drones

MDPI

Article

Tiny-Object Detection Based on Optimized YOLO-CSQ for Accurate Drone Detection in Wildfire Scenarios

Tian Luan, Shixiong Zhou *, Lifeng Liu and Weijun Pan *

Faculty of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; luantian@cafuc.edu.cn (T.L.); lfliu@cafuc.edu.cn (L.L.)
* Correspondence: zsx@cafuc.edu.cn (S.Z.); panatc@cafuc.edu.cn (W.P.)

Abstract: Wildfires, which are distinguished by their destructive nature and challenging suppression, present a significant threat to ecological environments and socioeconomic systems. In order to address this issue, the development of efficient and accurate fire detection technologies for early warning and timely response is essential. This paper addresses the complexity of forest and mountain fire detection by proposing YOLO-CSQ, a drone-based fire detection method built upon an improved YOLOv8 algorithm. Firstly, we introduce the CBAM attention mechanism, which enhances the model's multi-scale fire feature extraction capabilities by adaptively adjusting weights in both the channel and spatial dimensions of feature maps, thereby improving detection accuracy. Secondly, we propose an improved ShuffleNetV2 backbone network structure, which significantly reduces the model's parameter count and computational complexity while maintaining feature extraction capabilities. This results in a more lightweight and efficient model. Thirdly, to address the challenges of varying fire scales and numerous weak emission targets in mountain fires, we propose a Quadrupled-ASFF detection head for weighted feature fusion. This enhances the model's robustness in detecting targets of different scales. Finally, we introduce the WIoU loss function to replace the traditional CIoU object detection loss function, thereby enhancing the model's localization accuracy. The experimental results demonstrate that the improved model achieves an mAP@50 of 96.87%, which is superior to the original YOLOV8, YOLOV9, and YOLOV10 by 10.9, 11.66, and 13.33 percentage points, respectively. Moreover, it exhibits significant advantages over other classic algorithms in key evaluation metrics such as precision, recall, and F1 score. These findings validate the effectiveness of the improved model in mountain fire detection scenarios, offering a novel solution for early warning and intelligent monitoring of mountain wildfires.

Keywords: wildfire; ShuffleNetv2; CBAM; quadrupled ASFF

Citation: Luan, T.; Zhou, S.; Liu, L.; Pan, W. Tiny-Object Detection Based on Optimized YOLO-CSQ for Accurate Drone Detection in Wildfire Scenarios. *Drones* **2024**, *8*, 454. https://doi.org/10.3390/drones8090454

Academic Editors: Brian K. Gullett, Johanna Aurell, Pantelis Velanas, Diego González-Aguilera and Katerina Margariti

Received: 11 July 2024
Revised: 23 August 2024
Accepted: 31 August 2024
Published: 2 September 2024

1. Introduction

Forests, as the "green heart" and "ecological barrier" of the Earth, play an irreplaceable role in climate regulation, water conservation, biodiversity preservation, and carbon sequestration. However, in recent years, the increasing frequency of wildfires has not only caused large-scale destruction of forest resources and loss of animal and plant habitats but also triggered various secondary disasters such as soil erosion and water loss, severely impacting local ecosystems and socioeconomic conditions. For instance, the Australian bushfires that began in late 2019 swept across approximately 20% of the country's land area, resulting in 33 fatalities, displacing tens of thousands of people, destroying over 3000 homes, and burning through 24 million hectares. More than one billion mammals, birds, and reptiles perished, with at least 34 species driven to extinction. This catastrophic event, which lasted for nearly 200 days, is conservatively estimated to have caused USD 5 billion in health and property damages. More recently, on 19 February 2024, a wildfire broke out in Huaga Village, Huaga Township, Shuicheng District, Liupanshui City, Guizhou Province, China, spreading to nearby villages in Pu'an County, Qianxinan Prefecture. Tragically, during the

firefighting efforts in Longyin Town, two young firefighters, both in their 20s, lost their lives.

2. Related Works

To mitigate the safety risks posed by mountain wildfires, researchers worldwide have been exploring detection and early warning methods using satellite remote sensing and unmanned aerial vehicle (UAV) aerial detection. Satellite remote sensing, with its advantages of wide coverage, periodicity, and multi-spectral capabilities, plays an indispensable role in mountain wildfire monitoring. Zhao et al. [1] proposed a new framework for near-real-time, early-stage mountain wildfire detection based on Himawari-8 satellite imagery, which outperformed the JAXA fire detection product by integrating spatiotemporal spectral information. Similarly, Zhang et al. [2] utilized Himawari-8 satellite data to construct a spatiotemporal spectral recursive neural network model, achieving accurate detection of small-scale, early-stage, daytime, and night-time mountain wildfires. In addition to geostationary satellites, polar-orbiting satellites such as MODIS and VIIRS have been widely applied in mountain wildfire monitoring. Ding et al. [3] developed an adaptive mountain wildfire detection algorithm called DBTDW based on MODIS data, which demonstrated high applicability under various spatiotemporal conditions. Ji et al. [4] coupled the Bidirectional Reflectance Distribution Function (BRDF) physical model with deep learning techniques to achieve near-real-time monitoring of mountain wildfires using geostationary satellite imagery. Although these methods have made progress in addressing wildfire early warning issues, they still face limitations in monitoring range, data processing delays, warning accuracy, high alert costs, and susceptibility to meteorological factors due to constraints in spatiotemporal resolution, weather conditions, and satellite transmission costs. These limitations hinder the early detection and timely response to wildfires. In contrast to satellite remote sensing, UAVs offer advantages such as flexibility, mobility, and the ability to acquire high-resolution images, leading to their increasing application in mountain wildfire detection. Mohapatra et al. [5] reviewed recent advances in UAV applications for mountain wildfire detection, focusing on monitoring systems based on sensor nodes, UAV aerial photography, and ground camera networks. Moghadasi et al. [6] proposed a method for continuous mountain wildfire detection and monitoring using rotary-wing UAV formations, optimizing UAV trajectory planning to achieve sustained observation of suspected fire areas and fire mapping. Qiao et al. [7] designed a UAV-based mountain wildfire detection system using visible light/infrared cameras, coupling algorithms for smoke and flame segmentation, camera pose estimation, and feature matching to achieve early detection and distance localization of mountain wildfires. Chuang et al. [8] proposed using UAV swarms carrying L-band SAR and optical sensors to obtain high-resolution real-scene images through tomographic imaging techniques, enabling early identification of mountain wildfire hazards by inverting changes in tree dielectric constants.

In recent years, the rapid development of artificial intelligence methods such as machine learning and deep learning, and their widespread application in forest fire detection, has significantly enhanced the capability to detect mountain wildfires. Machine learning methods primarily involve automatically mining multi-dimensional features of images, including spectral, textural, and spatiotemporal characteristics, to construct classification decision functions or rules for forest fires. Representative methods include decision trees [9], random forests [10], and support vector machines [11]. Research has shown that machine learning methods can significantly improve fire detection accuracy in complex mountainous terrain conditions. For example, Bar et al. [12] used Landsat-8 and Sentinel-2 medium-resolution optical satellite images from 2016 to 2019 to identify forest fire areas in the western Himalayan state of Uttarakhand, India, through the Google Earth Engine (GEE) platform. They applied unsupervised classification using the Weka clustering algorithm to identify the shape and pattern of fire areas, and employed supervised classification algorithms such as Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Machine (SVM). Results showed that CART and RF algorithms achieved

similarly high accuracy (97–100%) in identifying forest fire areas. Janiec et al. [13] employed two machine learning classification methods—Maximum Entropy (MaxENT) and Random Forest—to analyze satellite images and products of different spatial and spectral resolutions (Landsat TM, Modis TERRA, GMTED2010, and VIIRS), vector data (OSM), and bioclimatic variables (WORLDCLIM). They found that the Random Forest prediction model was more effective in improving accuracy and reducing risk areas, while the MaxENT method showed lower accuracy. Mohajane et al. [14] developed five new hybrid machine learning algorithms—Frequency Ratio-Multilayer Perceptron (FR-MLP), Frequency Ratio-Logistic Regression (FR-LR), Frequency Ratio-Classification and Regression Tree (FR-CART), Frequency Ratio-Support Vector Machine (FR-SVM), and Frequency Ratio-Random Forest (FR-RF)—for mapping forest fire susceptibility. The results demonstrated that these hybrid models significantly improved the accuracy and performance of forest fire susceptibility studies, with the FR-RF model performing best (AUC = 0.989).

Unlike machine learning methods that rely on manual feature extraction, deep learning methods primarily learn multi-scale, multi-level deep features directly from raw images, thereby enhancing model performance in complex scenarios for fire point identification. Currently, deep learning-based mountain wildfire detection methods can be broadly categorized into two main types. The first is based on convolutional neural networks (CNNs) for mountain wildfire detection. Deep CNNs, with their powerful feature extraction and semantic expression capabilities, have become a research hotspot in the field of mountain wildfire detection. Ahmad et al. [15] proposed the FireXNet model for wildfire detection, which adopts a lightweight structure similar to MobileNetV3 and introduces SHAP interpretability analysis, achieving performance superior to models such as VGG16 and DenseNet201 on resource-constrained devices. Wang et al. [16] proposed an efficient real-time forest fire detection model called FireDetn for complex scenarios, introducing multi-scale detection heads, transformer encoders, and multi-head attention mechanisms to enhance the ability to capture global feature information and contextual information, thereby improving average precision in complex scenarios. Johnston et al. [17] thoroughly investigated the performance of YOLOv5 for real-time mountain wildfire detection on embedded systems, particularly the Raspberry Pi 4. Through performance comparisons with YOLOv3 and YOLOv3-tiny, their results showed that the proposed system achieved high detection accuracy, low power consumption, and strong adaptability to real environments. Mukhiddinov et al. [18] proposed an improved YOLOv5-based UAV visual early mountain wildfire smoke detection system, enhancing network architecture and detection speed by adding a spatial pyramid pooling fast layer, applying a bidirectional feature pyramid network, and employing network pruning and transfer learning methods. Their experimental results demonstrated the effectiveness of the proposed method and its superiority over other single-stage and two-stage object detectors. Casas et al. [19] conducted a comprehensive comparison of YOLO series models in smoke and mountain fire detection, utilizing multiple performance metrics including recall, precision, F1 score, and mean average precision. Their findings indicate that YOLOv5, YOLOv7, and YOLOv8 demonstrate relatively balanced performance across all metrics, while YOLO-NAS variants excel in recall but underperform in precision. This underscores the importance of considering specific model performance in relation to practical application requirements when selecting an appropriate model. He et al. [20] proposed two improved mountain fire detection models based on YOLOv5, reducing model parameters by simplifying the original network structure's neck and head, and eliminating backbone modules. Experimental results demonstrate that these lightweight models maintain high accuracy and recall while adapting to embedded devices, enabling real-time fire monitoring. Li et al. [21] introduced LEF-YOLO, a lightweight mountain fire detection model. By incorporating MobileNetv3's bottleneck structure and depth-wise separable convolutions, they reduced model complexity. Multi-scale feature fusion strategies, coordinate attention, and spatial pyramid pooling-fast blocks were employed to enhance feature extraction and improve detection accuracy. The LEF-YOLO model exhibited superior detection performance on

an extreme forest fire dataset, achieving 2.7 GFLOPs, 61 FPS, and 87.9% mAP. Gonçalves et al. [22] compared the performance of multiple models, including YOLOv7 and YOLOv8, in wildfire smoke detection for both ground-level and aerial imagery. They also discussed the impact of complex scene factors on detection accuracy.

Semantic segmentation approaches for mountain fire detection aim to assign semantic labels to each pixel in an image, enabling precise delineation of fire-affected areas. Valero et al. [23] proposed an accurate wildfire area segmentation method based on UAV thermal infrared videos. Their approach enhances video registration accuracy through trajectory stabilization, foreground histogram equalization, and multi-reference frame strategies. The KAZE feature matching algorithm is employed to achieve stable and accurate frame-by-frame segmentation of wildfire videos, supporting fire behavior analysis. Bouguettaya et al. [24] reviewed recent deep learning algorithms applied to UAV wildfire smoke segmentation, focusing on methods based on semantic segmentation networks such as FCN, U-Net, and SegNet. They systematically summarized key indicators, including accuracy and computational efficiency. Muksimova et al. [25] proposed a wildfire segmentation method based on a dual encoder-decoder structure. By improving residual modules and attention gate mechanisms, they enhanced the network's multi-scale feature extraction capabilities, outperforming existing methods in terms of accuracy, speed, and robustness. To further improve wildfire segmentation accuracy and real-time performance, some researchers have introduced novel network structures such as transformers to this field. Ghali et al. [26] employed TransUNet and TransFire, two transformer-based semantic segmentation networks, to achieve precise segmentation of wildfire areas in UAV aerial images, with F1 scores exceeding 99%. Garcia et al. [27] proposed a multi-layer wildfire smoke segmentation method based on level set theory, optimizing contour smoothness and segmentation confidence to enhance model detection performance.

AI-based fire detection methods using RGB image recognition and thermal imaging have been widely adopted. However, thermal imaging-based fire detection methods are often limited to infrared camera imaging areas and primarily use a single standard deviation as a distinguishing feature, which weakens their early fire detection capability when disturbed. Additionally, the high cost of thermal imaging equipment makes it challenging to widely implement in large-scale environments such as forests. Moreover, thermal imaging systems require substantial data processing power to analyze the vast amounts of data collected, resulting in high power consumption and necessitating high-performance computing resources. In contrast, AI-based fire detection using RGB images employs standard cameras already widely used in most surveillance systems, making it cost-effective and easy to integrate. RGB-based fire detection also leverages modern deep learning models, enabling fast and accurate real-time monitoring with high inference speed and flexibility. Therefore, this paper focuses on in-depth research into deep learning-based fire detection using RGB images.

Fire image detection technologies based on satellite or UAV vision have achieved notable success in mountain fire monitoring. However, their practical application still faces several major obstacles: Complex terrain and environmental factors significantly impact fire detection, making it challenging. The varied mountain terrain and diverse surface coverage, combined with atmospheric interference from clouds, fog, and smoke, make it difficult to accurately isolate fire signals from complex backgrounds, leading to frequent missed detections or false alarms. The rapid evolution of forest fire scales requires improved adaptability of detection models. In the early stages of mountain fires, smoke and flame areas are small and undergo rapid spatiotemporal changes, easily blending with complex backgrounds. Detecting small targets in the early stages of a fire is crucial for preventing large-scale wildfires and protecting the ecological environment from damage. Subsequently, detection algorithms need enhanced adaptability to overcome limitations in UAV visual angles, changes in fire area scale, and visual obstructions. Large models' high computational resource demands pose deployment challenges for onboard equipment. Complex deep learning algorithms result in high computational resource consumption

and long processing delays, making it difficult to meet the emergency response needs of mountain fire disasters. In light of these challenges, this paper proposes the YOLO-CSQ rapid mountain fire detection model based on YOLOv8n for quick identification and early warning of mountain fires in complex scenarios. The main contributions are as follows:

(1) By adding the P2 layer output to YOLOv8n and introducing the CBAM (Convolutional Block Attention Module) attention mechanism in the four scales of the neck network, the model's ability to independently extract and fuse feature information at small scales is improved. This enables effective capture of multi-scale features of smoke and flame targets and enhances information interaction across scales, fully utilizing semantic information at different levels to improve small target detection capabilities and localization accuracy for larger targets.

(2) The ShuffleNetV2 backbone network structure is improved by introducing depth-wise separable convolutions (DWConv), CBAM attention mechanism, and h-swish activation function, replacing the original CSPDarknet53 backbone network structure. This reduces the model's parameter count and computational complexity while maintaining high feature extraction capabilities in a more lightweight head network, effectively reducing the model size and making it more suitable for deployment on detection devices with limited computational resources. Most importantly, it can enhance the model's ability to detect small targets.

(3) A Quadrupled-ASFF detection head is proposed, and the loss function is optimized to enhance the model's understanding of complex scenes (especially those with high background noise or small-scale targets with occlusions), improving the balance between positional accuracy and detection precision. Additionally, the WIoU loss function is introduced to address the inability of the original CIoU loss function to provide effective gradients in certain situations (e.g., non-overlapping bounding boxes) and to consider the distance between center points and aspect ratios of bounding boxes, thereby enhancing the model's localization accuracy.

The remainder of this paper is organized as follows: Section 2 introduces the theoretical background of the original YOLOV8 detection method; Section 3 describes the network and structural improvements; Section 4 presents the dataset preparation, classification, and experimental results; and Section 5 discusses the results and provides conclusions.

3. YOLOv8 Detection Algorithm

YOLO (You Only Look Once), as a classic real-time object detection algorithm, has played a significant role in defect identification, protection warning, and other fields since its proposal in 2015. Currently, the relatively classic iterative version, YOLOV8 [28], is an improvement based on YOLOV5 [29] proposed by Ultralytics. It also adopts a single-stage detection strategy, integrating object localization and classification tasks into an end-to-end convolutional neural network. Compared with traditional two-stage object detection algorithms such as R-CNN [30], both detection speed and efficiency are significantly improved. The YOLOV8 model includes five versions: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Among them, YOLOv8n has the lowest complexity, maintaining high detection accuracy while having the fastest inference speed, making it convenient for deployment on mobile or embedded devices. Considering the requirements of lightweight deployment on airborne platforms for mountain fire detection and the need for real-time and high-precision detection, this paper selects YOLOV8n as the baseline model.

The YOLOV8n network structure mainly consists of three parts: Backbone, Neck, and Head, as shown in Figure 1. The Backbone uses the CSPDarknet53 network, which replaces the original CSP (Cross Stage Partial) module with the C2f (Cross Stage Partial Network Fusion) module based on YOLOv5. The C2f module adopts gradient flow linking, effectively improving the model's nonlinear representation ability while keeping the model lightweight, thereby better handling complex image features. In addition, the SPPF module is retained in the Backbone, which converts the input features into adaptive-size outputs through mapping pooling operations to better capture multi-scale features in the image. The

Neck part adopts the Path Aggregation Network with a Feature Pyramid Network (PAN-FPN) [31] structure to further fuse the features transmitted by the Backbone. Compared with the PAN-FPN structure in YOLOv5, YOLOv8n removes the convolutional structure computation after upsampling in PAN and replaces the original C3 module with the C2f module, constructing a top-down and bottom-up network structure. This improves the model's feature fusion efficiency while complementing shallow location information and deep semantic information, ensuring the completeness and diversity of the output feature maps. The Head part adopts the same decoupled head structure as YOLOX [32], separating the classification and detection heads, allowing each part of the model to focus on its specific task. By further processing the feature maps output by the Neck part, it predicts the location, category, and confidence of the target. Moreover, the Anchor Free method used describes the detection target using multiple key points or center points and boundary information, which is more suitable for detecting dense obstacles and targets with large-scale variations in mountain smoke and fire detection.

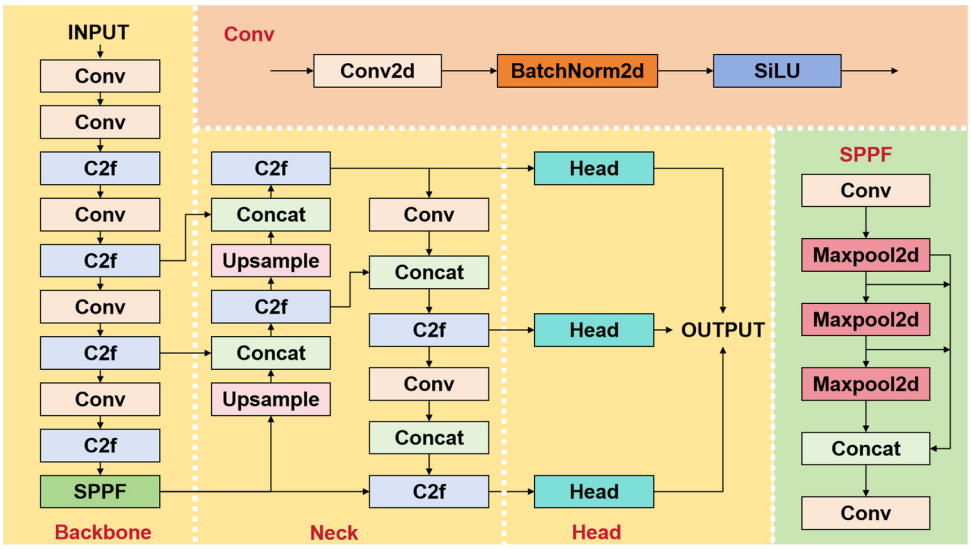

Figure 1. The network structure of original YOLOV8n.

Although the YOLOv8 model has balanced overall performance and performs well in many general scenarios, it still faces some challenges in the actual detection process due to the complex environment of mountain fires. Firstly, the background information in mountain fire scenes is redundant, and factors such as smoke, trees, and terrain can interfere with the identification of fire conditions, requiring the model to have strong robustness and generalization ability. Secondly, the shape and size of mountain fires vary, and there is a problem of large-scale differences, requiring the model to take into account the detection of targets at different scales. Moreover, under night conditions or low visibility conditions, the image quality is poor, requiring the model to adapt to low light, blur, and other interference factors. Finally, real-time performance is one of the important requirements for mountain fire detection, requiring the model to maintain a high inference speed while ensuring detection accuracy to support real-time detection and early warning of mountain fires.

4. Methods

To further improve the detection efficiency of mountain fire images on UAV-borne visual platforms, this paper proposes a YOLO-CSQ object detection algorithm based on the traditional YOLOV8n for dense small-target mountain fire detection in complex scenes. The network structure is shown in Figure 2. Firstly, by increasing the output of the P2 layer in YOLOv8n and introducing the CBAM (Convolutional Block Attention Module) attention mechanism in the neck network, the model's ability to independently extract and fuse feature information at small-scale levels is improved. This enables effective capture of

multi-scale features of smoke and flame targets, enhances information interaction between cross-scale features, and fully utilizes semantic information at different levels, thereby improving the detection ability of small targets and the localization accuracy of large-range targets. Secondly, an improved lightweight ShuffleNetV2 network is employed to replace the original backbone network. While introducing depthwise separable convolution and h-swish activation function, the CBAM attention mechanism is added before the P2 layer output to improve the model's detection performance. Subsequently, an improved four-head ASFF (Adaptive Spatial Feature Fusion) detection head is introduced, which enhances the detection capability of multi-scale targets by adaptively adjusting the fusion weights of features at different scales. Finally, the WIoU loss function is introduced to replace the original CIoU loss function, considering the specific weight of each pair of bounding boxes. This provides a more flexible and refined evaluation mechanism for object detection tasks in complex scenes or extreme conditions, improving the model's learning efficiency and generalization ability.

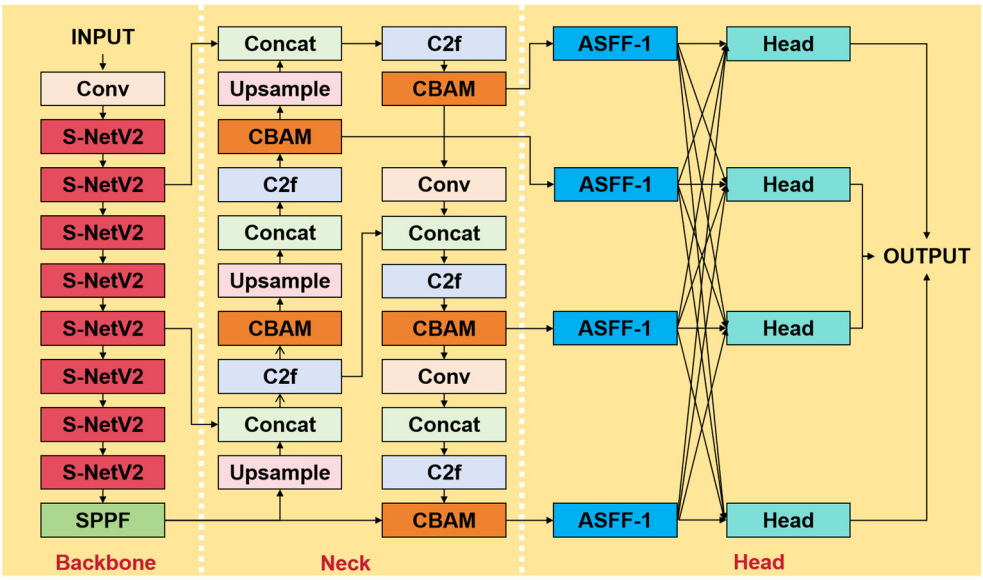

Figure 2. The network of YOLO-CSQ structure.

4.1. Improved Attention Mechanism

Due to the influence of illumination direction and the dynamic characteristics of flame and smoke targets, precise target detection in complex mountain fire scenes still faces enormous challenges, mainly reflected in the following three aspects. First, under the influence of the environment, there are fluctuations in the range and intensity of illumination, leading to uneven brightness distribution of flame and smoke images, weakening the contrast between the target and the background, and increasing the difficulty of positive target detection for the model. Second, due to the rapid movement and deformation of flames and smoke in the fire scene, the target edge contours become unclear, affecting the model's localization accuracy. Third, the complex terrain and rich vegetation cover in mountainous areas generate a large amount of redundant background information, making it difficult to distinguish smoke and fire targets from the background. To address these issues, some scholars have introduced attention mechanisms into object detection models to enhance useful feature information, suppress useless feature information, and enable the model to adaptively focus on key regions in the image, improving the model's detection accuracy.

CBAM (Convolutional Block Attention Module) [33], as an attention mechanism widely used in computer vision tasks, mainly consists of two sub-modules: Channel Attention Module and Spatial Attention Module. The Channel Attention Module obtains global information of the feature map through global average pooling and global max pooling operations, and then uses a multi-layer perceptron (MLP) [34] to learn the interdependencies

between channels, generating channel weights to identify the importance of each channel in the feature map. The Spatial Attention Module generates spatial weights by applying convolutional operations, effectively focusing on the importance of each spatial region in the feature map. CBAM combines the channel attention and spatial attention weights with the original feature map through element-wise multiplication to enhance the features of important channels and spatial regions while suppressing the unimportant parts. This process significantly improves the quality of the feature map, providing richer and more useful information for subsequent tasks such as mountain fire detection. The structure of the CBAM attention module is shown in Figure 3.

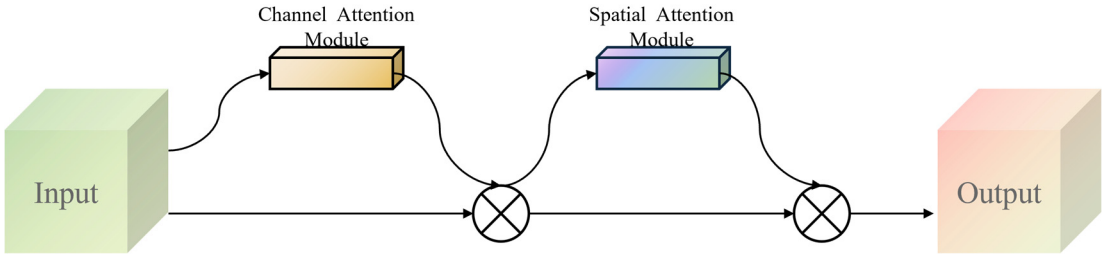

Figure 3. Structure of the CBAM attention mechanism.

Assuming the input feature map has dimensions $C \times H \times W$, where C represents the number of channels, and H and W represent the height and width of the feature map, respectively. In the CBAM module, the channel attention module first applies Global Average Pooling (GAP) and Global Max Pooling (GMP) operations to obtain the global average and maximum values for each channel, generating two feature vectors F_{avg}^{C} and $F_{\max}^{C}$ with dimensions of $C \times 1 \times 1$. The structure of the channel attention mechanism is illustrated in Figure 4.

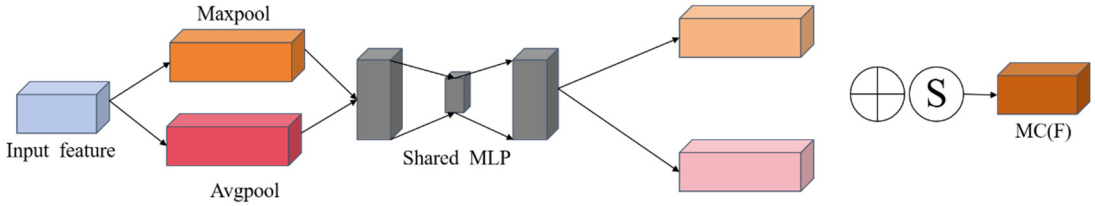

Figure 4. Structure of the channel attention mechanism.

The global average pooling operation can be expressed as:

$$F_{avg}^{C} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

The global maximum pooling operation can be expressed as:

$$F_{\max}^{C} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{2}$$

where F_{avg}^{C} represents the output result after applying average pooling to the $C-th$ channel, $F_{\max}^{C}$ denotes the output result after performing max pooling on the $C-th$ channel, and $u_c(i,j)$ represents the feature vector at spatial position (i,j) in the $C-th$ channel of the input feature map.

After obtaining the two feature vector descriptors, the Channel Attention Module uses two shared multilayer perceptions (MLPs) to transform F_{avg}^{C} and $F_{\max}^{C}$, learning the interdependencies between channels. To reduce the number of model parameters,

the number of neurons in the hidden layer is set to a reduced value $R^{C/r \cdot 1+1}$, where the reduction rate r determines the degree of neuron reduction. After processing each descriptor through the shared network, the outputs are aggregated through element-wise addition to obtain a unified output feature vector. The MLP output generates channel weights through the Sigmoid activation function, and the calculation formula M_C is as follows:

$$\begin{aligned} M_C(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{nax}))\right) \end{aligned} \tag{3}$$

where σ represents the Sigmoid activation function, and $W_0 \in R^{C/r*C}$ and $W_1 \in R^{C*C/r}$ are the weight parameters of the multi-layer perceptron MLP.

The Spatial Attention Module is mainly a complement to the Channel Attention Module. Its structure is shown in Figure 5. Firstly, average pooling and max pooling operations are performed on the channel dimension to reduce the dimensionality of the channel itself and generate two feature maps F^S_{avg} and $F^S_{\max}$ with a size of $1 \times H \times W$, respectively. The two feature maps are concatenated and sent to a convolutional layer for learning.

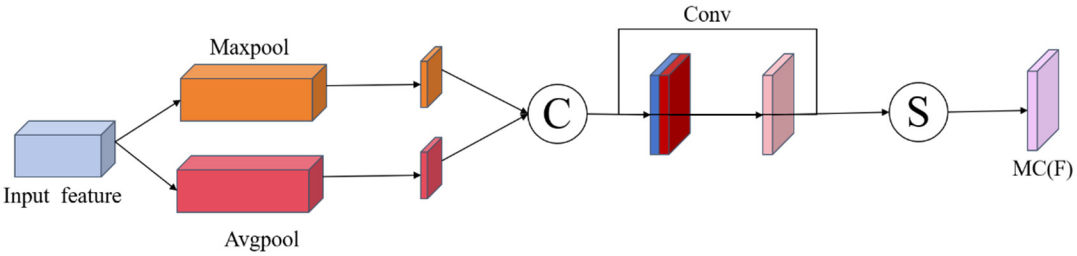

Figure 5. Structure of the spatial attention mechanism.

After the average pooling operation on the channel dimension, it can be represented as:

$$F^S_{avg} = \frac{1}{C}\sum_{i=1}^{C} x_i \tag{4}$$

The max pooling operation on the channel dimension can be represented as:

$$F^s_{max} = max(x_i), i \in [1, C] \tag{5}$$

where $F^s_{\max}$ represents the feature map of the $i-$th channel of the input feature map.

Subsequently, F^S_{avg} and $F^S_{\max}$ are concatenated on the channel dimension to obtain a feature map with a size of $2 \times H \times W$, and a 7×7 convolutional layer is applied to learn the interdependencies between different spatial locations, ultimately obtaining the attention features on the spatial dimension. The calculation process is as follows:

$$M_S(F) = \sigma\left(f^{7\times7}\left(F^s_{avg}; F^s_{max}\right)\right) \tag{6}$$

where is the Sigmoid activation function σ, $f^{7\times7}(\ldots)$ represents processing using a convolutional kernel of size 7×7, and $(;)$ represents the concatenation operation on the channel dimension.

The original feature map is element-wise multiplied with the generated channel feature weights and spatial feature weights, respectively, to obtain the final weighted feature map:

$$F'' = M_C(F) \otimes M_S(F) \otimes F \tag{7}$$

The CBAM attention mechanism effectively enhances the model's feature extraction ability [35]. Through adaptive adjustment of the weights of different channels and spatial

locations in the feature map, the model's representation ability of mountain fire region features is enhanced, the interference of complex backgrounds and noise is suppressed, and the feature utilization efficiency is improved, enabling the model to focus more on wildfire-related features. Concurrently, the model's generalization ability is enhanced, enabling it to adapt to diverse wildfire types and shooting conditions. This results in more accurate and robust wildfire detection results. Furthermore, the computational complexity of the CBAM attention mechanism is relatively low, facilitating integrated operations. While improving wildfire detection performance, it maintains the model's detection efficiency, meeting the requirements of practical applications.

4.2. Improved Backbone Network Architecture

The shapes of smoke and flames in mountain fire images are diverse, with blurred boundaries, often requiring a larger receptive field to capture their contextual information. Furthermore, the influence of illumination changes and background interference necessitates a more robust model for feature extraction. Although the traditional YOLOV8n's CSPDarknet53 backbone network performs well in general scenarios, it still struggles to adapt well to mountain fire target detection scenarios due to limitations in its receptive field range and feature extraction capabilities. In comparison to the traditional CSPDarknet53, ShuffleNetV2 [36], as a lightweight backbone network, exhibits distinctive advantages in the context of mountain fire detection. The structure of the network is depicted in Figure 6.

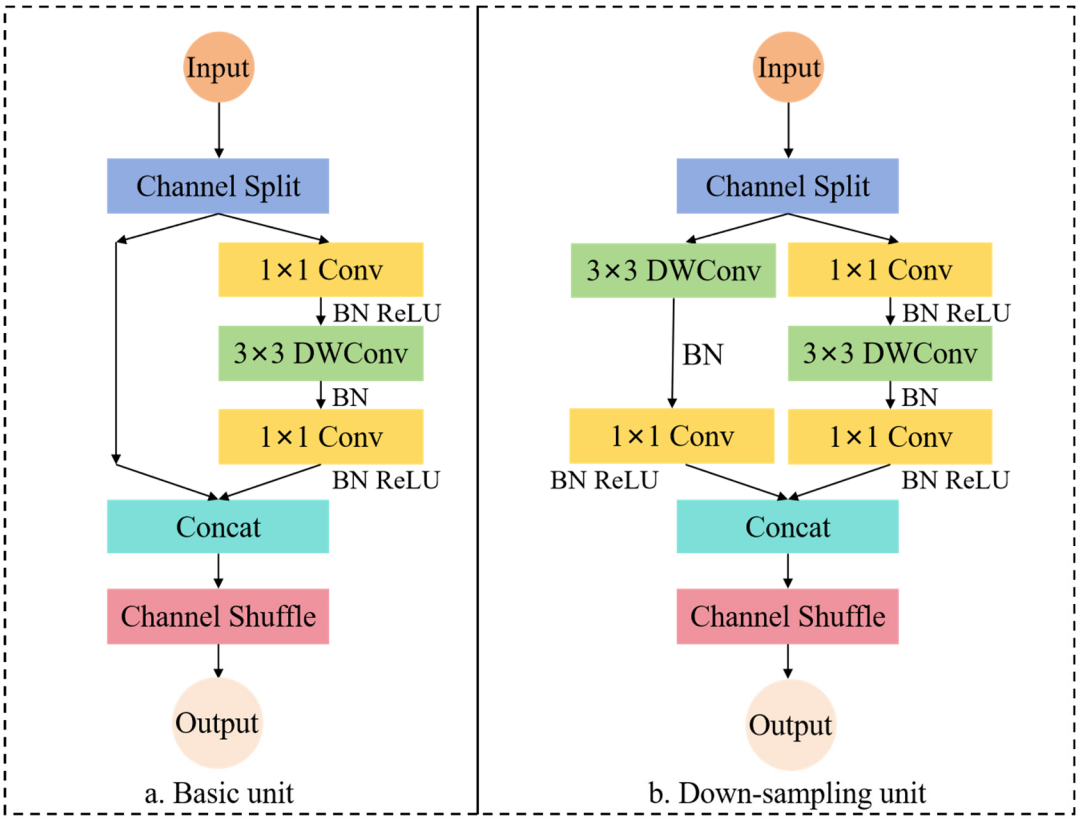

Figure 6. Structure of Original ShuffleNetv2.

In channel splitting, the input feature channels are divided into two branches, each with an equal number of channels, with the objective of reducing memory access costs and improving computational efficiency. The formula representation is as follows:

$$c_{in} \rightarrow c_{out\ 1} = c_{out\ 2} = \frac{c_{in}}{2} \tag{8}$$

where c_{in} represents the number of input channels, and $C_{\text{out }1}$ and $C_{\text{out }2}$ represent the number of output channels for the two branches, respectively.

Subsequently, the two branches undergo processing through a series of grouped convolutions, such as 1×1 and 3×3, in order to achieve a balance between computational complexity and model capacity. Finally, the results of the two processed branches are merged, and channel shuffling is performed to enhance information flow between different channels, thereby improving the model's representational power. In spatial downsampling operations, ShuffleNetV2 employs convolutions with a stride greater than 1 to achieve feature map spatial size reduction while maintaining feature richness.

Nevertheless, ShuffleNetV2 still exhibits certain deficiencies in the context of mountain fire detection. Firstly, the network depth and receptive field of ShuffleNetV2 remain relatively limited, which presents a challenge in fully capturing long-range dependencies and global contextual information in mountain fire scenes. Secondly, ShuffleNetV2 is deficient in sufficient scale invariance, rendering it incapable of effectively handling smoke and fire targets of varying sizes. To address these issues, this paper proposes an enhanced ShuffleNetV2 backbone network structure, as illustrated in Figure 7. The incorporation of depth-wise separable convolution (DWConv) [37], the CBAM attention mechanism, and the h-swish activation function has led to a notable enhancement in the model's performance with regard to mountain fire detection.

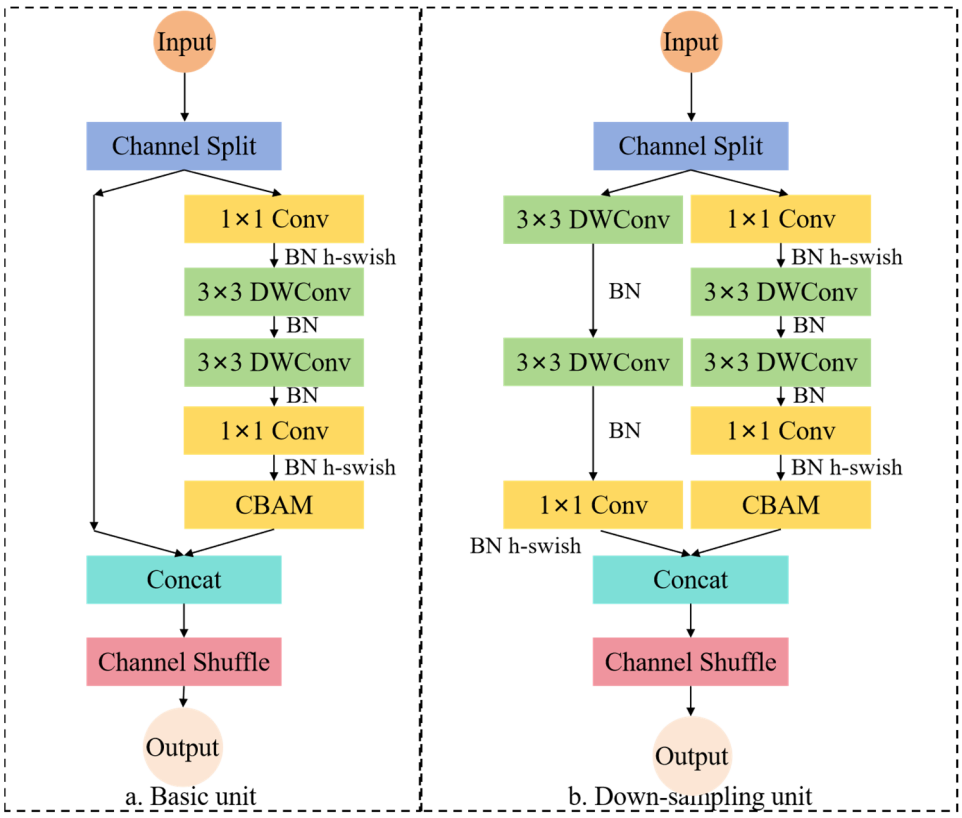

Figure 7. Structure of the Improved ShuffleNetv2.

Firstly, in the feature extraction stage of the model's backbone network, by introducing a set of 3×3 depthwise separable convolution (DWConv) kernels combined with the original convolutional layers, the model's receptive field is expanded. This allows the network to capture richer contextual information while only slightly increasing the computational load. It enhances the model's ability to recognize irregular and boundary-blurred targets in mountain fire scenes, strengthens the robustness of feature extraction, and enables more precise handling of the complexity in mountain fire scenes.

Secondly, in the final stage of the backbone network, by integrating the CBAM attention mechanism, multi-scale parallel learning and cross-spatial information interaction are realized, establishing connections between features of different scales. This improves the model's detection performance for multi-scale targets, particularly in capturing features of smoke and flames with varying scales, effectively increasing the detection rate of small and weak targets and the localization accuracy of large targets.

Finally, to further optimize the model's performance, we replace the traditional ReLU activation function with the h-swish activation function. By leveraging its superior nonlinear expression ability and smoothness characteristics, it alleviates the gradient vanishing problem that may be caused by the ReLU function, enabling the model to converge quickly and enhancing its generalization ability. Moreover, introducing the h-swish activation function reduces the model's dependence on cross-channel correlation and spatial correlation, allowing the model to focus more on channel information recognition and extraction, thereby improving the model's accuracy and reliability in the mountain fire detection task.

4.3. Improved Quadrupled-ASFF (Adaptive Spatial Feature Fusion) Detection Head Structure

Traditional object detection networks typically employ a fixed feature fusion strategy, combining feature maps of different scales with preset weights to generate final detection results. However, real-world mountain fire scenes are complex and varied, with smoke and fire targets often exhibiting significant multi-scale features. Using a "one-size-fits-all" fusion approach can ignore the specificity of targets at different scales, making it challenging to adapt to the multi-scale characteristics of mountain fire scenes, resulting in limited detection performance. To address this issue, researchers have proposed introducing an Adaptive Spatial Feature Fusion (ASFF) [38] module to adaptively adjust the fusion weights of features at different scales, enhancing the model's detection capability for multi-scale targets. The core idea of ASFF is to dynamically adjust the fusion weights of different feature maps based on the scale features of the targets, allowing the model to adaptively focus on targets of various scales.

Figure 8 depicts the ASFF module, which initially adjusts the feature maps F_1, F_2, and F_3 from disparate convolutional layers to a uniform spatial resolution through up-sampling or down-sampling. Subsequently, the system learns to generate weight maps (W_1,W_2, and W_3) for each feature map, which are employed to dynamically adjust the fusion weight of each feature map at each spatial location. The weight maps are normalized through a soft-max layer to ensure that the sum of weights is 1. For each location (x, y), the weight maps must satisfy the following conditions and weight calculation formula:

$$\sum_{i=1}^{3} W_i(x, y) = 1 \tag{9}$$

$$W_i(x, y) = \frac{e^{W_i(x,y)}}{\sum_{j=1}^{3} e^{W_j(x,y)}} \tag{10}$$

where W_i represents the weight map of the $i - th$ feature map.

After learning the weight maps, each feature map is weighted according to its corresponding weight map, and all the weighted feature maps are summed to form the final fused feature map F_{fused} . The calculation process of the fused feature map can be represented as:

$$F_{\text{fused}} = \sum_{i=1}^{3} W_i \cdot F_i \tag{11}$$

where $\cdot$ represents element-wise multiplication; F_i represents the adjusted feature map of the $i - th$ scale; W_i is the learned weight map of the $i - th$ feature map, determining its contribution to the fusion process; and F_{fused} represents the final fused feature map.

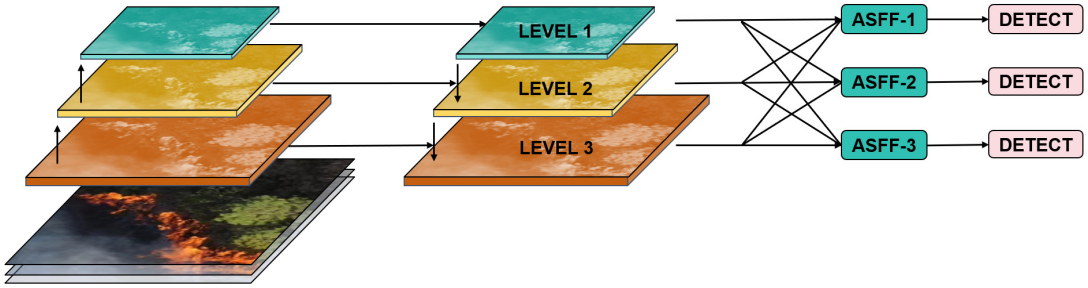

Figure 8. The overall structure of original ASFF.

Although ASFF can adaptively adjust the weights of feature fusion and enhance the model's detection performance for multi-scale targets, mountain fire scenes contain a considerable number of minute targets (e.g., distant fire points, thin smoke), rendering it challenging to accurately localize and identify these targets solely through the original three scales of feature maps. To further enhance the model's detection capability for tiny targets, this paper proposes an improved Quadrupled-ASFF detection head, as shown in Figure 9.

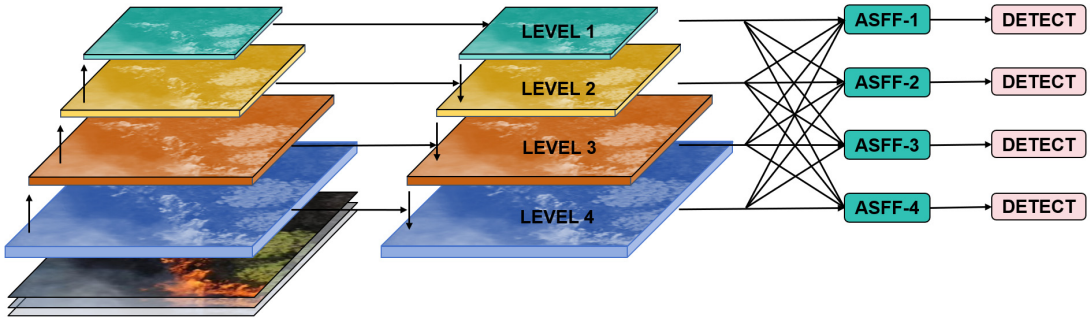

Figure 9. The overall structure of Quadrupled-ASFF.

The Quadrupled-ASFF detection head introduces an additional feature map F_4, corresponding to the P_2 layer in the backbone network, on top of the original three scales of feature maps, F_1, F_2, and F_3. By increasing the output of the P_2 layer, Quadrupled-ASFF is able to obtain richer detail information, which enables the detection of minute targets. Following the addition of the novel prediction head, the weight calculation formula and feature fusion calculation formula are presented as follows:

Weight calculation formula:

$$W_i(x,y) = \frac{e^{W_i(x,y)}}{\sum_{j=1}^{4} e^{W_j(x,y)}} \tag{12}$$

Feature fusion calculation formula:

$$F_{\text{fused}} = \sum_{i=1}^{4} W_i \cdot F_i \tag{13}$$

The improved Quadrupled-ASFF prediction head effectively improves the model's detection capability for tiny targets while maintaining high recognition performance for medium and large-sized targets by increasing the output of the P_2 layer. It also enhances the model's understanding of complex scenes, particularly in cases with high background noise or occlusion between targets, enabling more accurate localization and identification of tiny targets in complex mountain fire environments.

4.4. Improved Loss Function

The YOLOv8n algorithm employs the CIoU loss function as the default loss function for bounding box regression. The rationale behind the utilization of CIoU loss is to address the shortcomings of the IoU loss function, which is unable to provide effective gradients in specific instances, such as when two bounding boxes are not in spatial overlap. Furthermore, CIoU loss takes into account the distance between the center points of the bounding boxes and their aspect ratios, thereby enhancing the model's localization accuracy. The calculation formula for CIoU loss is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{14}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{15}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{16}$$

where IoU represents the intersection over the union of the predicted box and the ground truth box; $\rho(b, b^{gt})$ is the Euclidean distance between the center points of the predicted and ground truth boxes; c is the diagonal length of the smallest enclosing area containing the two boxes; v considers the aspect ratio of the predicted and ground truth boxes; w and h are the width and height of the predicted box, respectively, while w^{gt} and h^{gt} are the width and height of the ground truth box; and α is a weight parameter.

Despite the enhancements brought about by the CIoU loss function in bounding box regression performance, it still exhibits certain limitations. In the event of a significant discrepancy between the aspect ratio of the predicted and ground truth boxes, the model may be unduly penalized, which could have a detrimental impact on its learning efficiency and ultimate performance. Moreover, the CIoU loss function may lack the capacity to generalize in complex mountain fire scene conditions.

To further enhance the model's performance in the mountain fire detection task, this paper introduces the WIoU loss function [39] as a replacement for the original loss function. In calculating the IoU score, the WIoU assigns differential importance to each pair of predicted and ground truth boxes by incorporating specific weights for the bounding boxes. This weighted strategy enables the model to evaluate the overlap quality between different bounding boxes in a more flexible and meticulous manner, rendering it effective in handling object detection tasks in complex scenes. The WIoU calculation formula is as follows:

$$WIoU = \frac{\sum_{i=1}^{n} w_i \times \mathrm{IoU}(b_i, g_i)}{\sum_{i=1}^{n} w_i} \tag{17}$$

where n represents the number of annotated defect boxes; b_i represents the coordinates of the $i - th$ predicted box; g_i represents the coordinates of the $i - th$ ground truth box; $\mathrm{IoU}(b_i, g_i)$ represents the IoU value between the corresponding predicted and ground truth boxes; and w_i represents the weight value.

5. Experiment

5.1. Datasets

To further verify the effectiveness of the improved model in mountain fire detection, this paper targets forest mountain fires. In order to achieve this, relevant datasets must be collected, organized, and labeled. Schematic of forest wildfire image acquisition is shown in Figure 11. These datasets mainly consist of three parts:

(1) A dataset of a real fire scene is presented here. This portion of the data was derived from video image data collected during on-site inspections and rescue support pro-

cesses for a number of real mountain fires that occurred in the southwestern forest region, China's second-largest natural forest region, since 2018. The videos are converted into frames in order to capture the natural state of the mountain fires and their impact on the surrounding environment.

(2) The simulated fire scene dataset is as follows: In order to more comprehensively simulate different types of mountain forest fire scenarios, this study employs the use of dry tree branches and special smoke devices to simulate fire scenes under safe and controlled conditions. High-definition cameras mounted on drones are utilized to collect image and video data of fire smoke diffusion and flame spread from multiple angles and heights, thereby increasing the diversity and complexity of the dataset. During the data collection process, we adhered to the requirements of the Technical Specifications for Drone Surveying of Forest Fires. For simulated general forest fires, we maintained a drone flight altitude of no less than 80 m above the fire scene. Aerial images were captured from four directions: east, south, west, and north. The reconnaissance time, fire line, fire point, smoke point, wind direction indicator, and other key elements were annotated in the images.

(3) Integration of the public dataset: In order to further enrich the data foundation of this research and verify the generalization ability of the improved YOLOv8n algorithm in different fire scenarios, this paper selects and integrates public datasets such as FLAME [40] and Alert Wildfire [41]. Some datasets examples are shown in Figure 10.

Due to the random nature of mountain fires and the variability in the collection of image backgrounds, this dataset primarily encompasses bright backgrounds, night-time backgrounds, and similar scenes of fire and smoke. The specific distribution is presented in Table 1.

In order to enhance the accuracy and robustness of the fire monitoring model, this research collects a substantial quantity of mountain fire image data and employs the LABLEIMG tool to meticulously annotate flames and smoke. Furthermore, the model is capable of distinguishing various fire scene situations, including those with multiple targets, small targets, targets under occlusion, and images that may be similar to flames and smoke. Furthermore, in order to enhance the representativeness and comprehensiveness of the dataset, a series of data augmentation methods have been introduced. These include image cropping, rotation, flipping, and scaling operations, as well as the use of mosaic techniques to randomly stitch and combine images. This expands the diversity of mountain fire image samples and improves the model's adaptability and generalization ability in practical applications.

The specific preprocessing steps are as follows:

(1) The original RGB image is converted to a single-channel grayscale image, and Gaussian filtering is applied to the grayscale image in order to reduce the impact of noise through smoothing.

(2) The image size is randomly changed in order to enable the model to learn flames and smoke at different scales.

(3) The image is rotated at multiple different angles in order to increase the model's ability to identify target orientations.

(4) Mosaic techniques are applied in order to randomly stitch four different images together in order to enhance the diversity of the dataset.

After the screening process, 25,026 images were obtained, encompassing 52,694 annotation points. The data were divided into three sets: a training set (17,518 images), a validation set (5005 images), and a test set (2503 images). The ratio of the three sets was 7:2:1. This division was intended to facilitate comprehensive training, testing, and validation of the improved YOLOv8 model. Additionally, it allowed for the evaluation of the model's effectiveness and reliability in fire detection.

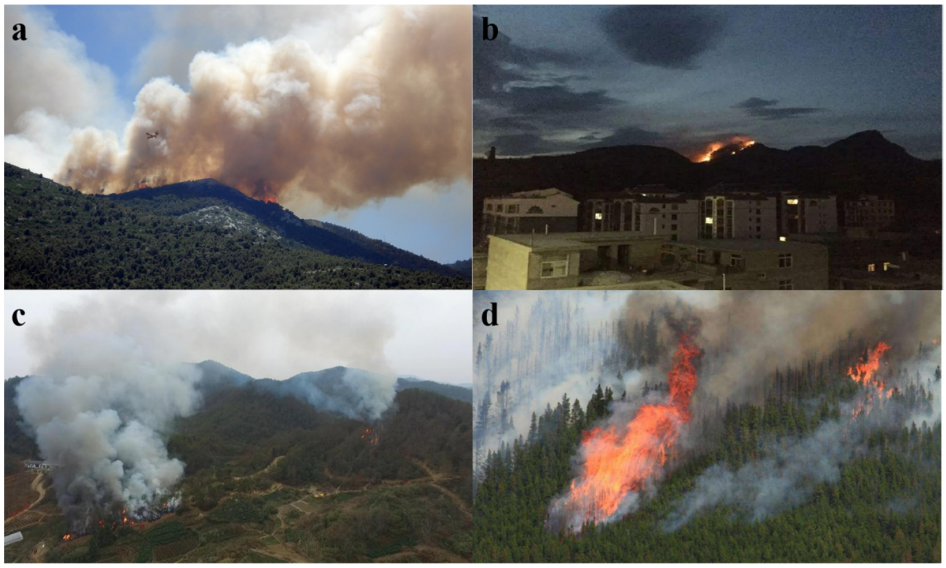

Figure 10. Dataset examples (**a**–**d**).

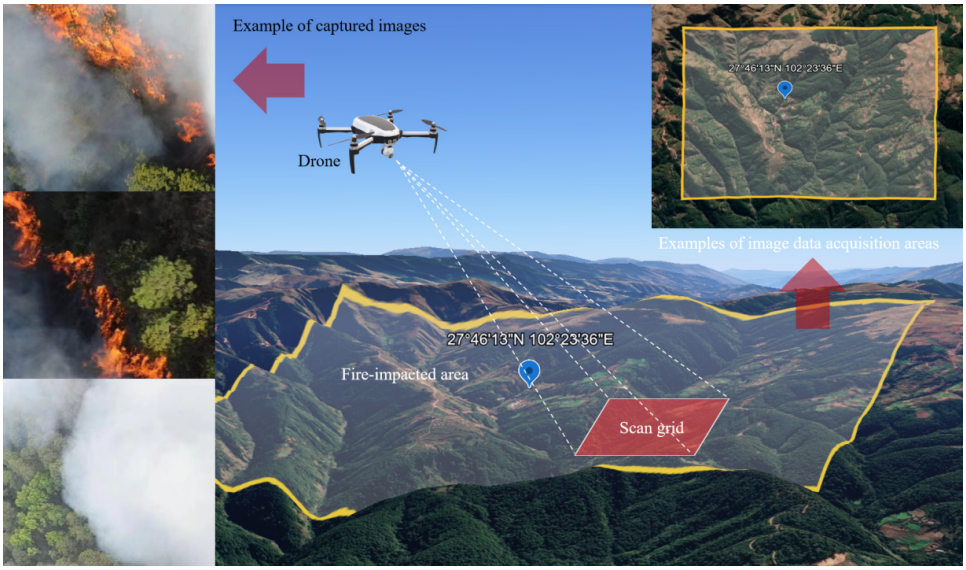

Figure 11. Schematic of forest wildfire image acquisition.

Table 1. Dataset distribution.

Application Scenario	Quantity
Daylight	10,265
Dark	9761
Similar to fire	2157
Similar to smoke	3294

5.2. Experimental Environment

The model was trained using the Windows 10 operating system, an NVIDIA GeForce GTX 3060 graphics processing unit (GPU), the PyTorch 17.0 deep learning framework, the Python 3.8 programming language, and the CUDA 10.2 parallel computing platform. The experiments were conducted with the YOLO8n pre-trained weights. The initial learning rate was set to 0.01, and the batch size was set to 64.

5.3. Evaluation Metrics

In order to provide an intuitive and comprehensive evaluation of the performance of the improved network, this paper employs a range of indicators, including Precision, Recall, and F1-score, to assess the model's performance. The calculation formula and its significance are presented below in Table 2 for clarity.

Table 2. Evaluation Indications.

Indicators	Notations	Meanings
Precision	$P = \frac{TP}{TP+FP}$	Measures the ratio of correctly identified objects to all objects identified by the model. A higher precision means fewer false positives among the objects identified by the model.
Recall	$R = \frac{TP}{TP+FN}$	Measures the ratio of correctly identified objects to all actual objects. A high recall means that the model can capture more true objects and reduce false negatives.
F1-Score	$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$	A comprehensive indicator reflecting the model's precision and robustness, the harmonic mean of precision and recall, used to comprehensively evaluate the model's accuracy.
mAP	$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$	The average of AP values for all categories, providing an overall measure of the model's performance across all categories.
Parameters	——	The number of parameters that need to be learned in the model, an important indicator for measuring model complexity.
GFLOPS	——	The number of floating-point operations used to measure the complexity of the algorithm/model.

5.4. Experiment Results

5.4.1. Overall Comparative Analysis of Models

A comparison of the loss curves reveals in Figure 12 that the enhanced YOLOv8 model demonstrates a notable advantage over the original YOLOv8n model in the context of mountain fire image detection. In comparison to the original model, the improved model demonstrates a more rapid rate of convergence during the initial stages of training. This enables the model to efficiently learn and extract crucial features of mountain fire images, as well as to swiftly optimize its parameters. Concurrently, the enhanced model exhibits enhanced stability in the latter stages of training, with diminished fluctuations in the loss function curve. In contrast, the loss curve of the original model continues to exhibit certain fluctuations and instability. This indicates that the improved model is capable of continuous optimization and fine-tuning of parameters, resulting in a stabilized loss function at a lower level. This, in turn, leads to enhanced robustness and generalizability. Furthermore, the loss function of the enhanced model reaches a lower value, indicating that it is more adept at accurately locating and identifying mountain fire targets, thereby reducing the occurrence of missed detections and false alarms. In conclusion, the enhanced YOLOv8 model exhibits notable advantages in terms of convergence speed, training stability, and detection accuracy, rendering it more suitable for the early warning and real-time monitoring of mountain fires.

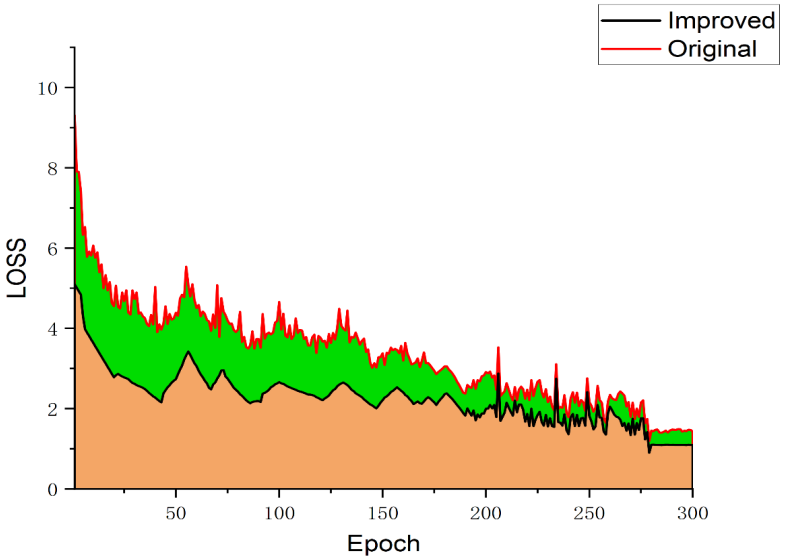

Figure 12. Total loss curve for the model training process.

Figure 13 illustrates the change in mAP@50 with the number of iterations (epochs) during the training process of the enhanced YOLOv8 model (YOLO-CSQ) and the original YOLOv8 model. It can be observed that at the outset of model training, both models exhibit a rapid increase in mAP@50. However, the original YOLOv8 model exhibits considerable fluctuations, with a tendency to stabilize only after 100 rounds. In contrast, the improved model stabilizes and gradually increases in performance after 50 rounds. In the middle and later stages, the improved model exhibits a relatively stable trend without significant fluctuations, indicating enhanced accuracy, stability, and robustness. Following the unfreezing of the training, both the original model and the improved model exhibited a slight increase. However, throughout the training process, the MAP@50 value of the improved model remained consistently higher than that of the original model, ultimately reaching 96.87%, a notable improvement compared to the original model. This suggests that the enhanced model exhibits enhanced detection accuracy and is better able to identify fire targets in images.

Figure 14 illustrates the mAP@50-95 change curve of the enhanced YOLOv8 model and the original YOLOv8 model throughout the training process. It is evident that during the initial training stage, both models exhibit a rapid rise, accompanied by varying degrees of fluctuations. However, after 50 training rounds, the enhanced model outperforms the original model, demonstrating a consistent and significant improvement in mAP@50-95, reaching a final value of 76.6%. This indicates that the enhanced model exhibits enhanced robustness and superior object detection performance.

As shown in Figure 15, a precision–recall curve analysis reveals that the enhanced YOLOV8 model demonstrates exceptional detection efficacy for flame targets. Throughout the entire recall interval, precision remains consistently high, exceeding 0.9. Even when recall approaches 1.0, the model can still maintain a precision of approximately 0.8. This indicates that while the model is highly accurate in detecting true flame targets, it is also capable of effectively controlling the false detection rate, thereby ensuring the accuracy of the detection results. This high-precision detection is of paramount importance for fire warning and emergency response, providing the most reliable data support for related decision-making and avoiding the serious consequences that could result from missed or false alarms.

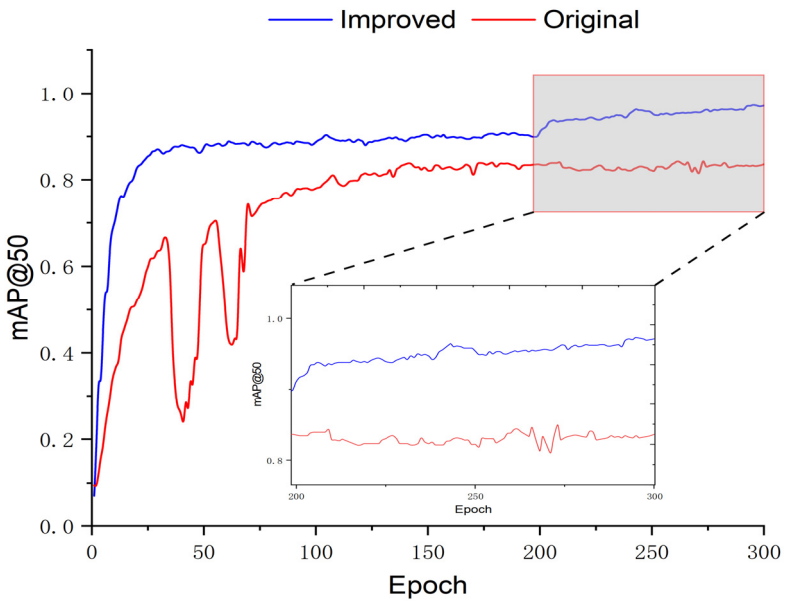

Figure 13. mAP@50 curve.

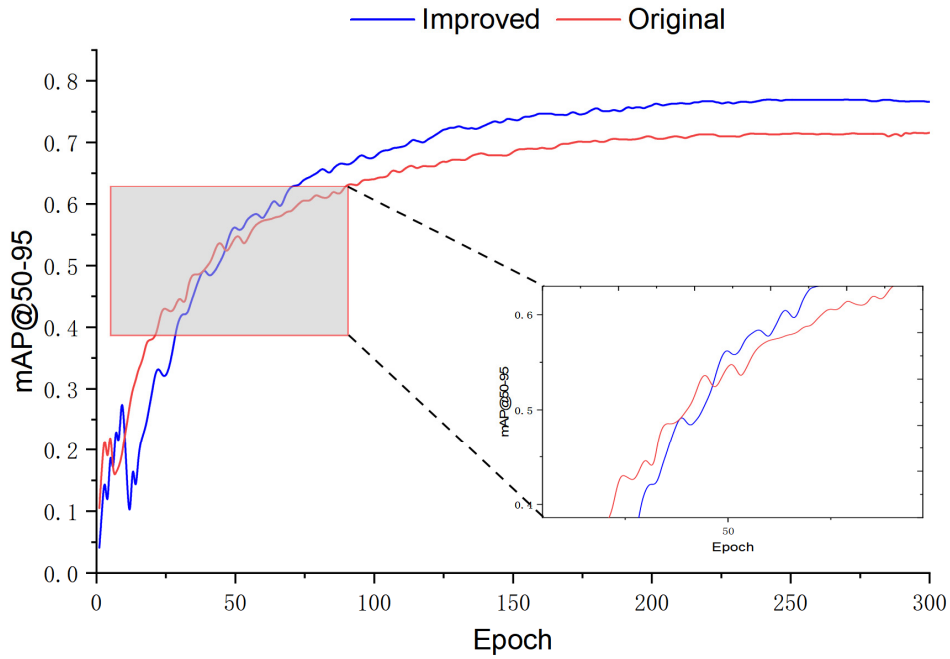

Figure 14. mAP@50-95 curve.

With regard to smoke targets, although the overall detection accuracy of the model is not as high as that of flame targets, the precision–recall curve still presents a relatively full and smooth shape, without obvious jumps or collapses. This indicates that the model is capable of maintaining relatively stable detection accuracy for smoke targets at different recall levels, thereby demonstrating robust performance. In practical applications, even in the face of complex and changing environmental factors, the model can provide relatively consistent and credible detection results. In particular, when the recall rate exceeds 0.8, the precision rate can still be maintained at approximately 0.7, which meets the usage requirements in the majority of scenarios.

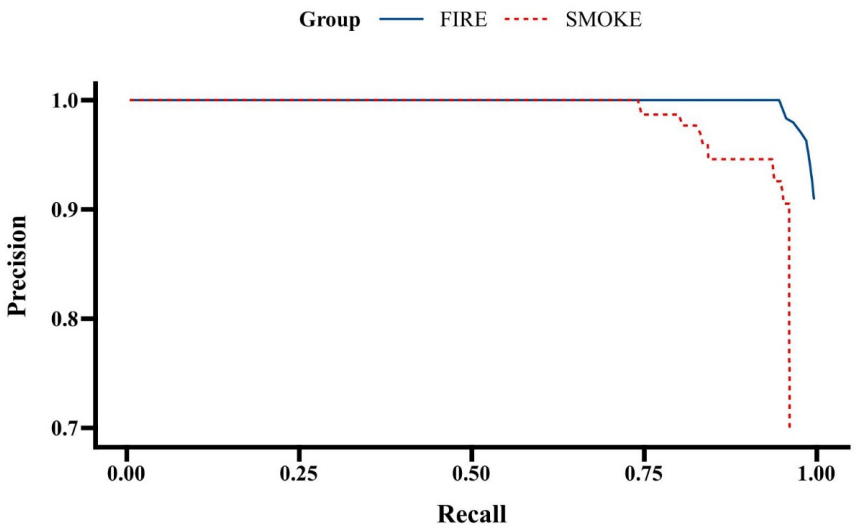

Figure 15. Precision–recall curve.

In conclusion, the precision–recall curve for the improved model demonstrates a high level of precision and a favorable shape, with a considerable area under the curve. This indicates that the improved YOLOv8 model achieves high average precision values for both targets and exhibits excellent detection performance. The curve for flame targets is consistently above that of smoke targets, primarily due to the visual feature differences between flames and smoke. Flames possess more pronounced and discernible visual characteristics, including vivid colors, clear textures, and edges, which facilitate the model's learning and recognition. In contrast, the visual features of smoke are relatively indistinct and uncertain, with lower distinguishability from the background, posing greater difficulties and challenges for detection.

In conclusion, a comparison and analysis of the performance of the original YOLOv8 model and the improved YOLOv8 model on different evaluation metrics reveals the following conclusions: the improved YOLOv8 model demonstrates significant advantages in precision, recall, and other aspects, thereby substantiating the effectiveness and feasibility of the improvement strategies in enhancing mountain fire detection performance.

5.4.2. Ablation Experiment

In order to gain a more nuanced understanding of the influence of each improvement module on the model's detection performance, this paper presents four sets of ablation experiments, conducted over 300 iterations under identical parameter settings. The results obtained are presented in Table 3.

A comparison of the results of the ablation experiments reveals that the detection performance of the enhanced YOLOv8 model on mountain fire images has been significantly enhanced. With the original YOLOv8 model serving as a benchmark, the enhanced model's mAP is 85.97%, F1 score is 81.34%, model parameter quantity is 11.2M, and computational complexity is 8.9 GFLOPS.

Table 3. Results of ablation experiments.

Model	mAP@50	mAP@50-95	Precision	Recall	F1-Score	Parameters	GFLOPs
Original YOLOv8	85.97	75.97	80.34	81.69	81.34	11.2	8.9
+CBAM	87.95	76.12	86.57	84.87	81.35	13.1	7.2
+CBAM + Improved ShufflenetV2	90.85	76.19	87.79	85.76	83.82	5.2	19.3
+CBAM + Improved ShufflenetV2 +Quadrupled ASFF	94.34	76.21	89.95	87.75	87.81	5.7	16.7
YOLO-CSQ	96.87	76.60	93.91	88.87	88.35	5.7	15.9

The introduction of the CBAM attention mechanism into the model resulted in an increase of 1.98% in mAP@50 and 6.23% in precision. This indicates that by adaptively

adjusting weights in the channel and spatial dimensions, the model can focus more on the salient features of fire targets and suppress background interference, thereby improving detection accuracy. Concurrently, the expansion in parameters is relatively modest, suggesting that the integration of the CBAM module markedly enhances performance without imposing undue computational burden. Subsequently, the replacement of the original backbone network with the improved ShuffleNetV2 resulted in an increase in mAP@50 by 2.9%, reaching 90.85%, while the number of parameters was reduced by 60.3% to only 5.2 M. This is due to the unique channel split and shuffle operations of ShuffleNetV2, which can significantly reduce the number of parameters and computational complexity while ensuring feature extraction capabilities, thus making the model more lightweight and efficient. However, it should also be noted that the introduction of grouped convolution and channel reordering strategies in the improved module has led to additional computational overhead, significantly increasing the algorithm's GFLOPS to 19.3, indicating an increase in computational complexity. Subsequently, the introduction of the enhanced Quadrupled-ASFF detection head resulted in a 3.49% increase in the model's mAP@50, reaching 94.34%. This effectively enhanced the model's ability to detect small targets and enhanced the model's understanding of complex scenes. Finally, the replacement of the CIoU loss with the WIoU loss resulted in the model's overall performance reaching its optimum, with mAP@50 at 96.87% and precision also rising to 93.91%. Furthermore, recall and F1-score remained high at over 88%. The introduction of weight coefficients in the WIoU loss function enables the model to focus its learning on targets with ambiguous features, such as small targets and blurred targets. This results in faster loss convergence and further improves detection accuracy. It is noteworthy that although the computational complexity of the model with the WIoU loss function is slightly higher than that of the model with the improved ShuffleNetV2, it still maintains relatively low parameters, meeting the requirements for deployment on terminals.

The results of the ablation experiments demonstrate that the model has achieved notable improvements in accuracy (mAP) and overall performance (F1-score) through the introduction of the CBAM attention mechanism, an enhanced ShuffleNetV2 backbone network, an optimized ASFF detection head, and a refined loss function. These enhancements have resulted in a reduction in the model's complexity while maintaining its efficacy. This suggests that the enhanced YOLOv8 model exhibits notable advantages in mountain fire image detection.

5.4.3. Comparison Experiment

To further verify the performance of the improved YOLOV8 network for mountain fire detection, representative single-stage object detection models, such as SSD and other YOLO series models, as well as the two-stage object detection model Faster R-CNN, were selected for comparison experiments. The results of the experimental analysis are presented in Table 4.

Table 4. Summary of multi-model horizontal comparison experimental results.

Model	mAP50	mAP50-95	Precision	Recall	F1-Score	Parameters	GFLOPs
Faster RCNN	63.40	51.22	57.00	65.91	63.15	137.1	8.1
SSD	60.78	53.54	59.17	55.32	67.29	26.2	9.2
YOLOV5	75.96	68.14	67.26	66.93	69.36	3.2	24.3
YOLOV7tiny	78.21	71.23	71.39	73.14	72.00	13.3	13.7
YOLOV8	85.97	75.97	80.34	81.69	81.34	11.2	8.9
YOLOV9	85.21	74.21	80.68	81.22	80.19	7.1	6.3
YOLOV10	83.54	72.16	79.27	80.57	79.32	15.4	21.6
YOLO-CSQ	96.87	76.60	93.91	88.87	88.35	5.7	15.9

A comparison of the performance of the enhanced YOLOV8 model with that of other established object detection models in the context of mountain fire detection reveals that the enhanced YOLOV8 model exhibits certain advantages in terms of accuracy,

speed, and model size. Specifically, the mean average precision (mAP) at 50% of the improved YOLOV8n model reaches 96.87%, an increase of 10.9% compared to the original YOLOV8 and 11.66% higher than the best-performing YOLOV9, reflecting a high level of detection accuracy. The improved YOLOV8 model also maintains a leading advantage in mAP@50-95, at 76.60%, indicating that the model performs exceptionally well at different IoU thresholds and has stronger robustness. In terms of precision, the improved YOLOV8 achieves a score of 93.91%, representing an increase of 13.57% compared to the original. In contrast, the precision of other models is generally below 82%. These results demonstrate that the aforementioned optimization measures have a significant impact on the model's localization and classification capabilities, resulting in enhanced accuracy in the detection of fire targets in complex backgrounds. In terms of recall and F1-score, the enhanced YOLOV8 also maintains a high level of approximately 88%, achieving an optimal balance between precision and recall, with optimal overall performance.

By comparing model efficiency, it can be observed that the improved YOLOV8 has only 5.7 M parameters, which is significantly lower than the two-stage Faster R-CNN and the traditional single-stage SSD. It also outperforms the YOLOV9 and YOLOV10 models in the same series, making it more lightweight and easier to deploy. The improved model's GFLOPs is 15.9, which, although higher than the original YOLOV8 and similar models, is still lower than YOLOV5 and YOLOV10, placing it at a moderate level. This demonstrates that the improved YOLOV8 model maintains a high inference speed while enhancing accuracy, achieving a relatively balanced trade-off between accuracy and detection speed. The optimization of the model's performance is mainly attributed to the introduction of the CBAM attention mechanism, the WIOU loss function, the improved ShuffleNetV2 network, and the use of the Q-ASFF detection head, which enhance feature utilization while maintaining a certain level of computational complexity.

In conclusion, the series of improvements implemented in the YOLOV8 model in this paper, including attention mechanisms, backbone networks, detection heads, and loss functions, have demonstrably enhanced the model's performance in the mountain fire detection task. The model outperforms traditional object detection models in key performance indicators, including accuracy, speed, and model size. It offers an efficient, accurate, and lightweight solution for the development of a mountain fire monitoring and early warning system.

5.4.4. Multi-Model Scenario Application Comparison

In order to intuitively demonstrate the superiority of the improved YOLOV8 model for mountain fire target detection, three groups of typical scenario images were selected: bright multi-target, complex and dim multi-target, and occluded weak multi-target. A total of nine models were employed for the detection of mountain fire targets. The models included in the study were Faster R-CNN, SSD, YOLOV5, YOLOV7, YOLOV8, YOLOV9, YOLOV10, and the improved YOLOV8. The results are presented in Figures 16 and 17. The detection results of each algorithm include the recognition of flames and smoke, as well as the corresponding confidence scores, which are presented in the form of bounding boxes.

Through analysis, it can be found that in the bright multi-target mountain fire scene, all nine models effectively detect the fire and smoke areas, and the improved YOLOV8 model maintains a relatively high confidence level. For the small target fire areas in the scene, Faster R-CNN and SSD do not make effective identifications. Although YOLOV5-V9 can partially identify the areas where small target fires are located, their range delineation is relatively rough, and there are omissions. Only the improved YOLOV8 model achieves precise detection of smoke areas and all weak fire points. In the complex and dim multi-target scene, due to environmental factors, each model exhibits varying degrees of false detection and missed detection. For example, the YOLOV8, YOLOV9, YOLOV7tiny, and YOLOV5 models mistakenly identify the residential area lights on the right side of the image as flame targets. SSD and Faster RCNN fail to effectively identify the smoke areas in the image. Although the recently released YOLOV10 model achieves precise detection

of dim and weak mountain fire targets, its detection range is relatively rough. Only the improved YOLOV8 model effectively identifies the dim and weak mountain fire targets. Finally, in the multi-weak target scene with dense smoke occlusion, all models effectively identify the mountain fire targets. However, YOLOV8 and YOLOV9 have more precise identification ranges, and the improved YOLOV8 model is the only one that identifies the weak fire point in the upper right corner based on these results, verifying the detection performance of the improved YOLOV8 model for partially occluded targets.

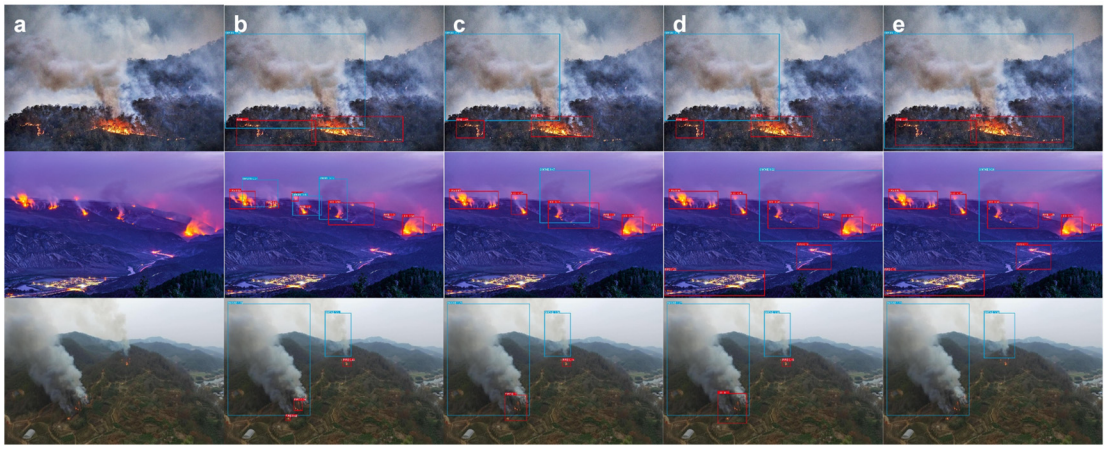

Figure 16. Multi-model scenario application comparison results: (**a**) original image, (**b**) Improved YOLOV8, (**c**) YOLOV10, (**d**) YOLOV9, (**e**) YOLOV8.

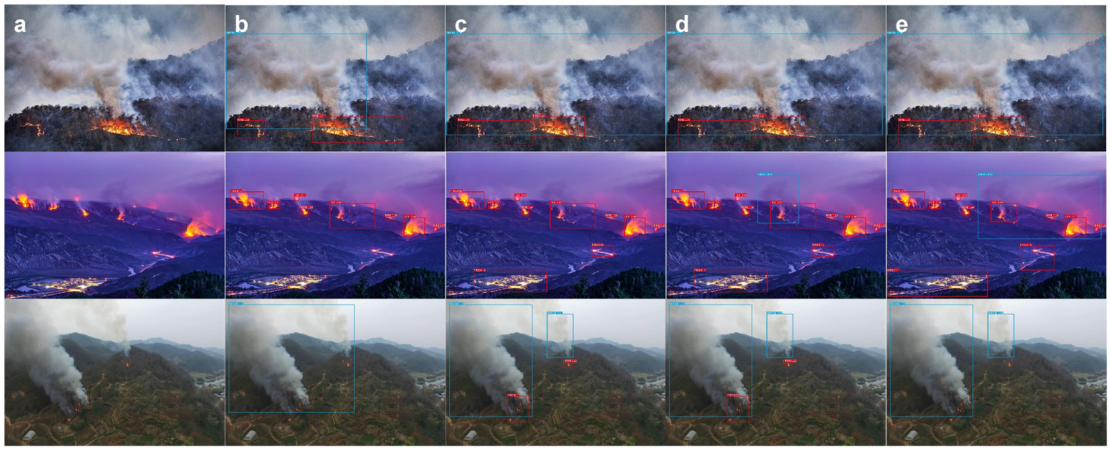

Figure 17. Multi-model scenario application comparison results: (**a**) original image, (**b**) YOLOv7tiny, (**c**) YOLOv5, (**d**) SSD, (**e**) Faster RCNN.

The comparative analysis presented above reveals that the bounding boxes of the enhanced YOLOV8 model exhibit a greater degree of overlap with the boundaries of the flame and smoke regions, thereby enhancing the precision of the detection process. In contrast, other models, such as Faster R-CNN, SSD, and other YOLO series models, generate bounding boxes that are broader or positionally deviated in complex scenarios, which can easily lead to missed and false detections. Moreover, the improved YOLOV8 model exhibits high confidence scores in multiple complex scenarios, thereby demonstrating its superior reliability and efficacy in reducing false alarms and missed detections in practical applications. This enhances the model's practicality in mountain fire detection tasks. It is noteworthy that the enhanced YOLOV8 model exhibits superior performance in the detection of small target fires in images, particularly in scenarios with multiple bright targets and complex, dim, and weak targets. The improved YOLOV8 model is the only one

that successfully marks all small fire points, thereby enabling early warning of mountain fires.

6. Discussion

6.1. Results Discussion

The results of multiple comparative experiments demonstrate that the YOLO-CSQ model proposed in this paper exhibits certain performance advantages in forest fire image recognition tasks. By optimizing the model structure and training strategies, the enhanced YOLOV8 model is capable of precise and comprehensive detection of flame and smoke targets within complex forest fire scenarios. In particular, the YOLO-CSQ model addresses the limitations of the original model in identifying and detecting small and weak targets. This results in significantly enhanced detection accuracy and robustness of small ignition points in fire scenes, providing crucial technical support for early warning and rapid response to forest fires. In comparison to other prevalent object detection models, the YOLO-CSQ model exhibits notable advantages in terms of detection accuracy, real-time performance, and scene adaptability. This evidence substantiates the practical value and extensive applicability of the YOLO-CSQ model in the domain of forest fire monitoring and prevention.

6.2. Limitations and Room for Improvement

Although the enhanced YOLOv8 model demonstrates remarkable detection efficacy in mountain fire detection, it is important to acknowledge that there are still certain limitations and potential for further improvement. First, in terms of data collection, according to the requirements of the Technical Specifications for Drone Surveying of Forest Fires [42], for general forest fires, the flight altitude should be no less than 80 m above the fire scene. Aerial images should be captured from the east, south, west, and north directions, with annotations for frontline reconnaissance time, fire points, smoke points, wind direction indicators, and other key elements. For large and above forest fires, the flight altitude should be no less than 120 m above the fire scene. Panoramic images of the fire scene should be captured upon arrival at the fire scene and after the fire is handled, which will be used for subsequent case studies. For major and especially significant forest fires, the flight altitude should be no less than 200 m. Oblique photography and 3D modeling of the fire scene should be conducted to assist the command department in developing firefighting strategies. However, the data used in this paper, although integrating both publicly available datasets and self-collected image data, are limited by environmental conditions. The simulated fire experiments cannot fully cover all UAV fire data collection scenarios. This limitation indirectly affects the generalization capability and adaptability of the dataset, which still needs further enhancement. In the future, the integration of the improved YOLOv8 model with other technologies (such as semantic segmentation and trajectory prediction) will be considered to achieve more comprehensive and intelligent mountain fire monitoring and early warning.

7. Conclusions

This paper proposes a multi-scale fire detection algorithm based on the improved YOLOV8 network as a means of achieving intelligent and rapid inspection of forest fires. YOLO-CSQ effectively addresses the challenges of severe external interference, the potential for false alarms, and the occurrence of missed detections in mountain fire detection. The main work of this paper is as follows:

(1) The CBAM attention mechanism was introduced to enhance the interaction between cross-scale features, improve the model's ability to detect multi-scale targets, and enhance localization accuracy. The ShuffleNetV2 backbone network was improved to increase the model's convergence speed and generalization ability. The Quadrupled-ASFF detection head was introduced to adaptively adjust the fusion weights of different scale features. The enhanced detection ability of multi-scale targets, particularly

those of a small and weak nature, is achieved through the adoption of the WIoU loss function in place of the original CIoU loss function. This results in an improvement in the model's learning efficiency and generalization ability, while simultaneously providing a more flexible and detailed evaluation mechanism for target detection tasks in complex scenes or extreme conditions. In comparison to the traditional YOLOV8 network, the enhanced YOLOV8 network exhibits superior detection performance, with mAP and F1 score improvements of 8.84% and 9.92%, respectively.

(2) In a series of tests conducted on a range of complex scenarios, including bright multi-target scenes, complex and dim small target scenes, and occluded scenes, the improved YOLOV8 network demonstrated superior detection performance when compared with Faster R-CNN, MobileNetV2, SSD, and other YOLO series models. This network is particularly well-suited to the detection of multi-form fire information in complex forest mountain fire scenarios, offering a high degree of practicality.

Author Contributions: Conceptualization, S.Z. and W.P.; Formal analysis, L.L.; Funding acquisition, W.P.; Methodology, T.L. and S.Z.; Resources, L.L.; Software, T.L. and S.Z.; Supervision, W.P.; Validation, T.L. and S.Z.; Visualization, W.P.; Writing—original draft, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Civil Aircraft Fire Science and Safety Engineering Key Laboratory of Sichuan Province, grant number MZ2024JB01, and the Program of China Sichuan Science and Technology, grant number 2021YFS0319.

Data Availability Statement: All the data can be available request for correspondence author.

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

References

1. Zhao, Y.; Ban, Y. GOES-R time series for early detection of wildfires with deep GRU-network. *Remote Sens.* **2022**, *14*, 4347. [CrossRef]
2. Zhang, Q.; Zhu, J.; Huang, Y.; Yuan, Q.; Zhang, L. Beyond being wise after the event: Combining spatial, temporal and spectral information for Himawari-8 early-stage wildfire detection. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103506. [CrossRef]
3. Ding, Y.; Wang, M.; Fu, Y.; Zhang, L.; Wang, X. A wildfire detection algorithm based on the dynamic brightness temperature threshold. *Forests* **2023**, *14*, 477. [CrossRef]
4. Ji, F.; Zhao, W.; Wang, Q.; Chen, J.; Li, K.; Peng, R.; Wu, J. Coupling physical model and deep learning for Near real-time wildfire detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6009205. [CrossRef]
5. Mohapatra, A.; Trinh, T. Early wildfire detection technologies in practice—A review. *Sustainability* **2022**, *14*, 12270. [CrossRef]
6. Moghadasi, N.; Kulkarni, A.; Crayton, D.; Grissom, R.; Lambert, J.H.; Feng, L. Formal methods in unmanned aerial vehicle swarm control for wildfire detection and monitoring. In Proceedings of the 2023 IEEE International Systems Conference (SysCon), Vancouver, BC, Canada, 17–20 April 2023; pp. 1–8.
7. Qiao, L.; Li, S.; Zhang, Y.; Yan, J. Early Wildfire Detection and Distance Estimation Using Aerial Visible-Infrared Images. *IEEE Trans. Ind. Electron.* **2024**. [CrossRef]
8. Chuang, H.Y.; Kiang, J.F. High-Resolution L-Band TomoSAR Imaging on Forest Canopies with UAV Swarm to Detect Dielectric Constant Anomaly. *Sensors* **2023**, *23*, 8335. [CrossRef]
9. Ba, R.; Song, W.; Li, X.; Xie, Z.; Lo, S. Integration of multiple spectral indices and a neural network for burned area mapping based on MODIS data. *Remote Sens.* **2019**, *11*, 326. [CrossRef]
10. Gigović, L.; Pourghasemi, H.R.; Drobnjak, S.; Bai, S. Testing a new ensemble model based on SVM and random forest in forest fire susceptibility assessment and its mapping in Serbia's Tara National Park. *Forests* **2019**, *10*, 408. [CrossRef]
11. Tang, X.; Machimura, T.; Li, J.; Liu, W.; Hong, H. A novel optimized repeatedly random undersampling for selecting negative samples: A case study in an SVM-based forest fire susceptibility assessment. *J. Environ. Manag.* **2020**, *271*, 111014. [CrossRef]
12. Bar, S.; Parida, B.R.; Pandey, A.C. Landsat-8 and Sentinel-2 based Forest fire burn area mapping using machine learning algorithms on GEE cloud platform over Uttarakhand, Western Himalaya. *Remote Sens. Appl. Soc. Environ.* **2020**, *18*, 100324. [CrossRef]
13. Janiec, P.; Gadal, S. A comparison of two machine learning classification methods for remote sensing predictive modeling of the forest fire in the North-Eastern Siberia. *Remote Sens.* **2020**, *12*, 4157. [CrossRef]
14. Mohajane, M.; Costache, R.; Karimi, F.; Pham, Q.B.; Essahlaoui, A.; Nguyen, H.; Laneve, G.; Oudija, F. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. *Ecol. Indic.* **2021**, *129*, 107869. [CrossRef]
15. Ahmad, K.; Khan, M.S.; Ahmed, F.; Driss, M.; Boulila, W.; Alazeb, A.; Alsulami, M.; Alshehri, M.S.; Ghadi, Y.Y.; Ahmad, J. FireXnet: An explainable AI-based tailored deep learning model for wildfire detection on resource-constrained devices. *Fire Ecol.* **2023**, *19*, 54. [CrossRef]

16. Wang, X.; Pan, Z.; Gao, H.; He, N.; Gao, T. An efficient model for real-time wildfire detection in complex scenarios based on multi-head attention mechanism. *J. Real-Time Image Process.* **2023**, *20*, 66. [CrossRef]
17. Johnston, J.; Zeng, K.; Wu, N. An evaluation and embedded hardware implementation of yolo for real-time wildfire detection. In Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June 2022; pp. 138–144.
18. Mukhiddinov, M.; Abdusalomov, A.B.; Cho, J. A wildfire smoke detection system using unmanned aerial vehicle images based on the optimized YOLOv5. *Sensors* **2022**, *22*, 9384. [CrossRef] [PubMed]
19. Casas, E.; Ramos, L.; Bendek, E.; Rivas-Echeverría, F. Assessing the effectiveness of YOLO architectures for smoke and wildfire detection. *IEEE Access* **2023**, *11*, 96554–96583. [CrossRef]
20. He, H.; Zhang, Z.; Jia, Q.; Huang, L.; Cheng, Y.; Chen, B. Wildfire detection for transmission line based on improved lightweight, YOLO. *Energy Rep.* **2023**, *9*, 512–520. [CrossRef]
21. Li, J.; Tang, H.; Li, X.; Dou, H.; Li, R. LEF-YOLO: A lightweight method for intelligent detection of four extreme wildfires based on the YOLO framework. *Int. J. Wildland Fire* **2023**, *33*, WF23044. [CrossRef]
22. Gonçalves, L.A.O.; Ghali, R.; Akhloufi, M.A. YOLO-Based Models for Smoke and Wildfire Detection in Ground and Aerial Images. *Fire* **2024**, *7*, 140. [CrossRef]
23. Valero, M.M.; Verstockt, S.; Butler, B.; Jimenez, D.; Rios, O.; Mata, C.; Queen, L.; Pastor, E.; Planas, E. Thermal infrared video stabilization for aerial monitoring of active wildfires. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2817–2832. [CrossRef]
24. Bouguettaya, A.; Zarzour, H.; Taberkit, A.M.; Kechida, A. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. *Signal Process.* **2022**, *190*, 108309. [CrossRef]
25. Muksimova, S.; Mardieva, S.; Cho, Y.I. Deep encoder–decoder network-based wildfire segmentation using drone images in real-time. *Remote Sens.* **2022**, *14*, 6302. [CrossRef]
26. Ghali, R.; Akhloufi, M.A.; Mseddi, W.S. Deep learning and transformer approaches for UAV-based wildfire detection and segmentation. *Sensors* **2022**, *22*, 1977. [CrossRef]
27. Garcia, T.; Ribeiro, R.; Bernardino, A. Wildfire aerial thermal image segmentation using unsupervised methods: A multilayer level set approach. *Int. J. Wildland Fire* **2023**, *32*, 435–447. [CrossRef]
28. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO (Version 8.0.0) [Computer Software]. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 6 September 2020).
29. Jocher, G. YOLOv5 by Ultralytics (Version 7.0) [Computer Software]. 2020. Available online: https://zenodo.org/records/7347926 (accessed on 6 September 2020).
30. Turan, M.; Almalioglu, Y.; Araújo, H.; Konukoglu, E.; Sitti, M. Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots. *Neurocomputing* **2017**, *275*, 1861–1870. [CrossRef]
31. Portenoy, R.; Burton, A.; Gabrail, N.; Taylor, D. A multicenter, placebo-controlled, double-blind, multiple-crossover study of Fentanyl Pectin Nasal Spray (FPNS) in the treatment of breakthrough cancer pain. *Pain* **2010**, *151*, 617–624. [CrossRef] [PubMed]
32. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
33. Woo, S.; Park, J.; Lee, J.; Kweon, I. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521. [CrossRef]
34. Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numer.* **1999**, *8*, 143–195. [CrossRef]
35. Su, H.; Wang, X.; Han, T.; Wang, Z.; Zhao, Z.; Zhang, P. Research on a U-Net Bridge Crack Identification and Feature-Calculation Methods Based on a CBAM Attention Mechanism. *Buildings* **2022**, *12*, 1561. [CrossRef]
36. Ran, H.; Wen, S.; Wang, S.; Cao, Y.; Zhou, P.; Huang, T. Memristor-Based Edge Computing of ShuffleNetV2 for Image Classification. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2021**, *40*, 1701–1710. [CrossRef]
37. Guillet, J.; Valdez-Nava, Z.; Golzio, M.; Flahaut, E. Electrical properties of double-wall carbon nanotubes nanocomposite hydrogels. *Carbon* **2019**, *146*, 542–548. [CrossRef]
38. Hu, B.; Gao, B.; Woo, W.; Ruan, L.; Jin, J.; Yang, Y.; Yu, Y. A Lightweight Spatial and Temporal Multi-Feature Fusion Network for Defect Detection. *IEEE Trans. Image Process.* **2020**, *30*, 472–486. [CrossRef]
39. Cho, Y. Weighted Intersection over Union (wIoU): A New Evaluation Metric for Image Segmentation. *arXiv* **2021**, arXiv:2107.09858.
40. Shamsoshoara, A.; Afghah, F.; Razi, A.; Zheng, L.; Fulé, P.; Blasch, E. Aerial Imagery Pile burn detection using Deep Learning: The FLAME dataset. *Comput. Netw.* **2020**, *193*, 108001. [CrossRef]
41. El-Madafri, I.; Peña, M.; Olmedo-Torre, N. The Wildfire Dataset: Enhancing Deep Learning-Based Forest Fire Detection with a Diverse Evolving Open-Source Dataset Focused on Data Representativeness and a Novel Multi-Task Learning Approach. *Forests* **2023**, *14*, 1697. [CrossRef]
42. *DB 43/T 2512-2022*; Technical Specification for Investigation of Forest Fire of Unmanned Aerial-Vehicle. Administration for Market Regulation of Hunan Province: Changsha, China, 2022.