancy. The dimensionality-reduced feature map then undergoes three consecutive Max Pooling operations with a 5×5 kernel and a stride of 1. To achieve comprehensive multi-scale perception, the original projected feature is concatenated channel-wise with the outputs of the three pooling layers, effectively fusing information from four distinct receptive field scales. Finally, the aggregated features are processed through a concluding 1×1 composite convolutional layer to restore the channel count to the target dimension. This design enables GSPPF to capture finer contextual features of targets while preserving computational efficiency, thereby bolstering the model's detection performance in complex backgrounds.

3. Experiments and Results

3.1. Experimental Setup

All experiments were conducted on a Linux system equipped with a single NVIDIA A800 GPU, while inference latency and FPS were evaluated on a CPU (AMD Ryzen 7 6800H) using ONNX Runtime with a single thread to simulate resource-constrained environments. For latency measurement, each model was first warmed up for 20 iterations, followed by 200 inference runs, and the reported latency was computed as the average over all runs. We utilized the Stochastic Gradient Descent (SGD) optimizer for all models, configured with a momentum of 0.937, a weight decay of 0.0005, and a uniform initial learning rate of 0.01. Given that the dataset contains a large number of small targets, including sub-pixel ignition points and diffused smoke with blurred boundaries, the input image resolution for all models was standardized to 1024×1024 pixels to preserve fine-grained target details and improve small object detection accuracy. This unified resolution setting was applied equally to all compared models to ensure strictly consistent experimental conditions and guarantee

the fairness and reproducibility of the comparative evaluation. The training duration was set to 200 epochs with a batch size of 16. To bolster model generalization, a suite of data augmentation strategies—including Mosaic, random erasing, and multi-scale scaling—was employed during the training phase. Specifically, to facilitate the model's ultimate convergence to the real-world data distribution, Mosaic augmentation was deactivated for the final 10 epochs. Conversely, to ensure the objectivity and reliability of the evaluation, no data augmentation was applied during the validation phase, thereby accurately measuring the model's true generalization performance.

To comprehensively evaluate the proposed method, we employed a total of five metrics categorized into performance and complexity domains. For detection performance, the Mean Average Precision at an IoU threshold of 0.5 (mAP50) and across the 0.5–0.95 range (mAP50-95), alongside the F1-score, were utilized to quantify detection accuracy and the equilibrium between precision and recall. Regarding model complexity, the number of Parameters and Giga Floating-point Operations (GFLOPs) were calculated to measure the model's parameter scale and hardware computational overhead, respectively. Furthermore, to evaluate real-world deployment feasibility in resource-constrained environments, Average Inference Latency (ms) and Frames Per Second (FPS) were additionally measured on a single-core CPU using ONNX Runtime (v1.19.2) with a single thread. All comparative models were trained under these identical configurations to ensure a fair and rigorous comparison.

3.2. Detailed Performance Analysis of the Proposed Method

To rigorously validate the operational reliability of YOLO-Fire, we conducted comprehensive quantitative evaluations on the urban fire test set. Table 1 corroborates that the proposed model exhibits robust detection capabilities across distinct fire categories. The model achieved an overall mAP50 of 75.7%, an mAP50-95 of 53.3%, and an F1-score of 73.7%. Notably, these metrics were attained with a parameter count of only 10.02 M, indicating that the model strikes an optimal balance between high-precision perception and lightweight structural efficiency, making it suitable for resource-constrained monitoring devices. Specifically, performance on the "Fire" category is particularly outstanding, yielding an mAP50 of 84.5% and an F1-score of 80.4%. This high precision validates the model's ability to accurately lock onto incipient ignition points characterized by distinct visual signatures. Conversely, while smoke detection inherently presents greater physical challenges due to its semi-transparent and amorphous nature, the model still maintained a robust mAP50 of 67.0% and an F1-score of 66.9%. This demonstrates that the introduced improvement modules effectively enhance feature extraction capabilities for weak-boundary targets, significantly reducing the risk of missed detections in early fire stages.

Table 1. Detection results of the YOLO-Fire model.

Model	Object	mAP50 (%)	mAP50-95 (%)	F1-Score (%)
YOLO-Fire	Fire	0.845	0.589	0.804
	Smoke	0.670	0.477	0.669
	Fire and Smoke	0.757	0.533	0.737

In safety-critical urban environments, detection faces substantial challenges: sub-pixel fire spots often suffer from feature paucity, while tenuous smoke exhibits low contrast against chaotic backgrounds, easily triggering missed detections or false alarms. To intuitively demonstrate the model's generalization capability across diverse operational scenarios, representative detection results are visualized in Figure 6.

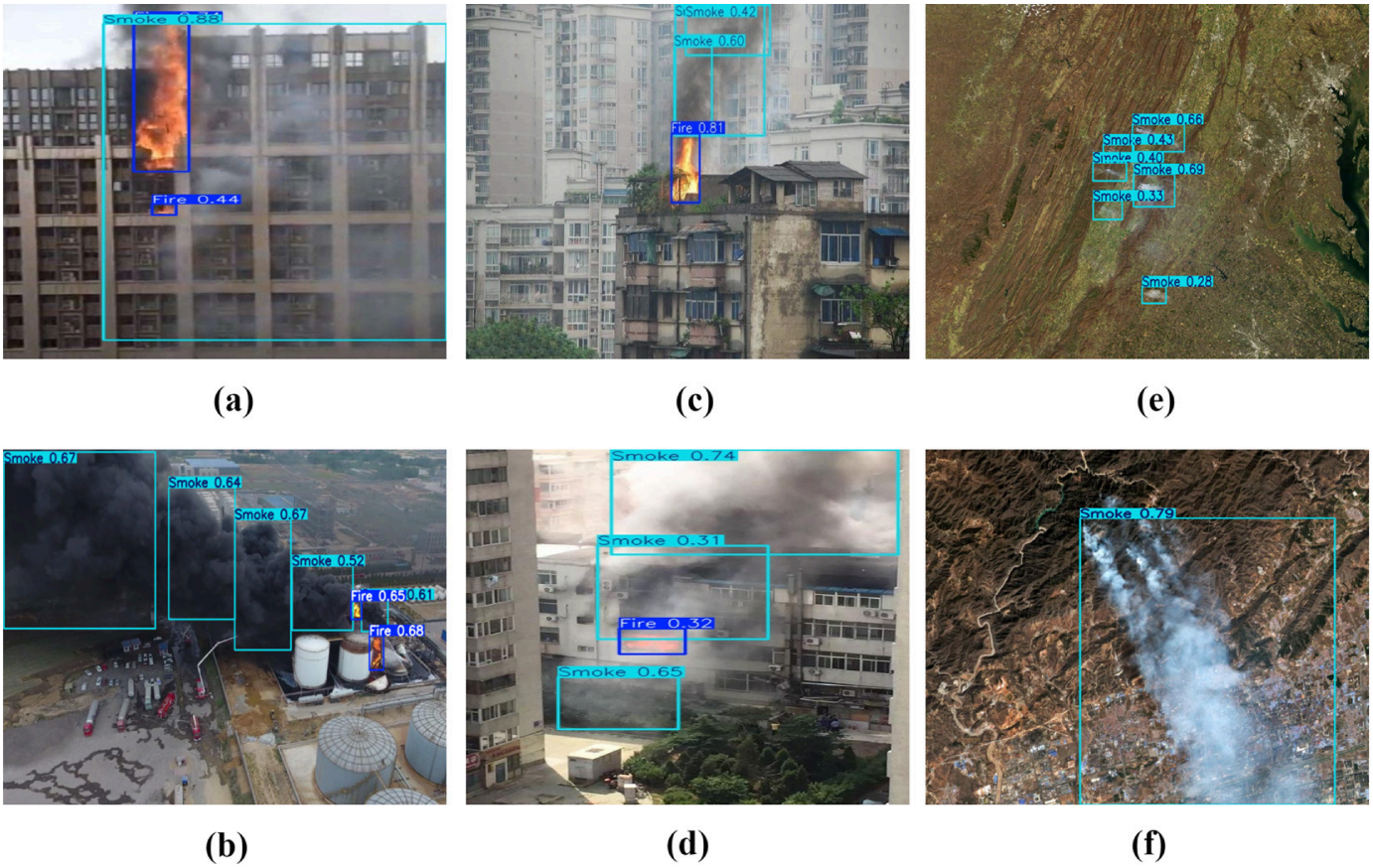

Figure 6. Visualization of detection results across diverse scenarios. (**a**,**b**) Sub-pixel ignition points in industrial zones. (**c**,**d**) Tenuous smoke under complex backgrounds. (**e**,**f**) Fire spread in satellite remote sensing imagery.

As illustrated in Figure 6a,b, the model successfully identified small-scale ignition points within high-rise buildings and industrial zones. Despite these targets occupying minimal pixel areas, the model accurately localized the anomalies, thereby validating the effectiveness of the HFFM in handling multi-scale variations. Subsequently, Figure 6c,d showcase the detection performance for thin, lingering smoke. Even when confronted with low-contrast interference from complex backgrounds (e.g., gray concrete structures), the model effectively overcame the semi-transparent characteristics to precisely segment smoke plumes. Furthermore, Figure 6e,f confirm the applicability of the proposed method in satellite remote sensing. Confronted with complex terrain textures and macro-scale backgrounds, the model accurately identified fire spread and smoke diffusion from an overhead perspective, demonstrating its immense potential in wide-area disaster early warning tasks.

3.3. Quantitative Comparison with Mainstream Models

To comprehensively evaluate the superiority of the proposed method, we compared it against a variety of advanced object detection models under identical experimental conditions. The comparative baselines encompass SSD, RetinaNet, the classic two-stage detector Faster R-CNN, the Transformer-based RT-DETR, RF-DETR, the mainstream YOLO series models, as well as the state-of-the-art (SOTA) models for fire detection, YOLO-FireAD and FireSmoke-YOLO. For the YOLO series, with the exception of YOLOv3, we uniformly selected the lightweight "s" (small) variants for a fair comparison. The detailed experimental results are presented in Tables 2 and 3.

Table 2. Comparison of integrated detection performance of different models.

Model	All			Params (M)	FLOPs (G)	Average Latency (ms)	FPS
	mAP50	mAP50-95	F1-Score				
SSD [50]	0.372	0.148	0.442	3.68	15.62	11.85	84.42
RetinaNet [51]	0.524	0.237	0.548	37.97	191.42	431.56	2.32
Faster RCNN [52]	0.440	0.196	0.484	136.71	401.71	—	—
RT-DETR [53]	0.655	0.398	0.641	31.99	103.40	209.52	4.77
RF-DETR [54]	0.676	0.453	0.649	27.23	97.30	256.42	3.90
YOLOv3-tiny [55]	0.677	0.421	0.665	12.13	18.90	25.27	39.57
YOLOv5 [56]	0.685	0.424	0.662	9.11	23.80	37.01	27.02
YOLOv8 [34]	0.730	0.482	0.704	11.13	28.40	48.23	20.73
YOLOv9 [57]	0.686	0.447	0.658	7.17	26.70	60.86	16.43
YOLOv10 [58]	0.718	0.49	0.686	7.22	21.40	45.20	22.12
YOLOv11 [33]	0.733	0.488	0.708	9.41	21.30	45.73	21.87
YOLOv12 [59]	0.677	0.422	0.649	9.07	19.30	63.97	15.63
YOLOv13 [60]	0.725	0.474	0.696	9.00	20.70	105.50	9.48
YOLO26 [61]	0.716	0.447	0.689	9.47	20.50	44.54	22.45
YOLO-FireAD [62]	0.730	0.480	0.699	9.07	32.40	138.69	7.21
FireSmoke-YOLO [8]	0.733	0.494	0.715	29.21	87.90	448.30	2.23
Ours	**0.757**	**0.533**	**0.737**	10.02	22.00	57.86	17.28

Note: bold indicates the best result. The same applies to subsequent tables.

Table 3. Comparison of smoke and fire detection performance of different models.

Model	Fire			Smoke		
	mAP50	mAP50-95	F1-Score	mAP50	mAP50-95	F1-Score
SSD	0.543	0.230	0.590	0.202	0.067	0.293
RetinaNet	0.708	0.335	0.699	0.339	0.139	0.398
Faster RCNN	0.599	0.271	0.604	0.281	0.121	0.364
RT-DETR	0.804	0.495	0.752	0.505	0.300	0.527
RF-DETR	0.695	0.496	0.654	0.657	0.410	0.644
YOLOv3-tiny	0.803	0.503	0.756	0.551	0.339	0.573
YOLOv5	0.841	0.538	0.781	0.529	0.311	0.539
YOLOv8	0.848	0.566	0.792	0.611	0.398	0.615
YOLOv9	0.843	0.555	0.775	0.529	0.339	0.538
YOLOv10	0.844	0.577	0.782	0.593	0.404	0.590
YOLOv11	0.844	0.572	0.794	0.622	0.404	0.621
YOLOv12	0.836	0.536	0.770	0.519	0.308	0.522
YOLOv13	0.843	0.557	0.786	0.608	0.391	0.606
YOLO26	0.829	0.529	0.777	0.603	0.366	0.600
YOLO-FireAD	**0.847**	0.568	0.791	0.613	0.392	0.606
FireSmoke-YOLO	0.845	0.574	0.798	0.621	0.415	0.633
Ours	0.845	**0.589**	**0.804**	**0.670**	**0.477**	**0.669**

Overall Performance Analysis As demonstrated in the comprehensive performance evaluation in Table 2, the proposed model comprehensively surpasses all comparative models across key metrics. Specifically, compared to FireSmoke-YOLO, which ranked second in overall performance, our method achieved all-around improvements: mAP50 increased by 2.4%, mAP50-95 by 3.9%, and the F1-score by 2.2%. Notably, our model is the only one among all comparative experiments to break the 50% threshold for mAP50-95. This consistent superiority across all indicators fully demonstrates that the improved modules not only elevate the upper bound of feature extraction but also significantly optimize the equilibrium between precision and recall.

Regarding fire detection performance shown in Table 3, further improving performance is inherently challenging as the baseline metrics of most advanced models have approached performance saturation. Despite this bottleneck, our model still achieves the best performance across all metrics except mAP50, reaching 84.5%, which is only 0.2% lower than that of YOLO-FireAD. In addition, it attains an mAP50-95 of 58.9% and an F1 score of 80.4%, these results outperform the second-ranked YOLOv10 and FireSmoke-YOLO by margins of 1.2% and 0.6%, respectively. This proves that even in high-precision scenarios, our method performs bounding box regression more precisely than existing models and effectively reduces false positives.

In the significantly more challenging task of smoke detection, our advantage is even more pronounced. The mAP50 and F1 score are improved by 1.3% and 2.5%, respectively, compared with the second-ranked RF-DETR, while the mAP50-95 is improved by 6.2% compared with FireSmoke-YOLO. Meanwhile, compared with the baseline model, the three metrics have improved by 4.8%, 7.3%, and 4.8% respectively. This substantial performance leap indicates that our model successfully overcomes the feature extraction difficulties caused by the semi-transparency and blurred edges of smoke.

Computational Complexity Analysis In terms of computational complexity (as shown in Table 2), our model maintains highly competitive efficiency while achieving the highest accuracy. Although the parameter count (10.02 M) is marginally higher than that of the lightweight YOLOv11 (9.41 M), this slight increase is well justified by the significant breakthroughs in accuracy. Regarding comparisons with specialized fire detection models, while our parameter count is slightly higher than that of YOLO-FireAD (9.07 M), our model achieves substantially lower computational overhead with FLOPs of 22.0 G, reducing computational cost by 10.4 G compared to YOLO-FireAD, and both metrics remain significantly below those of FireSmoke-YOLO (29.21 M parameters and 87.9 G FLOPs), demonstrating a more favorable efficiency profile. Furthermore, compared to the widely used YOLOv8 (11.13 M), RT-DETR (31.99 M) and RF-DETR (27.23 M), our model exhibits a superior Parameter-Accuracy Trade-off, making it more suitable for deployment in real-world scenarios with stringent performance requirements.

Regarding inference efficiency, all models were evaluated on a single-core CPU to simulate resource-constrained deployment environments. Faster R-CNN, due to its two-stage architecture involving region proposal generation and subsequent classification, is inherently unsuitable for real-time deployment and was therefore excluded from latency comparison. Among the remaining models, YOLO-Fire achieves an average latency of 57.86 ms and an inference speed of 17.28 FPS, maintaining competitive real-time performance comparable to mainstream lightweight models such as YOLOv11 (45.73 ms, 21.87 FPS) and YOLOv8 (48.23 ms, 20.73 FPS), while delivering substantially superior detection accuracy. Compared to specialized fire detection models, YOLO-Fire demonstrates a decisive advantage in inference efficiency, as YOLO-FireAD incurs a latency of 138.69 ms (7.21 FPS) and FireSmoke-YOLO reaches as high as 448.30 ms (2.23 FPS). Transformer-based models similarly exhibit prohibitively high latency, with RT-DETR at 209.52 ms and RF-DETR at 256.42 ms, rendering them impractical for resource-constrained scenarios. These results demonstrate that YOLO-Fire strikes a favorable balance between detection accuracy and inference efficiency, making it well-suited for practical real-time urban fire safety surveillance applications.

3.4. Qualitative Comparison of Detection Results

To further qualitatively validate the superiority of the proposed method, we conducted a comprehensive visual comparison against mainstream object detection models using

identical test samples. As illustrated in Figure 7, five representative and challenging scenarios were selected for this analysis.

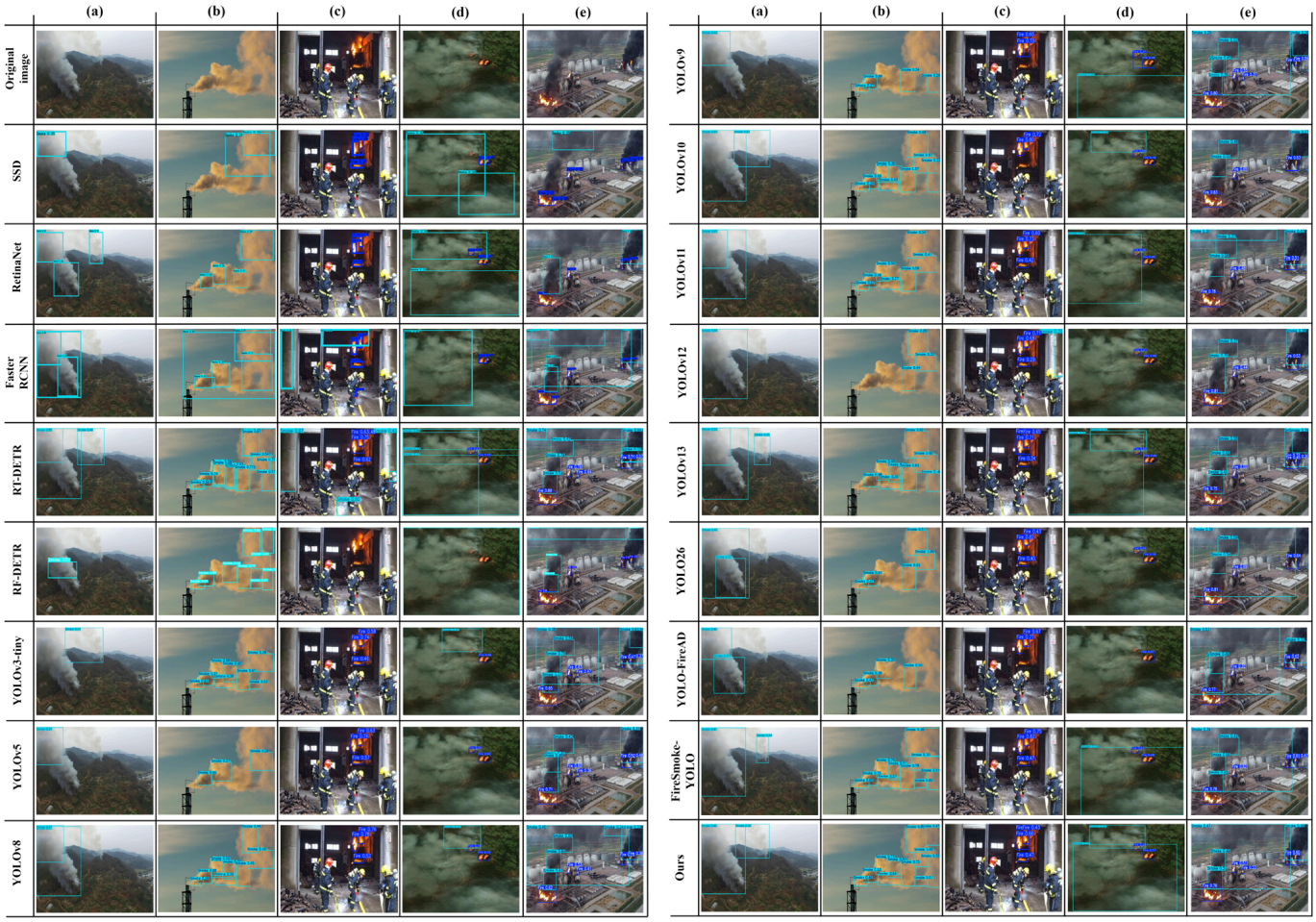

Figure 7. Comparison of detection of each model. (**a**,**b**) Tenuous and semi-transparent smoke captured by UAVs. (**c**) Fire under complex background interference. (**d**) Diffused smoke in satellite remote sensing imagery. (**e**) Multi-scale industrial fire.

Analysis of Early and Lightweight Detectors Observing the visualization results reveals that early or lightweight detectors (e.g., SSD, RetinaNet, Faster-RCNN, YOLOv3-tiny, and YOLOv5) exhibit significant limitations in feature recall. Constrained by weaker feature extraction capabilities, these models are prone to severe missed detections. Specifically, in Figure 7a, all the above-mentioned models failed to detect the target, while in Figure 7b, except for Faster RCNN, the rest of the models were unable to fully detect the thin and semi-transparent smoke, demonstrating their difficulty in effectively perceiving targets under low-contrast backgrounds. Furthermore, in the complex industrial fire scene of Figure 7e, SSD, RetinaNet and Faster-RCNN struggled to capture ground-level scattered small-scale fire points, while YOLOv3-tiny and YOLOv5 displayed a lack of sensitivity to the dense black smoke generated by industrial fires, leading to the loss of critical targets. Meanwhile, in Figure 7c, Faster RCNN shows a serious case of false detection.

Although RT-DETR and RF-DETR, these two types of Transformer-based models, have relatively high detection sensitivity, they still show severe under-detection and false detection in the complex scenarios of Figure 7. In Figure 7a,d,e, RF-DETR fails to detect smoke and small fire points, while in Figure 7c, RT-DETR is prone to misidentifying non-fire objects (such as fire-fighting equipment or complex background textures) as targets,

indicating that both models have obvious limitations in background suppression and precise prediction.

The latest YOLO series (from YOLOv8 to YOLO26) has effectively addressed the common issues of missed detections and false alarms in the Transformer-based models; however, they still face challenges regarding bounding box regression consistency when dealing with non-rigid targets. Due to the amorphous nature and gradient edges of smoke, the bounding boxes generated by these models in Figure 7a (ground-level smoke) and Figure 7d (satellite remote sensing smoke) are often not sufficiently refined—manifesting as boxes that are too loose or fractured, failing to completely cover drifting smoke edges. Additionally, in Figure 7e where multi-scale fire points coexist, they occasionally exhibit localization drift on tiny peripheral targets, suggesting room for improvement in regression stability under extreme scale variations.

Regarding the specialized fire detection SOTA models, YOLO-FireAD exhibits varying degrees of missed detections across all five test images. FireSmoke-YOLO performs comparatively well in Figure 7a–c; however, in Figure 7d, its smoke detection is incomplete, failing to capture the smoke regions on the left and upper portions of the scene, and in Figure 7e, it misses small-scale fire points in the industrial scenario. These results indicate that despite being specifically designed for fire detection tasks, both models still exhibit notable limitations in generalizing across diverse and complex scenarios, particularly in detecting semi-transparent smoke and small-scale fire sources under challenging backgrounds.

In contrast, YOLO-Fire method demonstrates the most superior detection performance across all comparative experiments. It not only maintains an exceptionally high recall rate but also achieves a qualitative leap in localization precision. As shown in Figure 7b,d, the bounding boxes generated by our model are highly compact and well-fitted, accurately outlining the contours of irregular smoke without including excessive background. Simultaneously, in Figure 7c, which is prone to false positives, our model exhibits excellent discriminative power with zero false alarms; in Figure 7e, it precisely captures all dispersed, minute fire points. This ample evidence proves that the proposed improvement modules endow the model with both extreme robustness and pixel-level localization accuracy in complex environments.

To further visually demonstrate the advantages of YOLO-Fire over the baseline and specialized fire detection SOTA models in challenging scenarios, we provide an additional qualitative comparison using satellite remote sensing imagery, as shown in Figure 8. Across all five images, the competing models exhibit varying degrees of missed detections and false alarms. The baseline YOLOv11 shows limited robustness against cloud interference, misclassifying cloud regions as smoke in multiple images, while also missing smoke targets in certain scenes. YOLO-FireAD demonstrates poor generalization in satellite remote sensing scenarios, failing to detect any target in images 3–5. FireSmoke-YOLO persistently misidentifies cloud regions and complex background terrain as smoke or fire points across multiple images, generating substantial false alarms that would severely compromise the reliability of early warning systems in practice. In contrast, YOLO-Fire successfully suppresses cloud and terrain interference across all five images while accurately detecting all fire and smoke targets without false alarms, demonstrating significantly stronger robustness and generalization capability in satellite remote sensing fire monitoring tasks.

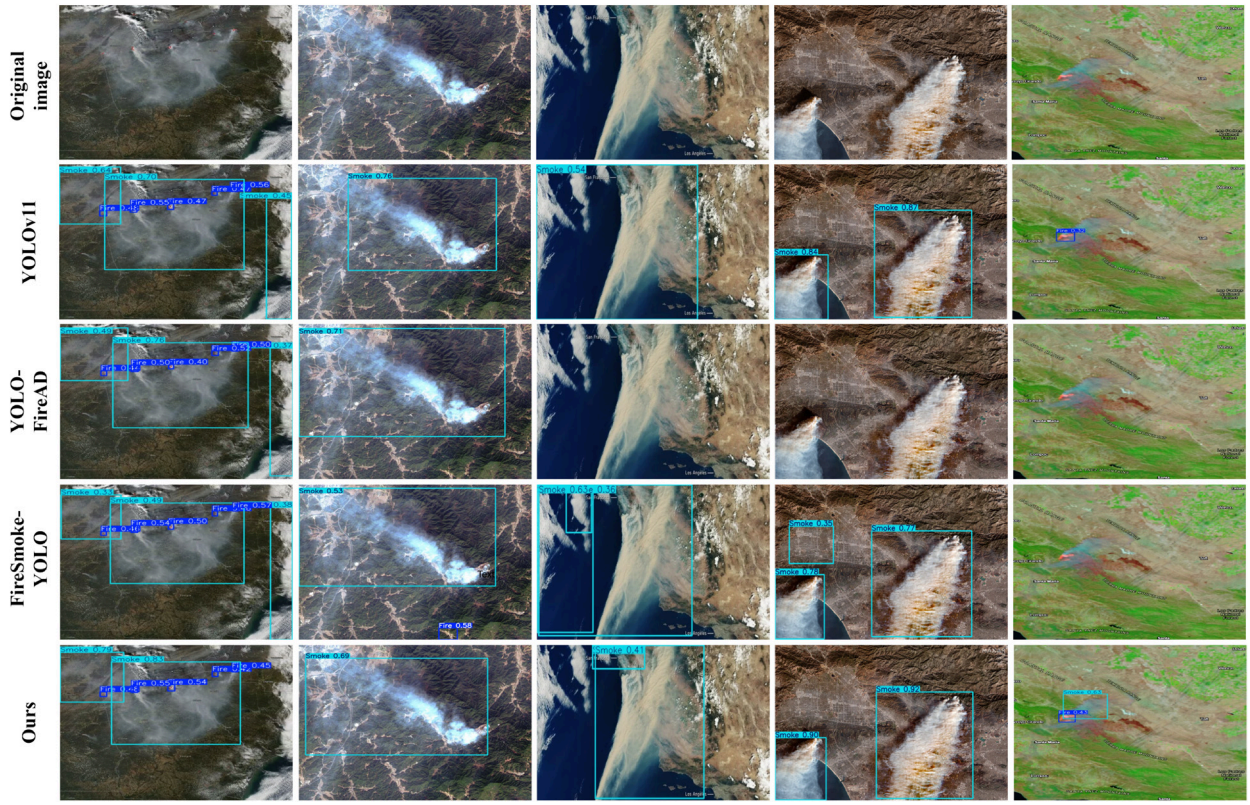

Figure 8. Detection comparison among the baseline, fire detection SOTA models, and YOLO-Fire on satellite remote sensing imagery.

4. Discussion

4.1. Effectiveness Analysis of the Proposed Framework

4.1.1. Impact Analysis of Modules and Internal Mechanisms

Contribution Analysis of Core Components. To systematically evaluate the independent contributions and synergistic effects of HFFM, C2f-DCD, and GSPPF on model performance, we conducted a stepwise ablation study, with quantitative results summarized in Table 4. From the perspective of stepwise module addition, starting from the baseline YOLOv11 (mAP50: 73.3%, mAP50-95: 48.8%, F1-score: 70.8%), the progressive introduction of each module yields consistent performance improvements. Upon introducing HFFM alone, mAP50 improves to 74.9%, mAP50-95 to 51.5%, and F1-score to 72.6%, demonstrating the significant contribution of frequency-domain structural decoupling to feature representation, particularly for the smoke category where mAP50 improves from 62.2% to 65.1%. Further incorporating C2f-DCD raises the overall mAP50 to 75.2% and mAP50-95 to 51.6%, confirming that the cascade-style contextual diffusion mechanism effectively complements HFFM by suppressing background interference and enhancing semi-transparent smoke features. Finally, the addition of GSPPF completes the full YOLO-Fire model, achieving the optimal overall performance with mAP50 of 75.7%, mAP50-95 of 53.3%, and F1-score of 73.7%, with the most pronounced gain observed in mAP50-95, confirming the necessity and cumulative effectiveness of all three modules operating in concert. From the perspective of individual module contributions, all single-module variants surpass the baseline, rigorously validating the independent effectiveness of each proposed module. Notably, although the GSPPF-only variant lags behind in overall metrics, it demonstrates superior sensitivity in capturing fire textures thanks to its expanded receptive field, achieving the highest single-category mAP50 of 85.1% for the Fire class. In the more challenging smoke detection task, the full model achieves the most substantial gains,

with mAP50 and mAP50-95 improving by 4.8% and 7.3% respectively over the baseline, confirming that the synergistic operation of all three modules is particularly critical for detecting semi-transparent and amorphous targets.

Table 4. Ablation study of each component in the proposed YOLO-Fire model.

Category	HFFM	C2f_DCD	GSPPF	mAP50	mAP50-95	F1-Score
All	√	√	√	**0.757**	**0.533**	**0.737**
	√	√	×	0.752	0.516	0.726
	√	×	×	0.749	0.515	0.726
	×	√	×	0.746	0.504	0.720
	×	×	√	0.738	0.493	0.716
	×	×	×	0.733	0.488	0.708
Fire	√	√	√	0.845	**0.589**	**0.804**
	√	√	×	0.850	0.584	0.801
	√	×	×	0.848	0.583	0.802
	×	√	×	0.845	0.572	0.798
	×	×	√	**0.851**	0.572	**0.804**
	×	×	×	0.844	0.572	0.794
Smoke	√	√	√	**0.670**	**0.477**	**0.669**
	√	√	×	0.653	0.447	0.651
	√	×	×	0.651	0.446	0.649
	×	√	×	0.647	0.436	0.642
	×	×	√	0.626	0.414	0.627
	×	×	×	0.622	0.404	0.621

Note: √ indicates "added", and × indicates "not added". The same applies to subsequent tables.

Analysis of HFFM Internal Mechanism and Position. We further investigated the internal substructure design of the HFFM module and its optimal integration position within the network, with comparative data presented in Table 5. The experiment primarily compared the module's performance at different feature levels (P4 vs. P5) and analyzed the impact of the integrity of its internal components (DAGB and PDB). Results show that compared to single-branch structures, the complete design combining DAGB and PDB achieved the best performance, validating the necessity of simultaneously capturing attention and multi-scale receptive fields. Regarding position selection, deploying HFFM at the P5 high-level semantic feature layer yielded improvements of 1.1%, 2.4%, and 1.1% in mAP50, mAP50-95, and F1-score, respectively, compared to deployment at the P4 layer. This discrepancy suggests that DAGB and PDB are more efficient when processing high-level features rich in semantic information, thereby better guiding the model to utilize deep semantics to distinguish background interference from target features.

Analysis of DCD Fusion Strategy and Dilation Rate. Table 6 details the impact of key design parameters in the DCD module on detection performance. Regarding the feature fusion mechanism, experiments revealed the clear superiority of the Element-wise Addition strategy. Its mAP50 reached 74.6%, which not only significantly outperforms the 73.3% achieved by multiplicative fusion but, notably, exceeds the 73.8% of the dual-branch configuration with no interaction. This result strongly suggests that complex multiplicative gating introduces an information bottleneck at early stages, suppressing the transmission of critical features. Conversely, the adopted addition strategy functions similarly to a non-linear residual injection: it integrates contextual information while maximally preserving the magnitude and gradient flow of the original features, thereby achieving lossless feature enhancement.

Table 5. Ablation study on the internal mechanism and position of the HFFM module.

Category	DAGB	PDB	P5	P4	mAP50	mAP50-95	F1-Score
All	√	√	√	×	**0.749**	**0.515**	**0.726**
	√	√	×	√	0.738	0.491	0.715
	√	×	√	×	0.743	0.50	0.721
	×	√	√	×	0.738	0.49	0.715
Fire	√	√	√	×	**0.848**	**0.583**	**0.802**
	√	√	×	√	0.842	0.565	0.800
	√	×	√	×	0.842	0.57	0.798
	×	√	√	×	0.843	0.563	0.797
Smoke	√	√	√	×	**0.651**	**0.446**	**0.649**
	√	√	×	√	0.633	0.416	0.631
	√	×	√	×	0.644	0.43	0.644
	×	√	√	×	0.633	0.417	0.633

Table 6. Ablation study on the fusion strategy of DCD and dilation rate selection.

Category	Dilation Rate (d = 2)	Addition	Multiply	mAP50	mAP50-95	F1-Score
All	√	√	×	**0.746**	**0.504**	**0.720**
	√	×	√	0.733	0.488	0.713
	√	×	×	0.738	0.492	0.715
	× (d = 1)	√	×	0.733	0.484	0.706
Fire	√	√	×	**0.845**	**0.572**	0.798
	√	×	√	0.843	0.566	0.799
	√	×	×	0.844	0.567	**0.800**
	× (d = 1)	√	×	0.843	0.563	0.794
Smoke	√	√	×	**0.647**	**0.436**	**0.642**
	√	×	√	0.623	0.411	0.626
	√	×	×	0.632	0.418	0.630
	× (d = 1)	√	×	0.624	0.406	0.619

Simultaneously, regarding receptive field configuration, we compared the adaptability of different convolution settings to multi-scale targets. For small Fire objects, the configuration with a dilation rate of d = 2 achieved an mAP50-95 of 57.2%, slightly outperforming the 56.3% of standard depthwise convolution. This corroborates that appropriate receptive field expansion does not negatively impact small object detection. More critically, depthwise dilated convolution exhibited stronger global modeling capabilities for large-scale environmental information, successfully overcoming detection difficulties caused by the amorphous nature and blurred boundaries of smoke targets. Compared to standard depthwise convolution, the use of d = 2 dilated convolution drove significant leaps of 2.3% and 3.0% in mAP50 and mAP50-95 for the "Smoke" category, respectively.

4.1.2. Visualization and Response Analysis

To provide a more intuitive evaluation of the comprehensive performance of various models across different confidence levels, we plotted the Precision-Recall (P-R) curves and F1-score curves generated from the test set in Figure 9.

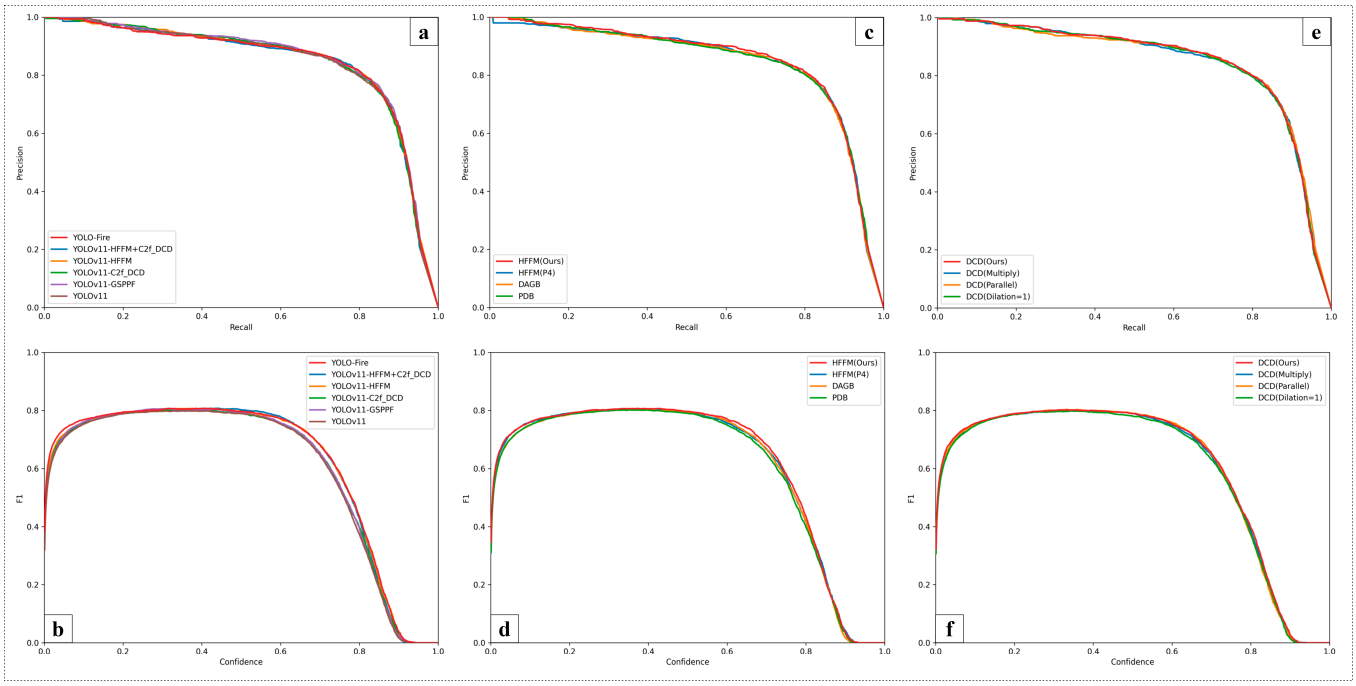

Figure 9. F1 Score curve and PR curve: (**a**,**c**,**e**) display the PR curves, while (**b**,**d**,**f**) show the F1 Score curves.

Superiority of the Overall Architecture: As illustrated in Figure 9a,b, the red curve represents the proposed YOLO-Fire and the brown curve represents the YOLOv11 baseline. It can be clearly observed that the red P-R curve completely envelopes the curves of other variants including the stepwise combinations. This implies that at the same Recall level, the complete model consistently maintains the highest Precision. Meanwhile, in the F1-score curve, the complete model exhibits the highest peak and the widest coverage range, while the baseline curve declines earliest, proving that the progressive introduction of each module contributes incrementally to the model's overall robustness.

Effectiveness of the HFFM Fusion Strategy: Figure 9c,d compare the performance of the HFFM module against other feature fusion designs. The HFFM adopted in this paper demonstrates a larger Area Under the Curve (AUC) in the P-R space, and its F1-score shows a more gradual decline in the high-confidence interval. This indicates that, compared to other fusion mechanisms, HFFM efficiently aggregates multi-scale features, thereby reducing both missed detections and false positives.

Validation of C2f-DCD Internal Design: Regarding the internal parameter design of the DCD module, Figure 9e,f provide compelling visual evidence. The scheme employing addition fusion significantly outperforms both multiplicative fusion and the parallel non-interaction approach, confirming the lossless nature of the addition operation during feature injection. Furthermore, compared to standard depthwise convolution with d = 1, our scheme introducing depthwise dilated convolution maintains higher precision at the tail end of the P-R curve. This revalidates the necessity of expanding the receptive field to improve the detection capability for complex samples.

To intuitively investigate the impact of different modules on feature extraction, we visualized the intermediate feature maps of critical network layers (as shown in Figure 10). These visualizations reveal the internal response mechanisms of the model when handling complex textures and background interference.

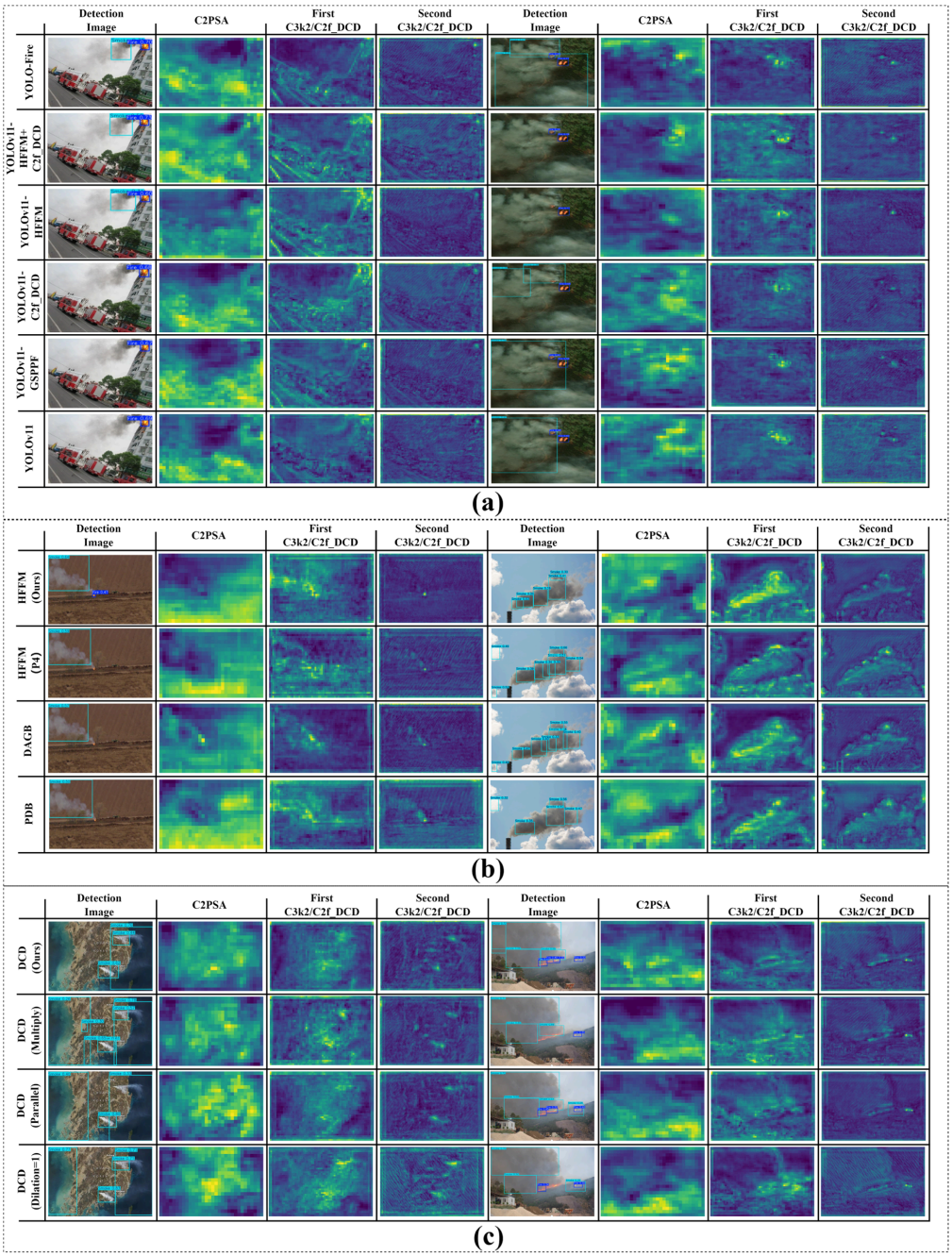

Figure 10. Visualization of feature maps. Note: "First C3k2/C2f_DCD" is the C3k2 or C2f_DCD after the first concat operation in the Neck module, and "Second C3k2/C2f_DCD" is the C3k2 or C2f_DCD after the second concat operation in the Neck module. The YOLO-Fire and DCD series models use C2f_DCD, while the rest of the models use C3k2. (**a**) Feature map of the core module ablation experiment; (**b**) Feature map of the HFFM ablation experiment; (**c**) Feature map of the DCD ablation experiment.

Figure 10a presents a comparison of feature responses between the proposed YOLO-Fire model and its variants, including stepwise combinations and the baseline. The YOLO-Fire model, which integrates all core components, effectively identifies thin smoke and tiny fire points in both real-world scenes and satellite imagery, generating feature maps with a significantly higher Signal-to-Noise Ratio (SNR). In smoke and fire regions, feature responses exhibit clear structured textures, whereas background regions are effectively suppressed. The YOLOv11-HFFM+C2f_DCD variant demonstrates notably improved feature responses compared to single-module variants and the baseline, confirming the

complementary effectiveness of these two modules in jointly suppressing background interference and enhancing target saliency. However, without GSPPF, its feature maps still exhibit slightly weaker responses in multi-scale target regions compared to the complete model. In contrast, the baseline YOLOv11 and single-module variants are often accompanied by substantial background noise or exhibit blurred target contours, with the baseline showing the weakest target discrimination capability. This indicates that the progressive introduction of each module incrementally enhances feature quality, and the synergistic operation of all three modules provides the highest-quality feature inputs for the subsequent detection head.

To further visually validate the effectiveness of the synergistic internal components of the HFFM module and the rationality of its deployment strategy, we analyzed intermediate feature maps under different configurations, as shown in Figure 10b. Notably, only the proposed HFFM module successfully detected the tiny fire points in the image and avoided misclassifying clouds as smoke. By comparing the feature maps, it is clearly observable that when relying solely on a single sub-module, the model's response to diffused smoke targets often exhibits distinct sparsity, failing to effectively cover the target's complete morphology. Conversely, the feature maps generated by the HFFM with its dual-branch integrated design demonstrate significant superiority: activations in target regions present high continuity and structural integrity, while preserving rich texture details. Furthermore, a comparison of background regions reveals that deploying HFFM at the high-level semantic layer (P5) more efficiently suppresses environmental noise interference and significantly boosts the SNR of target regions. This visually verifies the necessity of leveraging deep semantics for feature reconstruction.

To intuitively validate the design rationality of DCD, we compared feature maps under different parameters, as shown in Figure 10c. Our designed DCD module effectively distinguishes smoke from background in satellite imagery, whereas other comparative models consistently produced false detections. Regarding fusion strategies, feature maps in the Multiplication group appear significantly darker, confirming the suppressive effect of multiplicative gating on early signals. The Parallel group retains feature intensity but results in looser target texture structures due to the lack of information interaction between branches. In contrast, the Addition strategy effectively integrates the advantages of both branches, injecting context while maintaining high response levels, thus achieving lossless feature enhancement. Regarding receptive field, although the feature response of standard depthwise convolution (d = 1) captures smoke regions, the distribution is relatively scattered with excessive attention paid to surrounding non-smoke targets. However, the DCD successfully expands the effective receptive field, completely depicting the diffusive morphology of smoke. This provides a powerful visual explanation for the performance advantages of the proposed design in complex scenarios.

4.2. Discussion on Validity

The empirical dominance of YOLO-Fire underscores a critical insight for safety monitoring: aligning network architecture with the inherent physical properties of targets is far more effective than generic feature learning. While UFS-YOLO [63] enhances detection via standard CBAM attention, it oversimplifies the spectral heterogeneity between fire and smoke, treating them as uniform targets. Unlike classical attention mechanisms such as SE and CBAM, HFFM achieves structural decoupling at the architectural level. Incipient flames present high-frequency, sharp-boundary point features, while diffused smoke manifests as low-frequency, amorphous diffusive textures; processing both within a unified stream inevitably leads to mutual feature interference. By contrast, the parallel dual-stream design of HFFM structurally isolates the extraction pathways of these two

heterogeneous targets, ensuring that high-frequency flame boundaries and low-frequency smoke textures are processed independently. This design also offers distinct advantages over serial structures and multi-head attention mechanisms: serial structures suffer from cumulative feature contamination, where mixed representations in early layers constrain the expressiveness of subsequent layers; multi-head attention mechanisms, while introducing multiple perspectives, still operate on a shared input feature map, incurring inter-head competition and substantial computational overhead. HFFM's two streams originate from the same input yet compute entirely independently, achieving genuine complementarity through final fusion—a structural isolation that is particularly critical where high-frequency flame features and low-frequency smoke features inherently coexist, while maintaining significantly lower computational cost for lightweight real-time deployment.

Interestingly, IFS-DETR [64] also recognizes the significance of frequency-domain analysis, employing 2D-DCT for channel attention. However, HFFM achieves structural decoupling rather than merely re-weighting channels, offering superior interpretability. Furthermore, C2f-DCD functions as a semantic regularizer, fundamentally differing from late-fusion strategies such as Inception and ASPP, which extract features independently across parallel branches and aggregate them only at the terminal stage, severing inter-scale feature interaction. In contrast, C2f-DCD introduces a cascade-style feature diffusion mechanism that injects global contextual information into the local branch prior to local texture extraction, ensuring that local feature extraction is conditioned on global environmental knowledge. Unlike multiplicative gating mechanisms, which suppress feature transmission by scaling activations and risk losing weak signals such as thin smoke or sub-pixel fire points, additive injection preserves the magnitude and gradient flow of the original features, achieving lossless feature enhancement. This additive context injection facilitates a dual-enhancement strategy: by integrating dilated global context, it simultaneously prevents the fragmentation of semi-transparent smoke features and imposes an environmental consistency check to suppress isolated high-brightness noise, ensuring robust perceptibility for both amorphous smoke and sub-pixel fire sources.

In the realm of real-time safety surveillance, a dichotomy exists between two dominant paradigms: heavyweight Transformers and lightweight CNNs. The recent IFS-DETR and the improved DETR framework proposed of Li et al. [65] represent the former, utilizing end-to-end architectures to capture global dependencies. While IFS-DETR attempts to optimize speed via a lightweight backbone (LeanNet), it relies heavily on specialized TensorRT quantization to reach high frame rates, creating a deployment barrier. In contrast, YOLO-Fire demonstrates that pure CNN architectures can achieve comparable global perception without the stringent hardware dependencies of Transformers. By simulating global attention through DCD and GSPPF, YOLO-Fire maintains 10.02 M parameters—significantly lighter than standard DETR variants—while running efficiently on standard hardware without complex pre-compilation. This native efficiency renders it more suitable for resource-constrained edge devices where computational resources and quantization support may be limited.

Despite its superior quantitative performance, we must critically acknowledge the boundaries of YOLO-Fire's current capability, particularly in uncontrolled industrial or wilderness environments. Its reliance on the visible light spectrum creates an inherent spectral blindness that cannot be overcome by architectural optimization alone. Like other RGB-based architectures, it extracts features based on color and texture gradients; consequently, in nighttime scenarios or concealed smoldering events—such as underground peat fires or overheating process equipment—the model faces a definitive physical detection limit. Regarding environmental robustness, while the DCD mechanism effectively suppresses static urban noise, it remains sensitive to dynamic high-frequency perturbations.

Under extreme conditions—such as dense fog, industrial steam causing Tyndall effects, or strong winds driving chaotic vegetation movement—the background may generate transient high-frequency patterns mimicking flames, potentially causing false positives. Additionally, sub-pixel tiny fire sources remain challenging in wide-area remote sensing; despite receptive field expansion by GSPPF, deep convolutional downsampling inevitably causes feature erosion for objects occupying minimal pixel areas. While this study primarily contributes to the field of deep learning-based digital image processing for fire detection, the incorporation of UAV and satellite imagery demonstrates its potential applicability to remote sensing scenarios. Furthermore, it should be acknowledged that the UAV and satellite imagery incorporated in the dataset was obtained in RGB image format from publicly available platforms and the internet, rather than through conventional remote sensing data acquisition workflows. As a result, standardized remote sensing metadata such as sensor specifications, spatial resolution, spectral resolution, and geographic coordinate information are not uniformly available across all imagery sources, and the dataset does not follow a conventional remote sensing data structure with georeferenced imagery. This limits the ability to present detection results within a defined geographic coordinate reference system, which represents a gap relative to the standards of conventional remote sensing research.

To address these limitations, future research will proceed along the following directions. First, to overcome spectral blindness, a visible-light and thermal infrared dual-mode fusion framework will be explored. By incorporating mid-wave or long-wave infrared thermal imaging alongside the existing RGB stream, the model will detect heat signatures invisible to RGB sensors—including smoldering underground fires, concealed overheating industrial equipment, and nighttime fire sources. Cross-modal feature alignment strategies and modality-specific pretraining schemes will be investigated to bridge the domain gap between visible and infrared representations. Second, to mitigate dynamic background interference, spatiotemporal attention mechanisms will be introduced. Dense optical flow fields computed from consecutive video frames will provide motion-level cues, enabling the model to distinguish periodic vegetation swaying from the irregular turbulent motion of real fire. Background dynamic modeling will also be explored to establish temporal consistency constraints, suppressing transient high-frequency patterns under extreme conditions that may trigger false positives in spatial-only frameworks. Third, to address sub-pixel tiny fire sources in wide-area remote sensing, a dedicated small-object detection branch operating on high-resolution shallow feature maps will be introduced within an enhanced feature pyramid. By preserving fine-grained spatial information at the P2 feature level, this head will directly target objects occupying minimal pixel areas, improving recall for extremely small ignition points in satellite and UAV imagery.

Future work will also consider fine-grained annotation schemes for smoke concentration levels and diffusion phases—including initial, diffusion, and dissipation stages—to enable more detailed stage-wise performance evaluation and identify scenarios requiring further improvement. Although YOLO-Fire has been preliminarily validated in resource-constrained environments through single-core CPU simulation, achieving 17.28 FPS, further optimization is required for edge devices and UAV platforms. Specifically, model compression through INT8 and FP16 quantization, structured channel pruning, and knowledge distillation will be investigated, targeting NVIDIA Jetson Orin and ARM Cortex-based processors, with the goal of achieving over 30 FPS while maintaining detection accuracy within an acceptable margin. Finally, future work will prioritize constructing a more rigorous remote sensing fire dataset incorporating georeferenced imagery with complete metadata from specific satellite missions, enabling detection results to be presented within a defined

geographic coordinate reference system and enhancing the geospatial applicability of the proposed method.

5. Conclusions

Addressing the critical safety imperative of identifying incipient ignition points and segmenting diffused smoke in complex urban environments, this paper proposes YOLO-Fire, a lightweight yet high-precision deep learning detection framework built upon YOLOv11. To overcome the persistent bottlenecks of existing models in simultaneously handling tiny high-frequency flame features and low-frequency semi-transparent smoke textures, three targeted modules are strategically integrated into the architecture. The HFFM achieves structural frequency-domain decoupling between flame boundaries and smoke textures through a parallel dual-stream design, effectively eliminating the mutual feature interference that arises when heterogeneous targets are processed within a unified stream. The C2f-DCD module introduces a cascade-style contextual diffusion mechanism that injects global environmental context into local feature extraction, simultaneously suppressing fire-like background clutter and enhancing the contrast of semi-transparent smoke features. The GSPPF module optimizes multi-scale receptive field aggregation at the end of the backbone through GELU-activated pooling, expanding the model's perceptual range across extreme scale variations from sub-pixel ignition points to large-area smoke plumes.

Extensive experiments on a self-constructed large-scale multi-source urban fire dataset comprising 12,272 annotated images demonstrate that YOLO-Fire achieves an overall mAP50 of 75.7%, mAP50-95 of 53.3%, and F1-score of 73.7% with only 10.02 M parameters. Specifically, the model attains an mAP50 of 84.5% and an F1-score of 80.4% for the fire category, while maintaining a competitive mAP50 of 67.0% and F1-score of 66.9% for the inherently more challenging smoke category. Compared to the YOLOv11 baseline, YOLO-Fire achieves improvements of 2.4%, 4.5%, and 2.9% in mAP50, mAP50-95, and F1-score, respectively. Quantitative comparisons against a broad range of mainstream detection models and specialized fire detection SOTA models further confirm that YOLO-Fire comprehensively outperforms all compared methods across key overall performance metrics while maintaining a favorable balance between detection accuracy and computational efficiency. Furthermore, inference evaluation on a single-core CPU achieves 17.28 FPS, validating the practical deployment feasibility of YOLO-Fire in resource-constrained environments.

The proposed method effectively addresses the core challenges of detecting small-scale ignition points and semi-transparent diffused smoke in complex urban environments, offering an efficient and lightweight solution for practical fire safety surveillance. At a higher level, this study proposes a modular design paradigm of first decoupling (HFFM), then suppressing (DCD), and finally aggregating (GSPPF), and experimentally demonstrates that this paradigm effectively enhances model performance in detecting targets with small scale, semi-transparent appearance and high background interference, offering valuable insights for analogous fine-grained visual perception tasks in safety-critical domains.

Author Contributions: Conceptualization, L.M., M.W. and J.G.; methodology, L.M. and M.W.; software, L.M. and S.W.; validation, L.M.; formal analysis, L.M.; investigation, L.M. and M.W.; resources, L.M., M.W., J.G., S.W., X.S., J.Z., and H.L.; data curation, L.M.; writing—original draft preparation, L.M.; writing—review and editing, L.M., M.W., J.G., S.W., X.S., and J.Z.; visualization, L.M., J.G., X.C., L.L.,G.C. and J.L.; supervision, M.W., X.S. and J.Z.; project administration, M.W.; funding acquisition, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (42071385), the National Science and Technology Major Project of High-Resolution Earth Observation System (79-Y50-G18-9001-22/23), Ant Financial Group Industry–Academia–Research Cooperation Project and

the Shandong Science and Technology SMEs Technology Innovation Capacity Enhancement Project (2022TSGC2371). Research topics of Yantai City Smart City Innovation Lab (202310-04-ZHCS-01).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Tian, Y.Y.; Tsendbazar, N.E.; van Leeuwen, E.; Fensholt, R.; Herold, M. A global analysis of multifaceted urbanization patterns using Earth Observation data from 1975 to 2015. *Landsc. Urban Plan.* **2022**, *219*, 104316. [CrossRef]
2. Zhang, Y.; Shen, L.Y.; Ren, Y.T.; Wang, J.H.; Liu, Z.; Yan, H. How fire safety management attended during the urbanization process in China? *J. Clean. Prod.* **2019**, *236*, 117686. [CrossRef]
3. Pei, Z.X.; Li, J.W.; Guo, J.; Li, Q.; Chen, J. Using local co-location quotient and niche-based model to assess fire risk in urban environments: A case study of Beijing, China. *Sustain. Cities Soc.* **2023**, *99*, 104989. [CrossRef]
4. Shi, L.; Wang, J.; Li, G.; Chew, M.Y.L.; Zhang, H.; Zhang, G.; Dlugogorski, B.Z. Increasing fire risks in cities worldwide under warming climate. *Nat. Cities* **2025**, *2*, 254–264. [CrossRef]
5. Xiang, M.T.; Xiao, C.W.; Feng, Z.M.; Ma, Q. Global distribution, trends and types of active fire occurrences. *Sci. Total Environ.* **2023**, *902*, 166456. [CrossRef] [PubMed]
6. Zhang, X.X.; Yao, J.; Sila-Nowicka, K.; Jin, Y.H. Urban Fire Dynamics and Its Association with Urban Growth: Evidence from Nanjing, China. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 218. [CrossRef]
7. Kumar, V.; Bandyopadhyay, S.; Ramamritham, K.; Jana, A. Pinch analysis to reduce fire susceptibility by redeveloping urban built forms. *Clean Technol. Environ. Policy* **2020**, *22*, 1531–1546. [CrossRef]
8. Wang, S.K.; Wu, M.Q.; Wei, X.H.; Song, X.D.; Wang, Q.T.; Jiang, Y.C.; Gao, J.K.; Meng, L.Y.; Chen, Z.P.; Zhang, Q.Y.; et al. An advanced multi-source data fusion method utilizing deep learning techniques for fire detection. *Eng. Appl. Artif. Intell.* **2025**, *142*, 109902. [CrossRef]
9. Geetha, S.; Abhishek, C.S.; Akshayanat, C.S. Machine Vision Based Fire Detection Techniques: A Survey. *Fire Technol.* **2021**, *57*, 591–623. [CrossRef]
10. Huang, P.; Chen, M.; Chen, K.; Zhang, H.; Yu, L.; Liu, C. A combined real-time intelligent fire detection and forecasting approach through cameras based on computer vision method. *Process Saf. Environ. Prot.* **2022**, *164*, 629–638. [CrossRef]
11. Festag, S. False alarm ratio of fire detection and fire alarm systems in Germany—A meta analysis. *Fire Saf. J.* **2016**, *79*, 119–126. [CrossRef]
12. Martin, G.; Boehmer, H.; Olenick, S.M. Thermally-Induced Failure of Smoke Alarms. *Fire Technol.* **2020**, *56*, 673–692. [CrossRef]
13. Liu, G.; Yuan, H.Y.; Huang, L.D. A fire alarm judgment method using multiple smoke alarms based on Bayesian estimation. *Fire Saf. J.* **2023**, *136*, 103733. [CrossRef]
14. Zheng, L.X.; Wu, M.Q.; Xue, M.Y.; Wu, H.; Liang, F.; Li, X.P.; Hou, S.M.; Liu, J.Y. Power of SAR Imagery and Machine Learning in Monitoring Ulva prolifera: A Case Study of Sentinel-1 and Random Forest. *Chin. Geogr. Sci.* **2024**, *34*, 1134–1143. [CrossRef]
15. Liu, L.X.; Wu, M.Q.; Mao, Y.F.; Zheng, L.X.; Xue, M.Y.; Bing, L.; Liang, F.; Liu, J.Y.; Liu, B.W. Offshore wind energy potential in Shandong Sea of China revealed by ERA5 reanalysis data and remote sensing. *J. Clean. Prod.* **2024**, *464*, 142745. [CrossRef]
16. Zhang, G.Z.; Wu, M.Q.; Zhou, M.; Zhao, L.J. The seasonal dissipation of *Ulva prolifera* and its effects on environmental factors: Based on remote sensing images and field monitoring data. *Geocarto Int.* **2022**, *37*, 860–878. [CrossRef]
17. Chen, S.K.; Cao, Y.C.; Feng, X.Q.; Lu, X.B. Global2Salient: Self-adaptive feature aggregation for remote sensing smoke detection. *Neurocomputing* **2021**, *466*, 202–220. [CrossRef]
18. Yao, J.; Raffuse, S.M.; Brauer, M.; Williamson, G.J.; Bowman, D.M.J.S.; Johnston, F.H.; Henderson, S.B. Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. *Remote Sens. Environ.* **2018**, *206*, 98–106. [CrossRef]
19. Liu, L.X.; Wu, M.Q.; Chen, G.; Tang, Y.J.; Song, X.D.; Zou, M.; Xu, Y.D.; Zhang, X.; Wang, S.K.; Lv, J.Y.; et al. Satellite observations and deep learning unveil the rapid expansion of offshore wind turbines in China. *Resour. Conserv. Recycl.* **2026**, *227*, 108706. [CrossRef]
20. Lv, J.Y.; Wu, M.Q.; Liu, S.B.; Liu, L.X.; Song, X.D.; Wang, S.K.; Chen, G.; Tang, Y.J.; Liu, B.W.; Gao, J.K.; et al. Accurate mapping of mariculture areas and their carbon sink assessment based on deep learning and satellite imagery. *Aquaculture* **2026**, *612*, 743076. [CrossRef]
21. Liu, L.X.; Wu, M.Q.; Zhao, J.; Bing, L.; Zheng, L.X.; Luan, S.P.; Mao, Y.F.; Xue, M.Y.; Liu, J.Y.; Liu, B.W. Deep learning-based monitoring of offshore wind turbines in Shandong Sea of China and their location analysis. *J. Clean. Prod.* **2024**, *434*, 140415. [CrossRef]

22. Barmpoutis, P.; Dimitropoulos, K.; Kaza, K.; Grammalidis, N. Fire Detection from Images Using Faster R-CNN and Multidimensional Texture Analysis. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8301–8305.
23. Guan, Z.H.; Miao, X.Y.; Mu, Y.J.; Sun, Q.; Ye, Q.L.; Gao, D.M. Forest Fire Segmentation from Aerial Imagery Data Using an Improved Instance Segmentation Model. *Remote Sens.* **2022**, *14*, 3159. [CrossRef]
24. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
25. Wang, H.; Fayaz, M.; Ahmad, A.; Li, Y.; Nguyen, T.N.; Dang, L.M. Masked autoencoder-based vision framework for robust fire detection in complex environments. *Process Saf. Environ. Prot.* **2025**, *203*, 108019. [CrossRef]
26. Liu, L.; Ouyang, W.L.; Wang, X.G.; Fieguth, P.; Chen, J.; Liu, X.W.; Pietikainen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
27. Zou, Z.X.; Chen, K.Y.; Shi, Z.W.; Guo, Y.H.; Ye, J.P. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
28. Li, P.; Zhao, W.D. Image fire detection algorithms based on convolutional neural networks. *Case Stud. Therm. Eng.* **2020**, *19*, 100625. [CrossRef]
29. Zheng, H.T.; Wang, G.Y.; Xiao, D.; Liu, H.; Hu, X.Y. FTA-DETR: An efficient and precise fire detection framework based on an end-to-end architecture applicable to embedded platforms. *Expert Syst. Appl.* **2024**, *248*, 123394. [CrossRef]
30. Parvathy, S.; Jayachandran, T.P.; Reji, N.M.; Pranev, K.; Binu, G.S.; Manikandan, M.S. YOLO Architectures-Based Forest Fire Detection Under Different Background Scenarios. In Proceedings of the 2025 International Conference on Communication, Computing, Networking, and Control in Cyber-Physical Systems (CCNCPS), Dubai, United Arab Emirates, 10–12 June 2025; pp. 295–300.
31. Lin, Z.Y.; Yun, B.S.; Zheng, Y.N. LD-YOLO: A Lightweight Dynamic Forest Fire and Smoke Detection Model with Dysample and Spatial Context Awareness Module. *Forests* **2024**, *15*, 1630. [CrossRef]
32. Wang, L.; Guo, L.; Li, H.; He, B.; Yang, J.; Huang, Y. Enhanced forest fire detection via dynamic multiscale fusion and contextual partial cross features. *Eng. Appl. Artif. Intell.* **2025**, *162*, 112531. [CrossRef]
33. Jocher, G.; Qiu, J. Ultralytics YOLO11 (Version 11.0.0). 2024. Available online: https://docs.ultralytics.com/zh/models/yolo11 (accessed on 1 March 2026).
34. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8 (Version 8.0.0). 2023. Available online: https://docs.ultralytics.com/zh/models/yolov8 (accessed on 1 March 2026).
35. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 16–18 June 2020; pp. 390–391.
36. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
37. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
38. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
39. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
40. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
41. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Trans. Cybern.* **2022**, *52*, 8574–8586. [CrossRef]
42. Hendrycks, D. Gaussian Error Linear Units (Gelus). *arXiv* **2016**, arXiv:1606.08415.
43. Chen, Y.; Ye, Z.; Sun, H.; Gong, T.; Xiong, S.; Lu, X. Global–Local Fusion With Semantic Information Guidance for Accurate Small Object Detection in UAV Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4701115. [CrossRef]
44. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
45. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Munich, Germany, 8–14 September 2018; pp. 7132–7141.
46. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

47. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
48. Ma, N.; Zhang, X.; Sun, J. Funnel activation for visual recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 351–368.
49. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
50. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
51. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
52. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
53. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
54. Robinson, I.; Robicheaux, P.; Popov, M.; Ramanan, D.; Peri, N. RF-DETR: Neural architecture search for real-time detection transformers. *arXiv* **2025**, arXiv:2511.09554.
55. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
56. Jocher, G. Ultralytics YOLOv5 (Version 7.0). 2020. Available online: https://docs.ultralytics.com/zh/models/yolov5 (accessed on 1 March 2026).
57. Wang, C.-Y.; Yeh, I.-H.; Mark Liao, H.-Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 1–21.
58. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
59. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
60. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
61. Jocher, G.; Qiu, J. Ultralytics YOLO26 (version 26.0.0). 2026. Available online: https://docs.ultralytics.com/zh/models/yolo26 (accessed on 1 March 2026).
62. Pan, W.; Xu, B.; Wang, X.; Lv, C.; Wang, S.; Duan, Z.; Tian, Z. YOLO-FireAD: Efficient Fire Detection via Attention-Guided Inverted Residual Learning and Dual-Pooling Feature Preservation. *arXiv* **2025**, arXiv:2505.20884.
63. Zhang, K.F.; Mao, B.B.; Liu, H.; Zhang, Y. UFS-YOLO: A real-time small fire target detection method incorporated hybrid attention in underground facilities. *Measurement* **2025**, *254*, 117948. [CrossRef]
64. Chen, J.S.; Han, H.Z.; Liu, M.; Su, P.; Chen, X. IFS-DETR: A real-time industrial fire smoke detection algorithm based on an end-to-end structured network. *Measurement* **2025**, *241*, 115660. [CrossRef]
65. Li, Y.; Zhang, W.; Liu, Y.; Jing, R.; Liu, C. An efficient fire and smoke detection algorithm based on an end-to-end structured network. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105492. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.