remote sensing

MDPI

Article

LMDFS: A Lightweight Model for Detecting Forest Fire Smoke in UAV Images Based on YOLOv7

Gong Chen [1,†], Renxi Cheng [1,†], Xufeng Lin [1], Wanguo Jiao [1], Di Bai [2] and Haifeng Lin [1,*]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; wgjiao@njfu.edu.cn (W.J.)
[2] College of Information Management, Nanjing Agricultural University, Nanjing 210095, China; baidi000@njau.edu.cn
* Correspondence: haifeng.lin@njfu.edu.cn; Tel.: +86-25-8542-7827
† These authors contributed equally to this work.

Abstract: Forest fires pose significant hazards to ecological environments and economic society. The detection of forest fire smoke can provide crucial information for the suppression of early fires. Previous detection models based on deep learning have been limited in detecting small smoke and smoke with smoke-like interference. In this paper, we propose a lightweight model for forest fire smoke detection that is suitable for UAVs. Firstly, a smoke dataset is created from a combination of forest smoke photos obtained through web crawling and enhanced photos generated by using the method of synthesizing smoke. Secondly, the GSELAN and GSSPPFCSPC modules are built based on Ghost Shuffle Convolution (GSConv), which efficiently reduces the number of parameters in the model and accelerates its convergence speed. Next, to address the problem of indistinguishable feature boundaries between clouds and smoke, we integrate coordinate attention (CA) into the YOLO feature extraction network to strengthen the extraction of smoke features and attenuate the background information. Additionally, we use Content-Aware Reassembly of FEatures (CARAFE) upsampling to expand the receptive field in the feature fusion network and fully exploit the semantic information. Finally, we adopt SCYLLA-Intersection over Union (SIoU) loss as a replacement for the original loss function in the prediction phase. This substitution leads to improved convergence efficiency and faster convergence. The experimental results demonstrate that the LMDFS model proposed for smoke detection achieves an accuracy of 80.2% with a 5.9% improvement compared to the baseline and a high number of Frames Per Second (FPS) - 63.4. The model also reduces the parameter count by 14% and Giga FLoating-point Operations Per second (GFLOPs) by 6%. These results suggest that the proposed model can achieve a high accuracy while requiring fewer computational resources, making it a promising approach for practical deployment in applications for detecting smoke.

Keywords: deep learning; forest fire smoke detection; Ghost Shuffle Convolution; coordinate attention; CARAFE; SIoU; Yolov7

Citation: Chen, G.; Cheng, R.; Lin, X.; Jiao, W.; Bai, D.; Lin, H. LMDFS: A Lightweight Model for Detecting Forest Fire Smoke in UAV Images Based on YOLOv7. *Remote Sens.* **2023**, *15*, 3790. https://doi.org/10.3390/rs15153790

Academic Editors: Javaan Chahl, Huajian Liu, Asanka Perera and Ali Al-Naji

Received: 24 June 2023
Revised: 24 July 2023
Accepted: 28 July 2023
Published: 30 July 2023

1. Introduction

Forests, as one of the most valuable resources in nature, play a crucial role in ecological functions, such as preventing wind erosion and conserving water and soil. On the other hand, forests also have enormous economic value for humans. Forest fires often lead to severe consequences such as soil erosion, air pollution, and threats to animal survival, causing significant ecological and economic damage [1]. Therefore, the early detection and control of forest fires are crucial. Smoke, as an important precursor to forest fires, can be effectively monitored for their detection and control, which is significant for their suppression [2].

The detection of forest smoke has gone through various developmental stages, including manual inspections, instrument-based detection, and detection based on computer vision. Manual inspections require a high level of manpower and material resources and have a low efficiency. Moreover, detection results often fail to meet expectations. Instrument-based detection mainly depends on various detectors and sensors from the past two decades. However, instruments are prone to interference from small particles, such as dust in the environment [3]. Additionally, they only trigger an alarm when the concentration of smoke reaches a threshold. Due to the complexity of outdoor air flow and other environmental factors, a fire may become difficult to control by the time the alarm goes off. Therefore, this method has gradually been abandoned. In the phase of detection based on computer vision, pattern recognition is used for feature extraction and classification to achieve the identification of forest smoke [4]. Gubbi et al. [5] used wavelets to extract the features of smoke and then classified smoke using a support vector machine (SVM). H. Cruz et al. [6] proposed a new color detection index for detecting the colors of flames and smoke. This method enhances the color by normalizing the RGB channel color and mainly combines the features of the motion and color of smoke to obtain the regions of flames and smoke through thresholding. Prema et al. [7] used a comprehensive approach to detect smoke, which included the YUV color space and wavelet energy, taking the relationship and contrast of smoke into account. However, due to the limitations of human experience, it is subject to various forest environments. In summary, although some progress has been made compared to instrument-based detection, traditional image detection methods have difficulty extracting the intrinsic features of smoke. The time required for detection is also too long, and the rate of false alarms is high, with poor generalization ability.

In recent years, with the rapid development of artificial intelligence, drones with deep learning have injected strong development momentum into detection via computer vision [8]. Due to their high accuracy, real-time performance, strong robustness, and low cost, deep-learning-based detection algorithms of smoke are widely applicable in various complex scenarios and hold great research value. Convolutional neural networks (CNNs) can achieve the high-precision recognition of the data of a two-dimensional image, and researchers have attempted to apply CNNs in the detection of smoke. Salman Khan et al. [9] comprehensively studied various detection algorithms and proved that the CNN has a high accuracy in smoke detection tasks. Additionally, the detection of smoke is often prone to errors due to the complexity of the background. In outdoor environments, such as forests, interferences such as clouds in the sky, reflections in lakes, and changes in lighting can easily cause false alarms [10]. Therefore, many scholars have proposed algorithms for improvement. Xuehui Wu et al. [11] used algorithms of background subtraction and achieved good results in the detection of dynamic smoke. The rate of false detection for classifying clouds reflected from sunlight was reduced, but the rate of false detection for newly formed objects remained high. Yin et al. [12] adjusted the parameters according to changes in the actual environment and thus could accurately detect smoke in different conditions. Zhang, Q. et al. [13] constructed a simulated smoke dataset and trained it using the proposed deep convolutional generative adversarial network. They effectively monitored smoke areas and reduced false alarms, but their method was demanding in terms of hardware and difficult to widely deploy to meet real-time requirements. Lightweight models are widely used in practical tasks by virtue of their lower energy consumption and faster inference speeds. Guo, Y. et al. [14] used the constructed S-Mobilenet module to realize the lightweight YOLO model for the problem of the real-time detection of ship targets of a smaller size and evaluated its effectiveness on hardware devices. However, there is still the problem of weak applicability in real tasks. Li, W. et al. [15] developed the lightweight WearNet based on a novel convolutional block, which can be deployed with embedded devices for the detection of scratches. Although all of the above achieved good results, there are still problems in existing research on smoke detection. Sheng, D. et al. [16] used a CNN network and linear iterative clustering (SLIC) for smoke

image segmentation and applied density-based spatial clustering of applications with noise (DBSCAN), which can achieve faster detection. However, their proposed method has a low FPR rate, which indicates high model sensitivity and needs further improvement.

In summary, the deep-learning-based detection algorithms of smoke mentioned above have achieved considerable success, but there are three problems when it comes to actually using edge equipment for detection. Firstly, models of a large network have a huge number of parameters and high hardware requirements, making it difficult to deploy them for practical tasks and meet real-time requirements for the detection of smoke. Secondly, existing lightweight models can detect smoke more quickly under the same conditions, but their accuracy of detection is often far lower than that of models with a large network. For detecting things with thin features, such as smoke, the fusion of the features is often incomplete, which leads to a lower accuracy of detection. Therefore, there is the problem of a performance imbalance between the accuracy and speed of detection. Thirdly, so-called small smoke is a type of smoke produced in the early stages of a forest fire and is characterized by a small volume and thinness. Thin and small smoke cannot effectively extract information due to the small number of features it can extract. It is more difficult to detect than typical smoke that has already taken shape and is susceptible to disturbances, such as lens impurities. This leads to the problem of UAVs obtaining noisy images during detection missions, which can cause missed detections [17] as well as false detections caused by interfering objects, such as cloud cover [18]. These make the detection of forest smoke a major challenge.

In order to solve the problems above, a lightweight model for detecting forest fire smoke based on YOLOv7 [19] is proposed in this paper. (1) To address the problem of the original model being large in size and difficult to deploy in real edge devices, we use GSConv to replace the standard convolution in the neck layer and construct fast pyramid pooling modules by using GSELAN and GSSPPFCSPC, based on GSConv. This can speed up the model convergence and fuse the features of smoke at a faster rate with less computation when dealing with images of smoke. (2) Considering the blurred feature boundaries of smoke-like objects and smoke, it is very easy to confuse the detection of clouds and that of smoke from forest fires in a forest environment. There is the problem of low interclass heterogeneity, and the foreground and background of images of smoke are difficult to effectively distinguish, which can cause false detection. In response, we embed multilayer coordinate attention in the backbone network, which improves its ability to distinguish between the smoke and background by effectively fusing the channel relations and location information, focusing on the location of interest to the network, suppressing useless information, and improving the separation of clouds and smoke. (3) Thin and fine smoke cannot carry sufficient information due to its inconspicuous features, which also weakens the accuracy of smoke detection. Moreover, the use of the CARAFE upsampling operator allows us to extract information more fully from the image by expanding the sensory fields in order to effectively improve the detection accuracy of small targets. The SIoU loss function is used to improve the speed and accuracy of inference during model training.

2. Materials and Methods for Experiments

2.1. YOLOv7

YOLOv7 is the latest version of the series of YOLO [20–23]. Compared to the previous YOLOv5, it surpasses all known detectors in terms of both speed and accuracy. This is because it uses faster convolution operations and a smaller model, which allows it to achieve higher accuracy and faster speed at the same computational cost when detecting.

The model of YOLOv7 mainly consists of four main parts: input, backbone, neck, and head. The backbone is mainly composed of multiple modules, including CBS, ELAN [24], and MPConv, and is used for feature extraction in image analyses. ELAN enhances feature

aggregation by connecting the outputs of multiple layers of convolutions. For neck, the traditional FPN + PAN [25,26] structure is still used to achieve the integration of high-resolution information and high-level semantic information through the fusion of high-level features and underlying features. The head utilizes a reparameterization structure, Rep, which enhances the representational capability when training and facilitates faster inference when testing.

The official network framework based on YOLOv7 contains YOLOv7, YOLOv7-d6, and YOLOv7-e6, etc., which are different from each other. We choose the original framework of YOLOv7. The structure of the network model of YOLOv7 is shown in Figure 1. Considering the problem of huge number of parameters in YOLOv7, we modified the depth multiple of YOLOv7 to 0.33 and the width multiple to 0.5, following the practice of YOLOv5, in order to maintain the original inference effect as much as possible while reducing the number of parameters in the model.

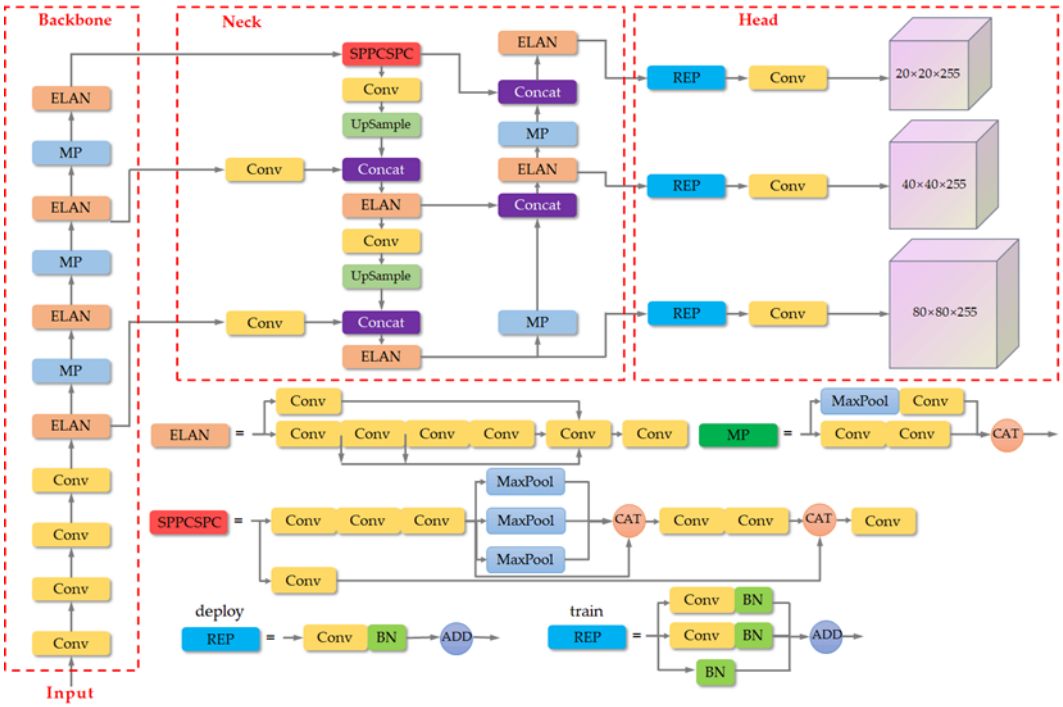


Figure 1. The network architecture of YOLOv7.

2.2. Improvements to Lightweighting

2.2.1. Ghost Shuffle Convolution

Standard convolution (SConv) applies different convolutional kernels to multiple channels simultaneously, leading to an increase in the number of parameters required and a decrease in the speed of network as the network tries to extract more features. Conversely, depth-wise separable convolution (DWConv) stitches the outputs of separate depth-wise convolutions by using a 1 × 1 convolution kernel after convolution of the channels, allowing it to save a significant number of parameters as the features to be extracted increase, resulting in faster inference. However, as a trade-off, DWConv also leads to the loss of some semantic information during operation, which reduces accuracy of the model.

The process of GSConv [27] is shown in Figure 2, which combines the advantages of standard convolution and depth-separable convolution. It uses SConv and DWConv together when handling the input images of forest fire smoke, and it does not completely cut off the links between channels as DWConv does, but tries to preserve the links as much as possible in order to ensure the high accuracy of the model. The results are combined and shuffled to enhance the nonlinear representation. For smoke targets that change with

fire and environmental conditions, these nonlinear features can better represent the processes of deformation and expansion of smoke, providing more information for model to learn and thus enhancing its generalization ability and robustness. The mathematical expression is calculated as follows:

$$X_c = \sigma(bn(Conv2d(X_{input}))) \quad (1)$$

$$X_{out} = \delta(X_c \oplus DWConv(X_c)) \quad (2)$$

where $Conv2d$ represents the two-dimensional convolution of the input image X_{input}, bn represents the normalization operation, σ represents the activation function, $\oplus$ represents the concating of the two kinds of convolution, and the final δ represents shuffle, aiming to obtain the final output X_{out} by shuffling this result.

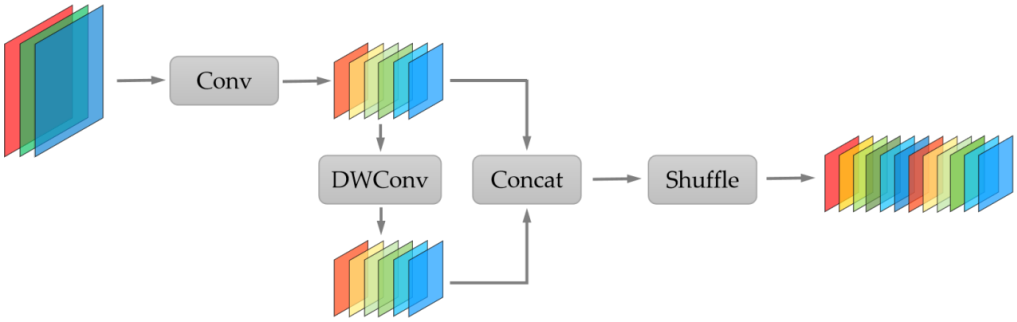


Figure 2. Structure of the model Ghost Shuffle Convolution.

However, if GSConv is used in all phases of the model, the number of layers of the model will increase significantly, which will increase the inference time for fast detection of smoke targets. Therefore, it would be better to use GSConv only in a single stage. In the network framework of YOLOv7, for the backbone layer, which requires a large amount of convolution to extract enough smoke features, a great degree of correlation between the channels that the standard convolution has is necessary. Therefore, we only perform convolutional operation replacement in the neck layer. This will reduce the redundant and repetitive information, thus reducing the computational cost and achieving a lightweight model. In this paper, to further exploit the role of GSConv, we make a further improvement in ELAN by using GSELAN module to replace the W-ELAN block in the neck layer. The constructed GSELAN structure is shown in Figure 3.

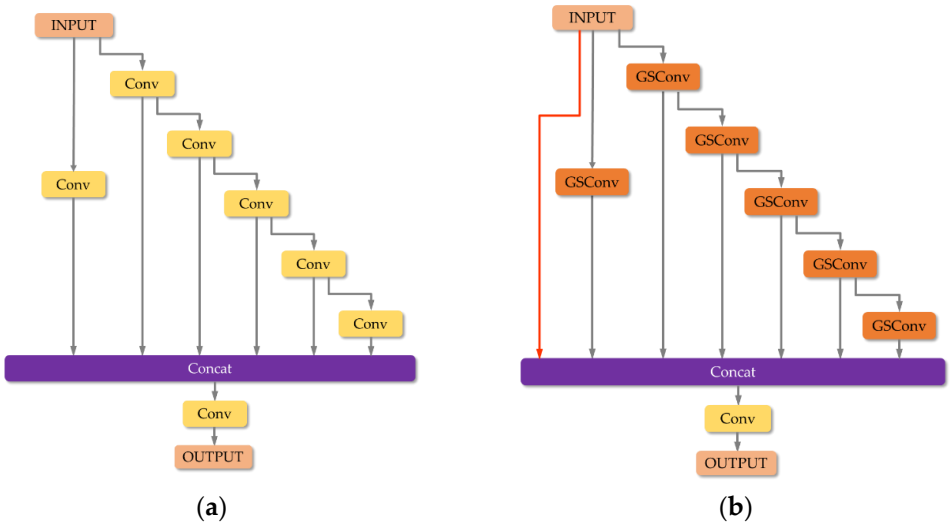


Figure 3. ELAN model before and after improvement: (**a**) Structure of W-ELAN; (**b**) Structure of GS-ELAN.

By replacing the standard convolution with GSConv, which aims to reduce the computational complexity and the number of parameters, a larger speedup can be obtained during the actual run. The input smoke image is convolved by successive GSConv convolutions, and each shuffle operation is able to effectively fuse the smoke feature maps of different channels with a reduced number of parameters, thus approximating the result of the standard convolution. This allows the final output smoke image to take advantage of DSC while mitigating the negative impact of DSC deficiencies on the model. In addition, we add identity mapping [28] to the module, which can effectively avoid information loss of smoke features during transmission and enhance the robustness of the model by mapping the input directly to the output.

In addition, in the original YOLOv7, we note that it uses the SPPCSPC module to expand the perceptual field of the model by combining a Maxpool branch and a convolution branch, which has a better feature fusion effect compared with that of SPPF when dealing with targets of different scales, but introduces a large number of parameters and a large amount of computation. In this paper, we borrow the idea of SPPF and improve the SPP [29] branch in SPPFCSPC to be similar to the SPPF method by replacing the original parallel pooling with successive max-pooling operations, which can eliminate more redundant information and noise in the smoke image and make the obtained feature maps of the smoke image have better coherence. Its computational speed is further optimized while preventing overfitting. In addition, after the feature extraction of backbone layer, the attribute information and multidimensional channel information of the target image are obtained. We replace the convolution of SPPFCSPC with GSConv, which not only can reduce the cost of computation, but also can preserve the connection between each channel as much as possible. The improved GSSPPFCSPC module is shown in Figure 4.

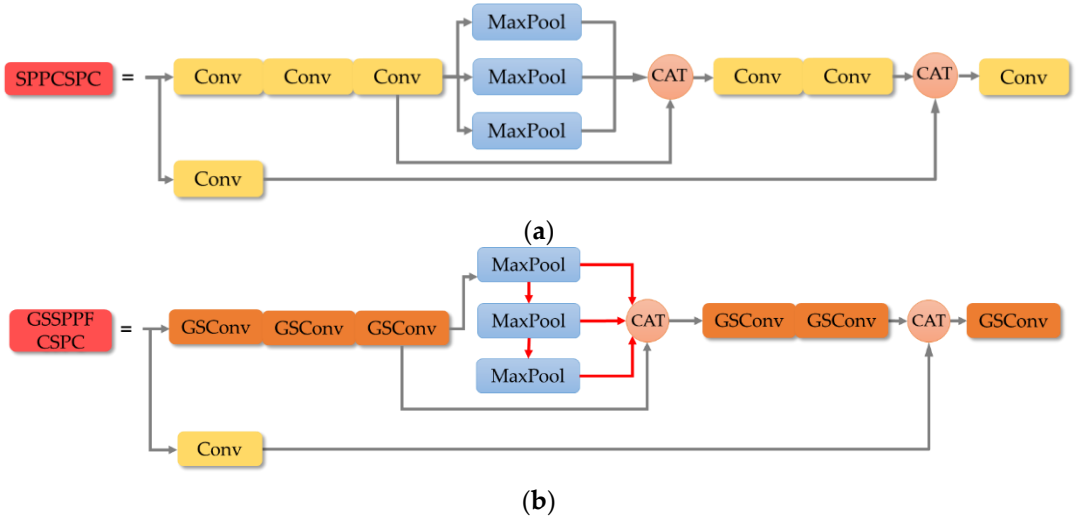


Figure 4. Spatial pyramid pooling module before and after improvement: (**a**) Structure of SPPCSPC; (**b**) Structure of GSSPPFCSPC.

2.2.2. Improvements in the Activation Function

In YOLOv7, the activation function SiLU is still used. We aim to obtain a lightweight model for detection of smoke that can maintain high accuracy, and activation functions are crucial for the computational accuracy and speed of the model. The Hardswish function is defined as shown in Equation (2), and it has the characteristics of having no upper bound, having a lower bound, smoothness, and non-monotonicity, which makes the processing of the detection for activation layer more diverse. The advantages of the Hardswish function are two-fold: firstly, it uses linear interpolation and has good numerical stability and fast calculation speed, which help to make the expression of the model

for smoke detection more diverse. Secondly, the Hardswish function has a simple derivation and can effectively prevent the phenomenon of neurons being difficult to activate by any data due to gradients approaching zero during the training of the model for smoke detection.

$$Hardswish(x) = \begin{cases} 0, & \text{if } x \le -3, \\ \dfrac{x(x+3)}{6}, & \text{if } -3 < x < 3, \\ x, & \text{if } x \ge 3. \end{cases} \tag{3}$$

2.3. Improvements in Accuracy

2.3.1. Coordinate Attention

The use of attention mechanism in the field of image recognition can effectively help the network to better address the attention preference of the network and focus on the region of its interest. In this paper, we try to embed the coordinate attention (CA) [30] mechanism, which is able to obtain a better description of the smoke target by re-modeling both channel and spatial dimensions to capture both orientation perception and location information simultaneously. The flowchart is shown below (Figure 5).

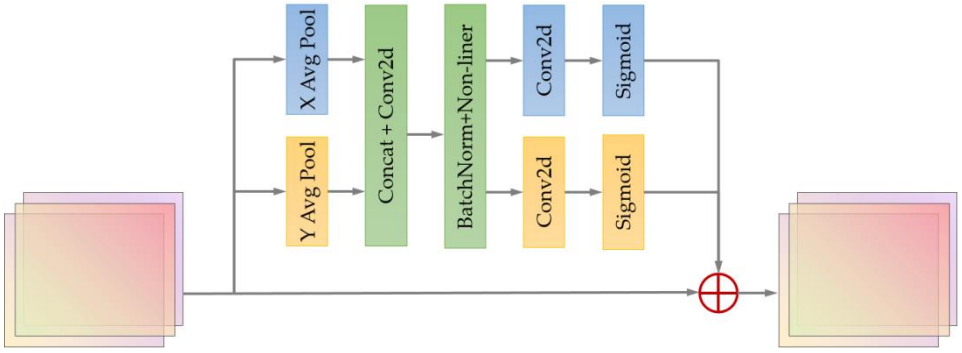


Figure 5. Flowchart of attention mechanism CA.

When the feature matrix X is input, the overall flow is shown below:

(1) Firstly, shift adaptive averaging pooling is performed simultaneously along the horizontal and vertical directions, respectively, and its mathematical expression is as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < \le W} x_c(h, i) \tag{4}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < \le H} x_c(j, w) \tag{5}$$

The above expression represents two averaging pooling operations using two one-dimensional global pooling kernels $(h, 1)$ and $(1, w)$ for the image $x_c(h,w)$ of the cth channel of the input along the vertical and horizontal directions, respectively, to obtain the pooling results $z_c^h(h)$ and $z_c^w(w)$ in both directions.

(2) The two outputs obtained above are then stitched together and 1×1 convolved, and the flow is shown below:

$$f = \delta(C_1[z^h; z^w]) \tag{6}$$

where *C*1 represents the 1 × 1 convolution, δ represents the nonlinear activation function, and *f* represents the result of aggregated coding from the two directions of feature vector.

(3) Next, *f* is expanded along two dimensions, h and w, to obtain the feature attention maps f^h and f^w in two directions, and convolution operations are performed to obtain g^h and g^w, respectively.

$$g^h = \sigma(C_h(f^h)) \tag{7}$$

$$g^w = \sigma(C_w(f^w)) \tag{8}$$

where C_h and C_w represent different convolution operations in two directions, respectively, and σ is activation function's sigmoid. Thus, the attention weights of the two directions are obtained.

(4) Finally, the attention weights and the original feature maps are multiplied and weighted to obtain the final output as follows:

$$y_{out} = x \times g^h \times g^w \tag{9}$$

where *x* is the graph of the original special diagnosis, g^h represents the attentional weight along the direction h, and g^w represents the attentional weight along the direction w.

For the thin and small smoke images, it is difficult to effectively extract information, so we add a multi-layer CA mechanism to the backbone network. For each input image, the feature weight of the h direction and the feature weight of the w direction are compared with the original. The weighted fusion of images strengthens the focus on the region of interest, and an output image is obtained focused on the smoke target to enhance the model's ability to capture and identify smoke. Meanwhile, if the model's attention is limited to some local areas, it will miss the grasp of the overall features of the smoke, thus increasing the rate of false alarms. The introduction of CA can focus the model's attention on the key feature areas, so that the network can grasp the overall features of the smoke from a global perspective, thus improving the recognition accuracy of the model.

2.3.2. Content-Aware Reassembly of Features

In YOLOv7, nearest-neighbor interpolation upsampling is used, which is widely applied due to its simplicity and low computational cost. However, nearest-neighbor interpolation only considers adjacent pixels, resulting in the failure to fully utilize the semantic information of the feature map. CARAFE [31], on the other hand, effectively extracts semantic information from the feature map and expands the receptive field under the premise of lightweight operation. CARAFE consists of the upsampling prediction module and the feature recombination module, as shown in Figure 6. For the upsampling prediction module, the first step is to process the input image with size $H \times W \times C$. When the upsampling factor is set to σ, a 1×1 convolutional layer is applied to compress the image channel. Then, convolutional kernels of size $k_{up} \times k_{up}$ are applied for convolutional operations, expanding the number of channels to $\sigma^2 \times k_{up}^2$ for content encoding. Finally, the output is normalized to reduce the number of parameters. In the feature recombination module, point-wise multiplication is performed on corresponding positions of the output feature map obtained through the above process and the feature map obtained through traditional upsampling, resulting in the final output value.

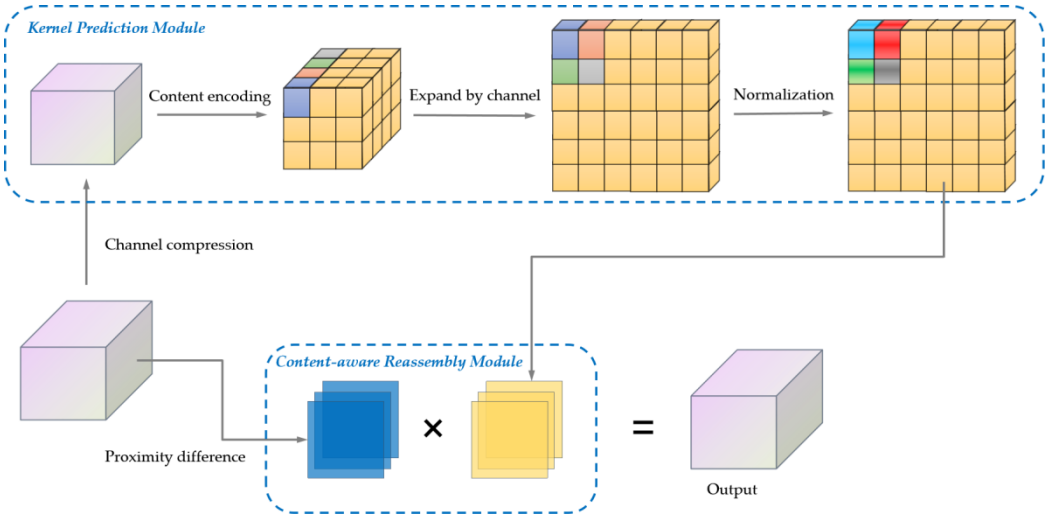


Figure 6. Sampling on CARAFE.

Smoke images have characteristics such as thinness and limited information. The upsampling mechanism of CARAFE can be utilized to interpolate low-resolution feature maps using learnable interpolation weights, reducing information loss, and restoring details. Additionally, CARAFE introduces a context-adaptive information fusion mechanism that dynamically adjusts interpolation weights based on local contextual information, enabling it to better capture fine features of smoke and reduce errors. In this paper, we replace the original upsampling method with lightweight CARAFE to obtain a better feature map of smoke.

2.4. SCYLLA-Intersection over Union

In YOLOv7, CIoU loss [32] is still used to compute the localization loss function, which was also utilized in YOLOv5, but the inherent properties of smoke require issue of mismatched angles between the predicted and ground-truth bounding boxes to be considered, which was not addressed in CIoU. SIoU [29] addresses this problem by introducing the vector angle between the predicted and ground-truth bounding boxes, fully taking into account the direction between them in the process of smoke detection, thereby speeding up the convergence rate of the model. The redefined loss function consists of four parts: the angle cost (which measures the difference in angles between two objects or shapes), distance cost (which evaluates the spatial separation or distance between two objects), shape cost (which assesses the dissimilarity in shape characteristics between two objects), and IoU cost (which calculates the intersection over union (IoU) value, representing the overlapping area between two objects divided by their combined area). The specific formulas are as follows.

(1) Angle cost:

$$\Lambda = 1 - 2 \times \sin^2 \left(\arcsin \left(\frac{c_h}{\sigma} \right) - \frac{\pi}{4} \right) \tag{10}$$

where

$$\frac{c_h}{\sigma} = \sin(\alpha) \tag{11}$$

$$\sigma = \sqrt{(b_{c_x}^{gt} - b_{c_x})^2 + (b_{c_y}^{gt} - b_{c_y})^2} \tag{12}$$

where α is the angle between side σ and side c_w; c_w, c_h, and σ are the three sides of the right triangle in Figure 7; $(b_{c_x}^{gt}, b_{c_y}^{gt})$ are the center point coordinates of the ground-truth box; and (b_{c_x}, b_{c_y}) are the center point coordinates of the predicted box.

(2) Distance cost:

$$\Delta = \sum_{t=x,y} \left(1 - e^{-(2-\Lambda)\times\rho_t}\right) \tag{13}$$

where

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2,\ \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_x}}{c_w}\right)^2 \tag{14}$$

The incorporation of angle cost and distance cost results in larger loss values when there is a greater difference between the angles of the two boxes, promoting a faster convergence rate.

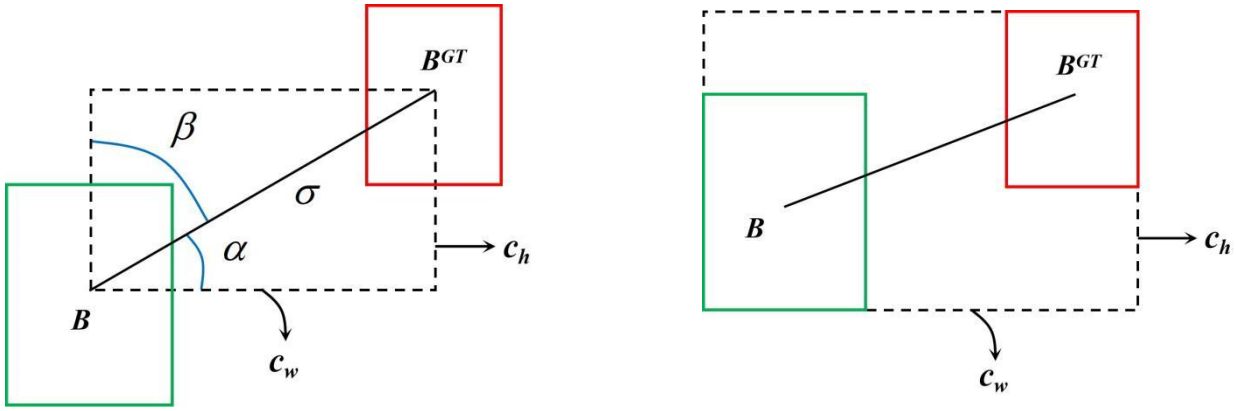


Figure 7. Schematic diagram of SIoU.

(3) Shape cost:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-w_t}\right)^{\theta} = \left(1 - e^{-w_w}\right)^{\theta} + \left(1 - e^{-w_h}\right)^{\theta} \tag{15}$$

where

$$w_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, h_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{16}$$

θ controls the degree of attention paid to the shape cost, where w^{gt} and h^{gt} are the width and height of the ground-truth box, and w and h are the width and height of the predicted box. In this paper, following the recommendation of our original paper, θ is set to 4, in order to have a more reasonable focus on the shape cost.

(4) IoU cost:

$$IoU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \tag{17}$$

where $B \cap B^{GT}$ represents the intersection between the predicted and ground-truth boxes, and $B \cup B^{GT}$ represents their union.

The final loss function is as follows:

$$SIoU\_Loss = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{18}$$

2.5. A Lightweight Model for Detecting Forest Fire Smoke Based on YOLOv7

In summary, the overall structure of the modified YOLOv7 used in this paper is shown in Figure 8, and the changes are framed by the solid green line.

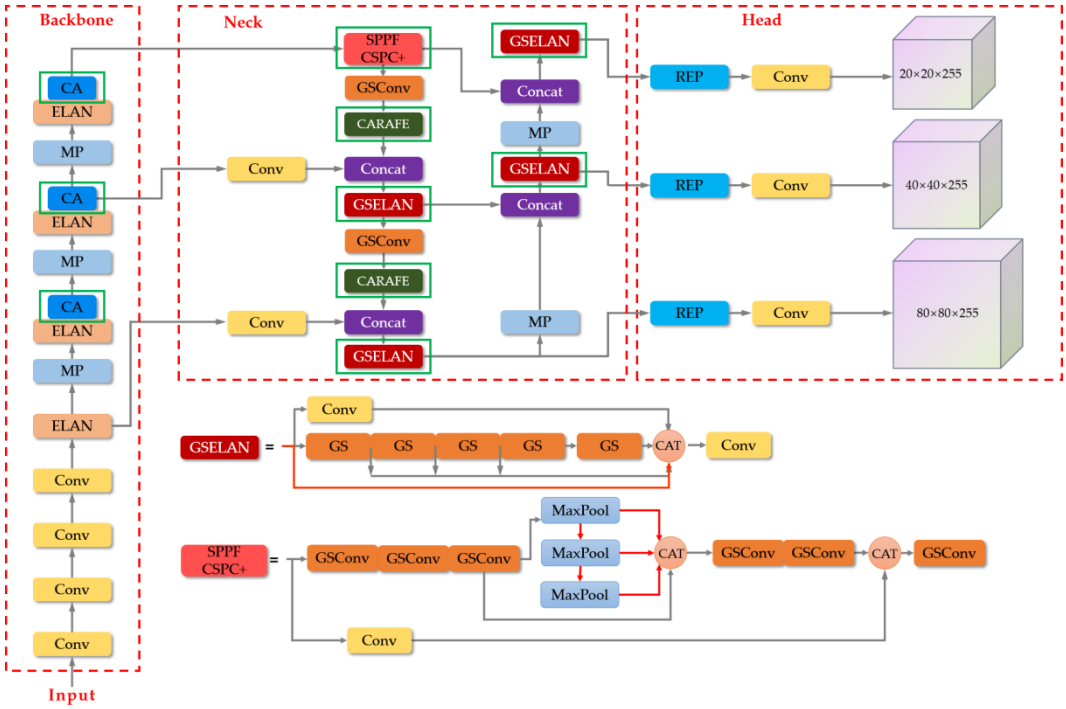


Figure 8. The network architecture of modified YOLOv7.

3. Methods for Evaluation

3.1. The Dataset

It is well known that in the field of deep learning, the quality of the training set is directly related to the performance of the detection results. After extensive information search and inquiry, we learned that there are no standardized and reasonable datasets of smoke in forest environment on the web. Therefore, a unique dataset is essential. In this thesis, firstly, we browsed photos of forest fire smoke from drones on the internet with a high point of view. Since the oblique view is the angle at which forest smoke drones usually perform detection and identify it faster, our dataset consists mainly of forest fire smoke images in oblique view, supplemented by a blend of some nadir images. These images comprise typical forest fire smoke in a forest background, thin and small smoke photographed from a distance, and smoke with disturbances such as clouds. In addition, we note that the use of synthetic smoke in [33] can increase the diversity of training data and improve the robustness of the model. Therefore, we used the method of synthesizing smoke to construct the dataset by adding some synthetic images. By copying smoke layers into different background environments or moving smoke layers to different locations in the same image, we can make full use of the limited background environment and smoke image resources to create more scenarios of forest fire smoke and thus effectively improve the generalization ability of the model.

Eventually more than 4019 images were grabbed from the web and some of them were synthesized using synthetic smoke. A total of 5311 images were integrated as the dataset for this study. Some of the images from the dataset are shown in Figure 9 below. It contains a variety of smoke images from the viewpoint of UAVs, such as typical forest fire smoke, small smoke, smoke with distractors, and synthetic smoke, which is mentioned above. The classification of various types of images with different viewpoints is shown in Table 1.

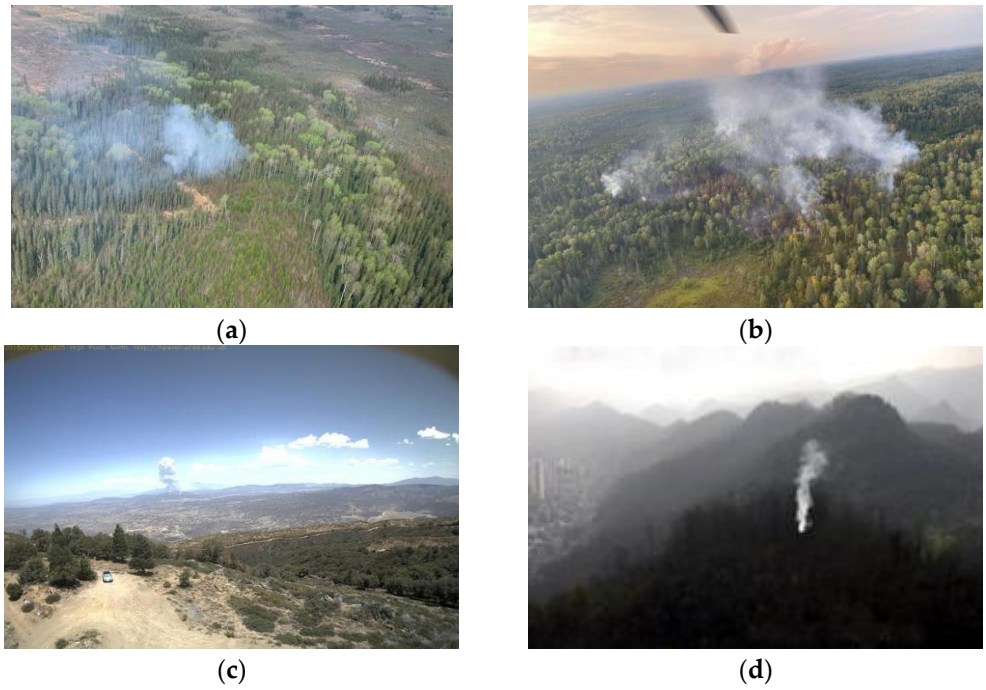


Figure 9. Typical representative images of a dataset of forest fire smoke. (**a**) Normal smoke; (**b**) Smoke with multiple scales; (**c**) Smoke with interference containing something similar to smoke in images; (**d**) Synthetic smoke.

Table 1. Details of the dataset.

Taken From an Overhead Perspective	Taken at an Oblique Angle			
Normal smoke	Normal smoke	Small smoke	Smoke with smoke-like interference	Synthetic smoke
377	1017	968	1657	1292

In the experiments of this paper, the dataset of forest fire smoke was randomly split into training, validation, and test sets in the ratio of 8:1:1. The specific number of images in each set is shown in Table 2.

Table 2. Number of images in each set.

Dataset	Train	Validation	Test	Summary
Number	4249	531	531	5311

3.2. Evaluation of the Model

This study evaluates the quality of the model from two aspects: accuracy of recognition and lightweight degree of the model. Therefore, AP@.5 and AP@.5:.95 are selected as two indicators to evaluate prediction accuracy of the model. Gigabit floating point operations per second (GFLOPs), frames per second (FPS), and parameters are chosen as three indicators to evaluate lightweight degree of the model.

(1) AP indicators: In the confusion matrix, TP refers to the number of smoke samples that are correctly predicted as smoke, FN refers to the number of smoke samples that are incorrectly predicted as non-smoke, FP refers to the number of non-smoke samples that are incorrectly predicted as smoke, and TN refers to the number of non-smoke samples that are correctly predicted as non-smoke. Based on these, precision rate (P) and recall rate (R) can be defined, where P reflects the accuracy of the smoke detection and R reflects the completeness of the smoke detection. The formulas for calculating P and R are shown in Equation (19) and Equation (20), respectively.

$$P = \frac{TP}{TP + FP} \tag{19}$$

$$R = \frac{TP}{TP + FN} \tag{20}$$

AP is the area under the PR curve and is used to describe the average accuracy of forest smoke detection. Its formula is shown in Equation (21).

$$AP = \frac{1}{r}\sum_{i=1}^{r} P_i \tag{21}$$

(2) GFLOPs: GFLOPs is used to describe the time complexity of the model, which is positively correlated with the performance of the hardware required. The formula for calculating GFLOPs is shown in Equation (22).

$$GFLOPs = (2C_i K^2 - 1)HWC_0 \tag{22}$$

where C_i and C_0 represent the number of input and output channels, K represents the size of the kernel, and H and W are used to describe the size of the feature map.

(3) FPS: FPS stands for the number of images that can be processed per second. Time refers to the amount of time required to process each image frame, including image preprocessing, inference, and non-maximum suppression. The formula for calculating FPS is shown in Equation (23).

$$Time = Pre - process + Inference + NMS \tag{23}$$

Therefore, FPS can be used to describe the speed of model detection. Its value is equal to the number of images the model processes per second. The formula for calculating FPS is as follows:

$$FPS = \frac{1}{Time} \tag{24}$$

(4) Parameters: Parameters represents the number of parameters the model uses, measured in millions. It affects the final size of the output model after training.

3.3. Comparison with Other Models

It is essential to compare the detection effects of our model with those of various other mainstream networks, so as to further verify the effectiveness of the network model we proposed. We chose the networks Faster R-CNN [34], EfficientNet [35], SSD [36], Retinanet [37], and YOLOv5, and a brief description of these models is shown below.

(1) Faster R-CNN: Faster R-CNN is a popular object detection model that combines region proposal network (RPN) and Fast R-CNN. It achieves high accuracy but has a slower inference speed compared to that of other models.
(2) EfficientDet: EfficientDet is a state-of-the-art object detection model that achieves high accuracy while being efficient in terms of computation. It uses a compound scaling method to balance accuracy and efficiency.
(3) SSD: SSD (Single Shot MultiBox Detector) is a fast object detection model that achieves real-time performance. It uses multiple layers for predicting bounding boxes and class probabilities but may have lower accuracy compared to that of some other models.
(4) RetinaNet: RetinaNet is an object detection model that addresses the problem of class imbalance during training by introducing a focal loss. It provides a good balance between accuracy and speed but may not achieve the highest accuracy compared to that of some other models.

(5) YOLOv5: YOLOv5 is part of the You Only Look Once (YOLO) series, known for its real-time object detection capabilities. YOLOv5 is lightweight and achieves a good balance between accuracy and speed. It has a smaller model size and is suitable for various applications.

4. Results

4.1. The Environment for Training and Hyper-Parameters

The runtime environment used in this article is shown in Table 3. The parameters related to training the model for detecting forest smoke are shown in Table 4.

Table 3. Experimental conditions.

Experimental Environment	Details
Programming language	Python 3.8
Operating system	Windows 10
Deep learning framework	PyTorch 1.10.0
GPU	NVIDIA GeForce GTX 3080
GPU acceleration tool	CUDA:11.0

Table 4. Training parameters of the forest fire detection model.

Training Parameters	Details
epochs	300
batch-size	16
img-size (pixels)	640 × 640
initial learning rate	0.01
optimization algorithm	SGD

4.2. Analysis of Module Effectiveness

To verify whether the introduction of individual modules in our model outperforms the baseline, we performed an analysis of the modules' effectiveness.

4.2.1. Effectiveness of Hardswish

Activation functions play a critical role in neural networks, but they can also lead to the well-known problem of vanishing or exploding gradients, which can have a significant impact on model training and accuracy. Therefore, a thorough analysis of the sensitivity of activation functions is necessary to evaluate their suitability in deep learning models. This can be achieved by examining the performance of the model under various activation functions and comparing the results to identify which function is the most effective. In order to verify whether the HardSwish activation function we selected has better advantages compared to other activation functions, we compared the results of different activation functions for the first 250 rounds of training. Figure 10 shows the experimental results.

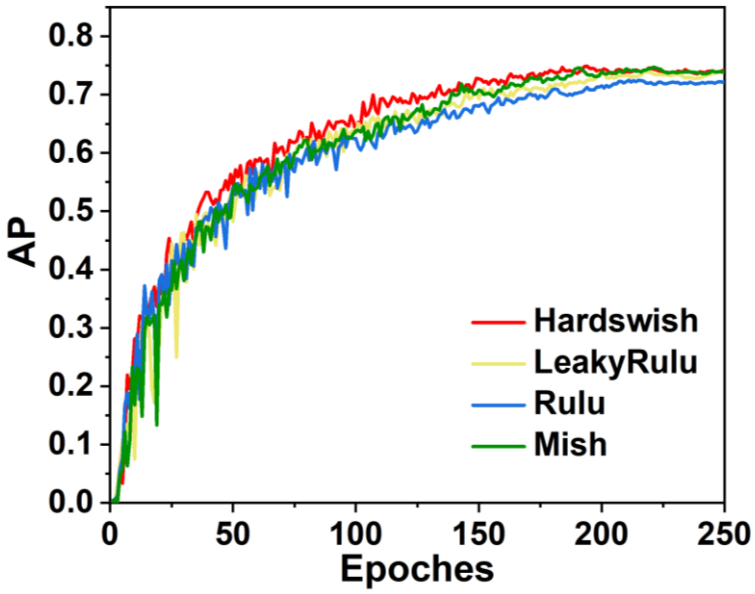


Figure 10. Comparison of different activation functions.

As indicated by the above figure, the HardSwish activation function exhibits a faster convergence speed, which can improve the efficiency and stability of gradient propagation, thereby accelerating the training of the network and leading to greater stability. Therefore, selecting HardSwish for the research of this paper is a wise choice.

4.2.2. Effectiveness of CA

In this section, to verify the effectiveness of the CA attention mechanism we selected, we compared it with the Squeeze-and-Excitation (SE) block [34] and the Convolutional Block Attention Module (CBAM) [35] by incorporating them into YOLOv7, and the results are shown in Table 5.

Table 5. Comparison of different attention mechanisms.

MODEL	P/%	R/%	AP@.5/%
YOLOv7	69.8	68.1	74.3
YOLOv7-CBAM	67.9	65.1	71.5
YOLOv7-SE	76.5	65.1	75.6
YOLOv7-CA	75.3	68.9	76.9

We can judge the quality of the attention mechanism by evaluating its AP@.5, P, and R for our own dataset. It can be seen that when CA is added to the model, the AP@.5 increases by 2.6%, P increases by 5.5%, and R increases by 0.8%, with good performance in all indicators. Therefore, we selected CA to enhance the feature extraction ability of YOLOv7.

4.2.3. Effectiveness of SIoU

In our experiments, we use four loss functions, GioU [36], DIoU, CIoU, and SIoU, respectively, on the basis of the baseline and observe their performance in identifying our dataset, respectively, and the results are shown in Figure 11. Observing the experimental results, we find that the use of SIoU can make the model converge faster than other loss functions, and the final stable loss value achieved is the lowest among the four loss functions, which fully illustrates that our choice of SIoU as the loss function is quite reasonable.

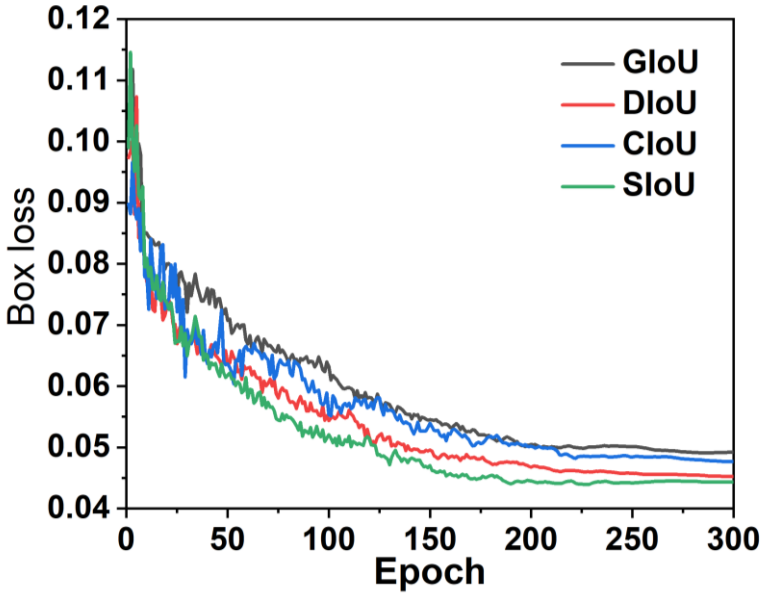


Figure 11. Comparison of different loss functions.

4.3. Ablation Experiments

The ablation experiment is essential to verifying the necessity of introducing each module in the final model and to exploring the impact of each module on the model. The effects of the model after introducing different modules were tested on the same test set, and the experimental results are shown in Table 6, where GSI represents the integration of GSELAN and GSConv.

Table 6. Results of ablation experiment.

MODEL	P/%	R/%	AP@.5/%	AP@.5:.95/%	Parameters/M	GFLOPs	FPS
YOLOv7	69.8	65.0	74.3	47.4	9.32	26.7	62.89
YOLOv7-CA	70.3	71.9	76.9	49.8	9.38	26.8	64.1
YOLOv7-SIoU	72.3	73.1	77.7	51.1	9.32	26.7	65.3
YOLOv7-GSELAN	69.8	73.2	75.5	48.8	8.67	25.4	67.3
YOLOv7-GSSPPFCSPC	73.1	67.7	74.8	48.1	8.45	26.0	64.1
YOLOv7-CARAFE	71.7	71.6	76.8	49.8	9.36	26.8	60.25
YOLOv7-CA-SIoU	74.0	71.2	78.2	51.2	9.38	26.8	64.5
YOLOv7-CA-SIoU-GSI	73.8	71.5	79.2	51.0	7.93	25.0	67.34
Ours	77.1	71.8	80.2	52.8	7.96	25.1	63.39

From the data in the table, it can be seen that the introduction of attention mechanisms, the improvement in loss functions, and the use of CARAFE upsampling mainly improve the indicators of AP@.5 and AP@.5:.95, while the other metrics used to measure the degree of lightness change little. This improves the accuracy of the model's prediction without increasing the cost of the model's calculation. On the other hand, the introduction of GSELAN and GSSPPFCSPC reduces the parameters by 0.65 M and 0.87 M, whereas the GFLOPs reduced them by 0.65 M and 0.92 M, and the FPS increased them by 4.41 and 1.21, respectively. The AP@.5 and AP@.5:.95 do not change significantly, indicating that the introduction of these two modules reduces the computational cost of the model without changing the accuracy, speeds up the convergence speed, and makes the model more lightweight.

As mentioned above, the introduction of the CA, SIoU, and CARAFE modules separately successfully improves the accuracy of the model's recognition, while the introduction of the GSConv and Hardswish modules makes the model more lightweight. In the following experiments, other modules are continuously introduced based on the model embedding CA. In experiment 7, the replacement of the CIoU loss function with SIoU allows for the model to better learn the location and size information of smoke, increasing the accuracy of smoke detection from 76.9% (experiment 2) to 78.2%. Then, after introducing the lightweight convolution GSConv, the parameters are reduced by 1.45 M, and there is a 1.8 reduction in the GFLOPs and a 2.84 increase in the number of FPS, because GSConv can help reduce the size and computational complexity of the model, making the system of smoke detection more rapid and efficient in processing data. Finally, adding CARAFE upsampling improved the model's accuracy by 2% without noticeable changes in the computational speed, indicating that CARAFE upsampling can adaptively increase the resolution of smoke images for different sizes and resolutions, helping the network better perceive smoke information in complex scenes, thus improving the generalization and accuracy of the model. The final model proposed in this paper achieves an AP@.5 of 80.2% and a number of FPS of 63.39, while the GFLOPs are only 25.1. Compared to the baseline, the indicators P and R are improved by 7.3 and 6.8, respectively, demonstrating the higher accuracy of the prediction. On the other hand, the AP@.5 is improved by 5.9% and the GFLOPs are reduced by 1.6, enabling better detection results to be achieved while using fewer computational resources.

To demonstrate the significant improvements of the enhanced model compared to the baseline in terms of its prediction accuracy and lightweight design, significance tests can be performed for several metrics, including the AP@.5, parameters, GFLOPs, and FPS. Assuming no significant differences exist between our model and the baseline, the corrected paired Student's *t*-test was chosen as the statistical test. The results of the significance tests are presented below (Table 7).

Table 7. Results of the significance tests.

Indicators	AP@.5	Parameters	GFLOPs	FPS
Null Hypothesis	There is no significant difference between our model and the baseline.			
Statistical Test Method	Corrected paired Student's *t*-test.			
p-value/%	2.21	0.93	0.55	0.86

According to the aforementioned test results, at a significance level of 5%, we reject the null hypothesis, indicating that our model shows a statistically significant difference compared to the baseline in terms of the specified metrics. Therefore, the improved model exhibits significant enhancements in both its prediction accuracy and lightweight performance compared to those of the baseline.

4.4. Comparison Experiments

In order to further verify the effectiveness of the network model proposed, we compared the detection effects of various mainstream networks, including Faster R-CNN [37], EfficientNet [38], SSD [39], Retinanet [40], and YOLOv5, etc., for the same dataset. The performance results are shown in Table 8.

Table 8. Results of comparison experiments.

MODEL	AP@.5/%	GFLOPs	Parameters/M	FPS
Faster R-CNN	81.1	206.66	41.12	38.8
EfficientDet	71.9	116.73	18.34	27.8
SSD	68.2	342.75	23.75	94
Retinanet	73.5	153.79	19.61	50.1

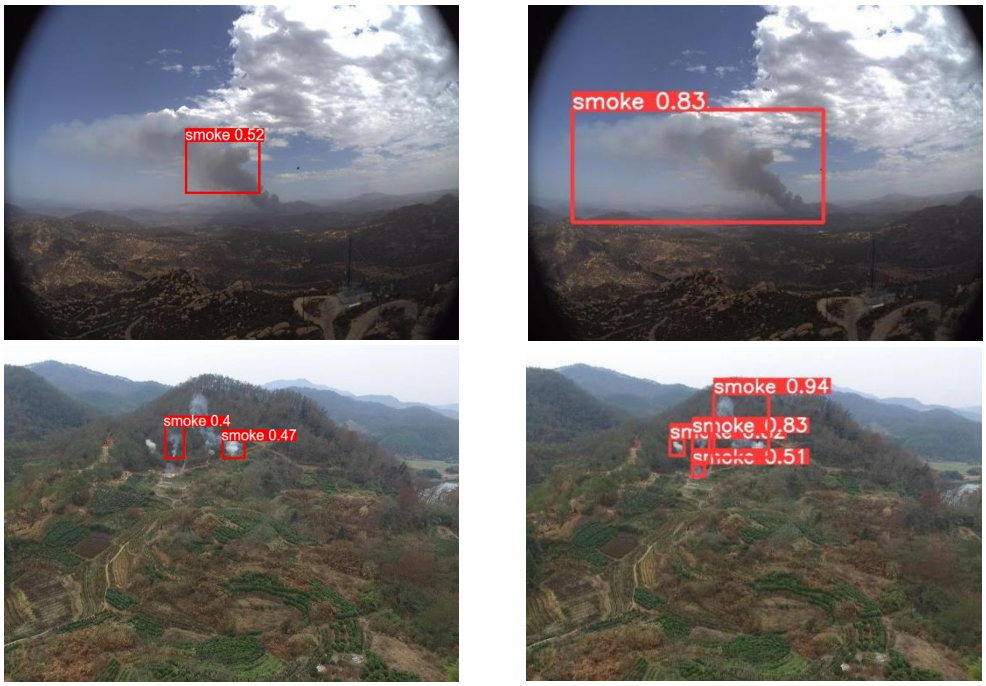
YOLOv5m	75.1	48.2	20.8	80.2
Ours	80.2	25.1	7.96	63.39

From the above table, it can be seen that Faster R-CNN, as a two-stage network, has an advantage of about 1% over the proposed network, but its number of parameters and amount of computation are much more than those of the algorithm proposed in this paper. Our model has the best results in terms of its accuracy and detection speed compared to other one-stage detection algorithms. Compared to the two-stage target detection network Faster R-CNN, there is a slight difference in accuracy, but there is a difference of about seven times in terms of the parameters. Our proposed improved algorithm has a broader application scenario with fewer parameters, a faster speed, and better detection accuracy, and can play a greater role in detecting forest fire smoke.

4.5. Testing in Different Scenarios

We tested the performance of the unimproved YOLOv7 model and our improved model in detecting forest fire smoke in different scenarios, and some test results are shown in Figure 12. Observing the test results, in the test of group (a), the original YOLOv7 model was unable to detect the smoke in the image, even for very obvious smoke with a large volume, while our model performed quite well in terms of detection; in the test of group (b), the original YOLOv7 model was unable to detect the complete smoke target and could only detect a part of the smoke, indicating its insufficient ability to extract and fuse the features of smoke. In contrast, our model was able to accurately identify the complete smoke target and could distinguish smoke and fog very well, even in cases in which smoke and fog were stuck together. This demonstrates the effectiveness of our proposed improvements.

(**a**)



(**b**)

Figure 12. Test results of the original YOLOv7 model and the improved YOLOv7 model in different scenarios: (**a**) The baseline was unable to detect smoke, while the improved model was able to detect smoke; (**b**) The baseline was unable to detect the complete smoke, while the improved model was able to accurately identify the complete smoke.

In addition, the model proposed in this article can recognize small smoke well, as shown in the (a) group of pictures in Figure 13. The model can detect the presence of small smoke early, so that forest fires can be detected and extinguished in a timely manner. At the same time, in the case of the existence of similar smoke, the model can eliminate the interference containing whatever is similar to the smoke in the images and achieve the high-precision detection of forest smoke in outdoor multi-environmental backgrounds, as shown in the (b) group of pictures in Figure 13.

(**a**) (**b**)

Figure 13. Recognition results of small smoke and smoke with interference containing something similar to smoke in images: (**a**) Small smoke; (**b**) Smoke with interference containing something similar to smoke in images.

5. Discussion and Conclusions

Predicting and preventing forest fires is crucial to protecting forests. On one hand, when comparing the development histories of the means of detecting forest fire smoke, manual detection is less effective and too costly, while detection by using instruments is easily disturbed by fine particles such as dust in the environment. Compared with these two methods, our method is based on computer vision, uses pattern recognition for feature extraction and classification, is able to detect smoke well, has low deployment costs, and is a good strategy for detecting forest fire smoke. On the other hand, in smoke detection based on deep learning, many scholars have proposed network structures, such as R-CNN or other algorithms [41], which do improve the accuracy of smoke detection to some extent, but they are more demanding in terms of hardware than the LMDFS proposed in this paper, making them difficult to deploy to meet real-time requirements. Moreover, they cannot provide an effective solution for detecting small smoke and smoke containing disturbances. Although FfireNet [42] provides a faster detection method, there is still a possibility to improve its accuracy. Our model takes both high accuracy and low computational costs into account and improves the detection accuracy of small smoke by aggregating larger sensory fields. Furthermore, our model can also more effectively separate the essential difference between forest fire smoke and smoke-like smoke, which solves this painful problem in the field of detecting forest fire smoke and provides a new idea for preventing and controlling forest fires.

YOLOv7, as the latest target detection model, has a high capability to extract and aggregate the features of images, thus achieving a high accuracy in target recognition. However, better detection results require a large computational expenditure, which is inconvenient for the model's deployment in edge devices. For this reason, we built the GS-ELAN module by using GSConv. GSConv is able to improve the effectiveness of convolution while enhancing the calculation efficiency through the effective combination of DWConv and SConv. So, it is an efficient means to lighten the model. Taking the GS-ELAN module constructed in this paper as an example, the problem of a possible lack of links for GSConv due to the replacement of convolution can be eliminated, and it is helpful for the transfer and flow of information in the model in that it introduces identity mapping. In addition, we borrow the structure of the SPPF to improve the SPPCSPC, which can have a higher computational efficiency and training efficiency with fewer parameters. Then, we add a multi-layer CA mechanism to the feature extraction network, because under a forest environment, there exist a large number of smoke-like disturbances, such as floating clouds, atmospheric fog, etc. Due to their similar characteristics to those of forest fire smoke, the traditional feature extraction network cannot accurately extract the features of forest smoke. The addition of CA significantly enhances the model's ability to extract smoke features and can more effectively separate the essential differences between forest

fire smoke and other clouds, thus reducing the false detection rate of non-smoke. In addition, in regards to the characteristics of the thinness and fineness of small smoke produced in the early stages of a forest fire, especially for images of forest smoke taken at long distances with long views, its shape is even smaller. It is more difficult for feature fusion to detect this than the typical smoke that is already formed, i.e., there is a possibility of smoke being filtered out. For this reason, we add CARAFE upsampling, which can help the network perceive a wider range of contextual information by expanding the perceptual field of the model, and improve the capability of feature representation by contextual fusion judgments in order to extract and fuse these fine features. Finally, we use the loss function SIoU to replace the original localization overlap loss function by judging the angular difference between judgement boxes, which not only allows for fast convergence during training to improve the model's accuracy, but also allows for the fast screening of NMS during detection to locate smoke locations more quickly and accurately, which is also essential for the fast detection of forest fire smoke. The final experimental results for the constructed dataset demonstrate that the model proposed achieves an AP@.5 of 80.2%, a number of FPS of 63.39, and a total number of parameters of 7.96 M. Compared to the baseline, the proposed model shows comprehensive improvements. Furthermore, when compared to other detectors of the same class, it achieves the best performance for all indicators. Its lighter weight and better detection performance make it more deployable in the practical tasks of detecting forest fire smoke. In addition, we note the important role of sensors in fire detection tasks. Abeer D. Algarni et al. [43] compare multiple sensors in wildfires. The advantages and limitations of detection have inspired us to consider using sensors, such as thermal infrared remote sensors, to improve the detection of forest fire smoke from a multidimensional direction in our later studies.

6. Future Work

Our experimental results demonstrate that the model proposed in this paper has a wide range of applications. On one hand, it can be installed on drones and watchtowers equipped with video surveillance, which can be used for the real-time prediction of incipient fires or fires that have not yet occurred; on the other hand, it can also be installed on fire cameras for observing and describing the development of fires that have already occurred, providing reference for the rescue work of firefighters. In future research, we will further explore its coherence with other monitoring equipment.

In the field of forest fire detection, wildfire detection based on satellite imagery has a deep research foundation [44,45], but it also has some shortcomings. For instance, it is easy to detect large-scale fire situations because satellite images usually cover a large area, while it is not easy to detect the features of smoke in the early stage of a fire, especially small smoke, and it is crucial for forest fires to be extinguished as early as possible. To address the above issues, our model has good potential for application. Firstly, our model performs excellent when detecting small smoke and smoke with smoke-like inference. Secondly, our model is designed to be lightweight and suitable for resource-constrained environments, such as emergency response sites or platforms such as UAVs. This makes our model easy to deploy and integrate into existing satellite-imagery-based wildfire detection systems.

Certainly, the model proposed in this article also has some limitations. The model mainly focuses on detecting forest smoke during the daytime, and the dataset used is mostly from the daytime. However, the risk of forest fires occurring at night is also high. Therefore, in our next study, we will incorporate data on forest smoke at night to improve the generalization ability and broad applicability of this model.

Author Contributions: G.C. and R.C. devised the programs and drafted the initial manuscript. X.L. helped with data collection and data analysis. W.J., H.L. and D.B. designed the project and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development plan of Jiangsu Province (No. BE2021716), the Jiangsu Modern Agricultural Machinery Equipment and Technology Demonstration and Promotion Project (NJ2021-19), and the Nanjing Modern Agricultural Machinery Equipment and Technological Innovation Demonstration Projects (No. NJ [2022]09).

Data Availability Statement: Not applicable.

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

References

1. Scholten, R.C.; Jandt, R.; Miller, E.A.; Rogers, B.M.; Veraverbeke, S. Overwintering fires in boreal forests. *Nature* **2021**, *593*, 399–404. https://doi.org/10.1038/s41586-021-03437-y.
2. Hu, Y.; Zhan, J.; Zhou, G.; Chen, A.; Cai, W.; Guo, K.; Hu, Y.; Li, L. Fast forest fire smoke detection using MVMNet. *Knowl.-Based Syst.* **2022**, *241*, 108219.
3. Zheng, R.; Zhang, D.; Lu, S.; Yang, S. Discrimination Between Fire Smokes and Nuisance Aerosols Using Asymmetry Ratio and Two Wavelengths. *Fire Technol.* **2019**, *55*, 1753–1770. https://doi.org/10.1007/s10694-019-00829-5.
4. Li, X.; Liu, J.; Huang, Y.; Wang, D.; Miao, Y. Human Motion Pattern Recognition and Feature Extraction: An Approach Using Multi-Information Fusion. *Micromachines* **2022**, *13*, 1205. https://doi.org/10.3390/mi13081205.
5. Gubbi, J.; Marusic, S.; Palaniswami, M. Smoke detection in video using wavelets and support vector machines. *Fire Saf. J.* **2009**, *44*, 1110–1115. https://doi.org/10.1016/j.firesaf.2009.08.003.
6. López-Naranjo, E.J.; Alzate-Gaviria, L.M.; Hernández-Zárate, G.; Reyes-Trujeque, J.; Cruz-Estrada, R.H. Effect of accelerated weathering and termite attack on the tensile properties and aesthetics of recycled HDPE-pinewood composites. *J. Thermoplast. Compos. Mater.* **2013**, *27*, 831–844. https://doi.org/10.1177/0892705712473625.
7. Emmy Prema, C.; Vinsley, S.S.; Suresh, S. Multi Feature Analysis of Smoke in YUV Color Space for Early Forest Fire Detection. *Fire Technol.* **2016**, *52*, 1319–1342. https://doi.org/10.1007/s10694-016-0580-8.
8. Rong, D.; Xie, L.; Ying, Y. Computer vision detection of foreign objects in walnuts using deep learning. *Comput. Electron. Agric.* **2019**, *162*, 1001–1010. https://doi.org/10.1016/j.compag.2019.05.019.
9. Khan, S.; Muhammad, K.; Mumtaz, S.; Baik, S.W.; de Albuquerque, V.H.C. Energy-Efficient Deep CNN for Smoke Detection in Foggy IoT Environment. *IEEE Internet Things J.* **2019**, *6*, 9237–9245. https://doi.org/10.1109/jiot.2019.2896120.
10. Minghua, J.; Yaxin, Z.; Feng, Y.; Changlong, Z.; Tao, P. A self-attention network for smoke detection. *Fire Safety J.* **2022**, *129*, 103547. https://doi.org/10.1016/j.firesaf.2022.103547.
11. Wu, X.; Lu, X.; Leung, H. A motion and lightness saliency approach for forest smoke segmentation and detection. *Multimed. Tools Appl.* **2019**, *79*, 69–88. https://doi.org/10.1007/s11042-019-08047-5.
12. Yin, H.; Wei, Y.; Liu, H.; Liu, S.; Liu, C.; Gao, Y. Deep Convolutional Generative Adversarial Network and Convolutional Neural Network for Smoke Detection. *Complexity* **2020**, *2020*, 6843869. https://doi.org/10.1155/2020/6843869.
13. Zhang, Q.; Lin, G.; Zhang, Y.; Xu, G.; Wang, J. Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. *Procedia Eng.* **2018**, *211*, 441–446.
14. Guo, Y.; Chen, S.; Zhan, R.; Wang, W.; Zhang, J. LMSD-YOLO: A Lightweight YOLO Algorithm for Multi-Scale SAR Ship Detection. *Remote. Sens.* **2022**, *14*, 4801.
15. Li, W.; Zhang, L.; Wu, C.; Cui, Z.; Niu, C. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1999–2015.
16. Sheng, D.; Deng, J.; Xiang, J. Automatic Smoke Detection Based on SLIC-DBSCAN Enhanced Convolutional Neural Network. *IEEE Access* **2021**, *9*, 63933–63942. https://doi.org/10.1109/access.2021.3075731.
17. Ilina, O.V.; Tereshonok, M.V. Robustness study of a deep convolutional neural network for vehicle detection in aerial imagery. *J. Commun. Technol. Electron.* **2022**, *67*, 164–170.
18. Marciniak, T.; Chmielewska, A.; Weychan, R.; Parzych, M.; Dabrowski, A. Influence of low resolution of images on reliability of face detection and recognition. *Multimed. Tools Appl.* **2015**, *74*, 4329–4349.
19. Wang, C.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
22. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
23. Bochkovskiy, A.; Wang, C.; Liao, H.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
24. Zhang, X.; Zeng, H.; Guo, S.; Zhang, L. Efficient long-range attention network for image super-resolution. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; Part XVII, pp. 649–667.
25. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

26. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
27. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
31. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
33. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, J. Deep domain adaptation based video smoke detection using synthetic smoke images. *Fire Saf. J.* **2017**, *93*, 53–59.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
35. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
38. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
39. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; Part I 14, pp. 21–37.
40. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
41. Zhang, L.; Wang, M.; Ding, Y.; Bu, X. MS-FRCNN: A Multi-Scale Faster RCNN Model for Small Target Forest Fire Detection. *Forests* **2023**, *14*, 616.
42. Khan, S.; Khan, A. Ffirenet: Deep learning based forest fire classification and detection in smart cities. *Symmetry* **2022**, *14*, 2155.
43. Hendel, I.; Ross, G.M. Efficacy of remote sensing in early forest fire detection: A thermal sensor comparison. *Can. J. Remote. Sens.* **2020**, *46*, 414–428.
44. Enoh, M.A.; Okeke, U.C.; Narinua, N.Y. Identification and modelling of forest fire severity and risk zones in the Cross–Niger transition forest with remotely sensed satellite data. *Egypt. J. Remote Sens. Space Sci.* **2021**, *24*, 879–887.
45. Wang, Y.; Xu, R.; Bai, D.; Lin, H. Integrated Learning-Based Pest and Disease Detection Method for Tea Leaves. *Forests* **2023**, *14*, 1012. https://doi.org/10.3390/f14051012.