

Article

Real-Time Detection of Smoke and Fire in the Wild Using Unmanned Aerial Vehicle Remote Sensing Imagery

Xijian Fan ¹, Fan Lei ^{2,*} and Kun Yang ¹
¹ College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China

² Key Laboratory of Natural Resources Monitoring and Supervision in Southern Hilly Region, Ministry of Natural Resources, Changsha 410118, China

* Correspondence: lf0213@img.net

Abstract: Detecting wildfires and smoke is essential for safeguarding forest ecosystems and offers critical information for the early evaluation and prevention of such incidents. The advancement of unmanned aerial vehicle (UAV) remote sensing has further enhanced the detection of wildfires and smoke, which enables rapid and accurate identification. This paper presents an integrated one-stage object detection framework designed for the simultaneous identification of wildfires and smoke in UAV imagery. By leveraging mixed data augmentation techniques, the framework enriches the dataset with small targets to enhance its detection performance for small wildfires and smoke targets. A novel backbone enhancement strategy, integrating region convolution and feature refinement modules, is developed to facilitate the ability to localize smoke features with high transparency within complex backgrounds. By integrating the shape aware loss function, the proposed framework enables the effective capture of irregularly shaped smoke and fire targets with complex edges, facilitating the accurate identification and localization of wildfires and smoke. Experiments conducted on a UAV remote sensing dataset demonstrate that the proposed framework achieves a promising detection performance in terms of both accuracy and speed. The proposed framework attains a mean Average Precision (mAP) of 79.28%, an F1 score of 76.14%, and a processing speed of 8.98 frames per second (FPS). These results reflect increases of 4.27%, 1.96%, and 0.16 FPS compared to the YOLOv10 model. Ablation studies further validate that the incorporation of mixed data augmentation, feature refinement models, and shape aware loss results in substantial improvements over the YOLOv10 model. The findings highlight the framework's capability to rapidly and effectively identify wildfires and smoke using UAV imagery, thereby providing a valuable foundation for proactive forest fire prevention measures.

Keywords: UAV remote sensing; smoke and fire detection; deep learning; object detection



Academic Editors: Haidi Abdullah, Elnaz Neinavaz and Margarita Huesca

Received: 18 November 2024

Revised: 13 January 2025

Accepted: 13 January 2025

Published: 22 January 2025

Citation: Fan, X.; Lei, F.; Yang, K. Real-Time Detection of Smoke and Fire in the Wild Using Unmanned Aerial Vehicle Remote Sensing Imagery. *Forests* **2025**, *16*, 201. <https://doi.org/10.3390/f16020201>

Copyright: © 2025 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Wildfires are considered one of the most destructive and hazardous natural disasters worldwide [1]. In early 2022, the United Nations Environment Programme (UNEP) published a report titled “Spreading like Wildfire—The Rising Threat of Extraordinary Landscape Fires”, which defines wildfires as “an abnormal combustion of vegetation that can be triggered by human malice, accidental causes, or natural factors, resulting in negative impacts on social, economic, or environmental values”. Each year, millions of acres of land are devastated by wildfires, causing significant destruction to human life, vegetation canopies, and forest resources [2,3]. Ecosystems such as peatlands and forests experience

wildfires that release substantial amounts of carbon dioxide into the atmosphere, significantly affecting the global carbon cycle. In addition to the direct loss of life, the large quantities of harmful particulate matter generated by wildfire smoke pose serious health threats to populations.

The speed at which a fire is detected and warnings are communicated to the relevant authorities is a crucial factor in effectively reducing wildfire risks. Therefore, the timely and accurate early detection of forest fires is key to ensuring that these incidents remain manageable [4]. Over the years, various technologies have been proposed to assist in identifying wildfires during their early stages, thereby facilitating the allocation of appropriate resources for extinguishing them [5–7]. Among these methods, ground-based watchtowers and satellite remote sensing monitoring represent two of the most prevalent approaches. However, watchtower observations are often constrained by topographical limitations, resulting in limited coverage, blind spots, and areas devoid of surveillance. In addition, they cannot be established in remote locations lacking basic living conditions. In contrast, satellite remote sensing technology can utilize captured imagery to compare background data with fire scenes to determine the presence of a wildfire. However, this monitoring technique also faces temporal and spatial limitations; it generally operates on longer cycles and cannot provide real-time monitoring, while the resolution of acquired images may be inadequate. In recent years, unmanned aerial vehicles (UAVs) have gained widespread application in wildfire detection due to their high flexibility, low cost, and ease of operation, demonstrating their promising performance in this field [8–10].

Traditional wildfire detection methods typically rely on image-derived color and texture features for fire recognition. Color feature-based methods have led to the development of models utilizing various color spaces, including RGB, HSI, and YCbCr. For instance, Celik and Demirel [11] created a classification model for flame pixels by leveraging the spectral characteristics of flames. Their research demonstrates that utilizing this model in the YCbCr color space significantly enhances the effectiveness of fire recognition. Similarly, Hamida et al. introduced a novel PJF color space, which enables the effective separation of flame and non-flame pixels, thereby improving the identification of flame pixels [12]. In the realm of texture feature-based methods, Dimitropoulos et al. [13] introduced a high-order linear dynamic system (h-LDS) descriptor for analyzing multidimensional dynamic texture features. They integrated this approach with particle swarm optimization techniques to merge multidimensional dynamic texture analysis with the spatiotemporal modeling of smoke, resulting in accurate flame recognition. Similarly, Prema et al. conducted flame recognition by leveraging edge and texture information from flames. Their methodology employed techniques such as color segmentation and wavelet analysis within the spatiotemporal domain to improve flame identification [14].

Recent advancements in deep learning-based computer vision techniques, such as image classification and object detection, have demonstrated applications in smart agriculture and forest [15–17]. These techniques also have emerged as promising solutions for the early detection of wildfires using imagery acquired from UAVs [18–21]. Among the various deep learning architectures, Convolutional Neural Networks (CNNs) are one of the most representative models due to their exceptional ability of nonlinear image representation, which has led to their remarkable performance in computer vision tasks. Building upon this foundation, some researchers have applied CNN models to the field of wildfire detection. For instance, Srinivas and Dua [20] utilized the foundational CNN model, AlexNet, to classify forest fire images, achieving a classification accuracy of 95%. Similarly, Lee et al. [22] implemented five different CNN frameworks to categorize UAV-acquired images into “fire” and “no fire” classes. While these studies effectively employed image classification methods, they were limited to image-level predictions of wildfires and lacked the ability to

accurately localize wildfire regions. To address the need for the precise localization of fire points, Barmpoutis et al. [23] adopted the classical two-stage object detection algorithm, Faster R-CNN, to detect flame targets within UAV images, achieving a detection accuracy of 70.6%. However, the complexity of two-stage algorithms like Faster R-CNN often leads to slower inference speeds, making them challenging to implement in real-time detection scenarios. In contrast, Goyal et al. [24] employed a one-stage object detection model, YOLO (You Only Look Once), as the primary framework for outdoor fire detection, successfully achieving high recognition accuracy while maintaining the real-time performance. Wang et al. [25] further enhanced the detection efficiency by utilizing a more lightweight version of the YOLO architecture, Light-YOLOv4, which significantly accelerated the inference speed. Considering that fire and smoke often coexist during wildfires, several studies have attempted to detect both simultaneously [26]. Sathishkumar et al. [27] investigated the transfer learning of pre-trained models for detecting forest fires and smoke, and Mamadaliev et al. [28] proposed a smoke and fire detection method based on the YOLOv8 model, incorporating several key architectural modifications.

Despite these advancements in wildfire detection, several challenges persist. Firstly, the current methods often utilize standard object detection algorithms for wildfire detection without accounting for the unique characteristics of wildfires, such as variations in shape due to different fire stages and the wind, as well as the relatively small size of wildfire incidents in their early stages. This oversight might impede the model's effectiveness in achieving optimal results. Secondly, smoke often takes on irregular shapes and has high transparency during wildfire events and may even appear before the flames. Therefore, detecting smoke is equally important. However, the aforementioned methods primarily focus on flame detection and do not integrate smoke detection into their objectives.

To address the challenges mentioned, this study proposes an integrated object detection framework based on YOLOv10 that simultaneously identifies both wildfires and smoke as detection targets. The proposed method first employs mixed data augmentation, including the Masaic augmentation, to enrich the dataset with small targets, thereby improving the detection performance for small smoke and fire instances in the wild. In addition, a backbone enhancement strategy is developed that incorporates the region convolution and feature refinement module, which improves the ability to localize features with high transparency within the complex background. Finally, by adopting shape aware IoU loss, the proposed method effectively captures irregularly shaped smoke and fire targets with complex edges, which enables the model to accurately identify and localize targets with non-standard shapes.

2. Materials and Methods

2.1. Dataset and Preprocessing

To validate the effectiveness of the proposed framework, a UAV-acquired dataset, namely SmokeFireUAV, was developed for the experimental analysis, which included bounding box labels for both fire and smoke targets. This dataset incorporates image samples from the publicly available UAV fire dataset, Flame [29], comprising fire images gathered by UAVs during prescribed burn pile treatments conducted in pine forests located in Arizona, USA. Given that the Flame dataset originally only contained bounding box labels for fire, we manually annotated the smoke targets using LabelImg software (<https://github.com/HumanSignal/labelImg>, accessed on 12 January 2025). To enhance diversity, the dataset further included UAV forest fire images captured by the authors themselves, as well as those acquired from online sources. Consequently, the final dataset consisted of 1489 UAV images for experimentation, with all samples uniformly resized to 1280×1280 pixels.

In our experiments, the labeled dataset was divided into training, validation, and testing sets at a ratio of 7:2:1. The samples from the dataset are illustrated in Figure 1.



Figure 1. Samples of SmokeFireUAV dataset.

As illustrated in the figure, smoke and fire targets display diverse sizes and sparse spatial distributions, presenting additional challenges for the detection network compared to natural images. Moreover, the presence of small fire spots within the smoke further complicates the task of distinguishing between smoke and fire.

2.2. Methods

YOLOv10, the latest iteration in the YOLO series, is a groundbreaking end-to-end object detection model that excels in various challenging scenarios, demonstrating exceptional efficiency, accuracy, and robustness. Therefore, we have chosen to adopt YOLOv10 [30] as the baseline framework for our research. YOLOv10 consists of three key components: the Backbone, Neck, and Detection Head. The Backbone is responsible for feature extraction using Convolutional (Conv) layers, Cross to Future (C2f), the C2f based on a Compact Inverted Block (C2fCIB), spatial-channel decouple down-sampling (SCDown) and Spatial Pyramid Pooling Fusion (SPPF) modules. The Conv layer extracts feature information from images through convolution operations, and SCDown is a down-sampling operation, while the C2f module merges feature maps of different scales and the SPPF module captures features at varying spatial scales. The Neck component integrates feature maps of different scales from the Backbone to fuse shallow and deep features, which thereby enhances the model's representational capacity. By employing a decoupled Head, YOLOv10 separates classification, bounding box regression, and confidence prediction tasks for independent processing before combining their outputs. Moreover, the implementation of a dual-label assignment strategy reduces the dependency on non-maximum suppression, leading to decreased inference latency and an improved performance.

In the proposed method, Mosaic data augmentation was implemented to enhance the diversity of small targets within the training data, thereby mitigating the issue of misidentifying small smoke and fire targets. By considering the distinct properties present in the smoke and fire features within UAV images, region attention convolutions were integrated into the YOLOv10 Backbone to improve the model's ability to capture contextual information, thus enhancing the localization of smoke and fire targets with varying shapes. The feature refinement module was positioned at the end of the Backbone, where an attention mechanism was combined with 2D convolutions to boost feature extraction capabilities in complex backgrounds while reducing the computational complexity. In addition, the adoption of Shape IoU aimed at addressing the detection challenges arising

from high transparency and irregular edges. The overall structure of the proposed method is shown in Figure 2.

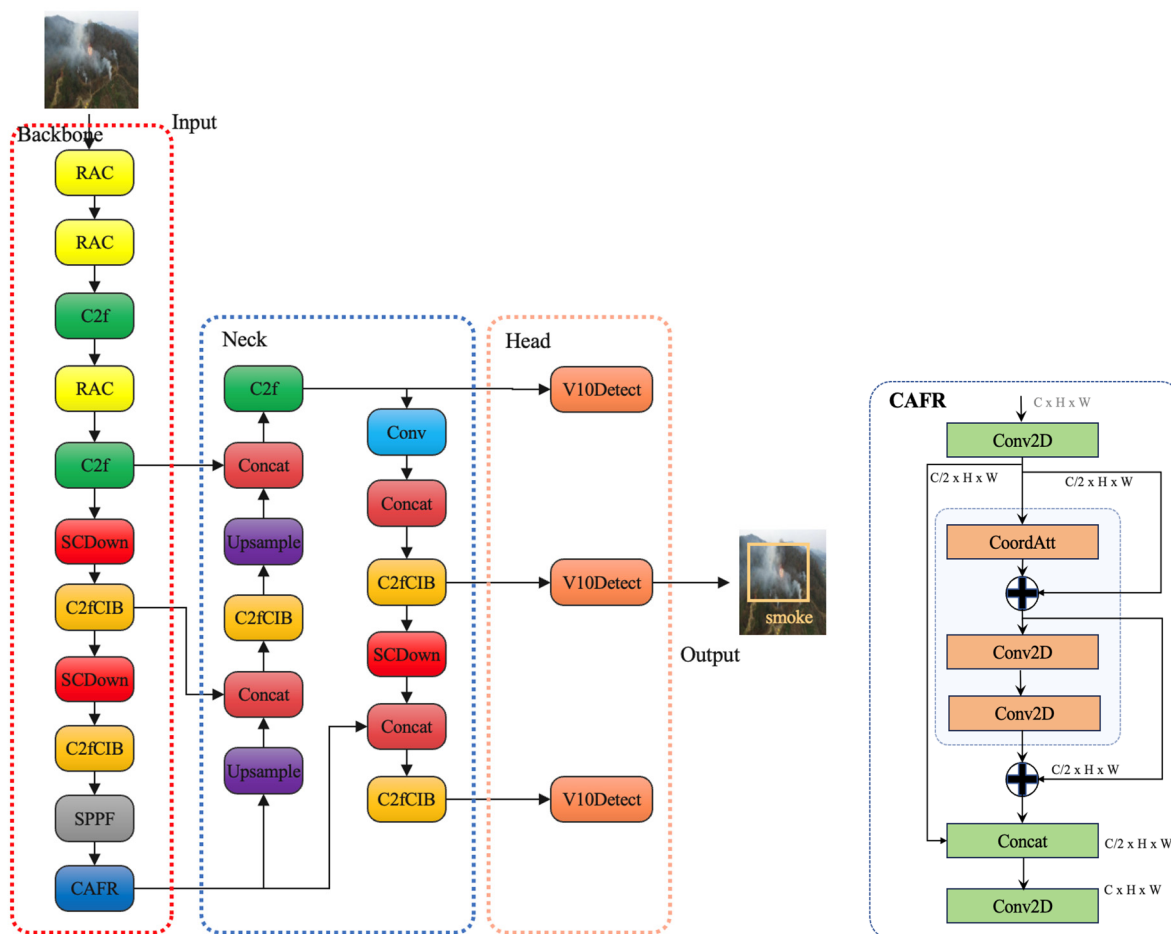


Figure 2. Overall structure of the proposed method for the detection smoke and fire.

2.2.1. Mixed Data Augmentation

To enhance the diversity of the dataset and improve the model's performance, various data augmentation techniques were employed. These included HSV channel color conversion, image horizontal flip, vertical flip, and contrast adjustment methods. Additionally, to tackle the challenges presented by small-sized smoke and fire instances in the UAV remote sensing images, the Mosaic data augmentation technique was utilized. This augmentation process involved selecting a batch of data samples from the overall dataset, then randomly choosing nine images depicting smoke and fire for random scaling and cropping. The corresponding segments were subsequently cross-stitched together. This procedure was repeated for each batch size, and the finalized stitched image was fed into the proposed object detection network model for training. The target boxes associated with each original image were constrained by cross-cropping and did not exceed the cropping boundaries of the original image. The specific workflow of the Mosaic data augmentation algorithm is illustrated in Figure 3. By introducing many small-sized fire and smoke targets, the Mosaic data augmentation enriches the detection dataset, which thereby enhances the network's robustness.



Figure 3. Samples using Mosaic data augmentation.

2.2.2. Vision Feature Enhancement

While the original YOLOv10 object detection framework utilizes the C2f design to strengthen the feature representation capability for the target, it might be insufficient for smoke and fire detection. This is because smoke and wildfire targets in UAV remote sensing images have unique characteristics, such as low contrast, dynamic shapes, and sometimes a translucent appearance, which significantly set them apart from other targets in remote sensing imagery. To facilitate the localization of smoke and wildfire targets, this framework aims to improve the feature extraction ability of YOLOv10 Backbone. First, the targets of smoke and fire in UAV remote sensing often exhibit two distinct properties: (1) a varying shape and (2) a scattered spatial distribution. The normal convolution operation in Backbone networks uses fixed kernels to extract information through the same parameters, making it difficult to accurately capture the specific smoke and fire features in remote sensing images. Thus, we proposed a region attention-based convolution (RAC), which combines spatial attention and receptive field features. Unlike traditional convolution, RAC focuses on capturing the local receptive field characteristics of input feature maps. It uses an attention mechanism to accurately evaluate the importance of each feature point within the receptive field, reducing the information sparsity caused by parameter sharing [31]. This operation helps ensure that co-related feature points in local regions are adequately represented, thereby enhancing the network's ability to capture important information, thus enhancing the ability to depict the smoke and fire feature regardless of the shape and distribution. The RAC mainly consists of two stages (as shown in Figure 4). Initially, a $K \times K$ group convolution is applied to the input feature map $F \in R^{C \times H \times W}$ to extract the feature representation of each feature point within the receptive field to obtain the feature of the dimension $CK^2 \times H \times W$, where H , W , and C denote the height, width, and channel numbers, respectively. According to ref. [31], K was set to 3 to create the 3×3 convolution kernel for local spatial representation. The process was followed by batch normalization and the activation function. An adjust shape operation was then performed [31], resulting in the receptive-field spatial features with the dimensions of $C \times KH \times KW$. Next, the spatial features extracted from the receptive field were utilized for the spatial attention computation. This involves employing max pooling and average pooling, followed by a convolution operation and a sigmoid activation function. At the same time, the Squeeze-and-Excitation [32] method was employed as a channel attention mechanism with the global average pooling, fully connected layer, and sigmoid activation functions. The spatial and channel attention maps were finally combined to allocate the receptive-field spatial

features as weight maps, where the feature maps were multiplied by the spatial attention map and then combined along the channel dimension using the channel attention map. The overall structure of the RAC is shown in Figure 4.

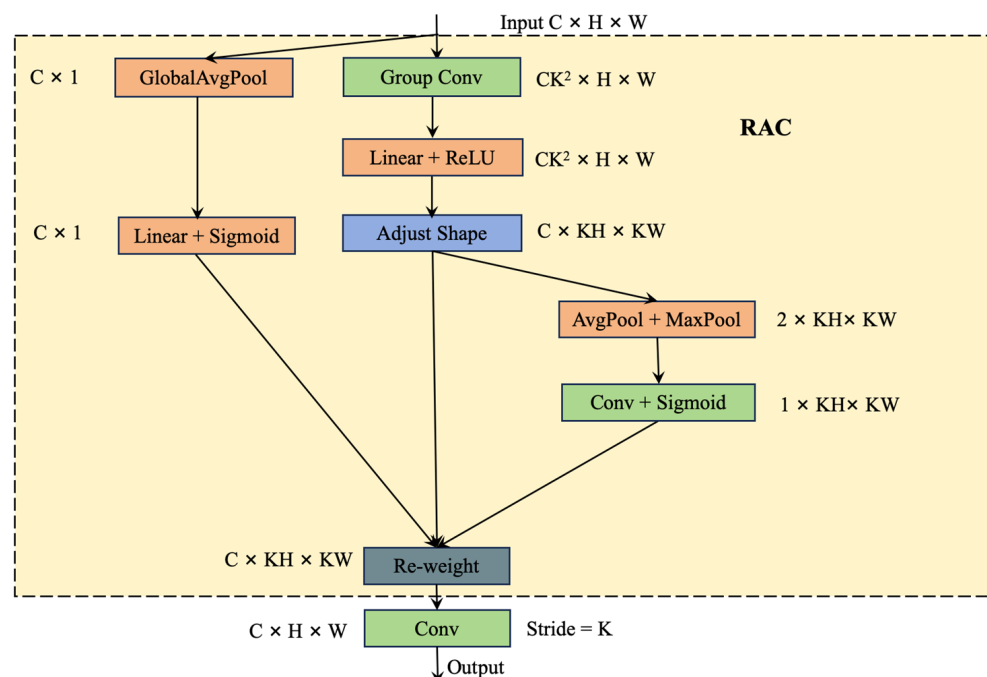


Figure 4. The structure of the RAC.

In addition, YOLOv10's Backbone incorporates a PSA module to enhance feature representation. However, the use of a self-attention strategy in PSA increases the computational complexity. To address this issue, we introduced a Coordinate Attention-based Feature Refinement (CAFR) module that enhances efficiency while preserving feature extraction capabilities. The CAFR module replaces multi-head self-attention with the CoordAtt [33] attention mechanism, which is adept at capturing smoke and fire targets of varying shapes and distributions (shown in Figure 5). CoordAtt is a simple and efficient attention mechanism that embeds positional information into channel attention, allowing for the acquisition of information over a larger range while avoiding significant computational overhead. To preserve positional information without resorting to 2D global pooling, this CoordAtt proposes the decomposition of channel attention into two parallel 1D feature encodings. These encodings aggregate input features along the vertical and horizontal directions, creating two separate directional-aware feature maps to efficiently integrate spatial coordinate information. Then, these two feature maps, each containing embedded direction-specific information, are further encoded into two attention maps. Each attention map captures long-range dependencies along a spatial direction in the input feature map, which enables the effective representation of the context of smoke and fire with varying shapes. By applying the element-wise multiplication of these two attention maps to the input feature map, the mechanism finally emphasizes the representation of the regions of interest. This approach effectively highlights the importance of spatial information.

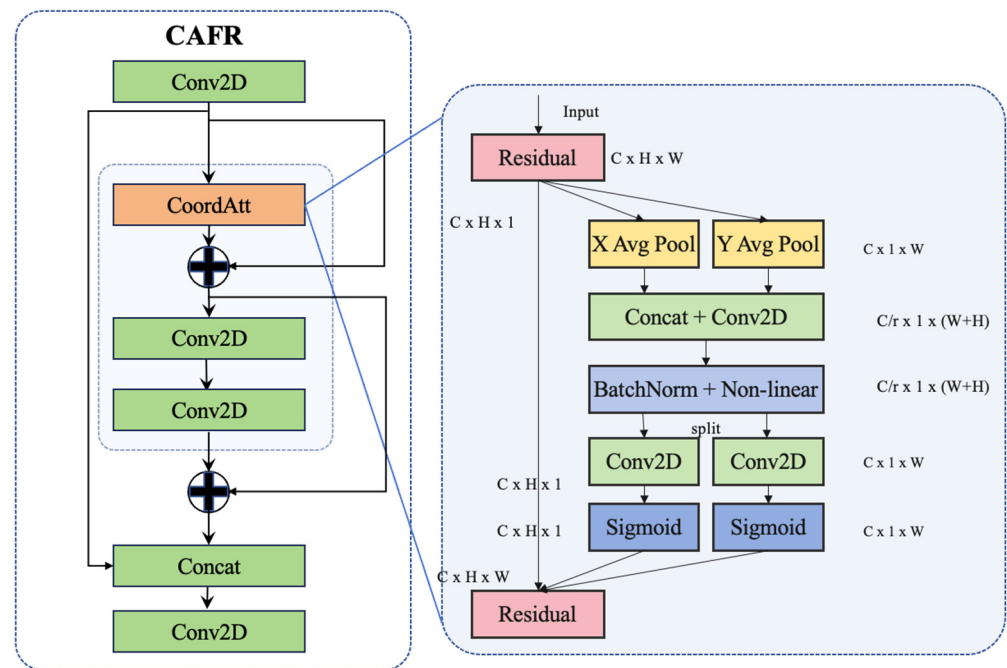


Figure 5. The structure of the CAFR.

Specifically, given an input, X , two spatial pooling kernels of size $(H, 1)$ or $(1, W)$ are used to encode each channel along the horizontal or vertical coordinates. Therefore, the output for channel, c , at height h can be formalized as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i). \quad (1)$$

Similarly, the output of the c -th channel with a width of w can be expressed as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (2)$$

The transformations described above aggregate features along two spatial directions, which produces a pair of directional-aware feature maps. Given the aggregated feature maps generated by Equations (1) and (2), they are first concatenated and then passed through a shared 1×1 convolutional transformation function, F_1 .

$$f = \delta \left(F_1 \left(\left[z^h, z^w \right] \right) \right) \quad (3)$$

where $\left[z^h, z^w \right]$ denotes the concatenation operation along the spatial dimensions, and δ is a nonlinear activation function. The intermediate feature map $f \in R^{C/r \times 1 \times (H+W)}$ encodes spatial information in both horizontal and vertical directions. Subsequently, f is split into two independent tensors along the spatial dimension: $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times w}$. Two 1×1 convolution operations, F_h and F_w , are then applied separately to convert f^h and f^w into tensors that possess the same number of channels as the input, X , and obtain.

$$g^h = \sigma \left(F_h \left(f^h \right) \right) \quad (4)$$

$$g^w = \sigma \left(F_w \left(f^w \right) \right) \quad (5)$$

The outputs g^h and g^w are unfolded and used as attention weights. Finally, the output, Y , of the coordinate attention mechanism can be expressed as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (6)$$

Enhanced features by using attention are fed into two successive convolution blocks to obtain the final refined features.

2.2.3. Loss Function

For the task of object detection, bounding box regression loss plays a crucial role. While Intersection over Union (IoU) is the most commonly used loss function in the field of object detection, it fails to accurately describe the positional relationship between the predicted box and the ground truth (GT) box when there is no overlap between them (i.e., $\text{IoU} = 0$). In the original YOLOv10, although the CIoU loss function minimizes the normalized distance between the center points of the predicted box and the GT box by adding a shape loss term, it still overlooks the impact of the shape and scale of the bounding box itself on regression, which fails to localize the targets of smoke and fire with an irregular shape and size. To improve the ability of localizing the smoke and fire targets with varying sizes, this proposed method adopts the Shape IoU as the loss function. This method focuses on calculating the loss based on the shape and scale of the bounding box itself, thereby achieving a more accurate bounding box regression [34]. The concept of Shape-IoU is visually depicted in Figure 6. The yellow area represents the GT box, with b^{gt} as the center of the GT box located at coordinates (x^{gt}, y^{gt}) , and h^{gt} and w^{gt} representing the height and width of the GT box, respectively. The blue area corresponds to the predicted box, with b as the center of the predicted box positioned at coordinates (x, y) , and h and w denoting the height and width of the predicted box, respectively.

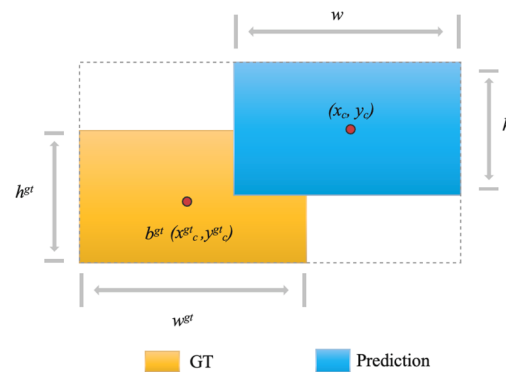


Figure 6. Shape-IoU between the GT and prediction boxes.

The formula for Shape-IoU is as follows:

$$\text{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (7)$$

$$ww = \frac{2 \times (w^{gt})^{\text{scale}}}{(w^{gt})^{\text{scale}} + (h^{gt})^{\text{scale}}} \quad (8)$$

$$hh = \frac{2 \times (h^{gt})^{\text{scale}}}{(w^{gt})^{\text{scale}} + (h^{gt})^{\text{scale}}} \quad (9)$$

$$\text{distance}^{\text{shape}} = hh \times (x_c - x_c^{gt})^2 / c^2 + ww \times (y_c - y_c^{gt})^2 / c^2 \quad (10)$$

$$\Omega^{\text{shape}} = \sum_{t=w,h} (1 - e^{-w_t})^{\theta}, \theta = 4. \quad (11)$$

$$\begin{cases} w_w = hh \times \frac{|w - w^{st}|}{\max(w, w^{st})} \\ w_h = ww \times \frac{|h - h^{st}|}{\max(h, h^{st})} \end{cases}, \quad (12)$$

where scale is the scaling factor that is related to the sizes of the targets in the dataset; ww and hh are weight coefficients for the horizontal and vertical directions, respectively, with their values dependent on the shape of the GT box. By combining Equations (7), (10), and (11), the bounding box regression loss for YOLOv10 can be obtained as follows:

$$L_{\text{Shape-IoU}} = 1 - \text{IOU} + \text{distance}^{\text{shape}} + 0.5 \times \Omega^{\text{shape}} \quad (13)$$

where $\text{distance}^{\text{shape}}$ denotes the shape distance, which measures the shape difference between the predicted box and the GT box, and $0.5 \times \Omega^{\text{shape}}$ is a penalty term utilized to correct the shape differences between the predicted box and the GT box.

3. Experimental Results and Analysis

3.1. Experimental Setting

The training in this study was conducted on an Ubuntu 18.04 system using an Intel(R) Xeon® Gold 5118 CPU @ 2.30GHz and an NVIDIA GeForce RTX 1080Ti GPU. The training framework employed was PyTorch 1.10.2, and parallel computing utilized CUDA version 11.6. During the experiments, several parameters were set. The image input size was set to 1280×1280 , the batch size was 2, and a total of 100 experimental rounds were performed. The initial learning rate was set to 0.01, momentum was set to 0.93, and the learning rate decay followed the cosine function. The optimization algorithm used was SGD.

3.2. Evaluation Metric

To evaluate the performance of the proposed smoke and fire detection method, various metrics were utilized. Precision measures the proportion of correctly predicted positive samples among all identified samples, while the Recall rate represents the proportion of correctly identified positive samples out of all actual positive samples. The F1 score combines both Precision and Recall into a single metric. The mean Average Precision (mAP) calculates the average Precision across different categories. These metrics are computed as:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (14)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

$$F1 = 2 * \frac{P * R}{P + R} \quad (16)$$

$$AP = \int_0^1 P(R) dR \quad (17)$$

$$mAP = \frac{\sum A_P}{N}, \quad (18)$$

where TP represents the number of positive samples correctly identified, FP represents the number of positive samples incorrectly recognized, and FN represents the number of positive samples not recognized. AP represents the average Precision for each category, while N represents the total number of categories. In addition, to evaluate the inference speed, frames per second (FPS) is utilized as a metric. FPS is used to indicate the number of

individual frames that the model can process in one second, with higher values indicating a faster processing speed.

3.3. Comparative Experiments and Analysis

To assess the effectiveness of the proposed model in detecting forest fires and smoke, this study performed a comparative analysis using six different one-stage real-time detection networks: YOLOv4 [35], YOLOv5, YOLOX [36], YOLOv7 [37], and YOLOv10 [30]. These models were tested under the same environmental configuration, ensuring fair experimentation. Consistent parameters were selected for all six networks to ensure fairness in the evaluation. The comparative results are shown in Table 1. From the table, it is evident that the proposed method significantly outperforms the other methods for accuracy metrics, such as Recall, F1 score, and mAP. This validates that the proposed method is advantageous for detecting targets with varying sizes and distributions in the UAV images. The mAP metric of the proposed method surpasses that of the original YOLOv10 and YOLOX by 4.27% and 7.38%, while the F1 score of the proposed method beats that of the original YOLOv10 and YOLOX by 1.96% and 10.44%, respectively. In terms of the detection efficiency, the proposed method performs faster than YOLOX by a large margin, and slightly faster than the original YOLO10. Overall, for the detection of fire and smoke in UAV images, the proposed method outperforms other one-stage object detection methods in terms of both detection accuracy and speed.

Table 1. Comparison of detection results using different models.

Model	Input Size	Precision (%)	Recall (%)	mAP (%)	F1 (%)	FPS
YOLOv4	1280 × 1280	64.22	71.71	71.33	67.72	11.8
YOLOv5	1280 × 1280	69.31	68.97	74.52	69.12	10.6
YOLOv7	1280 × 1280	85.14	63.41	69.23	72.77	16.9
YOLOX	1280 × 1280	60.4	71.9	71.9	65.7	20.1
YOLOv10	1280 × 1280	80.59	68.75	75.01	74.20	8.82
Proposed	1280 × 1280	76.73	75.56	79.28	76.14	8.98

To further demonstrate the effectiveness of the proposed method compared to the original YOLOv10, the qualitative detection results are shown in Figure 7. The first and third columns display images with ground truth bounding boxes, where the first column contains large smoke targets, while the third column shows some smoke and fire targets of very small sizes. The second and fourth columns display the detection results using the proposed method. It is evident from the images that the proposed method is capable of capturing smoke and fire targets of different sizes. Despite the smoke targets often having high transparency and irregular shapes (edges), the proposed method performs well in detecting these targets.

3.4. Ablation Study

To validate the effectiveness of each improvement strategy proposed in this paper, a series of ablation experiments were conducted based on the YOLOv10 model by individually adding the improved components. The experimental results are shown in Table 2. From the table, it is evident that the introduction of mosaic data augmentation results in an increase of 2.59% and 1.02% in Precision and F1 score metrics, respectively, compared to the original YOLOv10. By using the Shape IoU loss function to train the detection model, all accuracy metrics improved, indicating that the Shape IoU loss is particularly suitable for our task. This is because both smoke and fire targets have irregular shapes and edges, and the Shape IoU is sensitive to capturing such information, thereby enhancing the detection perfor-

mance. The inference speed also experienced a slight increase, with the FPS rising from 8.82 to 8.92. In addition, the refinement of the CAFR module by substituting the multi-head self-attention block with the CoordAtt convolution block led to a notable enhancement in the model's inference speed, increasing it from 8.82 to 11.76. This improvement in speed was achieved while maintaining a slight increase in the detection accuracy, with the mAP improving from 75.01% to 76.23%. Also, the introduced RAC module has shown promising results in enhancing the detection accuracy, as indicated by the mAP metric. The model's mAP score has improved significantly from 75.01% to 75.61% with the integration of the RAC module. Finally, the method proposed, which integrates various strategies and modules, has resulted in a significant improvement in the detection performance across multiple metrics. Specifically, there have been notable increases in the Recall, mAP, and F1 score, showcasing the effectiveness of the proposed method compared with the original YOLOv10. This demonstration highlights that the proposed method is well-suited for detecting wildfires and smoke targets in UAV remote sensing applications. More importantly, while achieving a promising detection performance, the proposed method operates slightly faster than the original YOLOv10, with a slight increase in the FPS from 8.82 to 8.98. This indicates its practical applicability for real-time wildfire monitoring.

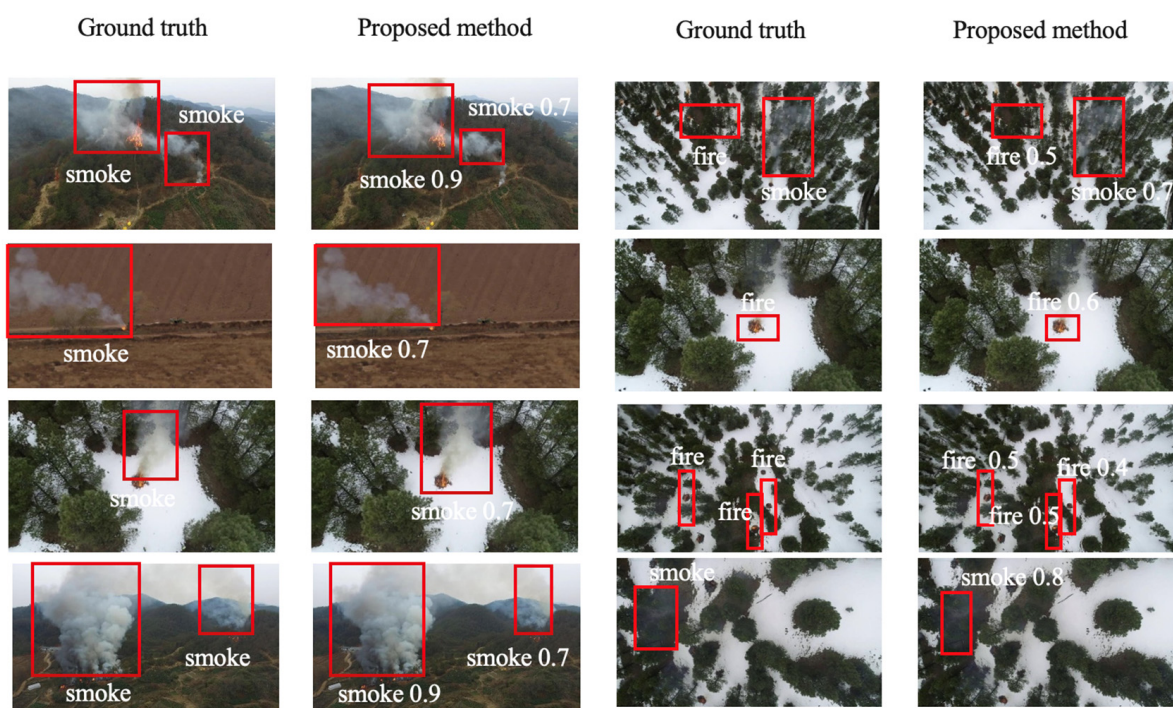


Figure 7. Detection results using the proposed method.

Table 2. Detection results by integrating different modules with YOLOv10.

Model	Input Size	Precision (%)	Recall (%)	mAP (%)	F1 (%)	FPS
YOLOv10	1280 × 1280	80.59	68.75	75.01	74.20	8.82
+Mosaic9	1280 × 1280	83.18	65.09	76.03	73.03	8.72
+Shape IoU	1280 × 1280	81.5	70.01	76.90	75.31	8.92
+RAC	1280 × 1280	75.73	69.22	75.61	72.33	7.14
+CAFR	1280 × 1280	83.18	65.09	76.23	73.03	11.76
Proposed	1280 × 1280	76.73	75.56	79.28	76.14	8.98

To better illustrate the advantages of the proposed method compared to the original YOLOv10, the detection results for both the original YOLOv10 and the proposed method are presented in Figure 8. The top row shows the detection results of the original YOLOv10, while the bottom row displays the prediction outcomes of the proposed method. It can be observed from the figures (two subfigures in the first column) that the proposed method achieves a high confidence score in detecting smoke targets. This indicates that the designed feature enhancement module effectively captures smoke targets with high transparency and indistinct edges. As depicted in the two subfigures of the second column, the proposed method correctly identifies the smoke target, whereas the original YOLOv10 mistakenly detects smoke as fire. This may be attributed to the design of the RAC, which enhances feature representation in the proposed method, thereby improving its ability to distinguish between smoke and fire. Similarly, the proposed method successfully detects the fire target within the smoke area (as seen in the subfigures of the third column), indicating that the Shape IoU loss is effective at capturing fires with irregular edges. Furthermore, the proposed model excels in detecting small fires compared to the original YOLOv10, as seen from the subfigures in the fourth column.

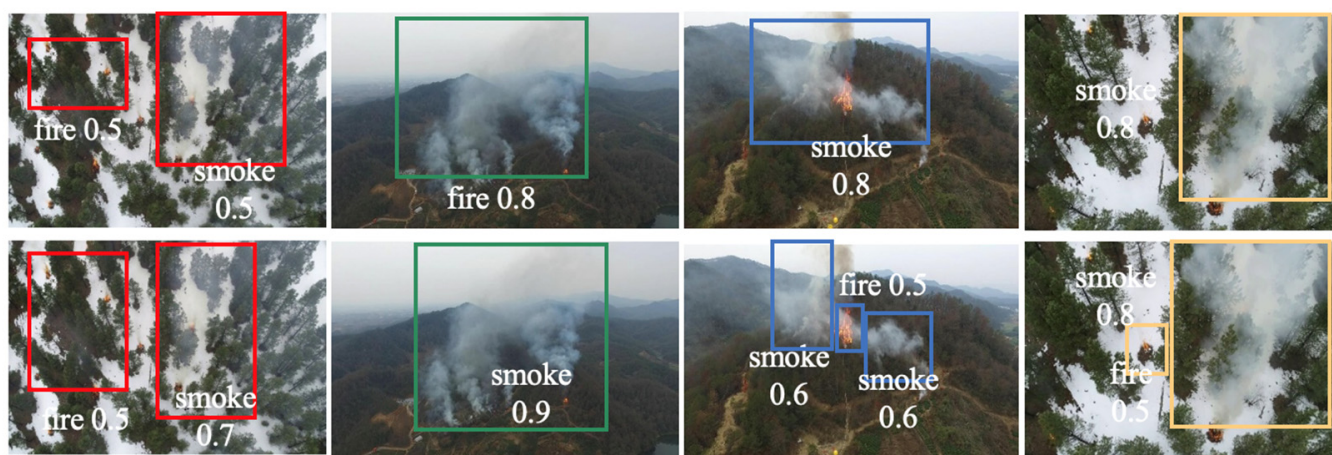


Figure 8. Detection results using the original YOLOv10 and proposed method, with the top row showing the detection results from the original YOLOv10 and the bottom row displaying the results from the proposed method.

To further demonstrate the effectiveness of the proposed model in accurately localizing salient targets, such as smoke and fire, GradCAM is utilized to generate the heatmap, as depicted in Figure 9. The first row displays the original sample, while the second to fourth rows present the heatmaps for the output feature maps from the three detection heads. The heatmaps on the left correspond to the original YOLOv10 method, whereas those on the right depict the results obtained using the proposed method. From the figure, it can be observed that the output features generated using the proposed method are more sensitive in detecting the regions of smoke and fire, where the targets can be captured accurately. In addition, the different heads demonstrate the ability to capture targets of varying sizes.

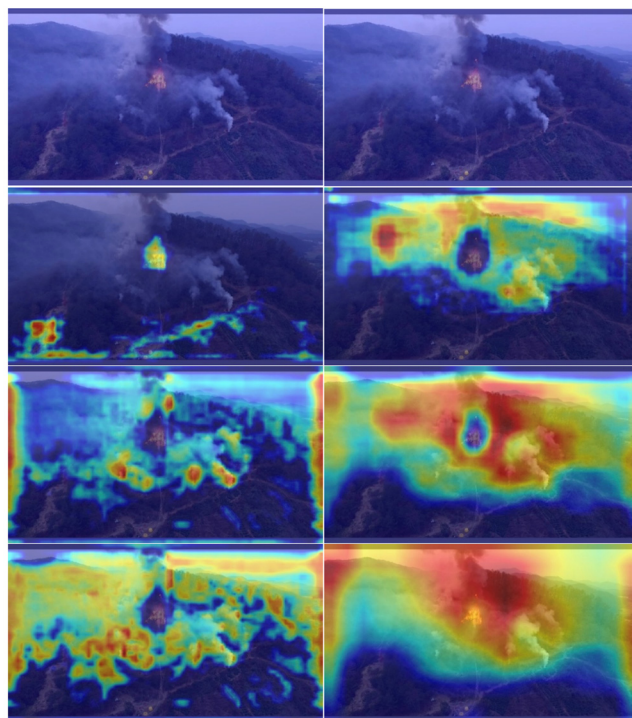


Figure 9. Heatmaps generated from the three detection heads using the original YOLOv10 and proposed methods, with the left column showing the heatmaps from the original YOLOv10 and the right column displaying the heatmaps from the proposed method.

4. Conclusions

Building upon the advantages of real-time object detection provided by the YOLO-based algorithm, this study is based on the recently proposed YOLOv10 method. It presents an integrated real-time object detection network for the simultaneous identification of wildfires and smoke in UAV remote sensing images. Considering the specific characteristics of smoke and fire targets in UAV imagery, enhancements have been made to the YOLOv10 network to enhance its detection accuracy. The utilization of a mosaic augmentation strategy has increased the representation of small smoke and fire targets within the dataset, thereby improving the model's ability to detect smaller targets effectively. Moreover, the introduction of a region attention-based convolution module and coordinate attention feature refinement module aims to boost feature extraction related to smoke and fire features. In addition, the integration of the Shape IoU loss function has enhanced the model's capacity for accurate data fitting, particularly addressing the challenges associated with identifying targets such as smoke and fire with unclear boundaries and edges. We believe that the proposed method could be an effective tool for the real-time monitoring of forest fires and smoke, and it can be deployed on edge devices or monitoring equipment.

Author Contributions: Conceptualization, X.F. and F.L.; methodology, X.F.; software, K.Y.; writing—original draft preparation, X.F.; writing—review and editing, F.L.; visualization, K.Y.; project administration, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Open Research Project of Key Laboratory of Natural Resources Monitoring and Supervision in Southern Hilly Region, grant number MRMSSHR202307.

Data Availability Statement: Some of the dataset samples can be accessed from the publicly available Flame dataset (<https://paperswithcode.com/dataset/flame>, accessed on 12 January 2025).

Acknowledgments: We would like to express our gratitude to the editors and reviewers for their dedicated efforts and valuable feedback.

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

References

1. Lambrou, N.; Crystal, K.; Anastasia, L.S.; Erica, A.; Charisma, A. Social drivers of vulnerability to wildfire disasters: A review of the literature. *Landsc. Urban Plan.* **2023**, *237*, 104797. [\[CrossRef\]](#)
2. Yang, W.; Jiang, X.L. Review on Remote Sensing Information Extraction and Application of the Burned Forest Areas. *Sci. Silvae Sin.* **2018**, *54*, 135–142.
3. Sousa, M.J.; Alexandra, M.; Miguel, A. Wildfire detection using transfer learning on augmented datasets. *Expert Syst. Appl.* **2020**, *142*, 142112975. [\[CrossRef\]](#)
4. Rashkovetsky, D.; Florian, M.; Martin, L.; Michael, S. Wildfire detection from multisensor satellite imagery using deep semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7001–7016. [\[CrossRef\]](#)
5. Barmpoutis, P.; Periklis, P.; Kosmas, D.; Nikos, G. A review on early forest fire detection systems using optical remote sensing. *Sensors* **2020**, *20*, 6442. [\[CrossRef\]](#)
6. Spadoni, G.L.; Moris, J.V.; Vacchiano, G.; Elia, M.; Garbarino, M.; Sibona, E.; Tomao, A.; Barbati, A.; Sallustio, L.; Salvati, L.; et al. Active governance of agro-pastoral, forest and protected areas mitigates wildfire impacts in Italy. *Sci. Total Environ.* **2023**, *890*, 164281. [\[CrossRef\]](#)
7. Maestas, J.D.; Joseph, T.S.; Brady, W.A.; David, E.N.; Matthew, O.J.; Casey, O.C.; Chad, S.B.; Kirk, W.D.; Michele, R.C.; Andrew, C.O. Using dynamic, fuels-based fire probability maps to reduce large wildfires in the Great Basin. *Rangel. Ecol. Manag.* **2022**, *89*, 33–41. [\[CrossRef\]](#)
8. Mohapatra, A.; Timothy, T. Early Wildfire Detection Technologies in Practice—A Review. *Sustainability* **2022**, *14*, 12270. [\[CrossRef\]](#)
9. Kasyap, V.L.; Sumathi, D.; Alluri, K.; Reddy Ch, P.; Thilakarathne, N.; Shafi, R.M. Early Detection of Forest Fire Using Mixed Learning Techniques and UAV. *Comput. Intell. Neurosci.* **2022**, *1*, 3170244. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Yang, X.; Chen, R.; Zhang, F.; Zhang, L.; Fan, X.; Ye, Q.; Fu, L. Pixel-level automatic annotation for forest fire image. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104353. [\[CrossRef\]](#)
11. Celik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [\[CrossRef\]](#)
12. Amal, B.H.; Chokri, B.A.; Yasser, A. A New Color Model for Fire Pixels Detection in PJF Color Space. *Intell. Autom. Soft Comput.* **2022**, *33*, 1607–1621.
13. Dimitropoulos, K.; Barmpoutis, P.; Grammalidis, N. Higher order linear dynamical systems for smoke detection in video surveillance applications. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 1143–1154. [\[CrossRef\]](#)
14. Prema, C.E.; Vinsley, S.S.; Suresh, S. Efficient flame detection based on static and dynamic texture analysis in forest fire detection. *Fire Technol.* **2018**, *54*, 255–288. [\[CrossRef\]](#)
15. Lu, X.; Wang, R.; Zhang, H.; Zhou, J.; Yun, T. PosE-Enhanced Point Transformer with Local Surface Features (LSF) for Wood–Leaf Separation. *Forests* **2024**, *15*, 2244. [\[CrossRef\]](#)
16. Wang, Q.; Fan, X.; Zhuang, Z.; Tjahjadi, T.; Jin, S.; Huan, H.; Ye, Q. One to All: Toward a Unified Model for Counting Cereal Crop Heads Based on Few-Shot Learning. *Plant Phenomics* **2024**, *6*, 0271. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wu, X.; Fan, X.; Luo, P.; Choudhury, S.D.; Tjahjadi, T.; Hu, C. From laboratory to field: Unsupervised domain adaptation for plant disease recognition in the wild. *Plant Phenomics* **2023**, *5*, 0038. [\[CrossRef\]](#)
18. Wang, J.; Fan, X.; Yang, X.; Tjahjadi, T.; Wang, Y. Semi-Supervised Learning for Forest Fire Segmentation Using UAV Imagery. *Forests* **2022**, *13*, 1573. [\[CrossRef\]](#)
19. Shamta, I.; Demir, B.E. Development of a Deep Learning-Based Surveillance System for Forest Fire Detection and Monitoring Using UAV. *PLoS ONE* **2024**, *19*, e0299058. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Srinivas, K.; Mohit, D. Fog computing and deep CNN based efficient approach to early forest fire detection with unmanned aerial vehicles. In *Inventive Computation Technologies 4*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 646–652.
21. Govil, K.; Morgan, L.W.; J-Timothy, B.; Carlton, R.P. Preliminary results from a wildfire detection system using deep learning on remote camera images. *Remote Sens.* **2020**, *12*, 166. [\[CrossRef\]](#)
22. Lee, W.; Kim, S.; Lee, Y.T.; Lee, H.W.; Choi, M. Deep Neural Networks for Wildfire Detection with Unmanned Aerial Vehicle. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics, Las Vegas, NV, USA, 8–10 January 2017. [\[CrossRef\]](#)
23. Barmpoutis, P.; Tania, S.; Kosmas, D.; Nikos, G. Early fire detection based on aerial 360-degree sensors, deep convolution neural networks and exploitation of fire dynamic textures. *Remote Sens.* **2020**, *12*, 3177. [\[CrossRef\]](#)
24. Goyal, S.; Shagill, M.; Kaur, A.; Vohra, H.; Singh, A. A yolo based technique for early forest fire detection. *Int. J. Innov. Technol. Explor. Eng.* **2020**, *9*, 1357–1362. [\[CrossRef\]](#)
25. Wang, Y.F.; Hua, C.C.; Ding, W.L.; Wu, R.N. Real-time detection of flame and smoke using an improved YOLOv4 network. *Signal Image Video Process.* **2022**, *16*, 1109–1116. [\[CrossRef\]](#)

26. Gaur, A.; Singh, A.; Kumar, A.; Kumar, A.; Kapoor, K. Video flame and smoke based fire detection algorithms: A literature review. *Fire Technol.* **2020**, *56*, 1943–1980. [[CrossRef](#)]
27. Sathishkumar, V.E.; Cho, J.; Subramanian, M.; Naren, O.S. Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecol.* **2023**, *19*, 9. [[CrossRef](#)]
28. Mamadaliev, D.; Touko, P.L.M.; Kim, J.-H.; Kim, S.-C. ESFD-YOLOv8n: Early Smoke and Fire Detection Method Based on an Improved YOLOv8n Model. *Fire* **2024**, *7*, 303. [[CrossRef](#)]
29. Shamsoshoara, A.; Afghah, F.; Razi, A.; Zheng, L.; Fulé, P.Z.; Blasch, E. Aerial imagery pile burn detection using deep learning: The FLAME dataset. *Comput. Netw.* **2021**, *193*, 108001. [[CrossRef](#)]
30. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
31. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3560–3569.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
33. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
34. Zhang, H.; Zhang, S. Shape-iou: More accurate metric considering bounding box shape and scale. *arXiv* **2023**, arXiv:2312.17663.
35. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
36. Ge, Z. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
37. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.

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.