fire

MDPI

Article

MFFDet: Enhancing Multi-Scale Forest Fire Detection in UAV Imagery

Zhengshen Huang [1], Rui Wang [2], Xin Li [3], Weili Kou [1,*], Qinyan Gu [1], Zengxing Li [1], Jiangxia Ye [4] and Qiuhua Wang [5,6]

1 College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China; hzs@swfu.edu.cn (Z.H.); guqinyan@swfu.edu.cn (Q.G.); lzx457883347@swfu.edu.cn (Z.L.)
2 Forestry Dual Centers of Yunnan Province, Kunming 650225, China; brouff@163.com
3 Yunnan Provincial Government Service Center, Kunming 650051, China; jackyxinli@hotmail.com
4 College of Forestry, Southwest Forestry University, Kunming 650224, China; yjx125@swfu.edu.cn
5 College of Civil Engineering, Southwest Forestry University, Kunming 650224, China; qhwang2010@swfu.edu.cn
6 Yunnan Key Laboratory of Forest Disaster Warning and Control, Kunming 650224, China
* Correspondence: kwl@swfu.edu.cn

Abstract

In Unmanned aerial vehicle (UAV) forest fire detection, flames and smoke exhibit dramatic scale variations. Existing methods often struggle with multi-scale feature extraction, fusion quality, and localization reliability, resulting in limited accuracy improvements. To address this issue, this study optimizes the backbone, neck, and head of YOLOv11n to propose a novel multi-scale forest fire detector (MFFDet), which consists of three key modules: (1) the Multi-Scale Feature Calibration Module (MFCM) is designed to improve multi-scale feature representation by context aggregation and detail calibration; (2) the Cross-Scale Semantic Alignment Module (CSAM) is proposed to enhance fusion quality by applying channel reorganization and local spatial refinement; and (3) the Location Quality Estimator Head (LQEH) is presented for reliable localization by mapping the statistical information of regression distributions into a localization quality score, which systematically boosts the accuracy and stability of multi-scale object detection. In addition, to alleviate the scarcity of UAV forest fire detection data, this study constructs a UAV Forest Fire Dataset (UF^2D), providing important data support for UAV-based fire detection. Experiments on UF^2D show that MFFDet achieves an mAP@0.5 of 70.1%, the best among all compared models, representing a 4.4% improvement over the baseline. Moreover, it attains the top performance on small, medium, and large objects, with AP_s of 20.3%, AP_m of 31.5%, and AP_l of 44.8%, highlighting MFFDet's robustness and accuracy for multi-scale flame and smoke detection in a complex forest fire environment, which bears important practical significance for the intelligent upgrade of forest fire prevention and control.

Keywords: forest fire detection; multi-scale objects; deep learning; YOLOv11n; UAV imagery; multi-scale detector

Check for updates

Academic Editors: Hongyong Yuan and Xiaole Zhang

Received: 12 March 2026
Revised: 30 June 2026
Accepted: 2 July 2026
Published: 4 July 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.

1. Introduction

Forest fires are a type of high-frequency and highly destructive ecological disaster. They not only lead to a decline in biodiversity but also release a large amount of carbon dioxide, causing severe impacts on human society [1,2]. In forest fires, flame and smoke images exhibit significant scale variation characteristics [3]. Early-stage fire points and

thin smoke plumes are typically small in scale, which then rapidly expand into large-scale, complex formations as the fire spreads. Such drastic variations pose a challenge in simultaneously achieving a high detection rate for small targets and stable localization for large targets [4]. Therefore, there is an urgent need to develop effective and reliable multi-scale object detection techniques specifically tailored for forest fires.

In recent years, UAV-based sensing technologies, benefiting from high mobility, real-time responsiveness, wide-area coverage, and low cost, have gradually become an important means for forest fire detection [5]. Meanwhile, the integration of deep learning with high-resolution imaging devices mounted on UAVs has significantly improved the automation level and recognition accuracy of forest fire detection [6–10]. As the foundation for model training and evaluation, the quality of UAV-view forest fire datasets plays a decisive role in detection performance [11,12]. However, real UAV-captured forest fire data remain scarce, and multi-scale samples of flames and smoke are insufficient, which limits model generalization and reliable detection in complex scenarios.

YOLO (You Only Look Once), as a real-time object detection technique, has been widely applied in many fields due to its high inference speed and low computational cost [13]. However, in certain detection scenarios, large variations in object scale can significantly degrade detection performance. Therefore, how to effectively handle scale changes and improve accuracy has become an important challenge in computer vision. In remote sensing scenarios, Ref. [14] significantly enhanced small-object detection performance through Efficient Multi-scale Attention (EMA) [15], while Ref. [16] proposed an edge-guided multi-scale foreground attention network to improve change detection in remote sensing images. In surface defect detection, Ref. [17] leveraged a scale-aware attention mechanism to improve multi-scale defect detection accuracy for steel bridge inspection, while Ref. [18] replaced standard convolutions with multi-scale convolutions to enhance the capture of edge information for conveyor-belt tear detection. In forest fire detection, Ref. [19] designed a decoupled multi-scale dilated convolution mechanism to strengthen the multi-scale feature extraction capability of the backbone network, achieving a precision of 88.6% and a recall of 78.3%. Nevertheless, stacked dilated convolutions are prone to grid artifacts that degrade edge features of tiny smoke, and the fixed dilation ratios fail to accommodate dynamic variations in fire object scales. Ref. [20] proposed a dynamic multi-scale fusion pyramid network that effectively fuses multi-scale features and strengthens the model's ability to detect flames and smoke of different sizes, achieving a precision of 74.0% for fire and 74.9% for smoke, outperforming the original model. However, its dynamic alignment strategy only corrects spatial sampling positions, lacking mechanisms to eliminate inter-level semantic discrepancies and refine features, which results in semantic misalignment between different feature levels. He et al. [21] redesigned the detection head to help the detector achieve stronger multi-scale localization and classification performance. Nevertheless, FireNet employs a generic CBAM attention paradigm and delivers limited improvement in mitigating localization drift.

These studies indicate that improving the YOLO backbone, neck, and head is feasible for enhancing multi-scale object detection accuracy, leading to an increase of 2.6% in mAP@0.5. Overall, the above algorithms only focus on the improvement of a single module and fail to achieve network-level global collaborative optimization, resulting in persistent deficiencies in multi-scale feature extraction, fusion, and localization.

To address these limitations, this study improves upon YOLOv11n and proposes a novel multi-scale forest fire detector(MFFDet). The main contributions of this study are summarized as follows:

- To address the scarcity of UAV-based forest fire detection data, the UAV Forest Fire Dataset (UF^2D) has been constructed. This dataset covers multi-scale fire scenarios

and can provide data support for UAV-based fire detection, generalization evaluation, and comparative studies of detection algorithms.

- To address the inadequate multi-scale detection capability in forest fire detection caused by existing methods that improve only a single module, this study coordinately optimizes the backbone, neck, and head of YOLOv11 to propose the MFFDet.
- To enhance multi-scale information representation, fusion quality, and localization reliability, MFFDet is designed with three key modules: the Multi-Scale Feature Calibration Module (MFCM) for multi-scale feature contextual aggregation and calibration, the Cross-Scale Semantic Alignment Module (CSAM) for cross-scale semantic alignment and refinement, and the Location Quality Estimator Head (LQEH), which maps the statistical information of regression distributions into a localization quality score, thereby systematically improving the accuracy and stability of multi-scale object detection.
- Ablation and comparison experiments on the UF^{2}D show that MFFDet achieves the highest mAP@0.5 of 70.1%, a 4.4% improvement over the baseline. It also excels in detecting small, medium, and large objects with AP_s of 20.3%, AP_m of 31.5%, and AP_l of 44.8%, demonstrating its advantage in multi-scale detection. Additionally, MFFDet exhibits strong generalization on the public M^{4}SFWD dataset.

The rest of this paper is organized as follows: Section 2 introduces the materials and methods, including the introduction of the self-built dataset and a detailed explanation of each module; Section 3 presents the experimental setup, ablation study, comparison results, and visualization; Section 4 discusses the limitations of this study and outlines future work; Section 5 summarizes the paper.

2. Materials and Methods

2.1. Construction of the UAV Forest Fire Dataset (UF^{2}D)

High-quality datasets provide stable training for the model, facilitating more effective convergence and improving detection performance [22,23]. However, due to the sudden and highly time-sensitive nature of forest fires, real UAV-captured forest fire imagery is difficult to obtain effectively, making such datasets extremely scarce [24]. To address this issue, this study constructs UF^{2}D from real forest fire scenarios. Some representative samples are shown in Figure 1. The dataset construction pipeline consists of three stages: data acquisition, data processing, and data annotation.

In the data acquisition stage, the research team conducts an aerial survey of a real forest fire on 11 April 2023, in Jiangchuan District, Yuxi City, Yunnan Province, using a DJI Matrice 300 RTK UAV (DJI, Shenzhen, China). Since the fire occurs in complex mountainous terrain and spreads rapidly, the data collection process faces challenges such as unstable weather conditions and heavy smoke occlusion. To ensure high data quality, the team monitors the scenario in real time and dynamically adjusts the flight trajectory, thereby comprehensively collecting high-resolution images and videos of flames and smoke.

In the data processing stage, the raw aerial videos are first segmented and sampled to generate static image samples for annotation and training. Then, the Structural Similarity Index (SSIM) [25] is used to measure the similarity between adjacent frames and remove highly redundant frames. In addition, image quality is screened using the signal-to-noise ratio (SNR) criterion [26], and samples with SNR $\leq$ 30 dB are discarded, ensuring negligible noise interference for reliable model training. As a result, 519 valid images are retained. The 519 valid UAV-captured images are randomly divided into training (368 images), validation (107 images), and test (44 images) sets in a 7:2:1 ratio at the frame level, ensuring balanced sample distribution and unbiased experimental evaluation.



Figure 1. Representative UAV-captured real forest fire samples from the UF^{2}D.

In the data annotation stage, three professional annotators manually labeled the filtered images using LabelMe (v5.1.1) for two categories: fire and smoke. All annotations are conducted following unified criteria. For the clear targets, bounding boxes are tightly aligned with the outer boundaries; for objects with blurred boundaries or slight occlusion, bounding boxes are drawn to cover the main recognizable regions while maintaining consistent annotation criteria. After annotation, all labels are uniformly reviewed and corrected to fix missing, incorrect, or inconsistent annotations. The annotations are then converted into YOLO-required txt format, resulting in 2750 bounding boxes in total.

In addition, to enrich the data diversity and enhance the robustness of the model to be trained, 2907 images are obtained from open-source platforms including Kaggle (https://www.kaggle.com, accessed on 25 January 2026) and Roboflow (https://roboflow.com, accessed on 25 January 2026), and 782 real-scene images are gathered from public network channels such as WeChat and Baidu. Light and tone of images from different sources are adjusted uniformly, and interfering elements such as watermarks are eliminated. Visual styles of all images are kept consistent to reduce domain deviations among different data sources. All collected images are strictly screened and annotated in accordance with the above unified standards. Ultimately, the constructed UF^{2}D dataset consists of a total of 4208 images, which are randomly divided into training (2945 images), validation (841 images) and test (422 images) sets at a 7:2:1 ratio to ensure uniform category and scene distribution.

UF^{2}D covers forest fire scenarios with objects at different scales, and this scale diversity helps improve the robustness of the model for forest fire detection across different distances and fire-development stages. Meanwhile, UF^{2}D partially alleviates the limitations of existing public forest fire datasets, such as limited data volume, insufficient scenario coverage, and the scarcity of real UAV forest fire data, providing a valuable resource for subsequent research on UAV-based forest fire detection, model generalization evaluation,

and algorithm comparison. The UF^{2}D is available at https://github.com/huang-61/UAV-Forest-Fire-Dataset-UF2D (accessed on 1 February 2026).

2.2. Overview of the Proposed MFFDet

Figure 2 illustrates the overall framework of the proposed MFFDet, which is mainly composed of three key modules. First, in the backbone, the designed MFCM replaces the original C3k2 blocks to better extract multi-scale feature information of flames and smoke in forest fire scenarios. Second, in the neck, a CSAM is proposed and inserted before the detection head to align semantic information and enhance feature refinement, thereby improving the quality of cross-scale feature fusion. Finally, a dedicated detection head, termed the LQEH, is presented for multi-scale forest fire objects, which incorporates localization-quality information into the computation of classification confidence, enabling more reliable and stable multi-scale object localization.

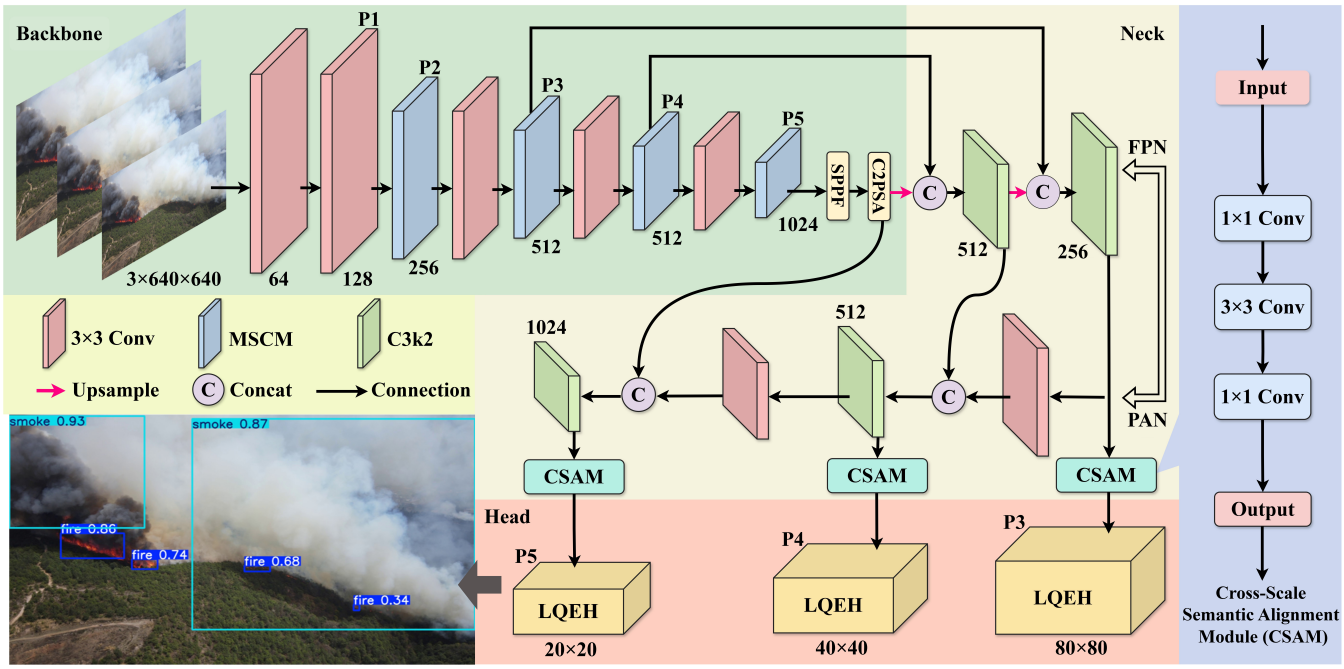


Figure 2. Overall framework of the proposed MFFDet. MSCM: Multi-Scale Feature Calibration Module; CSAM: Cross-Scale Semantic Alignment Module; LQEH: Location Quality Estimator Head. The rightmost part shows a schematic diagram of the CSAM structure.

2.3. Multi-Scale Feature Calibration Module (MFCM) for Improving Multi-Scale Feature

In a forest fire scenario, flame and smoke objects often exhibit significant scale variation: large-scale objects tend to have blurred boundaries and irregular shapes, while small-scale objects are easily overwhelmed by large objects or complex background information during feature learning, resulting in a higher miss-detection rate and consequently increasing the risk of fire escalation. To enhance the ability to extract multi-scale features, this study designs the Multi-Scale Feature Calibration Module (MFCM), the front-end feature enhancement base of MFFDet. It enables efficient aggregation of global contextual cues while precisely preserving and significantly enhancing critical local low-frequency features, thereby improving multi-scale representation and boosting detection accuracy and robustness for flames and smoke across scales, and providing reliable preliminary feature input for subsequent cross-scale semantic alignment and localization optimization.

In addition, a Feature Enhancement Unit (FEU) is designed to perform detail calibration before multi-scale feature fusion, ensuring that important local structural information in the feature maps is sufficiently preserved and reinforced. As illustrated in the detailed

structure of MFCM, the module consists of two complementary paths: a local feature extraction branch and a multi-scale context aggregation branch. Figure 3 shows the overall architecture of the proposed MFCM. A comprehensive description of the module is provided below.

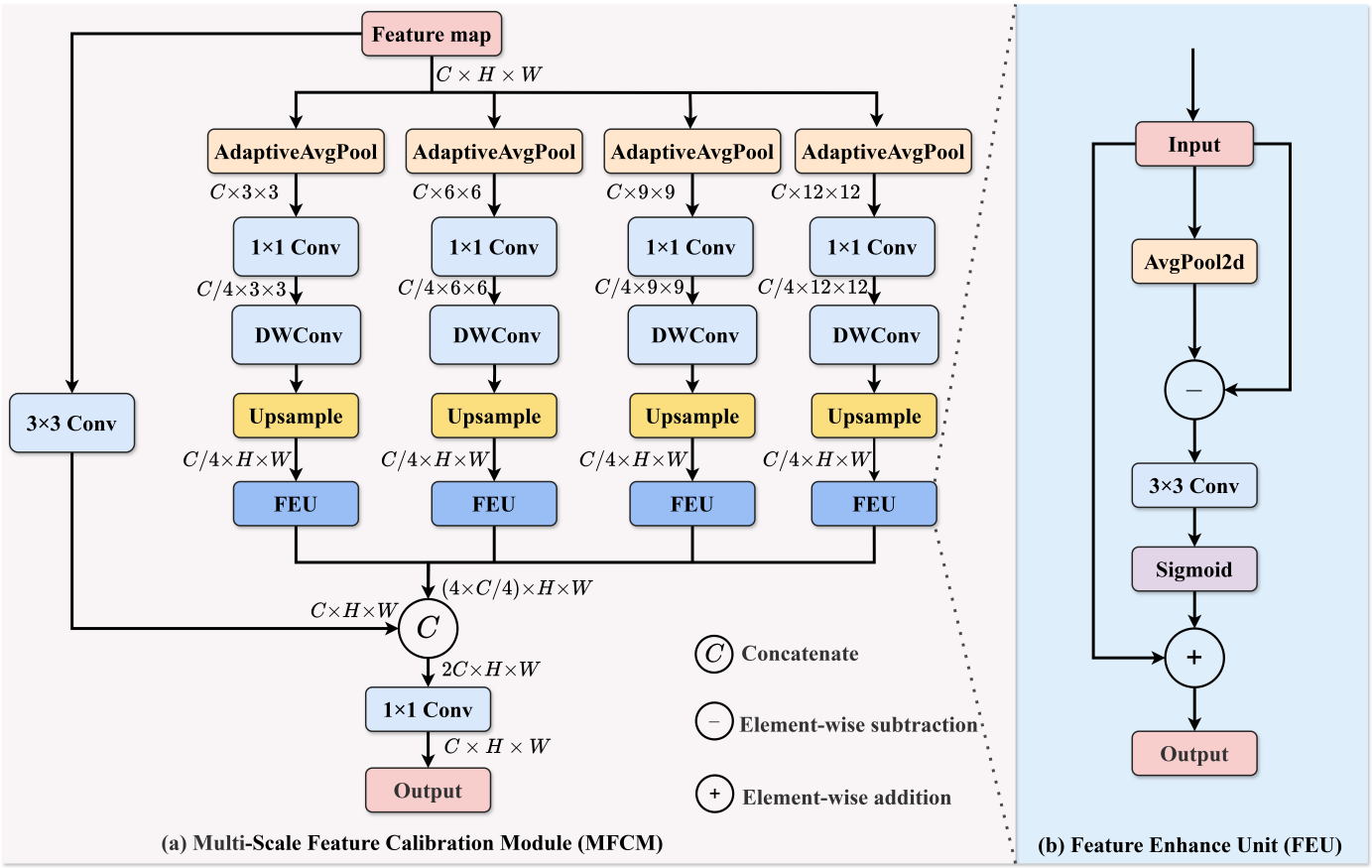


Figure 3. (**a**) Schematic diagram of the MFCM structure, (**b**) FEU structure.

The MFCM first takes an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the number of channels, height, and width, respectively. It then decomposes $\mathbf{X}$ into two parallel paths.

$$\mathbf{X}_{\text{local}} = \mathcal{L}_{3\times 3}(\mathbf{X}) \tag{1}$$

Here, $\mathbf{X}_{\text{local}}$ denotes the extracted high-resolution local feature map, and $\mathcal{L}_{3\times 3}(\cdot)$ represents a 3×3 convolution operation applied to the input feature map.

In the multi-scale context aggregation path, an adaptive average pooling operator $\mathcal{A}_{b_i}(\cdot)$ (AdaptiveAvgPool2d) is employed to spatially compress $\mathbf{X}$ into a fixed $b_i \times b_i$ grid

$$\mathcal{A}_{b_i}(\mathbf{X}) \in \mathbb{R}^{C\times b_i\times b_i}, \quad b_i \in \mathcal{B} \tag{2}$$

where $\mathcal{B} = \{3, 6, 9, 12\}$ denotes four pooling bins for capturing contextual cues at different spatial granularities; smaller b_i aggregates more global context, while larger b_i preserves finer local statistics.

For each bin $b_i \in \mathcal{B}$, a branch is constructed to generate a compact scale feature $\mathbf{X}_i$ via a 1×1 projection $\mathcal{F}_{1\times 1}(\cdot)$ followed by a 3×3 depthwise convolution $\mathcal{D}_{3\times 3}(\cdot)$

$$\mathbf{X}_i = \mathcal{D}_{3\times 3}\Big(\mathcal{F}_{1\times 1}\big(\mathcal{A}_{b_i}(\mathbf{X})\big)\Big) \tag{3}$$

where $\mathbf{X}_i$ denotes the resulting compact scale feature map for bin b_i.

Each $\mathbf{X}_i$ is then upsampled back to the original spatial resolution via UpSample$(\cdot)$

$$\mathbf{X}_i^{\uparrow} = \text{UpSample}(\mathbf{X}_i) \tag{4}$$

Here, $\mathbf{X}_i^{\uparrow}$ represents the upsampled feature map corresponding to the i-th scale.

Finally, each upsampled feature $\mathbf{X}_i^{\uparrow}$ is refined by the Feature Enhancement Unit (FEU) for cross-scale detail calibration

$$\tilde{\mathbf{X}}_i = \text{FEU}\left(\mathbf{X}_i^{\uparrow}\right) \tag{5}$$

where $\tilde{\mathbf{X}}_i$ denotes the FEU-calibrated multi-scale feature map.

For final fusion, MFCM concatenates the high-resolution local feature $\mathbf{X}_{\text{local}}$ with the FEU-calibrated multi-scale features $\{\tilde{\mathbf{X}}_i\}_{i=1}^{N}$ along the channel dimension to form $\mathbf{Z}$, where $N = 4$. Finally, $\mathbf{Z}$ is processed by a 3×3 fusion convolution $\mathcal{F}_{\text{Final}}$ to generate the final enhanced feature $\mathbf{Y}_{\text{final}}$, which improves robustness to multi-scale flame and smoke objects.

Feature Enhancement Unit (FEU)

Small-scale flames are characterized by low contrast and weak boundaries. Under the inherent low-pass filtering effect of upsampling, their high-frequency cues (e.g., flame edges and thin-smoke textures) are more likely to be smoothed out, which can lead to missed detections or localization drift. Therefore, the Feature Enhancement Unit (FEU) is designed for detail calibration.

Given an upsampled feature map $\mathbf{X}_i^{\uparrow}$, a smoothed component $\mathcal{P}(\mathbf{X}_i^{\uparrow})$ is obtained using a 3×3 average pooling. Compared with traditional Gaussian blur or hand-crafted low-pass filters, average pooling requires no manual parameter tuning and introduces no additional trainable parameters, making the network more lightweight. The high-frequency residual is computed as

$$\mathbf{E}_i = \mathbf{X}_i^{\uparrow} - \mathcal{P}(\mathbf{X}_i^{\uparrow}) \tag{6}$$

where $\mathbf{E}_i$ denotes the high-frequency residual, capturing edges and fine textures (e.g., flame contours, thin smoke) that are lost due to upsampling smoothing, and $\mathcal{P}(\mathbf{X}_i^{\uparrow})$ represents the smoothed version of the input feature map obtained by average pooling. Different from traditional fixed-template high-pass filtering and existing complex high-frequency attention methods relying on massive channel recalibration and weight learning, this residual subtraction strategy separates low-frequency background and valid target high-frequency features according to actual feature distribution, and avoids noise interference caused by indiscriminate texture enhancement.

The residual $\mathbf{E}_i$ is then processed by a convolution $\mathcal{F}_{\text{conv}}(\cdot)$ followed by a Sigmoid activation $\sigma(\cdot)$ to produce an enhancement term, yielding the FEU-calibrated feature

$$\tilde{\mathbf{X}}_i = \mathbf{X}_i^{\uparrow} + \sigma\left(\mathcal{F}_{\text{conv}}(\mathbf{E}_i)\right) \tag{7}$$

where $\tilde{\mathbf{X}}_i$ denotes the FEU-enhanced feature map, $\mathcal{F}_{\text{conv}}(\cdot)$ is a convolutional layer applied to the residual, and $\sigma(\cdot)$ is the Sigmoid activation that acts as a learnable gating function to adaptively enhance high-frequency responses.

The core mechanism of FEU is to explicitly treat small-scale objects as local high-frequency cues: it first extracts edge and texture information via the residual $\mathbf{E}_i$; then introduces a learnable gating ($\sigma(\cdot)$) to selectively amplify high-frequency responses; and finally injects the enhancement term back into the original features through a residual connection. This design strengthens boundary and texture responses with minimal computational overhead, thereby improving the representation capability and detection stability for small-scale, low-contrast objects.

2.4. Cross-Scale Semantic Alignment Module (CSAM) for Enhancing Multi-Scale Fusion

In forest fire scenarios, flame and smoke detection is often disturbed by complex backgrounds (e.g., strong glare, haze, and smoke mist). Effective information transfer across different feature levels and scales can improve the model's robustness to such interference. However, the Feature Pyramid Network (FPN) in YOLOv11n is prone to cross-scale semantic inconsistency during feature transmission and fusion, which degrades the detection stability for multi-scale objects. To alleviate this issue, this study proposes a Cross-Scale Semantic Alignment Module (CSAM). This module performs channel reorganization and local spatial refinement on the enhanced multi-scale features of different granularities output by MFCM, generating more consistent and discriminative feature representations, thereby improving the fusion quality and detection stability across scales.

After fusing high-level semantic features with low-level detail features, the fused representation is fed into C3k2, whose internal convolutions reshape the channels and produce $\mathbf{X}' \in \mathbb{R}'^{C_{\text{out}} \times H \times W}$. To achieve cross-scale semantic alignment with minimal overhead, CSAM adopts a lightweight refinement pipeline composed of a 1×1 channel reorganization, a 3×3 local spatial refinement, and a final 1×1 projection. Specifically, the first 1×1 convolution compresses the channel dimension to $C_{\text{out}}/2$, which can be regarded as a learnable linear combination at each spatial location. It mixes multi-source information into new semantic channels rather than simply stacking them, thereby suppressing redundant or conflicting responses and achieving the initial semantic alignment. Since a 1×1 convolution performs channel mixing only and does not model local neighborhoods, a subsequent 3×3 convolution is introduced to capture local spatial structures and refine details (e.g., boundaries and textures), alleviating detail dilution and localization instability caused by cross-scale fusion. Finally, another 1×1 convolution restores the channel dimension to C_{out} required by the detection head, yielding a more consistent and discriminative fused representation. This semantic alignment process suppresses interference noise introduced by cross-scale fusion and guides the model to focus on critical object regions, thereby improving robustness to complex backgrounds and enhancing detection stability. The overall process can be summarized as

$$\mathbf{Y} = \phi\Big(\mathcal{C}^{(2)}_{1\times1}\big(\phi\big(\mathcal{C}_{3\times3}\big(\phi\big(\mathcal{C}^{(1)}_{1\times1}(\mathbf{X}')\big)\big)\big)\big)\Big), \qquad \mathbf{Y} \in \mathbb{R}'^{C_{\text{out}} \times H \times W}. \tag{8}$$

where $\mathbf{X}'$ denotes the fused feature map fed into CSAM and $\mathbf{Y}$ denotes the refined output feature map. $\mathcal{C}^{(1)}_{1\times1}(\cdot)$ represents the first 1×1 convolution (channel reorganization and compression), $\mathcal{C}_{3\times3}(\cdot)$ represents the 3×3 convolution (local spatial refinement), and $\mathcal{C}^{(2)}_{1\times1}(\cdot)$ represents the second 1×1 convolution (channel restoration and projection). $\phi(\cdot)$ denotes the nonlinear activation function. C_{out}, H, and W denote the channel number, height, and width of the feature map, respectively; thus, $\mathbf{Y} \in \mathbb{R}'^{C_{\text{out}} \times H \times W}$ indicates that the output feature map has the size $C_{\text{out}} \times H \times W$.

2.5. Location Quality Estimator Head (LQEH) for Reliable Multi-Scale Localization

In the original YOLOv11n detection head, classification confidence is typically used directly for candidate ranking and NMS suppression. However, classification scores do not explicitly reflect localization quality, which can cause high-score but low-quality candidates to suppress truly high-quality boxes. This issue becomes more pronounced in multi-scale settings with dense proposals. To address this, inspired by the theoretical framework of the GFL series [27,28], this study presents a Location Quality Estimator Head (LQEH) based on statistical cues extracted from bounding-box distributions. This module takes the multi-scale enhanced features from MFCM and the semantically aligned features processed by CSAM as inputs, calibrates the classification confidence with localization-

quality information, thereby enabling more reliable multi-scale object localization. The architecture of the proposed LQEH is illustrated in Figure 4.

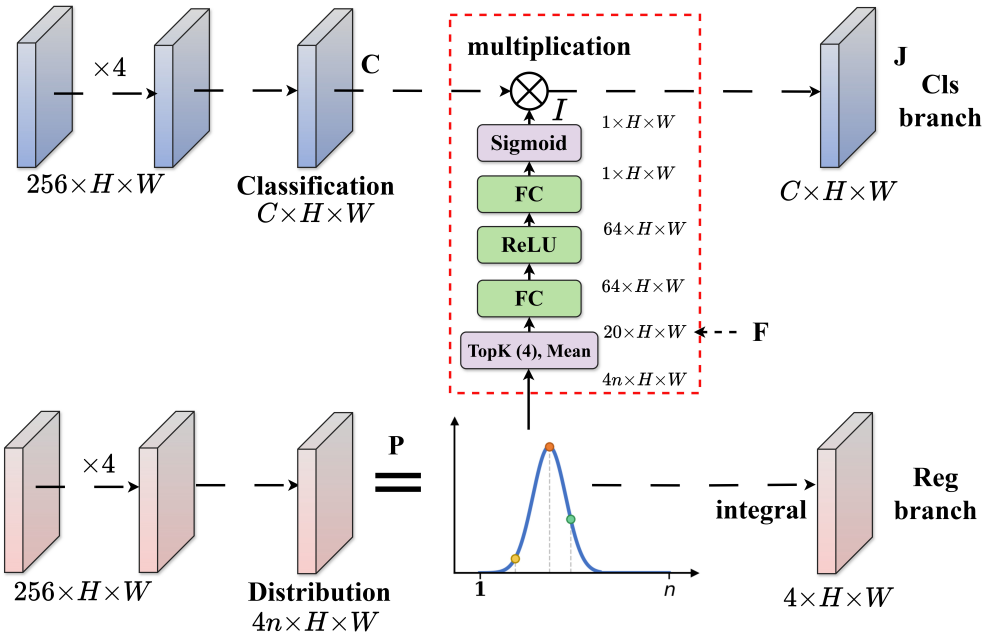


Figure 4. Schematic diagram of LQEHead structure, with the LQE module indicated by the dashed-line box. C: class-wise confidence map from the classification branch; P: four-direction DFL regression distributions $\{P^w\}_{w\in\{l,t,r,b\}}$; J: quality-aware ranking score map; F: a statistical feature vector extracted from the four-direction regression distributions.

This study constructs a classification-quality joint by fusing the classification response C and the estimated localization quality I to obtain the final score J

$$J = C \odot I \tag{9}$$

where $C \in [0,1]^{H\times W\times c}$ denotes the class-wise confidence map, $I \in [0,1]^{H\times W\times 1}$ denotes the localization-quality scalar inferred from the regression distribution, and $\odot$ represents element-wise multiplication.

To obtain I, this study leverages the four-direction discrete distributions P^w $(w \in \{l,t,r,b\})$ produced by the DFL regression branch and extracts compact statistical descriptions from the distribution shapes. Here, l,t,r,b represent the distance offsets from a prediction point to the left, top, right, and bottom sides of the object bounding box (left, top, right, bottom offsets), respectively. Accordingly, for each direction, the regression branch outputs a discrete probability distribution of length $n(= Reg_max)$ to characterize the uncertainty of the distance in that direction. Specifically, this study applies the Topkm$(\cdot)$ operator to each directional distribution and concatenates the results along the channel dimension to form a statistical feature

$$F = \text{Concat}\Big(\{\text{Topkm}(P^w)\}_{w\in\{l,t,r,b\}}\Big) \in \mathbb{R}^{4(k+1)} \tag{10}$$

The Topkm$(\cdot)$ operator denotes the joint operation of extracting the Top-k probabilities from a distribution vector and appending their mean. This statistical feature effectively captures the sharpness of the distribution, where a sharper distribution usually indicates higher localization certainty. In practical forest fire scenarios, when boundary localization is more accurate, such as when flame or smoke contours are clear, the probability mass tends to form a pronounced peak near the true position, yielding larger and more concentrated Top-k values. In contrast, when localization quality is poor, such as in blurred or overlapping

flame–smoke boundary regions, the distribution becomes flatter, and the corresponding Top-k values are smaller and more dispersed. Moreover, the Top-k values together with their mean make the statistical feature insensitive to relative shifts over the distribution domain (see Figure 5), thereby providing a more stable quality representation across objects of different scales.

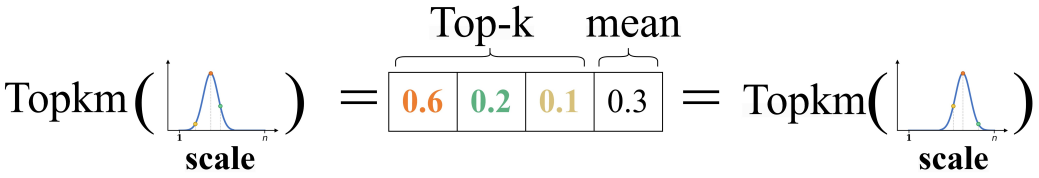


Figure 5. Topkm(·) feature is robust to object scales.

Finally, a lightweight two-layer MLP (FC-ReLU-FC) followed by a Sigmoid activation is adopted to output the localization quality I. Here, $\delta(\cdot)$ and $\sigma(\cdot)$ denote ReLU and Sigmoid, respectively.

$$I = \sigma\big(W_2\,\delta(W_1 F)\big), \tag{11}$$

By injecting I into the classification ranking score J (see Equation (9)), the model tends to preserve more reliably localized candidates under dense multi-scale proposals, thereby improving the localization accuracy and stability for flame and smoke objects.

3. Experiments and Results

3.1. Environment Configuration

The experiments are conducted on a Windows 11 workstation equipped with an AMD Ryzen 9 7945HX CPU (integrated Radeon GPU, 2.50 GHz), an NVIDIA GeForce RTX 4060 GPU with 8 GB video memory, and 32 GB system memory. During training, SGD is selected as the optimizer, with the following key hyperparameters: batch size of 16, learning rate of 0.01, and 200 training epochs. Before training, all input images are resized to a resolution of 640×640 pixels. All experiments are trained and evaluated under identical hyperparameter settings.No pre-trained weights are adopted during model training, and hyperparameters are not separately optimized for each comparative model. All comparison methods are reproduced based on their official open-source implementations, providing a reliable basis for validating the effectiveness of the proposed method.

3.2. Evaluation Metrics

To ensure an objective and comprehensive evaluation of detection performance, this study adopts multiple metrics from the perspectives of model scale, computational cost, detection accuracy, and inference efficiency:

(1) Parameters: The total number of trainable parameters, which reflects the network complexity and storage requirements. A larger parameter count usually indicates higher representation capacity, but it also increases memory consumption and may prolong training time. (2) FLOPs: The total number of floating-point operations required to process a single input image. This metric provides insight into computational complexity and helps estimate inference speed and hardware resource demands. (3) F1-score: The harmonic mean of precision and recall, offering a single measure that balances the two. (4) AP_s, AP_m, and AP_l: average precision for small, medium, and large-scale objects, respectively, computed from the corresponding precision-recall (PR) curves to evaluate detection performance across different object scales. (5) mAP@0.5 and mAP@0.5:0.95: mean Average Precision at an IoU threshold of 0.5 (mAP@0.5), and the average mAP over IoU thresholds from 0.5 to 0.95 with a step size of 0.05 (mAP@0.5:0.95), which jointly assess performance

under different localization strictness levels. (6) FPS (Frames Per Second): A measure of inference speed, indicating how many frames the model can process per second to evaluate real-time capability. The best result for each metric is highlighted in bold.

3.3. Experimental Results and Analysis

3.3.1. Performance Comparison Between MFFDet and Other Object Detection Models

To comprehensively evaluate the performance of the proposed MFFDet, extensive comparisons are conducted against a range of mainstream object detectors, including the CNN-based two-stage detector Faster R-CNN and one-stage detectors SSD and RetinaNet, the Transformer-based RT-DETR, as well as several representative variants of the YOLO family, so as to validate the overall effectiveness of MFFDet.

As reported in Table 1, MFFDet achieves the highest mAP@0.5 of 70.1%, improving the baseline YOLOv11n by 4.4% and ranking first among all compared models. In addition, MFFDet also outperforms the competing methods on mAP@0.5:0.95, F1-score, Recall, and Precision, demonstrating its strong detection capability. Although the lightweight advantage of MFFDet is not prominent when compared with some YOLO variants, it still delivers favorable results in terms of parameter count, computational complexity, and FPS, indicating a better trade-off between accuracy and efficiency and strong potential for UAV-side deployment. For objects at different scales (small, middle, and large), MFFDet surpasses the second-best method by 2.5%, 1.9%, and 1.7%, respectively, suggesting stronger multi-scale representation capability, with particularly more pronounced gains on small objects.

Table 1. Performance comparison of different models.

Models	Params (M)	FLOPs (G)	FPS	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95	AP_s (%)	AP_m (%)	AP_l (%)
Faster R-CNN [29]	41.8	134.4	32.7	69.8	61.2	63.9	64.7	35.1	17.8	28.5	40.5
SSD [30]	35.6	92.7	41.5	68.3	60.5	62.4	63.2	32.6	16.9	27.8	40.0
RetinaNet [31]	36.4	163.9	14.2	64.6	55.2	58.7	60.1	29.2	10.2	25.6	32.6
RT-DETR [32]	32.0	103.4	31.5	71.2	58.1	61.6	63.0	32.5	17.2	27.6	39.1
YOLOv5n [33]	2.6	7.2	199.2	72.7	60.6	65.8	67.1	36.4	15.8	29.6	41.9
YOLOv7-Tiny [34]	6.0	13.0	123.6	67.5	58.8	62.3	63.4	32.7	15.4	26.8	41.9
YOLOv8n [35]	3.0	8.1	168.1	73.0	59.9	66.1	66.9	36.2	18.6	28.3	42.7
YOLOv10n [36]	2.7	8.2	189.7	63.9	59.1	61.2	62.1	30.1	11.3	26.4	37.5
YOLOv11n [37]	2.6	6.3	193.5	70.8	62.0	65.1	65.7	35.5	16.5	28.8	42.4
YOLOv12n [38]	**2.5**	**5.8**	212.6	71.6	61.3	65.6	68.3	36.8	18.1	29.2	43.1
YOLOv13n [39]	**2.5**	6.1	**217.8**	68.0	60.5	64.1	64.5	34.9	16.3	27.7	42.2
ours	2.9	6.9	167.8	**74.6**	**63.5**	**69.0**	**70.1**	**38.7**	**20.3**	**31.5**	**44.8**

These results further indicate that MFFDet maintains high computational efficiency while remaining feasible for practical deployment. In UAV scenarios, it exhibits stable and excellent detection performance for early-stage forest fire small objects (e.g., small flames and thin smoke plumes). This advantage mainly stems from the targeted improvements to the YOLOv11 architecture: MFCM strengthens local low-frequency feature representation, CSAM facilitates cross-level feature fusion, and LQEH improves bounding-box localization quality. Overall, these modular innovations significantly enhance the model's multi-scale detection capability, which is especially critical for complex forest fire scenarios with severe background clutter and dramatic scale variations.

3.3.2. Ablation Experiments

To validate the effectiveness of each module and improvement strategy in the model, this study employed a control variable method, gradually introducing the MFCM, CSAM, and LQEH components to assess the contribution of these three key improvements to overall performance. As shown in Table 2, each modification to the baseline model YOLOv11 has had a positive impact on the results. When MFCM is introduced alone, mAP@0.5 and mAP@0.5:0.95 improved by over 2%, thanks to the effective design of multi-scale context aggregation and FEU feature enhancement. In the separate evaluation of CSAM, the most notable improvements are seen in mAP@0.5 (fire) and (smoke), which increased by 2.3% and 2.0%, respectively. This demonstrates its enhancement of cross-layer semantics, effectively reducing background interference and false positives. After improving the original detection head, LQEH increased the mAP@0.5:0.95, which is more sensitive to localization, by 2.1% to 37.4%, indicating more reliable and accurate localization of objects. When CSAM, MFCM, and LQEH are integrated individually, a further significant improvement is observed, with a notable increase in mAP, reflecting the adaptability between the modules. Finally, when all modules are combined, the precision metrics reached their peak, especially with mAP@0.5 achieving the highest value of 70.1%, a 4.4% increase over the baseline model. Although this integration increased the number of parameters and complexity, the overall improvement in detection performance is significant, enhancing the model's robustness in multi-scale and complex backgrounds.

Table 2. Ablation results of different module combinations in MFFDet.

MFCM	CSAM	LQEH	Params (M)	FLOPs (G)	FPS	mAP@0.5	mAP@0.5:0.95	$mAP@0.5_{fire}$	$mAP@0.5_{smoke}$
–	–	–	2.58	**6.35**	201.7	65.7	35.5	62.9	68.4
✓	–	–	**2.53**	6.46	**218.3**	67.7	36.7	64.4	70.0
–	✓	–	2.91	6.93	155.3	67.8	37.0	65.2	70.4
–	–	✓	2.59	6.51	200.8	67.5	37.4	63.7	69.4
–	✓	✓	2.91	6.95	157.2	68.6	37.2	66.5	70.7
✓	–	✓	2.54	6.49	217.7	69.4	37.8	66.4	72.5
✓	✓	–	2.86	6.91	168.6	69.6	38.0	66.5	71.7
✓	✓	✓	2.86	6.93	167.8	**70.1**	**38.7**	**68.4**	**72.7**

✓: Indicates that the corresponding module is adopted in the model configuration.

3.4. Visualization Analysis

3.4.1. Visualization of Detection Results

Figure 6 presents a visual comparison of YOLOv8n, YOLOv10n, YOLOv11n, YOLOv12n, YOLOv13n, and the proposed MFFDet across diverse real-world forest fire scenarios. As shown in Figure 6a, compared with other models that suffer from localization drift and redundant bounding boxes when detecting small-scale smoke, MFFDet can localize smoke more accurately and produce higher confidence scores, mainly benefiting from the LQEH design that retains higher-quality candidate boxes based on quality scoring. In Figure 6b, thin smoke is easily confused with background clouds, making it difficult for the YOLOv11n baseline to capture effective object cues. In contrast, MFFDet suppresses noise interference by enhancing semantic fusion, thereby focusing more on the true object region. In Figure 6c, when faced with multi-scale characteristics caused by tiny ignition points and dynamic smoke, MFFDet still achieves more precise detection, thanks to the efficient aggregation of multi-scale information provided by MFCM. As illustrated in Figure 6d, even under heavy fog occlusion, MFFDet can detect small fire spots, which is crucial for early forest fire warning by enabling timely discovery and suppression of fire spread, whereas other models fail to

deliver comparable results. Finally, in Figure 6e, under nighttime conditions, YOLOv8n and YOLOv11n mistakenly detect lights as flames, while MFFDet shows stronger discrimination against non-fire bright regions, demonstrating improved robustness to interference. Overall, MFFDet exhibits superior localization quality and detection stability under multi-scale objects and complex backgrounds, further validating its potential for practical deployment in complex forest fire environments.

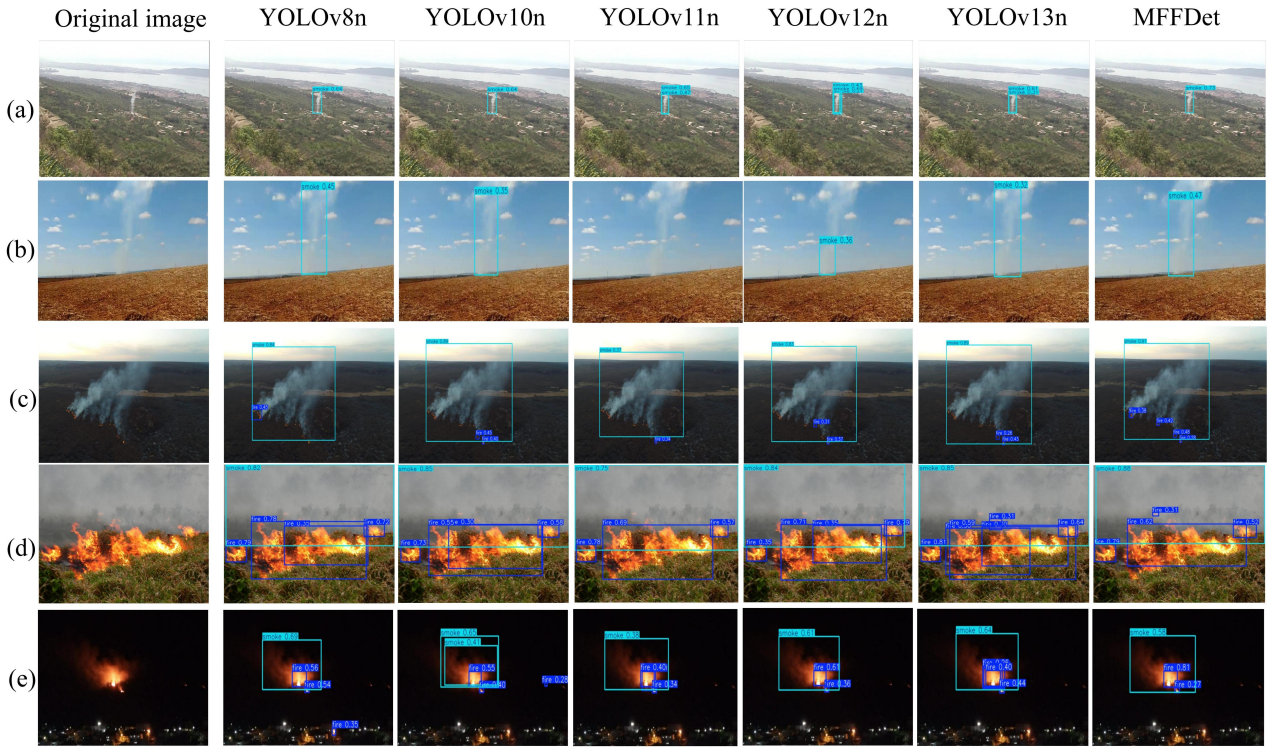


Figure 6. Visual comparison of detection results of different deep learning models on the UF^{2}D.

3.4.2. Visualization of Grad-CAM Results

To further investigate how the models localize fire and smoke regions in images, Grad-CAM is employed to generate attention heatmaps for MFFDet and the YOLO-series models. These heatmaps provide an intuitive visualization of the key regions that the models focus on during prediction. As shown in Figure 7, although YOLOv8n and YOLOv12n can activate some object-related areas, their responses are relatively scattered, resulting in unstable focus on the main bodies of flames and smoke. Moreover, YOLOv10 and YOLOv12n are able to capture object cues to some extent, but their coverage of critical regions remains insufficient, making it difficult to accurately delineate object contours. In contrast, compared with YOLOv11n, MFFDet exhibits more concentrated activation on the flame and smoke bodies as well as their boundary details, while effectively suppressing irrelevant background responses, demonstrating stronger object-attention capability and robustness to interference. These results further confirm that MFFDet can more effectively focus on discriminative regions under complex scenarios, thereby improving the localization reliability and robustness of fire and smoke detection.

3.5. Statistical Validation

To fully verify the performance stability of the proposed model, this study conducts multiple independent repeated experiments for statistical verification. Under unified experimental conditions including software and hardware environments as well as dataset settings, five independent training and testing trials are conducted. Key metrics such as mAP@0.5 and mAP@0.5:0.95 are recorded, and all detailed results are summarized in Table 3.

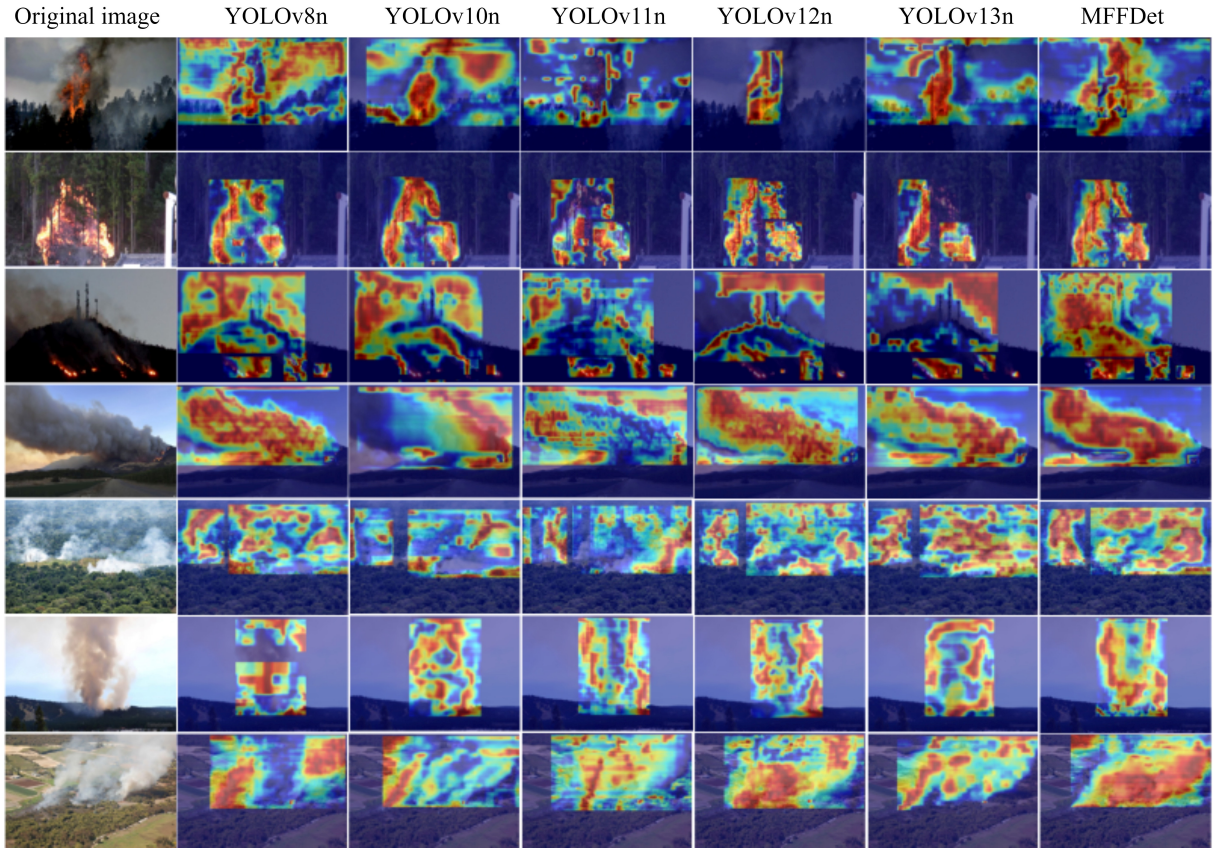


Figure 7. Grad-CAM visualizations comparing different deep learning models on UF^{2}D.

Table 3. Summary of model performance metrics over 5 independent runs.

Performance Metric	Mean (%)	Standard Deviation (%)	95% Confidence Interval (95% CI, %)
Precision	74.58	0.30	(74.21, 74.95)
Recall	63.46	0.57	(62.75, 64.17)
mAP@0.5	70.13	0.23	(69.85, 70.41)
mAP@0.5:0.95	38.71	0.35	(38.28, 39.14)

The statistical results demonstrate that the proposed model exhibits excellent performance stability across multiple independent runs. The average precision reaches 74.58% with a standard deviation of 0.30%, and its 95% confidence interval calculated by t-distribution is from 74.21% to 74.95%. The average recall is 63.46% with a standard deviation of 0.57%, and its 95% confidence interval ranges between 62.75% and 64.17%. In terms of detection accuracy, the mean value of mAP50 is 70.13% with a standard deviation of 0.23%, and the mean mAP@0.5 is 38.71% with a standard deviation of 0.35%.

Such slight performance fluctuations illustrate that the detection capability of the model is not generated by random experimental errors, which fully verifies its stable output and strong experimental reproducibility. In conclusion, the repeated experimental results effectively prove that the proposed model can achieve competitive detection performance and maintain outstanding stable operation ability.

3.6. Generalizability Analysis

The publicly available M^{4}SFWD dataset [40] is a multi-dimensional synthetic dataset designed for remote-sensing forest fire detection, providing diverse simulated forest fire scenarios. To further evaluate the generalization capability of MFFDet, comparative experiments are conducted on this dataset. By analyzing the performance of different models

on the dataset, the adaptability and effectiveness of MFFDet across diverse scenarios are validated. As shown in Table 4 shows that MFFDet achieves the highest mAP@0.5 (87.5%), precision (84.8%), and recall (82.3%) on the M^{4}SFWD dataset, indicating stronger generalization across diverse remote-sensing forest fire scenarios. In terms of efficiency, MFFDet requires only 2.9 M parameters and 6.9 GFLOPs, and maintains real-time inference at 271.5 FPS. Although it is not the most lightweight model, it delivers a better accuracy–efficiency trade-off overall. These results further highlight the robustness and adaptability of MFFDet across different scenarios.

Table 4. Performance comparison of MFFDet with YOLO-series models on the M^{4}SFWD dataset.

Model	mAP@0.5 (%)	Precision (%)	Recall (%)	Parameters (M)	GFLOPs	FPS
YOLOv8n	86.8	84.5	79.9	3.0	6.8	252.7
YOLOv10n	85.5	82.8	78.0	2.7	6.5	277.8
YOLOv11n	87.0	83.1	81.6	2.6	6.3	270.3
YOLOv12n	86.5	83.8	81.1	**2.5**	**5.8**	**292.6**
YOLOv13n	86.9	84.2	81.5	**2.5**	6.1	281.9
MFFDet	**87.5**	**84.8**	**82.3**	2.9	6.9	271.5

As shown in Figure 8, three representative images are selected from the M^{4}SFWD dataset to systematically evaluate the generalization performance of different models under complex conditions. In the daytime multi-ignition scenario (Figure 8a), although the baseline models YOLOv11n and YOLOv13n can roughly delineate the contours of fire spots and smoke, MFFDet can still precisely capture subtle ignition points in regions with blurred smoke boundaries. This capability is crucial for early wildfire warning, as timely detection can effectively reduce the risk of fire spread. In the multi-scale scenario with illumination interference (Figure 8b), YOLOv11n misses faint smoke plumes due to strong lighting, while YOLOv13n misclassifies faint smoke as flames. In contrast, MFFDet reliably detects both flames and smoke, maintaining high confidence scores. Under the nighttime condition without illumination (Figure 8c), YOLOv11n fails to detect small-scale smoke; although YOLOv13n can detect it, its description of smoke morphology is less complete, whereas MFFDet provides a more comprehensive characterization of smoke shapes. Overall, the results across these typical scenarios demonstrate that MFFDet achieves more robust detection in diverse environments, further confirming its adaptability and accuracy in complex forest settings.

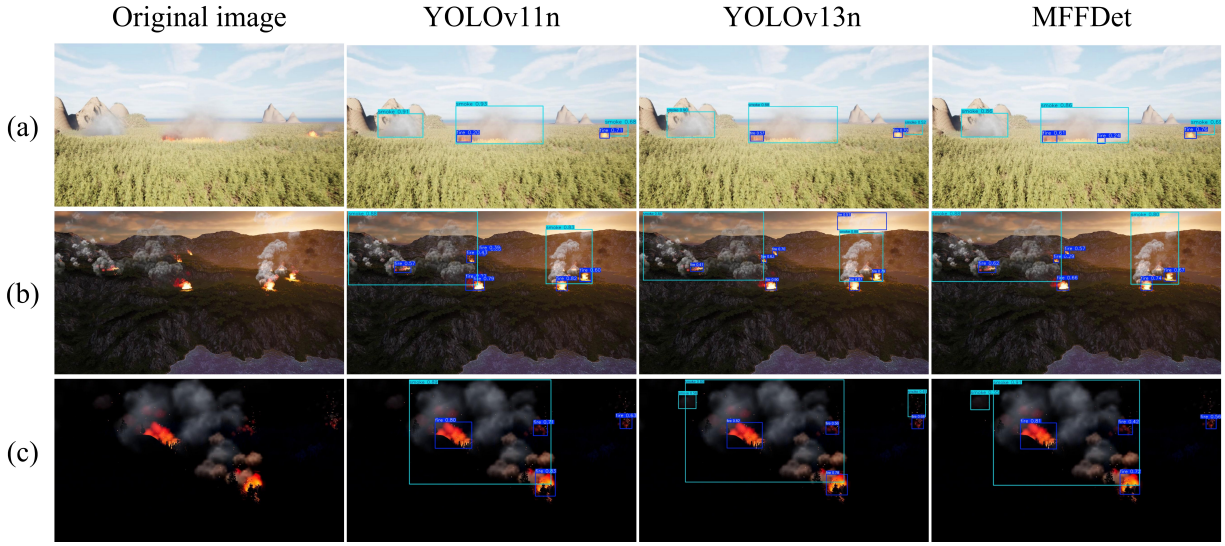


Figure 8. Visual comparison of detection results of different YOLO models on the M^{4}SFWD.

4. Discussion

Considering that flames and smoke in complex forest fire environments exhibit pronounced multi-scale characteristics and are difficult to detect accurately, existing studies have typically improved performance by modifying only a single component of the YOLO architecture. For example, some works optimize only the backbone to enhance feature extraction, improve only the neck to strengthen multi-scale feature fusion, or adjust only the head to refine classification and localization [9,10,19–21]. However, such localized enhancements often lack a unified, end-to-end collaborative design spanning from feature extraction to final prediction, making it difficult to achieve stable and reliable detection under complex backgrounds. To address this limitation, this study proposes MFFDet for multi-scale flame and smoke detection in complex forest fire scenarios.

First, in the backbone, unlike SPP [41] and SPPF [33], which rely on fixed-size max pooling kernels, and ASPP [42], which uses dilated convolutions with pre-defined dilation rates, the proposed MFCM adopts adaptive average pooling with four different target output sizes (3 × 3, 6 × 6, 9 × 9, and 12 × 12). This design dynamically adjusts pooling kernels for flexible multi-scale context capture, free from fixed receptive-field constraints. Different from FPN's [43] cross-level fusion in the neck, the proposed MFCM extracts features directly on feature maps at multiple backbone modules, delivering richer calibrated multi-scale details to enhance detection of fire and smoke targets in complex scenes. Nevertheless, irreversible detail loss still occurs for ultra-fine smoke textures during upsampling operations. Second, during multi-scale feature fusion in conventional FPN [43] and PAN [44] structures, low-level detailed information is prone to being overwhelmed by high-level semantic information, which in turn weakens the discriminative ability for objects of different scales. Unlike traditional bottleneck convolution modules embedded inside convolutions merely to compress the number of parameters, and also different from conventional attention mechanisms that merely perform feature recalibration via channel or spatial weights, the proposed CSAM is strategically deployed between the neck feature fusion layers and each detection head. This module leverages convolutional combinations to perform semantic alignment and fine-grained enhancement on features at different scales, effectively improving the quality of cross-scale feature fusion and strengthening the model's focus on critical flame and smoke regions. Its alignment performance degrades under heavy smoke occlusion and extremely strong illumination, and cross-scale semantic conflicts cannot be completely eliminated in such harsh conditions. Finally, conventional detection heads have limited capability in modeling localization quality and are therefore easily affected by low-quality predictions in dense candidate-box scenarios [25,45]. To overcome this limitation, LQEH maps the statistical characteristics of the regression distribution into localization quality scores, enabling the model to prioritize high-quality bounding boxes and thereby further improving the localization accuracy and detection stability of multi-scale objects. When object contours are extremely blurred, however, the corresponding probability distribution tends to flatten, reducing the reliability of quality evaluation and weakening the screening performance of high-quality candidate boxes.

In addition, this study has certain limitations in experimental comparisons. The mainstream algorithms such as Faster R-CNN, RetinaNet and SSD listed in the table are detection models with large parameter volumes and high computational costs, which differ greatly in scale from the lightweight model proposed in this paper. In future research, models will be grouped by model scale to conduct more fair and reasonable horizontal comparative experiments. To alleviate the critical scarcity of datasets suitable for UAV-based forest fire detection, this study constructs the UF^2D, providing a reliable data foundation for UAV-side fire detection, model generalization evaluation, and comparative algorithmic studies. Nevertheless, UF^2D still has room for expansion in terms of extreme-condition

samples (e.g., nighttime, rain/snow, heavy fog, strong backlighting) and cross-region scenario coverage. Future work will further enlarge the dataset scale and scenario diversity and incorporate multimodal data (e.g., visible, infrared, and thermal imagery), and explore the theoretical framework of few-shot learning under data scarcity, so as to more comprehensively validate the model's generalization ability in real-world complex environments. Meanwhile, to further facilitate its practical application in this field, future studies will optimize the MFFDet model via network lightweight design, structured pruning and knowledge distillation. It can well balance detection accuracy and inference speed while improving engineering adaptability, so as to realize the practical deployment of the model on real edge devices.

5. Conclusions

This study proposes the MFFDet, a novel multi-scale forest fire detector built upon YOLOv11n, aimed at improving multi-scale object detection accuracy in complex forest fire scenarios. By introducing three key modules, these modules work progressively and complementarily to improve detection accuracy. MFCM first aggregates multi-level features to characterize fires and smoke of various sizes. Next, CSAM corrects cross-scale semantic misalignment and filters redundant noise. Finally, LQEH optimizes candidate box positioning based on the aligned features to reduce localization errors from complex backgrounds and scale changes. This sequential collaboration strengthens MFFDet's multi-scale detection capacity and stabilizes feature fusion and localization. Furthermore, the UAV Forest Fire Dataset (UF^{2}D), a self-constructed dataset capturing real-world UAV-based fire images, provides valuable data support for forest fire detection and evaluation, addressing the issue of data scarcity in UAV fire detection. Extensive experiments on the UF^{2}D show that MFFDet achieves an mAP@0.5 of 70.1%, a 4.4% improvement over the baseline model YOLOv11n, surpassing all compared models. Additionally, it attains the top performance on small, medium, and large objects, with AP_s of 20.3%, AP_m of 31.5%, and AP_l of 44.8%, demonstrating MFFDet's strong capability in multi-scale object detection. Furthermore, MFFDet achieves the best performance across various metrics on the public remote-sensing wildfire dataset M^{4}SFWD, highlighting its exceptional generalization ability.

In summary, the UF^{2}D constructed in this study addresses the data shortage for UAV-based forest fire detection. The proposed MFFDet achieves remarkable performance for multi-scale fire detection, which can provide robust technical support for UAV fire monitoring and early warning applications.

Author Contributions: Z.H.: writing—original draft, data curation, methodology, validation, visualization. R.W.: conceptualization, formal analysis, methodology. X.L.: validation, visualization. W.K.: writing—review and editing, conceptualization, funding acquisition. Q.G.: formal analysis, visualization. Z.L.: formal analysis, methodology. J.Y.: writing—review and editing, conceptualization. Q.W.: writing—review and editing, methodology. 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 [grant numbers: 32260391]; Yunnan International Joint Laboratory of Natural Rubber Intelligent Monitor and Digital Applications [grant numbers: 202403AP140001]; Yunnan International Joint R&D Center (China-Malaysia) for Digital Monitoring, Management and Applications of Nature Reserves [grant numbers: 202503AP140040]; Yunnan Fundamental Research Projects [grant numbers: 2018FG001-059]; and the Xingdian Talent Support Program [grant numbers: XDYC-CYCX-2024-0021, YFGRC202532].Additional support was provided by the Academician Li Wei Workstation of Yunnan Province (grant numbers: 202505AF350082).

Data Availability Statement: The UF^{2}D can be found at the following links: https://github.com/huang-61/UAV-Forest-Fire-Dataset-UF2D (accessed on 1 February 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cunningham, C.X.; Williamson, G.J.; Bowman, D.M.J.S. Increasing frequency and intensity of the most extreme wildfires on Earth. *Nat. Ecol. Evol.* **2024**, *8*, 1420–1425. [CrossRef] [PubMed]
2. Zhao, J.; Zheng, B.; Ciais, P.; Chen, Y.; Gasser, T.; Canadell, J.G.; Zhang, L.; Zhang, Q. Global warming amplifies wildfire health burden and reshapes inequality. *Nature* **2025**, *647*, 928–934. [CrossRef] [PubMed]
3. Ding, C.; Li, H.; Liu, Y.; He, B.; Lang, X.; Wang, G. Multi-scale enhanced contextual transformer network for forest fire detection. *Digit. Signal Process.* **2025**, *172*, 105850. [CrossRef]
4. Chen, Y.; Morton, D.C.; Randerson, J.T. Remote sensing for wildfire monitoring: Insights into burned area, emissions, and fire dynamics. *One Earth* **2024**, *7*, 1022–1028. [CrossRef]
5. Bailon-Ruiz, R.; Bit-Monnot, A.; Lacroix, S. Real-time wildfire monitoring with a fleet of UAVs. *Robot. Auton. Syst.* **2022**, *152*, 104071. [CrossRef]
6. Chen, G.; Cheng, R.; Lin, X.; Jiao, W.; Bai, D.; Lin, H. LMDFS: A lightweight model for detecting forest fire smoke in UAV images based on YOLOv7. *Remote Sens.* **2023**, *15*, 3790. [CrossRef]
7. Han, Y.; Duan, B.; Guan, R.; Yang, G.; Zhen, Z. LUFFD-YOLO: A lightweight model for UAV remote sensing forest fire detection based on attention mechanism and multi-level feature fusion. *Remote Sens.* **2024**, *16*, 2177. [CrossRef]
8. Yang, H.; Wang, J.; Wang, J. Efficient detection of forest fire smoke in UAV aerial imagery based on an improved YOLOv5 model and transfer learning. *Remote Sens.* **2023**, *15*, 5527. [CrossRef]
9. Titu, M.F.S.; Pavel, M.A.; Michael, G.K.O.; Babar, H.; Aman, U.; Khan, R. Real-time fire detection: Integrating lightweight deep learning models on drones with edge computing. *Drones* **2024**, *8*, 483. [CrossRef]
10. Akhloufi, M.A.; Couturier, A.; Castro, N.A. Unmanned Aerial Vehicles for Wildland Fires: Sensing, Perception, Cooperation and Assistance. *Drones* **2021**, *5*, 15. [CrossRef]
11. Wang, M.; Yue, P.; Jiang, L.; Yu, D.; Tuo, T.; Li, J. An open flame and smoke detection dataset for deep learning in remote sensing based fire detection. *Geo-Spat. Inf. Sci.* **2025**, *28*, 511–526. [CrossRef]
12. Pesonen, J.; Raita-Hakola, A.M.; Joutsalainen, J.; Hakala, T.; Akhtar, W.; Koivumäki, N.; Markelin, L.; Suomalainen, J.; Alves de Oliveira, R.; Pölönen, I.; et al. Boreal Forest Fire: UAV-collected wildfire detection and smoke segmentation dataset. *Sci. Data* **2025**, *12*, 1419. [CrossRef]
13. Murat, A.A.; Kiran, M.S. A comprehensive review on YOLO versions for object detection. *Eng. Sci. Technol. Int. J.* **2025**, *70*, 102161. [CrossRef]
14. Wei, X.; Li, Z.; Wang, Y. SED-YOLO based multi-scale attention for small object detection in remote sensing. *Sci. Rep.* **2025**, *15*, 3125. [CrossRef] [PubMed]
15. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [CrossRef]
16. Lin, J.; Wang, G.; Peng, D.; Guan, H. Edge-guided multi-scale foreground attention network for change detection in high resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104070. [CrossRef]
17. Luo, Y.; Ling, J.; Wang, J.; Zhang, H.; Chen, F.; Xiao, X.; Lu, N. SFW-YOLO: A lightweight multi-scale dynamic attention network for weld defect detection in steel bridge inspection. *Measurement* **2025**, *253*, 117608. [CrossRef]
18. Wang, X.; Wan, S.; Li, Z.; Chen, X.; Zhang, B.; Wang, Y. ECL-Tear: Lightweight detection method for multiple types of belt tears. *Measurement* **2025**, *251*, 117269. [CrossRef]
19. Zhu, H.; Ling, W.; Yan, H.; Zhong, X.; Liao, F. YOLO-MP: A lightweight forest fire detection model. *Ecol. Inform.* **2025**, *92*, 103516. [CrossRef]
20. 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]
21. He, Y.; Sahma, A.; He, X.; Wu, R.; Zhang, R. FireNet: A lightweight and efficient multi-scenario fire object detector. *Remote Sens.* **2024**, *16*, 4112. [CrossRef]
22. Freire, A.; Silva, L.H.d.S.; de Andrade, J.V.R.; Azevedo, G.O.A.; Fernandes, B.J.T. Beyond clean data: Exploring the effects of label noise on object detection performance. *Knowl. Based Syst.* **2024**, *304*, 112544. [CrossRef]
23. Mohammed, S.; Budach, L.; Feuerpfeil, M.; Ihde, N.; Nathansen, A.; Noack, N.; Patzlaff, H.; Naumann, F.; Harmouch, H. The effects of data quality on machine learning performance on tabular data. *Inf. Syst.* **2025**, *132*, 102549. [CrossRef]
24. Mowla, M.N.; Asadi, D.; Tekeoglu, K.N.; Masum, S.; Rabie, K. UAVs-FFDB: A high-resolution dataset for advancing forest fire detection and monitoring using unmanned aerial vehicles (UAVs). *Data Brief* **2024**, *55*, 110706. [CrossRef] [PubMed]
25. Wang, Q.; Guan, S.; Lyu, S.; Cheng, G. Enhancing drone-based fire detection with flame-specific attention and optimized feature fusion. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104655. [CrossRef]

26. Welvaert, M.; Rosseel, Y. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. *PLoS ONE* **2013**, *8*, e77089. [CrossRef] [PubMed]
27. 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. *arXiv* **2020**, arXiv:2002.10187. [CrossRef]
28. Li, X.; Wang, W.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11632–11641. [CrossRef]
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
31. 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. [CrossRef]
32. 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. [CrossRef]
33. Jocher, G. YOLOv5 by Ultralytics. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 8 January 2026).
34. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475. [CrossRef]
35. Jocher, G.; Qiu, J.; Chaurasia, A. Ultralytics YOLOv8. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 9 January 2026).
36. Chen, H.; Chen, K.; Ding, G.; Han, J.; Lin, Z.; Liu, L.; Wang, A. YOLOv10: Real-Time End-to-End Object Detection. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*; Curran Associates, Inc.: Red Hook, NY, USA, 2024. [CrossRef]
37. Ultralytics. YOLOv11: Real-Time Object Detection Model. 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 January 2026).
38. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524. [CrossRef]
39. LI, P.; Chen, J. Improved YOLOv13 for Infrared Remote Sensing Small Object Detection. *Comput. Eng. Appl.* **2026**, *62*, 131.
40. Wang, G.; Li, H.; Li, P.; Lang, X.; Feng, Y.; Ding, Z.; Xie, S. M4SFWD: A Multi-Faceted synthetic dataset for remote sensing forest wildfires detection. *Expert Syst. Appl.* **2024**, *248*, 123489. [CrossRef]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
42. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv* **2016**, arXiv:1606.00915. [CrossRef]
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
44. 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, 18–23 June 2018. [CrossRef]
45. Lin, Z.; Yun, B.; Zheng, Y. LD-YOLO: A lightweight dynamic forest fire and smoke detection model with dysample and spatial context awareness module. *Forests* **2024**, *15*, 1630. [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.