forests

MDPI

Article

Multi-Scale Forest Fire Recognition Model Based on Improved YOLOv5s

Gong Chen [1], Hang Zhou [1], Zhongyuan Li [1], Yucheng Gao [1], Di Bai [2,*], Renjie Xu [3] and Haifeng Lin [1,*]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[2] College of Information Management, Nanjing Agricultural University, Nanjing 210095, China
[3] Department of Computing and Software, McMaster University, Hamilton, ON L8S 4L8, Canada
* Correspondence: baidi000@njau.edu.cn (D.B.); haifeng.lin@njfu.edu.cn (H.L.); Tel.: +86-25-8542-7827 (H.L.)

Abstract: The frequent occurrence of forest fires causes irreparable damage to the environment and the economy. Therefore, the accurate detection of forest fires is particularly important. Due to the various shapes and textures of flames and the large variation in the target scales, traditional forest fire detection methods have high false alarm rates and poor adaptability, which results in severe limitations. To address the problem of the low detection accuracy caused by the multi-scale characteristics and changeable morphology of forest fires, this paper proposes YOLOv5s-CCAB, an improved multi-scale forest fire detection model based on YOLOv5s. Firstly, coordinate attention (CA) was added to YOLOv5s in order to adjust the network to focus more on the forest fire features. Secondly, Contextual Transformer (CoT) was introduced into the backbone network, and a CoT3 module was built to reduce the number of parameters while improving the detection of forest fires and the ability to capture global dependencies in forest fire images. Then, changes were made to Complete-Intersection-Over-Union (CIoU) Loss function to improve the network's detection accuracy for forest fire targets. Finally, the Bi-directional Feature Pyramid Network (BiFPN) was constructed at the neck to provide the model with a more effective fusion capability for the extracted forest fire features. The experimental results based on the constructed multi-scale forest fire dataset show that YOLOv5s-CCAB increases AP@0.5 by 6.2% to 87.7%, and the FPS reaches 36.6. This indicates that YOLOv5s-CCAB has a high detection accuracy and speed. The method can provide a reference for the real-time, accurate detection of multi-scale forest fires.

Keywords: forest fire detection; YOLOv5; coordinate attention; CoT; BiFPN

Citation: Chen, G.; Zhou, H.; Li, Z.; Gao, Y.; Bai, D.; Xu, R.; Lin, H. Multi-Scale Forest Fire Recognition Model Based on Improved YOLOv5s. *Forests* **2023**, *14*, 315. https://doi.org/10.3390/f14020315

Academic Editor: Víctor Resco de Dios

Received: 13 January 2023
Revised: 3 February 2023
Accepted: 3 February 2023
Published: 6 February 2023

1. Introduction

As one of nature's most precious resources, forests not only provide many economic contributions to humankind [1] but are also important in ecological terms [2]. In the context of global climate change, the frequency of forest fires is increasing worldwide. In addition to causing serious economic losses and ecosystem damage, they also pose a serious threat to human life. Wildfires are characterized by suddenness, destruction, and danger [3]. Consequently, the timely monitoring and early warning of forest fires are critical for minimizing the damage caused.

Manual patrols and watchtowers were used in the early stages of traditional forest fire detection, but there were issues regarding the limited field of view and high labor expenses. Sensor monitoring [4–6] detects forest fires based on light, temperature, smoke, and other features. However, it is difficult to reliably detect forest fires, because the detection process is restricted by distance in dense forests and is easily disrupted by noise in the detection environment. Additionally, since sensors are expensive, it is impossible to estimate the cost of their deployment in the forest. Satellite remote sensing [7,8] uses low-orbit satellites for forest fire detection, and the images collected are not affected by terrain or other conditions. However, satellite remote sensing requires a long scanning time, and the pixel points of

forest fires collected are small and cannot be detected in time, that is, in the initial and middle stages of a fire. Therefore, satellite remote sensing is more suited to fire area assessment after forest fires than real-time rapid detection. In summary, it is difficult to identify fires in their early stages using the current procedures.

With the rapid development of computer technology, the usage of image-processing-based forest fire detection techniques has increased dramatically [9,10]. Progress has been made in the manual extraction of features and the design of identifiers to detect forest fires. Jerome Vicente et al. [11] proposed an automatic system for early forest fire smoke detection using a combination of pixel and spectral analyses. A color model [12] for the classification of flames based on YCbCr color space was presented in order to effectively distinguish between flame brightness and chromaticity. This model has a higher detection rate and may be utilized to implement flame detection. Using LBP operons and a graphical pyramid model, Huang [13] established an algorithm for texture analysis. This algorithm effectively reduced the false alarm rate during forest fire detection. In addition, in [14], the radiation flux from the forest was simulated using airborne LiDAR data and computer graphics to provide a method for subsequent forest fire detection. In conclusion, forest-fire-monitoring methods based on image processing not only rely heavily on the manual extraction of forest fire features but also involve cumbersome image-processing methods. With the rise of deep learning in recent years, researchers have proposed numerous network architectures, such as R-CNN [15] and Fast R-CNN [16] in the case of two-stage target detection models and one-stage target detection models such as SSD [17] and YOLO [18–20]. Muhammad proposed [21] an effective CNN fire detection structure for surveillance video based on CNN. R-CNN was employed by Alessio [22] to detect forest fire locations using the fire's spatial features, which increased the detection accuracy but reduced the detection speed. In [23], Shen et al. obtained a forest fire detection model with a high accuracy based on YOLO. To enhance the performance of image fire identification, image fire detection algorithms using Faster RCNN, R-FCN, SSD, and YOLO v3 [24] were developed. The algorithms can automatically extract complicated fire features and successfully apply them to various fire scenarios. In conclusion, deep-learning-based forest fire detection has a higher real-time performance and better robustness than the traditional approaches. Since forest fires are difficult to extract features from as dynamic targets, it is still a great challenge to balance the real-time performance and maintenance of a good recognition accuracy in forest fire detection at the same time.

In this study, we propose an improved YOLOv5s-based target detection algorithm called YOLOv5s-CCAB in order to increase the precision of multi-scale forest fire detection. Firstly, to address the problem regarding the insufficient extraction of available features from forest fire images, a CA attention mechanism is introduced. This mechanism causes the network to pay greater attention to the forest fire image features and suppress useless information. Then, to address the problem that traditional convolutional operations cannot integrate local feature details and global features when capturing forest fire images, the CoT3 module is constructed by integrating CoT into the C3 module to extract richer features. We then introduce the power operation on the CIoU loss to improve the accuracy of bbox regression. Finally, to enable the more efficient processing of multi-scale forest fire features, we adopt BiFPN as a replacement for PANET in the neck to enhance the ability for feature fusion during forest fire detection.

The rest of the paper is organized as follows. In Section 2, the proposed Yolov5s improvement model is described, while in Section 3, we introduce the dataset used for the experiments and the model evaluation metrics. Section 4 discusses the structure of the experiments and some of the training parameter settings. The experiments validated the capacity of the CA attention module, the CoT3 module, the a-CIoU loss function, and the BiFPN module for identifying forest fires. The experimental results are discussed and analyzed in Sections 5 and 6, which conclude the paper.

2. The Proposed Method

2.1. YOLOv5

The YOLO algorithm is a single-stage target detection algorithm that was proposed by Joseph Redmon, Santosh Divvala et al. in 2016, which converts the target boundary localization problem into a regression problem [18]. It is characterized by a fast inference speed and high flexibility. Compared with YOLOv4, YOLOv5 further improves on the YOLOv4 algorithm, with a faster inference time and higher detection accuracy. In addition, the YOLOv5s module is easy to deploy during subsequent practical use due to its small memory footprint. The network architecture of YOLOv5 consists of four parts: Input, Backbone, Neck, and Prediction.

The Input side makes use of adaptive image-scaling methods for reducing redundant information, such as black edges, to improve the speed of target detection and inference. Using mosaic data argumentation, rotation cropping, and scaling greatly enriches the dataset sample while also improving the network's robustness and ability to detect small targets. Backbone is the core structure of YOLOv5, consisting of Focus, Conv, C3, and SPP, which is utilized to extract multi-scale picture features. The Conv module is the basic convolution unit of YOLOv5, consisting of two-dimensional convolution, regularization, and Sigmoid activation. The Focus module first slices the feature map into four copies and then performs Concat and Conv operations, thus minimizing the loss of the original data. The C3 module is based on the design of the cross-stage local network CSPNet [25], consisting of three Conv modules and several Bottlenecks, thus allowing fine-grained features to be extracted. SPP, the final module of Backbone, maximizes the pooling of several pooling kernels to increase the model's perceived field. In the Neck layer, the feature fusion network uses a combination of FPN [26] and PAN [27] structures. The FPN conveys top-down semantic information, and the PAN conveys top-down localization information. Then, the two are combined to improve the features that the network extracts. Finally, the three detection heads of the YOLO Head are used to acquire target information in three sizes: large, small, and medium. The network model structure of YOLOv5s is shown in Figure 1.

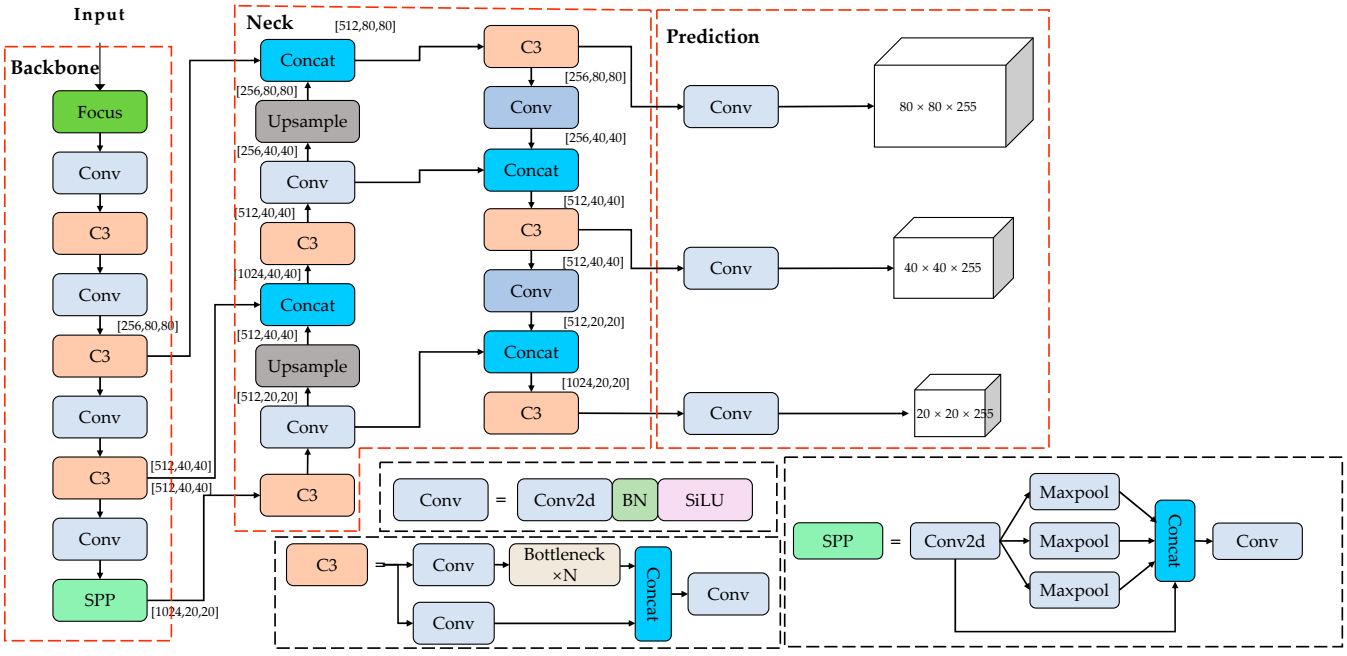


Figure 1. Model structure of YOLOv5s.

2.2. Coordinate Attention (CA)

In order to extract more critical feature information from the complex backgrounds of forest fire images, the CA attention module is embedded in YOLOv5s. The CA attention module [28] is a channel attention model that was proposed by Qibin Hou et al. The CA attention module decomposes channel attention into horizontal and vertical directional awareness in two different directions, one of which is responsible for capturing a large range of location dependencies and expanding the global receptive field of the network, while the other is responsible for retaining accurate location information. By applying them to the input feature maps, one can effectively enhance the representation of the regions of interest. The CA structure is shown in Figure 2.

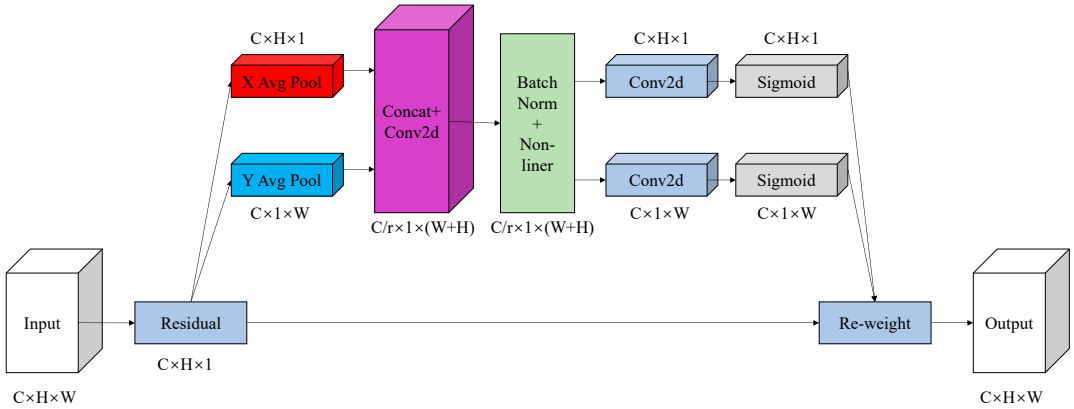


Figure 2. Coordinate attention module.

The input tensor is fed into two different average pooling layers to obtain two eigenvectors in two orthogonal directions. The two images are stitched together and then subjected to a convolution operation and passed through the Rectified Linear Unit function. They are then split in different directions and subjected to a convolution operation and the Sigmoid function, respectively. This results in output feature maps with attention weights in the width and height directions.

To address the problem of incomplete feature extraction due to the large-scale variation in forest fire images, the introduction of CA enables the convolutional networks to focus on the regions of the forest fire features that are important for learning. The preservation of remote dependency information enables the network to obtain a better grasp of the whole image. As a result, it is less likely to be confined to local areas, leading to misidentification. The preservation of location information allows the network to better capture the overall features of forest fires. This allows the network to focus more on the location of the forest fire rather than insignificant locations, thus improving the recognition accuracy of forest fires. In this paper, we fuse the CA module with the Backbone of YOLOv5s to effectively expand the receptive field and extract richer location information, which enables the model to accommodate multi-scale forest fire images.

2.3. Contextual Transformer (CoT3)

Images of forest fires involve a variety of flame patterns. Therefore, feature information and the associated scene information can be collected from large fields to improve the discrimination of flames in the images. In traditional convolutional neural networks, the convolution operator is efficient in extracting local features, but it is difficult to capture global features. While Transformer is a deep neural network based on a self-attention mechanism, its cascaded multi-head self-attention mechanism can capture long-distance dependencies [29]. To make full use of the rich contexts among the neighbor keys and solve the problem of missing transformer features, Li et al. designed a module known as the Contextual Transformer (CoT) block [30]. This design combines dynamic contextual

information aggregation with convolutional static contextual information, which it uses as an output. The block is depicted in Figure 3.

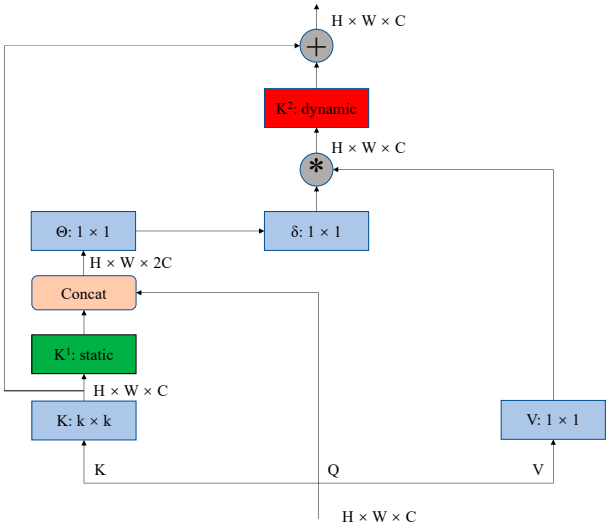


Figure 3. Contextual Transformer (CoT) block.

As shown above, unlike the self-attention mechanism of Transformer in traditional vision backbones, for the input image, CoT first uses a k × k group convolution on the neighbor keys to contextualize each key representation. We call the input K that has learnt the static contextual information K^1. Then, two consecutive 1 × 1 convolutions are performed on K^1 to obtain the attention matrix. Next, we obtain the feature map K^2 by multiplying the matrix with the value V after the feature mapping, which is used to capture the dynamic context among the inputs. Finally, the output is a linear fusion of the static context K^1 and the dynamic context K^2.

In this paper, we propose using CoT to replace the 3 × 3 convolution in the original C3 module in order to build the CoT3 module. The structure of CoT3 is shown in Figure 4.

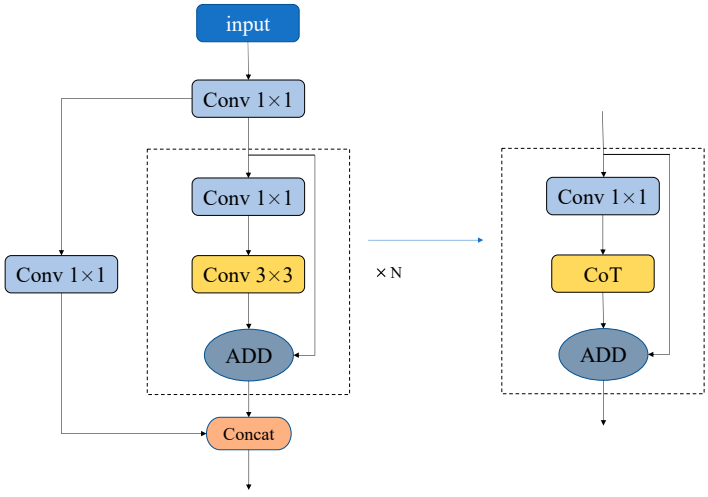


Figure 4. CoT3 module structure.

Since deep networks have stronger semantic information, enabling them to extract more forest fire features than shallow networks, the last C3 module of the original YOLOv5s was replaced with CoT3. This allows one not only to make full use of the convolution operation in order to obtain the static context of the forest fire images, retaining the local cues as feature maps, but also to combine the dynamic context of a transformer in order to aggregate the global representation between feature blocks [30]. As a result, the ability

to retrieve forest fire feature information is enhanced. Self-attention learning is enhanced by exploiting contextual information between input keys, ultimately strengthening the representational capacity of the network. In addition, CoT has fewer parameters and a higher performance, which enables it to detect forest fires faster than before, when it is mounted on a UAV.

2.4. Bi-Directional Feature Pyramid Network (BiFPN)

The FPN and PAN structures are adopted in the Neck part of the YOLOv5 network to enhance the feature fusion. As shown in Figure 5a, the FPN structure builds an up-sampling pathway in a top-down manner to perform feature fusion and deliver high-semantic information, while PAN delivers strong location information in a bottom-up manner on top of FPN, and the two are fused to improve the target detection accuracy, as shown in Figure 5b. However, due to obvious changes in the scale of forest fire targets, the original structure destroys the consistency of the informative features of forest fires. To solve this problem, the BiFPN structure employs a cross-scale-linking approach to improve the original PAN, as shown in Figure 5c. In Figure 5, P3, P4, and P5, respectively, represent the characteristic layers of the large, medium, and small targets.

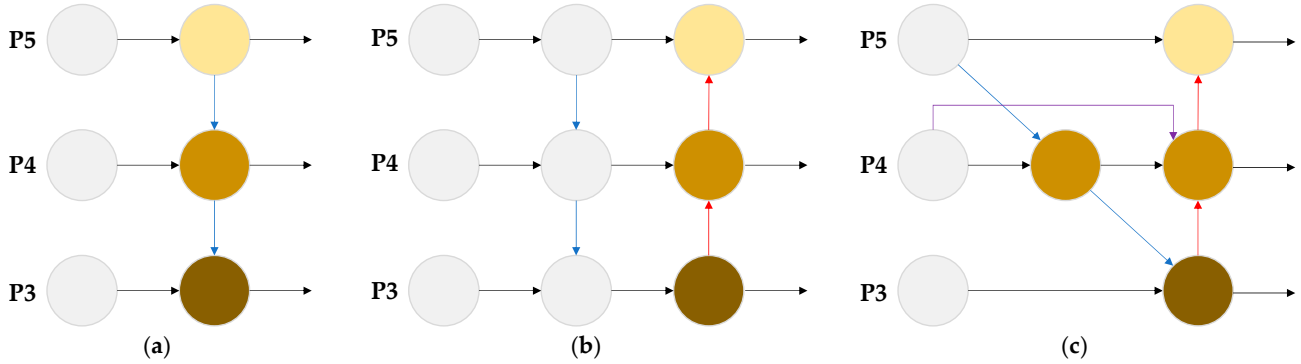


Figure 5. Feature fusion module, where red arrows represent downsampling and blue represents upsampling: (**a**) Feature Pyramid Network (FPN), a top-down unidirectional-flow network struct; (**b**) Path Aggregation Network (PAN), which adds a bottom-up aggregation link to FPN; (**c**) Bi-directional Feature Pyramid Network (BiFPN), which simplifies the nodes and adds a link between nodes at the same level for cross-scale linking, which is shown in the figure by the purple arrow.

The improved feature fusion network is more suitable for real-time forest fire detection, deleting the network nodes to simplify the bi-directional network. The BiFPN's unique bi-directional span linking allows the network to incorporate more features [31]. This effectively solves the problem of the inadequate fusion of multi-scale forest fire image features. The use of the weighted feature fusion technique also enables the output of each feature map to contain more complete forest fire information.

2.5. Complete-Intersection-Over-Union (CIoU)

The CIoU loss function used in YOLOv5 is an improvement on DIoU loss. The regression is more accurate, as it adds the length and width loss [32]. The formula is shown below.

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{\rho^2\left(\mathrm{b}, \mathrm{b}^{\mathrm{gt}}\right)}{\mathrm{c}^2} - \beta\upsilon \tag{1}$$

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

$$\beta = \frac{\upsilon}{(1 - \mathrm{IoU}) + \upsilon} \tag{3}$$

$$L_{CIoU} = 1 - CIoU \tag{4}$$

The three items in the CIoU correspond precisely to the calculation of the IoU, central point distance, the aspect ratio β, and the υ for the aspect ratio, calculated as shown above. w, h and w^{gt}, h^{gt} denote the height and width of the predicted frame and the real frame, respectively. a-CIoU [33] increases the loss and gradient weights of high IoU by performing a power operation on the IoU and its penalty term expressions, resulting in the improved regression accuracy of the model. The formula is as follows:

$$L_{a-CIoU} = 1 - IoU^a + \frac{\rho^{2a}\left(b, b^{gt}\right)}{c^{2a}} + (\beta\upsilon)^a \tag{5}$$

In this paper, we define the value of a as 3. Through the power operation, more attention is paid to the targets of high IoU, which not only accelerates the convergence of the network but also further improves the regression accuracy. Therefore, in this paper, we use a-CIoU loss function for the border regression.

2.6. Improved YOLOv5s-CCAB Structure

In summary, the overall structure of the improved YOLOv5 described in this paper is shown in Figure 6, where the changes are framed by black lines. Compared to the original structure, the CA module is inserted after the C3 module in Backbone, and the CoT3 module replaces the final C3 module in Backbone. Additionally, a cross-layer connected feature fusion is added to the Neck layer based on BiFPN. Finally, the issue of accuracy degradation caused by the overlapping target and prediction frames is resolved by a-CIoU loss. The newly improved YOLOv5 has a better detection performance.

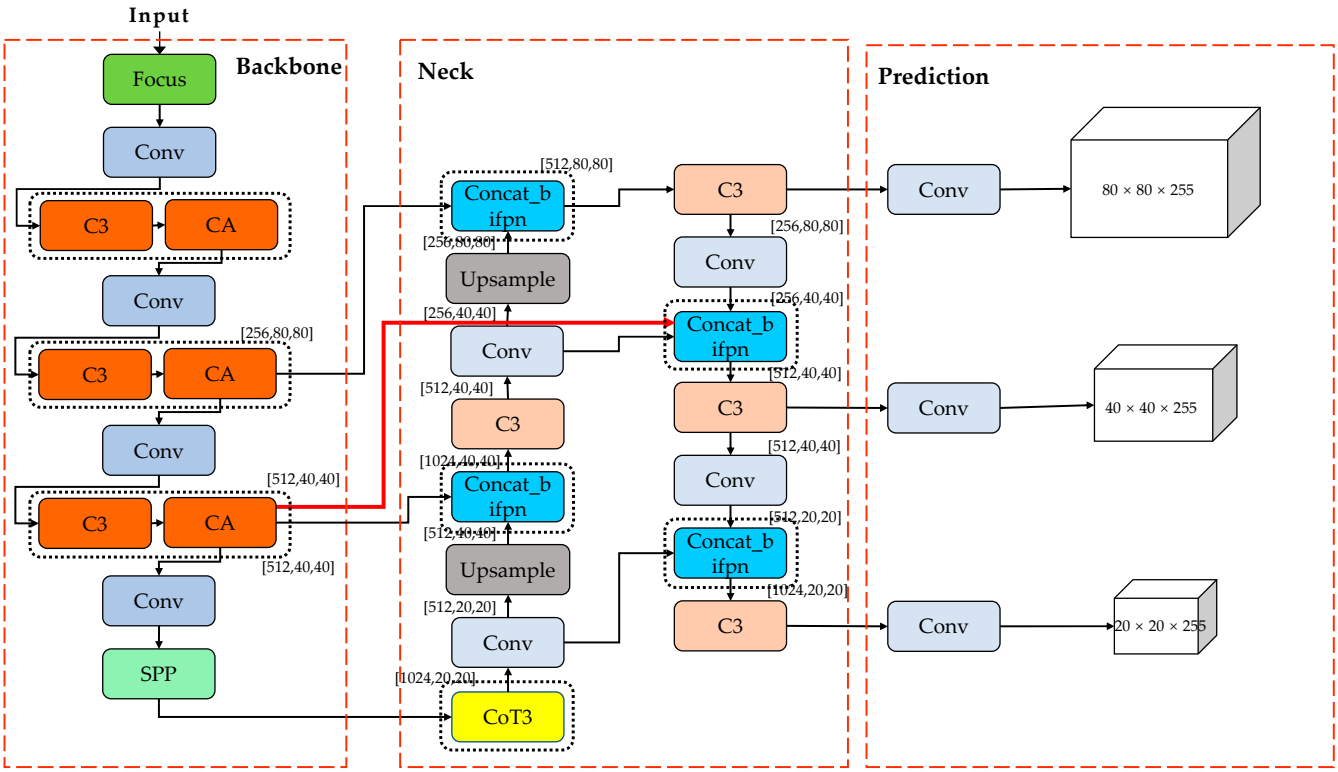
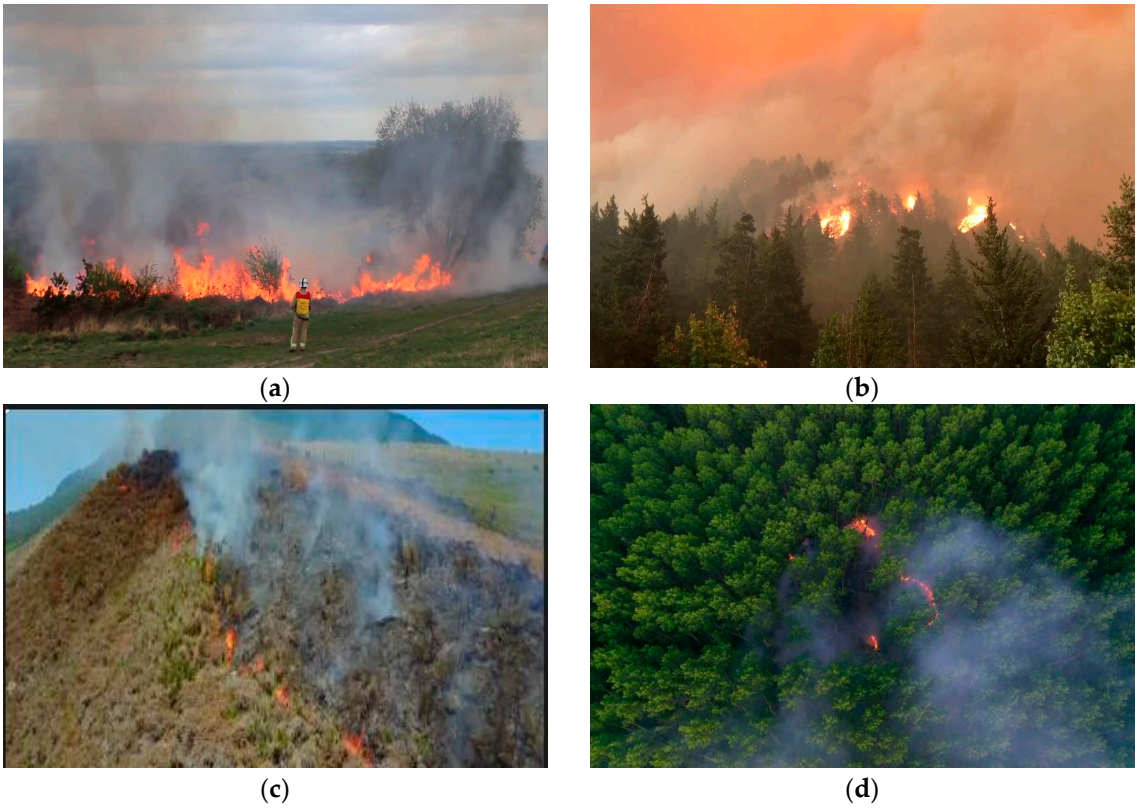


Figure 6. The improved YOLOv5s network architecture.

3. Evaluation Methodology

3.1. Datasets

In YOLOv5 fire detection, the quality of the dataset has a significant impact on the results of the training. The experimental dataset photos were collected from two sources: firstly, the public fire dataset BowFire [34] and others, and secondly, by searching the Internet to obtain photographs of forest fires. This allowed the model to extract more appropriate and effective features of the forest fires during training. According to the COCO dataset evaluation index [35], the selected forest fire images include typical and complex ground fire, canopy fire, and other medium to large forest fire images sized larger than 32^2 pixels. More importantly, small forest fire images sized smaller than 32^2 pixels from the former midterm set, which were taken from remote locations, were also included. Sample images from the dataset are shown in Figure 7.



(**a**) (**b**) (**c**) (**d**)

Figure 7. Representative forest fire images from the forest fire dataset: (**a**) ground fire; (**b**) canopy fire; (**c**) pre-midterm images of small forest fires; (**d**) remote shot of a forest fire.

The final dataset was produced in 2976 sheets. The dataset was randomly divided into a training set, validation set, and test set in a ratio of 8:1:1, and the exact number of images after partitioning is shown in Table 1. It should be noted that the default data enhancement methods used by YOLOv5 include rotation, scaling, cropping, mosaic data enhancement, etc. It should also be noted that because there are no real-life flames that rotate 180 degrees, we do not use the unnecessary data enhancement method of vertical rotation.

Table 1. Number of objects in the dataset.

Dataset	Train	Validation	Test	Summary
Number	2380	298	298	2976

3.2. Model Evaluation

Considering that the forest fire detection model is mainly deployed on mobile or embedded devices, the evaluation indicators used in this paper are the average precision (AP), giga floating-point operations per second (GFLOPs), frames per second (FPS), and inference time (Time). AP@0.5 is used to evaluate the accuracy of the prediction of forest fires and is calculated as shown below.

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \frac{1}{r} \sum_{i=1}^{r} P_i \tag{8}$$

In these equations, TP represents the number of forest fires marked and detected as forest fires, while FP represents the number of forest fires marked as non-forest fires but detected as forest fires or, in other words, the number of incorrect detections. FN represents the number of missed detections. Indicator P reflects whether the prediction of a forest fire is accurate, and indicator R reflects whether the forest fire is fully detected. AP is the area under the PR curve, representing the average precision of forest fire detection.

GFLOPs are used to measure the model's complexity. The fewer GFLOPs the model has, the less hardware performance is required. Therefore, it is better built into the device.

$$GFLOPs = \left(2 \times C_i \times K^2 - 1\right) \times H \times W \times C_0 \tag{9}$$

where C_i, C_0 represents the input and output channels, respectively, K represents the kernel size, and H, W represents the size of the feature map.

The three components of time—image pre-processing, inference, and non-maximum suppression—represent the time required to process each individual image frame.

$$\text{Time} = \text{Pre-process} + \text{Inference} + \text{NMS} \tag{10}$$

FPS represents a metric of the number of images that can be processed per second during target detection and is used to measure the detection speed of the model.

$$FPS = \frac{1}{\text{Time}} \tag{11}$$

4. Results

4.1. Training

The experimental environment of this study is shown in Table 2, and the training parameters related to the forest fire detection model are shown in Table 3. The dataset obtained previously (as shown in Table 1) is substituted into the training.

Table 2. Experimental conditions.

Experimental Environment	Details
Programming language	Python 3.8
Operating system	Windows 10
Deep learning framework	Pytorch 1.9.0
GPU	NVIDIA GeForce GTX 3080
GPU acceleration tool	CUDA:11.0

Table 3. 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
Pre-training weights file	None

It should be mentioned that, in network training, the activation function, as a crucial component of the network, can efficiently improve the expressiveness of the model. Additionally, an appropriate activation function can efficiently enhance the model's performance and training speed. In this paper, we compare several common activation functions, including Rectified Linear Unit (ReLU), Leaky Rectified Linear Unit (LeakyReLU), Mish, and Swish, whose images are shown in Figure 8, before the ablation experiment. The comparison of the results of the different activation functions, with respect to accuracy, are shown in Table 4.

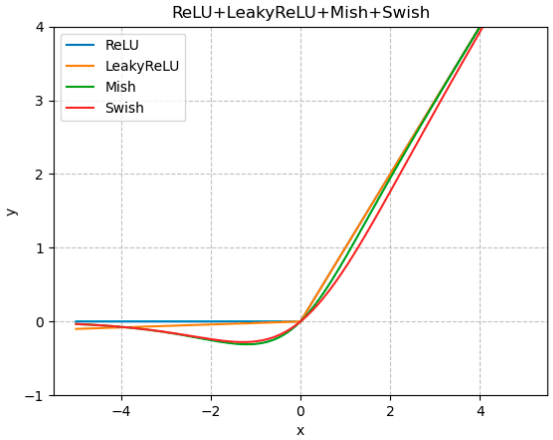


Figure 8. Image of the activation function.

Table 4. Results of training with different activation functions.

Activation Function	AP/%
Relu	84.8
LeakyRelu	86.7
Mish	82.5
Swish	84.6

By comparison, we can see that Swish has a better performance than the other functions in image processing. Therefore, in this paper, we use Swish as the activation function.

4.2. Ablation Experiments

To verify the necessity of each improvement module and the impacts of the different parts of the improvement on the performance of the forest fire detection model, the trained forest fire detection model was tested on the same test set in order to obtain its corresponding evaluation index, and the results were analyzed. The results of the experiments are shown in Table 5.

Table 5. The data of the ablation experiments.

MODEL	AP@0.5/%	GFLOPs	FPS	Time/ms
YOLOv5s	81.5	16.3	43.8	22.8
YOLOv5s-CA	85.4	16.3	41.6	24
YOLOv5s-CoT3	85.2	16.2	42.5	23.5
YOLOv5s-a-CIoU	83.4	16.3	42	23.8
YOLOv5s-BiFPN	83.7	17.6	44	22.7
YOLOv5s-CA-CoT3	86.3	16.4	33.5	29.8
YOLOv5s-CA-CoT3 -a-CIoU	86.5	16.5	34.9	28.6
YOLOv5s-CCAB	87.7	17.7	36.6	27.2

From the results of the ablation experiments, it can be seen that YOLOv5s is improved by each of these modules, including the addition of the CA attention mechanism, the introduction of CoT into the network residual structure, the improvement of the loss function, and the replacement of the neck FPN with BiFPN. In Experiment 2, AP@0.5 increased by 3.9%. This demonstrates that the introduction of CA into YOLOv5s can enable the more explicit feature extraction of dynamic such as forest fires. This can effectively boost the network accuracy. In Experiment 3, the addition of the CoT3 structure allowed the detection model to capture more global contextual information about the forest fire, which improved the performance of the forest fire detection model, ultimately resulting in a 3.7% improvement in AP@0.5 and a reduction in the GFLOPs. In the next experiment, the loss function achieved a higher regression accuracy by focusing on the high IoU targets, resulting in an improvement in AP@0.5 of 1.9%. The characteristics and location information of the flames can be captured more effectively for forest fires of varying scales, especially those with small flames, resulting in more precise forest fire identification. In Experiment 5, the use of BiFPN in the neck boosted AP@0.5 by 2.1% and led to an increase in FPS to 44, despite a slight increase in the computational effort required. This is due to the introduction of weighted feature fusion into BiFPN, which increases the network's feature fusion power. It can learn the characteristics of several forest fire types (canopy fire, trunk fire), which enhances the performance of the forest fire detection model.

In the subsequent ablation experiments, we fused these changes in turn. Combining the CoT module and the CA attention mechanism increased AP@0.5 in Experiment 6 by 4.8% while barely affecting the GFLOPs, which was much higher than the individual modules introduced in the previous four experiments. This proved that the model enhances the extraction of forest fire texture features with CoT and focuses more on its locations with CA. Note that although the FPS was also reduced to 33.5, the minimum frame rate required for real-time detection was still met. In Experiment 7, a-CIoU loss function was utilized to further improve the model's robustness in forest fire identification. In Experiment 8, we established a BiFPN on the neck based on the results of the previous experiments. The forest fire texture features learned by the network were further extracted and fused with the extracted flame features. In the case of small-target forest fires, the model may also have a higher detection effect. Overall, the average precision of our model reached 87.7%, and the detection speed reached 36.6 fps, providing the model with an advantage in terms of the recognition accuracy and speed.

4.3. Comparison

We selected a few pictures from the test set to better demonstrate the model's viability. A comparison of the detection results is shown below, where the left panel shows our detection model and the right panel shows the original model.

As shown in Figure 9, the original model is affected by missed detection issues when monitoring ground fires with different morphologies, whereas YOLOv5s-CCAB is better able to identify the flame targets. In Figure 10b, in the initial flame detection, we can see

that the original model detects only one forest fire target, while Yolov5s-CCAB can detect three forest fire targets (as shown in Figure 10a).

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

Figure 9. Detection results of typical ground flames. (**a**) Eight forest fire targets are detected by our model. (**b**) Six forest fire targets are detected by YOLOv5s.

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

Figure 10. Detection results of the initial forest fire. (**a**) Three forest fire targets are detected by our model. (**b**) Only one forest fire targets are detected by YOLOv5s.

Figure 11 compares the detection of images of a snowy scene taken with the UAV. YOLOv5s-CCAB can effectively identify images of fires that are obscured by trees. Figure 12 shows the detection of forest flames from a distance, where the original model appears to miss the detection targets, while YOLOv5s-CCAB can accurately identify small flames in the mountains.

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

Figure 11. Detection results of forest fire images from the viewpoint of the unmanned aerial vehicle. (**a**) The sheltered forest fire is well detected by our model. (**b**) The sheltered forest fire was not detected by YOLOv5s.

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

Figure 12. Detection results of canopy flames. (**a**) Three forest fire targets are detected by our model. (**b**) Two forest fire targets are detected by YOLOv5s.

5. Discussion

As forests are vital to human society, it is particularly important to identify forest fires in time so as to reduce the damage they cause. However, compared to other common objects, flames have a more varied morphology and more complex dynamic characteristics [36], combined with mutual occlusions throughout the forest [37], making it difficult to capture these features. The existing testing techniques have various drawbacks that make it difficult to address these problems. The traditional approach to detecting forest fires on video necessitates the use of a manually created recognizer. It does not extract the most essential features of the flame, resulting in a poor detection and slow detection process. In deep-learning-based detection approaches, two-stage target detection models such as Faster R-CNN require a lengthy training period and detection time, rendering them insufficient for real-time detection. In single-stage target detection, although the detection speed can meet the real-time requirements, there is a slight reduction in the accuracy. The typical examples are the SSD and YOLO series. The SSD commissioning process is quite complicated and primarily depends on manual experience. In contrast, the YOLO series, particularly YOLOv5, has the advantages of a compact model size, low deployment cost, and high detection speed, so that it stands out among methods for detecting forest fires. However, due to the large variation in the forest fire size, it is still difficult to obtain better recognition results in the case of multi-scale forest fire images. Consequently, forest fire detection remains a challenging research area.

In view of these issues, in this paper, we proposed an algorithmic model of YOLOv5s-CCAB based on YOLOv5 by adding a CA attention mechanism, replacing the backbone network module, improving the loss function, and optimizing the neck network. The experiments proved that the proposed model can be practically applied for the recognition of different types and sizes of forest fires by virtue of its high average precision and fast FPS.

Although YOLOv5s-CCAB can achieve a high accuracy in forest fire detection, it still has shortcomings. In further research, we will continue to optimize it in regard to the following aspects. Firstly, by fusing the modules in the network, YOLOv5s-CCAB's accuracy could be increased, but this would be at the expense of the detection speed. Therefore, we will conduct research on the lightweight version of the model to improve the real-time detection efficiency with the aim of ensuring its accuracy. Secondly, despite having 2976 photos of diverse sceneries and flame types in the forest fire dataset that was created in this research, the dataset used for this study was still very small and will be enlarged later to perform more precise fire detection. However, as dynamic targets, fires are too complex to be detected. Therefore, our next step will be to perform real-time video detection applications in order to improve the performance and ensure more precise flame detection.

The outcomes of the experiments demonstrate that YOLOv5s-CCAB has promising prospects for real-world use. In terms of actual use, it can be installed on drones and watchtowers with video monitoring capabilities for real-time detection or on fire cameras for the real-time monitoring of already-started fires, providing firefighters with easy access to fire references. However, as the devices rely on cameras as sensors, interference from elements such as sunlight reflections and flashes, airborne dust, etc., is unavoidable. This will significantly lower the image quality and result in false detections and missed detections. In order to further enhance the model detection efficiency, we will investigate methods to reduce error reporting in future studies, such as the use an infrared-transmitting hyperspectral sensor cameras to enhance the image contrast [38] in order to further improve the detection.

In our upcoming investigation, we will employ a UAV as a vehicle to test the forest fire detection capability of the model explored in this paper in order to verify the performance of the model in a realistic scenario. We observed that the authors of [39] developed an effective dual-spectral camera video surveillance system that can significantly enhance the detection of moving targets. This provides us with some ideas for our future work. Additionally, to address the issue of rapid battery consumption during the real-time detection of UAVs, edge computing devices and drones could be implemented together to optimize the cruise trajectory [40] used for sending the collected images to the server for processing, significantly reducing the level of computation required of the drones and lengthening their operational time. The effective detection of forest fires could be considerably increased by combining these two techniques.

6. Conclusions

Since forest fires are dynamic targets without a fixed shape, improving the accuracy and precision of forest fire detection remains a challenge. If the occurrence of forest fires can be accurately detected, the subsequent damage caused by forest fires can be greatly reduced. Therefore, it is of critical importance to study potential methods for detecting forest fires.

To address these problems, this paper was devoted to the construction of a better forest fire detection model. First, the CA attention module was added to the backbone network in YOLOv5 to enhance the focus on the forest fire feature information. CoT was introduced to improve the model's ability to gather fire feature information. The loss function was then improved to enhance the network convergence and achieve a more precise forest fire detection. Finally, the Bi-directional Feature Pyramid Network was added to the feature fusion layer, which combines the multi-scale images with the original information. In consequence, the feature fusion capability was improved. The experimental results show that the model proposed in this paper can reach 87.7% mAP and 36.6 FPS. Compared with the original YOLOv5, YOLOv5s-CCAB represents a better compromise between the accuracy, GFLOPs, and latency. These enhancements significantly improve the performance of the model. Since the real forest environment is complex, and it is difficult to identify forest fires of different scales, YOLOv5s-CCAB can detect forest fires in the appropriate time and has a better detection effect on forest fire targets of different scales, especially those in the early and middle stages. When conducting forest fire detection, it can effectively protect the forest.

Author Contributions: G.C. devised the programs and drafted the initial manuscript. H.Z. contributed to the writing and helped with the experiments. Z.L. and Y.G. helped with the data collection and data analysis. D.B. and R.X. revised the initial manuscript. H.L. 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 (grant 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. Li, Y.; Mei, B.; Linhares-Juvenal, T. The economic contribution of the world's forest sector. *Forest Policy Econ.* **2019**, *100*, 236–253. [CrossRef]
2. Sahoo, G.; Wani, A.; Rout, S.; Sharma, A.; Prusty, A.K. Impact and Contribution of Forest in Mitigating Global Climate Change. *Des. Eng.* **2021**, *4*, 667–682.
3. Ying, L.; Han, J.; Du, Y.; Shen, Z. Forest fire characteristics in China: Spatial patterns and determinants with thresholds. *Forest Ecol. Manag.* **2018**, *424*, 345–354. [CrossRef]
4. Tadic, M. GIS-Based Forest Fire Susceptibility Zonation with IoT Sensor Network Support, Case Study—Nature Park Golija, Serbia. *Sensors* **2021**, *21*, 6520.
5. Varela, N.; Díaz-Martinez, J.L.; Ospino, A.; Zelaya, N. Wireless sensor network for forest fire detection. *Procedia Comput. Sci.* **2020**, *175*, 435–440. [CrossRef]
6. Kizilkaya, B.; Ever, E.; Yekta, Y.H.; Yazici, A. An Effective Forest Fire Detection Framework Using Heterogeneous Wireless Multimedia Sensor Networks. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 1–21. [CrossRef]
7. Mae, A.; Uco, B.; Nyn, C. 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.
8. Tian, Y.; Wu, Z.; Li, M.; Wang, B.; Zhang, X. Forest Fire Spread Monitoring and Vegetation Dynamics Detection Based on Multi-Source Remote Sensing Images. *Remote Sens.* **2022**, *14*, 4431. [CrossRef]
9. Abid, F. A Survey of Machine Learning Algorithms Based Forest Fires Prediction and Detection Systems. *Fire Technol.* **2020**, *57*, 559–590. [CrossRef]
10. Cruz, H.; Eckert, M.; Meneses, J.; Martínez, J. Efficient forest fire detection index for application in unmanned aerial systems (UASs). *Sensors* **2016**, *16*, 893. [CrossRef]
11. Vicente, J.; Guillemant, P. An image processing technique for automatically detecting forest fire. *Int. J. Therm. Sci.* **2002**, *41*, 1113–1120. [CrossRef]
12. Celik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Safety J.* **2009**, *44*, 147–158. [CrossRef]
13. Huang, J.; Zhao, J.; Gao, W.; Long, C.; Xiong, L.; Yuan, Z.; Han, S. Local Binary Pattern Based Texture Analysis for Visual Fire Recognition. In Proceedings of the 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; pp. 1887–1891.
14. Xue, X.; Jin, S.; An, F.; Zhang, H.; Fan, J.; Eichhorn, M.P.; Jin, C.; Chen, B.; Jiang, L.; Yun, T. Shortwave radiation calculation for forest plots using airborne LiDAR data and computer graphics. *Plant Phenomics* **2022**, *2022*. [CrossRef] [PubMed]
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 23–28 June 2013.
16. Girshick, R. Fast R-CNN. Computer Science. In Proceedings of the Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
17. 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: Berlin, Germany, 2016; pp. 21–37.
18. 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.
19. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 21–26 July 2017; pp. 7263–7271.
20. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
21. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S.; Baik, S.W. Convolutional Neural Networks based Fire Detection in Surveillance Videos. *IEEE Access* **2018**, *6*, 18174–18183. [CrossRef]
22. Gagliardi, A.; Villella, M.; Picciolini, L.; Saponara, S. Analysis and Design of a Yolo like DNN for Smoke/Fire Detection for Low-cost Embedded Systems. In *Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES*; Springer: Cham, Switzerland, 2021; pp. 12–22.
23. Shen, D.; Chen, X.; Nguyen, M.; Yan, W.Q. Flame detection using deep learning. In Proceedings of the 2018 4th International Conference on Control, Automation and Robotics (ICCAR), Auckland, New Zealand, 20–23 April 2018; pp. 416–420.
24. Li, P.; Zhao, W. Image fire detection algorithms based on convolutional neural networks. *Case Stud. Therm. Eng.* **2020**, *19*, 100625. [CrossRef]
25. Wang, C.; Liao, H.M.; Wu, Y.; Chen, P.; Hsieh, J.; Yeh, I. CSPNet: A New Backbone that Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; Springer: Cham, Switzerland; pp. 390–391.

26. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *IEEE Comput. Soc.* **2017**, 2980–2988.
27. 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.
28. 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.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, A.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
30. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual Transformer Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef]
31. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
32. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Trans. Cybern.* **2022**, *52*, 8574–8586. [CrossRef]
33. He, J.; Erfani, S.; Ma, X.; Bailey, J.; Chi, Y.; Hua, X. α-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20230–20242.
34. Chino, D.Y.; Avalhais, L.P.; Rodrigues, J.F.; Traina, A.J. Bowfire: Detection of fire in still images by integrating pixel color and texture analysis. In Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Brazil, 26–29 August 2015; pp. 95–102.
35. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft Coco: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
36. Nagle, F.; Johnston, A. Recognising the dynamic form of fire. *Sci. Rep.* **2021**, *11*, 10566. [CrossRef]
37. Sun, C.; Huang, C.; Zhang, H.; Chen, B.; An, F.; Wang, L.; Yun, T. Individual tree crown segmentation and crown width extraction from a heightmap derived from aerial laser scanning data using a deep learning framework. *Front. Plant Sci.* **2022**, *13*, 914974. [CrossRef]
38. Zhao, H.; Ji, Z.; Li, N.; Gu, J.; Li, Y. Target detection over the diurnal cycle using a multispectral infrared sensor. *Sensors* **2016**, *17*, 56. [CrossRef]
39. Shi, B.; Gu, W.; Sun, X. XDMOM: A Real-Time Moving Object Detection System Based on a Dual-Spectrum Camera. *Sensors* **2022**, *22*, 3905. [CrossRef]
40. Cao, X.; Xu, J.; Zhang, R. Mobile edge computing for cellular-connected UAV: Computation offloading and trajectory optimization. In Proceedings of the 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Kalamata, Greece, 25–28 June 2018; pp. 1–5.