fire

Article

FFYOLO: A Lightweight Forest Fire Detection Model Based on YOLOv8

Bensheng Yun *, Yanan Zheng, Zhenyu Lin and Tao Li

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China; 1201004059@zust.edu.cn (Y.Z.); 1211004045@zust.edu.cn (Z.L.); 1211004021@zust.edu.cn (T.L.)
* Correspondence: yunbsh@zust.edu.cn

Abstract: Forest is an important resource for human survival, and forest fires are a serious threat to forest protection. Therefore, the early detection of fire and smoke is particularly important. Based on the manually set feature extraction method, the detection accuracy of the machine learning forest fire detection method is limited, and it is unable to deal with complex scenes. Meanwhile, most deep learning methods are difficult to deploy due to high computational costs. To address these issues, this paper proposes a lightweight forest fire detection model based on YOLOv8 (FFYOLO). Firstly, in order to better extract the features of fire and smoke, a channel prior dilatation attention module (CPDA) is proposed. Secondly, the mixed-classification detection head (MCDH), a new detection head, is designed. Furthermore, MPDIoU is introduced to enhance the regression and classification accuracy of the model. Then, in the Neck section, a lightweight GSConv module is applied to reduce parameters while maintaining model accuracy. Finally, the knowledge distillation strategy is used during training stage to enhance the generalization ability of the model and reduce the false detection. Experimental outcomes demonstrate that, in comparison to the original model, FFYOLO realizes an mAP$_{0.5}$ of 88.8% on a custom forest fire dataset, which is 3.4% better than the original model, with 25.3% lower parameters and 9.3% higher frames per second (FPS).

Keywords: forest fire detection; YOLOv8; CPDA; MCDH; MPD-IoU; GSconv

Citation: Yun, B.; Zheng, Y.; Lin, Z.; Li, T. FFYOLO: A Lightweight Forest Fire Detection Model Based on YOLOv8. *Fire* **2024**, *7*, 93. https://doi.org/10.3390/fire7030093

Academic Editors: Demin Gao, Shuo Zhang and Cheng He

Received: 25 January 2024
Revised: 11 March 2024
Accepted: 12 March 2024
Published: 16 March 2024

1. Introduction

A forest fire is a highly hazardous natural disaster that significantly impacts the ecosystem [1]. It not only devastates vast forest ecosystems, but also leads to an irreversible loss of biodiversity and to soil damage [2], disrupting the ecological balance. Additionally, forest fires can cause destruction to surrounding buildings, crops, and infrastructure, severely impacting economic development [3,4].

Current forest fire detection methods face several challenges. Primarily, these fires have a penchant for erupting in remote locales and are geographically widespread, where human resources are scant, making manual detection inefficient [5]. Secondly, forest fires spread at a rapid pace, and if not detected promptly, the blaze can quickly escalate and cause further devastation. Moreover, forest fires typically occur in complex and diverse natural environments such as mountainous regions, jungles, and wilderness, where detection is hindered by factors like terrain, vegetation, and weather [6], increasing the difficulty of detection. Early detection plays a crucial role in identifying and reducing response time before forest fires become uncontrollable or unmanageable [7]. Therefore, an effective detection method is of paramount importance.

Traditional forest fire detection methodologies predominantly employ digital image processing and pattern recognition techniques for image analysis. Chen et al. [8] delineated a fire early warning technique predicated on video processing, facilitating the extraction of fire and smoke pixels via chromaticity and disorder prediction based on the RGB model. This technique has achieved commendable fire accident detection while maintaining a low false alarm rate. Celik et al. [9] proffered an enhanced fire detection technique utilizing the YCbCr color space, which proficiently segregates chromaticity and luminance in comparison to the RGB color space. While traditional detection techniques suffice in adhering to real-time detection speed requisites, they exhibit subpar feature extraction capabilities in complex scenes, are prone to environmental interference, and have limited model generalization capacity.

With the development of deep learning technology, researchers have realized the strong feature extraction capabilities and minimal susceptibility to environmental interference of neural networks. Consequently, they have begun to investigate the use of convolutional neural networks for feature extraction in the context of fire and smoke detection. Applying deep learning technology to forest fire detection tasks can effectively address the limitations of traditional detection methods.

Object detection models in deep learning are generally divided into two categories: one-stage and two-stage. One-stage detection algorithms perform both candidate box generation and target classification simultaneously, resulting in fast detection but lower accuracy. Two-stage algorithms first determine the Region of Interest (ROI) to locate the target roughly, and then perform feature extraction within the region for classification and precise location. Although the two-stage model has higher accuracy, the inference speed is slower than the one-stage model. Emblematic one-stage detection algorithms encompass YOLO [10], SSD [11], and RetinaNet [12], whilst representative two-stage algorithms include Faster R-CNN [13], R-FCN [14] and Mask R-CNN [15].

On the one hand, Liu et al. [16] utilized the YOLOv5n model for forest fire detection, which can be easily deployed on low-power devices but fails to meet the accuracy requirements. On the other hand, Qian et al. [17] introduced the OBDS model, which combines CNN and Transformer to extract global feature information from forest fire smoke images. Li et al. [18] replaced the SPPF module with RFB in YOLOv5 to enable better focus on the global information of various forest fire and smoke. These strategies significantly enhance the ability of the model to capture contextual information, enable the model to acquire better feature information, and improve the detection accuracy of small-scale forest fires. However, they bring about an increase in both parameters and computational complexity.

In response to the aforementioned issue, this paper proposes an improved forest fire detection model based on YOLOv8, aiming to seek a balance between detection accuracy and speed. The main improvements are as follows:

- Through data analysis, an attention module with asymmetric dilated convolutions is designed, which allows the convolutional kernels to closely adhere to the target for feature extraction. And, detection head is improved to achieve a balance between accuracy and speed.
- The MPDIoU loss function, which utilizing the geometric properties of bounding box regression, is introduced to enhance the model's convergence speed and detection accuracy.
- The lightweight GSConv is used to replace standard convolution, alleviating the parameters increase caused by the attention module and make model more lightweight.
- Knowledge distillation strategy is adopted in the training stage, so that FFYOLO can learn more intrinsic connections between features, thereby improving the model's generalization ability and detection accuracy.

The rest of the paper is organized as follows. In Section 2, we provided a detailed explanation of the methods we improved and utilized. Section 3 presented the experimental environment and results. In Section 4, we summarized our approach and discussed future research directions.

2. Methods

2.1. YOLOv8

The YOLOv8 model manifests significant enhancements over its predecessors. A significant advancement is the shift from the anchor-based strategy to the anchor-free strategy for bounding box regression. Anchor-free strategy abandons the use of anchor boxes and instead employs a method of probability regression based on the center point for bounding box regression. The anchor-based strategy performs well in fixed datasets or scenes with consistent object distributions. However, its performance may falter in complex forest fire detection environments. In contrast, the anchor-free strategy is more suitable for forest fire detection. Furthermore, the anchor-free strategy significantly reduces the number of predicted boxes per grid compared to the anchor-based approach. This reduction expedites the Non-Maximum Suppression (NMS) process, thereby enhancing the inference speed. The model structure of the YOLOv8 is shown in Figure 1. The C2f structure in YOLOv8 significantly enhances the gradient flow within the model, while concurrently reducing the redundant connections inherent in the original C3 structure. Furthermore, YOLOv8 adopts the TaskAlignedAssigner [19] positive sample allocation strategy and utilizes the Distribution Focal Loss [20] combined with CIoU Loss for better bounding box regression.

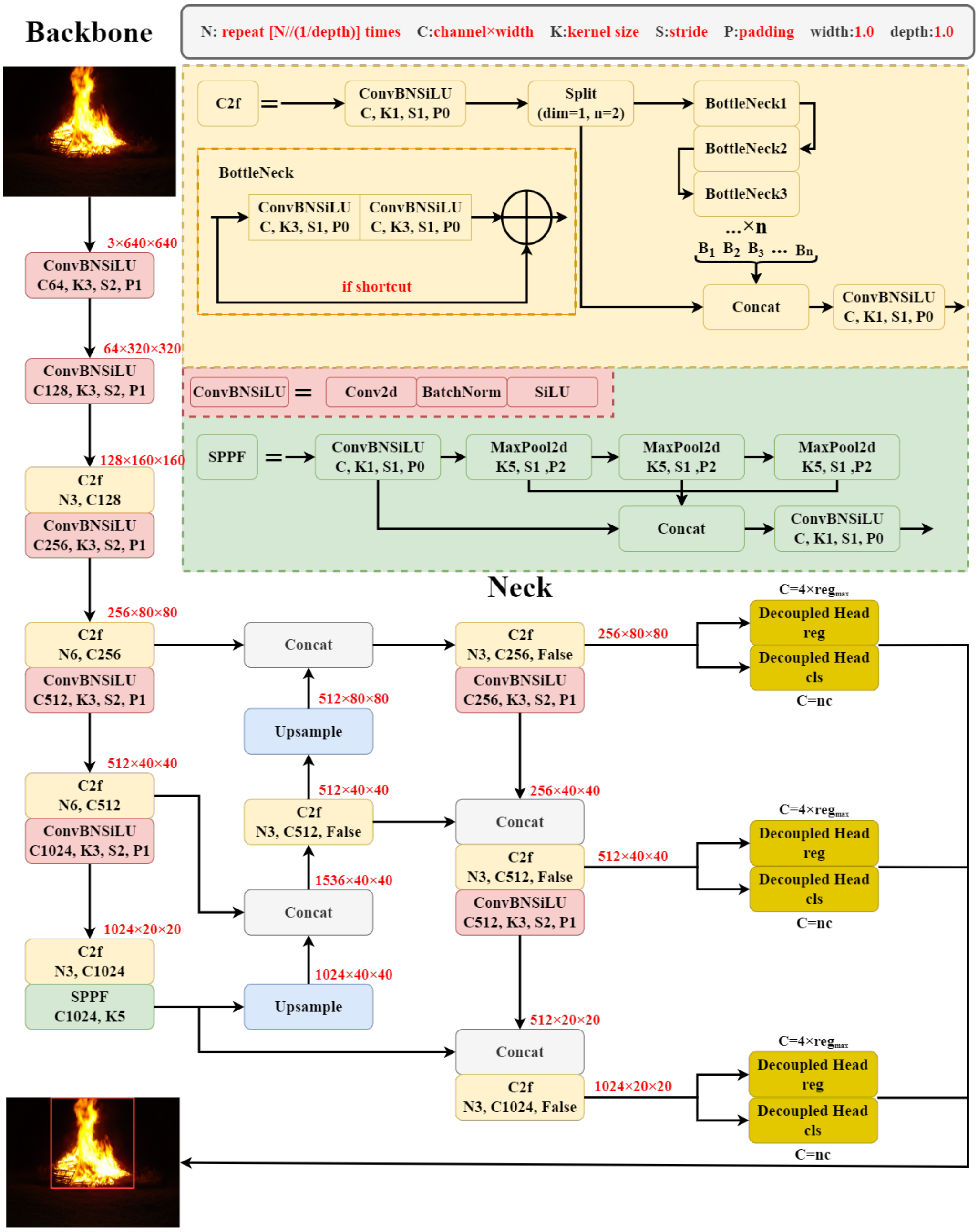


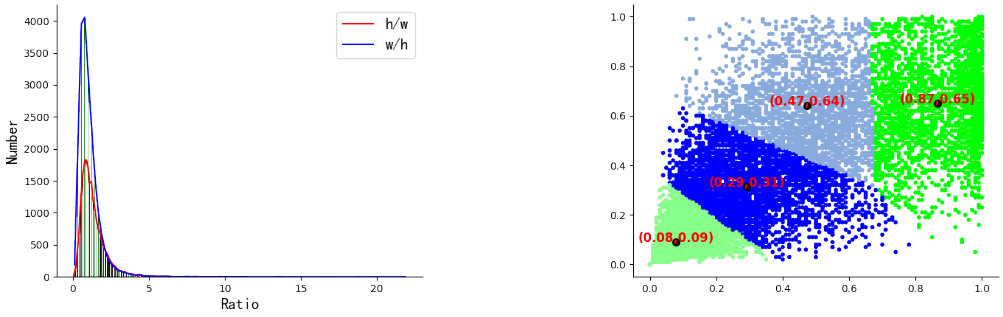
Figure 1. Model Structure of YOLOv8.

2.2. Channel Prior Dilatation Attention

The attention mechanism in computer vision endeavors to attenuate the influence of irrelevant features on the final output by simulating the manner in which humans perceive objects. This mechanism augments the model's ability to accurately identify key features in complex scenes, thus enhancing the model's robustness.

The attention mechanism is bifurcated into channel attention, spatial attention, and the combination of the two. A quintessential example of channel attention, denoted as Squeeze-and-Excitation (SE) attention [21], assesses the significance of each feature channel by compressing each 2D feature map and applying corresponding weights. Nonetheless, the SE module only concentrates on the importance of the channel, ignoring the spatial information contained in the feature map. In the subsequent study [22], a convolutional methodology was employed to engender spatial attention feature maps after channel attention, thereby markedly ameliorating the model's detection accuracy.

Given the intricacy of the forest fire scene, the global information encapsulated in the image substantially affects the final output, and the embedding of spatial information can mitigate misjudgment rates. Upon analyzing the priori fire and smoke features shown in Figure 2, Figure 2a shows the aspect ratio distribution of the ground truth boxes. By utilizing the K-means clustering algorithm shown in Figure 2b, this paper selected four aspect ratios and summarized them as 4:3 and 1:1 for spatial attention design.



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

Figure 2. Analysis of ground truth boxes: (**a**) ground truth boxes aspect ratio distribution; (**b**) K-means clustering result.

To augment the model's ability to focus on target features, this paper advocates for the incorporation of a spatial attention module, which is composed of multi-scale asymmetric dilated convolutions.

As shown in Figure 3, which is a comparison between standard convolution Figure 3a and dilated convolution Figure 3b, the dilated convolution method increases the receptive field of the convolutional kernel and reduces redundant information, without changing parameters.

The multi-scale asymmetric dilated convolution spatial attention module designed in this paper is shown in Figure 4. The bar convolutional kernels with a dilation rate of 2, which is marked in blue, allows for the implicit spatial encoding of feature maps in both width and height directions. In Figure 2b, this paper summarizes two common aspect ratios for the target bounding boxes, 1:1 and 4:3. For targets with a 1:1 aspect ratio, the standard rectangular convolutional kernel is sufficient to handle, so we do not make any special design. For targets with a 4:3 aspect ratio, using 3 × 4 and 4 × 3 convolutional kernels will change the size of the output feature map, which is not conducive to model construction. Therefore, for these targets, we designed a set of 3 × 5 and 5 × 3 convolutional kernels with no dilation rate, marked in green.

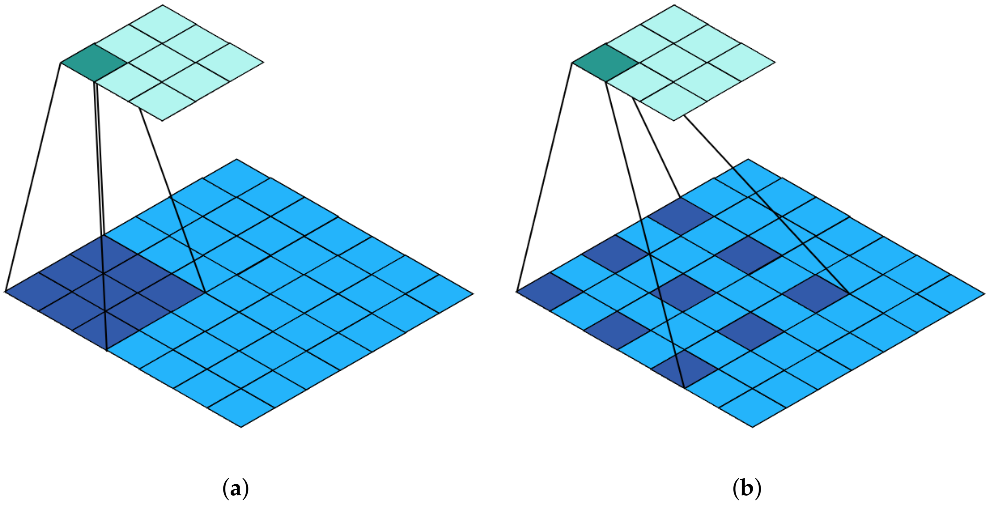


Figure 3. Comparison of standard and dilated convolution: (**a**) Standard Conv; (**b**) Dilated Conv.

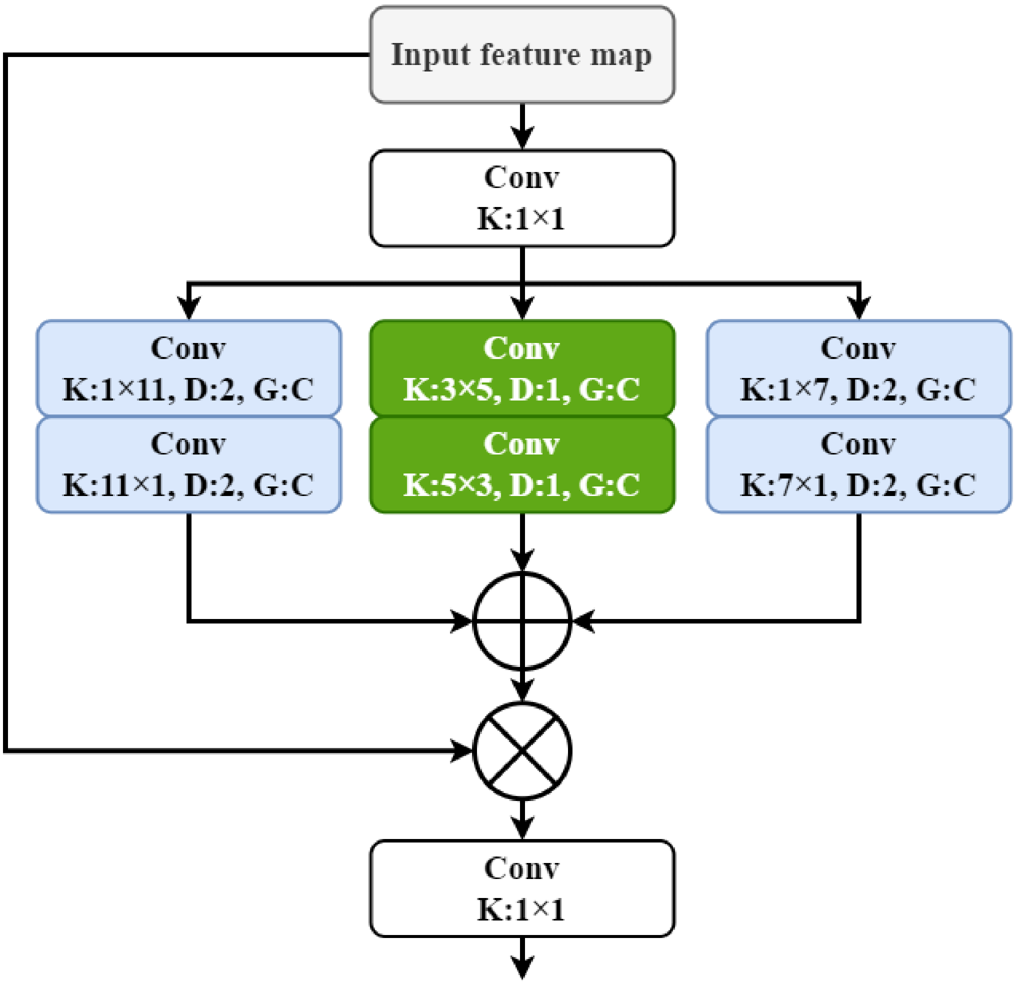


Figure 4. Multi-scale asymmetric dilated convolution spatial attention module.

By combining the channel attention module and multi-scale asymmetric convolution spatial attention, we designed a channel prior dilatation attention module (CPDA), as shown in Figure 5. The input feature maps go through a channel attention module. This channel attention module generates channel attention maps by combining the results of max-pooling and average-pooling, and evaluates the importance of each feature channel by multiplying it with the original feature map. After that, the feature maps enter the multi-scale asymmetric dilated convolution spatial attention module to obtain the final output feature maps.

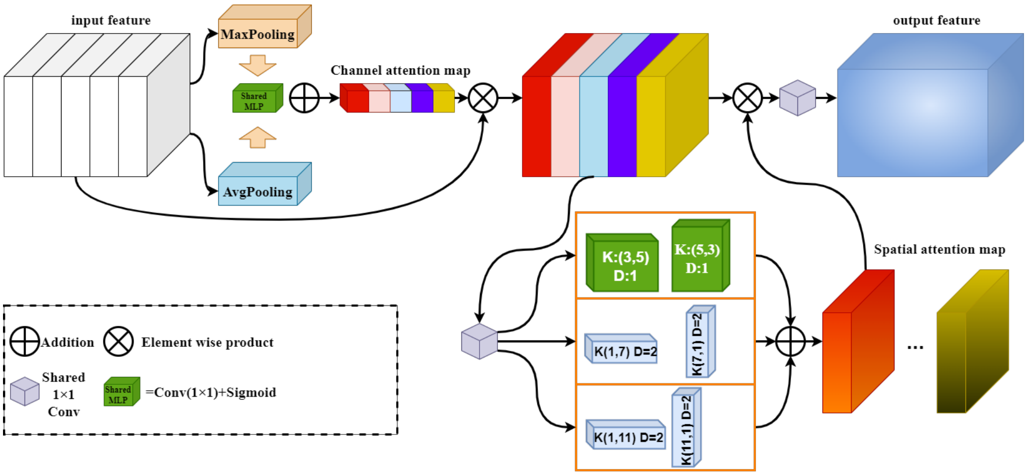


Figure 5. Structure of CPDA.

2.3. *Mixed-Classification Detection Head*

In the design of YOLOv8, the Decoupled Detection Head architecture employed two distinct convolutional modules to independently execute regression and classification tasks for bounding box and class(Cls) prediction. YOLOv7 [23] proposed a method of using an auxiliary detection head to guide the primary detection head to refine prediction outcomes (auxiliary head is not engaged during inference). Motivated by the detection head structure in YOLOv8 and the auxiliary detection head concept, the mixed-classification detection head (MCDH) is designed in this paper.

In the MCDH, both convolutional modules undertake the task of class prediction by appropriately weighting the outputs of these two convolutional layers, and the final class prediction outcome is derived. This architecture harnesses bounding box regression information to guide class prediction. In addition, we also redesign the convolutional kernel and channel to optimize the parameters. The final class prediction is derived from the following formula:

$$Cls^{pred} = (1 - \alpha)Cls_1 + \alpha Cls_2 \tag{1}$$

where Cls_1 represents the additional class prediction value in the original regression branch, and Cls_2 represents the class prediction value in the original classification branch. The value of α is determined to be 0.75 based on comparative experiments in Section 3.4.2.

The classification branch of the original detection head is composed of two 3 × 3 convolutional kernels and one 1 × 1 convolutional kernel, the same as the bounding box regression branch. Considering that our detection task involves only two classes, this has a lot of parameter redundancy. Therefore, through comparative experiments, this paper redesignes the classification branch to consist of three 1 × 1 convolutional kernels, ensuring both the accuracy and lightweightness of the detection head. The comparison between the original detection head and the FFYOLO detection head is shown in Figure 6.

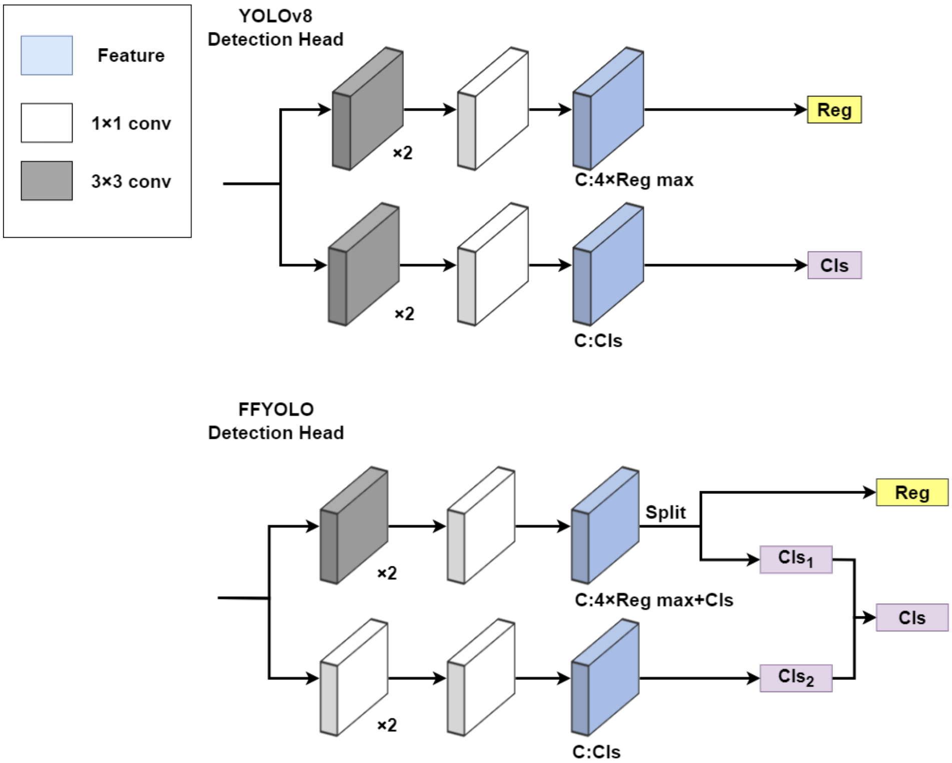


Figure 6. Comparative design of YOLOv8 detection head and FFYOLO detection head.

2.4. Lightweight GSConv

Xiao et al. [24] proposed the C3Ghost and GhostMP modules based on GhostNetv2 [25] to make the backbone of YOLOv7 lighter, thus reducing the computational cost and parameters of the model. In this paper, the standard convolution(SConv) of the Neck section is replaced by the lightweight GSConv [26].

To enable deep learning models to run on low-power devices, Megvii Technology used the Depthwise Separable Convolution (DSConv) to design ShuffleNet [27]. This approach significantly reduces the parameters while maintaining a certain level of accuracy. However, DSConv severs a large number of connections between neurons, inevitably leading to information loss during the backpropagation process. But, GSConv diligently retains these connections, thus delivering outputs more akin to SConv relative to DSConv.

Time complexity comparison (assuming that the convolution kernel size as $K_1 = K_2 = K$, W and H represent the size of the output feature map, while the number of input and output channel are represented by C_1 and C_2):

$$Time_{SConv} \sim O(W \times H \times K^2 \times C_1 \times C_2) \tag{2}$$

$$Time_{DSConv} \sim O[W \times H \times C_1(K^2 + C_2)] \tag{3}$$

$$Time_{GSConv} \sim O[W \times H \times K^2 \times \frac{C_2}{2}(C_1 + 1)] \tag{4}$$

The module structure of GSConv is shown in Figure 7. GSConv allows information from SConv to be mixed into DSConv. This methodology melds the low FLOPs advantage of DSConv with the smooth information exchange of SConv. Analytically, as shown in Figure 8, GSConv's time complexity is merely 50% of that of SConv (as the channel dimension increases, it gets closer to the theoretical value), yet the model retains a learning capability comparable to SConv.

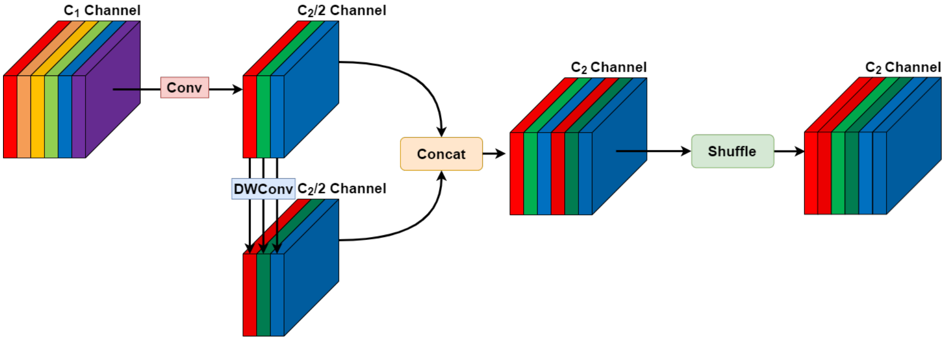


Figure 7. Structure of GSConv.

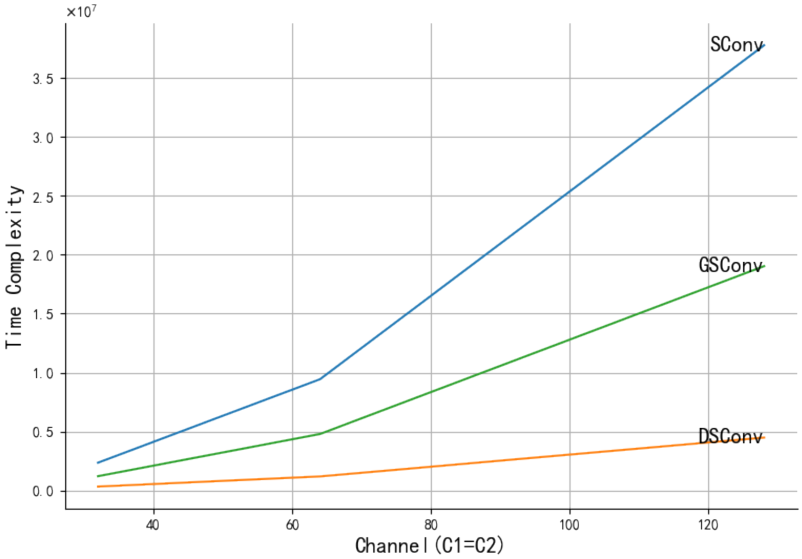


Figure 8. Time complexity of different convolution.

2.5. Soft Label Strategy Based on Knowledge Distillation

In detection tasks, the common practice is to employ the Binary Cross-Entropy (BCE) Loss function for computing the class loss. The BCE Loss measures the performance of a classification model. The formula of the BCE Loss is as follows:

$$y = Sigmoid(y^{pred}) = \frac{1}{1 + e^{y^{pred}}} \tag{5}$$

$$L_{class} = -[y^{*}log(y) + (1 - y^{*})log(1 - y)] \tag{6}$$

where y^{pred} denotes the predicted class value mapped to probability value y between 0 and 1 through the Sigmoid function, and y^{*} denotes the true label that encoded using one-hot encoding.

However, it is difficult for one-hot encoding to express the correlation between different classes. Employing one-hot encoding could engender overconfidence in the model's predictions, thereby inducing significant deviations from the true class label and potentially impairing the model's generalization performance [28].

In this paper, the method of knowledge distillation [29] is used to enhance the model's generalization ability. Knowledge distillation is a technique that compresses the model size and improves speed and efficiency by transferring the knowledge of complex models to simplified models. Complex teacher model has a strong feature extraction capability. The labels provided by the teacher model contain the inter-class relationships it has learned, which can guide the simplified student model for classification and regression, so that the student model can obtain a feature extraction capability equivalent to that of the teacher model. The knowledge distillation flowchart is shown in Figure 9.

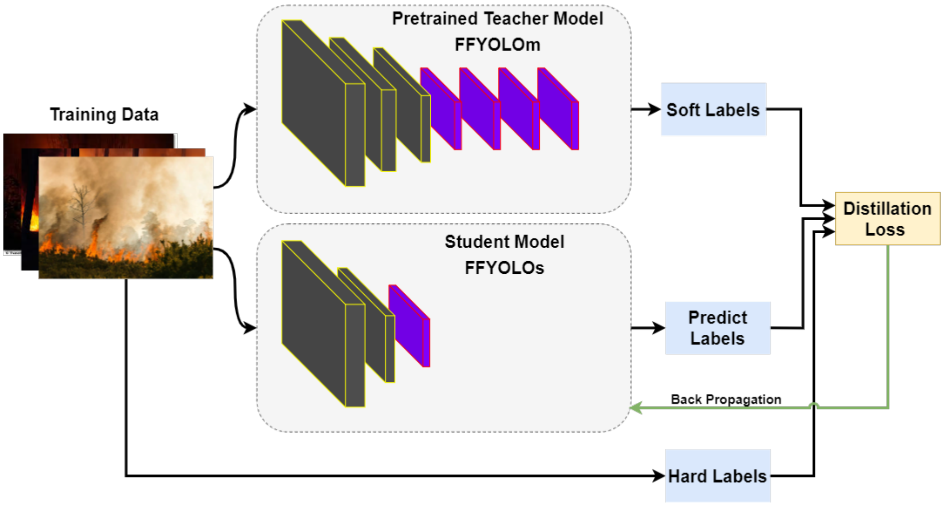


Figure 9. Knowledge distillation flowchart.

The final class loss function is defined as follows:

$$MSE(y^{cls}, y^{tcls}) = (y^{cls} - y^{tcls})^2 \tag{7}$$

$$Scale = Sigmoid(y^{tcls}) = \frac{1}{1 + e^{y^{tcls}}} \tag{8}$$

$$BCEWithLogitsLoss(y^{cls}, y^*) = -[y^* log[Sigmoid(y^{cls})] + (1 - y^*) log[1 - Sigmoid(y^{cls})]] \tag{9}$$

$$L_{class} = (1 - \alpha) BCEWithLogitsLoss(y^{cls}, y^*) + \alpha MSE(y^{cls}, y^{tcls}) \times Scale \tag{10}$$

where y^* represents the one-hot encoded label, and y^{cls} and y^{tcls} are the class values predicted by the student model and teacher model, respectively. $Scale$ measures the confidence of the teacher model's predictions for a certain detection target. It is used to reduce the impact of low-confidence objects on the model training. α determines which part of the loss is more inclined during the model training.

2.6. MPD-IoU Loss

The original YOLOv8 adopted CIoU loss and distributed focus loss for bounding box regression.

CIoU loss is defined as follows:

$$IoU = \frac{Inter\ area}{Union\ area} \tag{11}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{12}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{13}$$

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

where $\rho^2(b, b^{gt})$ measures the Euclidean distance between the center point of the predicted box and the ground truth box, and v measures the similarity between the predicted box and the ground truth box in aspect ratio, while c denotes the diagonal length of the minimum bounding rectangle between the predicted box and the ground truth box. IoU represents the ratio of the intersection and union of the predicted box and ground truth box.

Although CIoU takes into account the aspect ratio, intersection-over-union, and center point loss between predicted and ground-truth bounding boxes, the geometric properties of bounding box regression are not fully utilized in the existing loss functions. Therefore, this paper replaces CIoU with MPDIoU [30], which minimizes the distance between the top-left and bottom-right points of the predicted and ground-truth boxes for bounding box regression. MPDIoU regression method is shown in Figure 10.

The MPDIoU loss function is defined as follows:

$$d_1^2 = (x_1^{pred} - x_1^{gt})^2 + (y_1^{pred} - y_1^{gt})^2, d_2^2 = (x_2^{pred} - x_2^{gt})^2 + (y_2^{pred} - y_2^{gt})^2 \tag{15}$$

$$L_{MPDIoU} = 1 - IoU + \frac{d_1^2}{w^2 + h^2} + \frac{d_2^2}{w^2 + h^2} \tag{16}$$

where $(x_1^{pred}, y_1^{pred}), (x_2^{pred}, y_2^{pred})$ denote the top-left and bottom-right point coordinates of predicted box, $(x_1^{gt}, y_1^{gt}), (x_2^{gt}, y_2^{gt})$ denote the top-left and bottom-right point coordinates of ground truth box.

After knowledge distillation from the previous section, the bounding box regression loss function is defined as follows:

$$Scale = Sigmoid(y^{tcls}) = \frac{1}{1 + e^{y^{tcls}}} \tag{17}$$

$$L_{Box} = (1 - \alpha)L_{MPDIoU} + \alpha MSE(y^{box}, y^{tbox}) \times Scale \tag{18}$$

where $y^{tbox} = [x_1^{tpred}, y_1^{tpred}, x_2^{tpred}, y_2^{tpred}], y^{box} = [x_1^{pred}, y_1^{pred}, x_2^{pred}, y_2^{pred}]$.

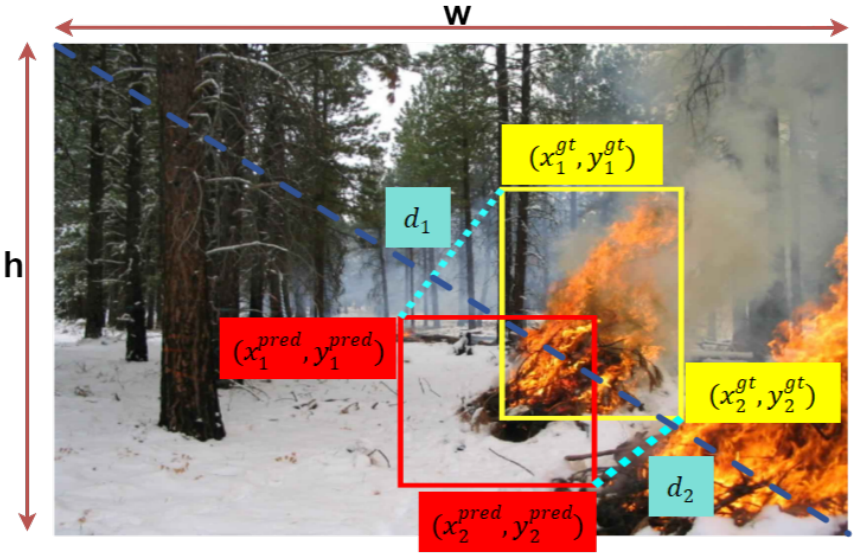


Figure 10. MPD-IoU regression method.

Through MPDIoU, all factors considered in the existing bounding box loss function can be determined by the coordinates of four points, and the conversion formulas are as follows:

$$|C| = (max(x_2^{gt}, x_2^{pred}) - min(x_1^{gt}, x_1^{pred})) \times (max(y_2^{gt}, y_2^{pred}) - min(y_1^{gt}, y_1^{pred})) \tag{19}$$

$$x_c^{gt} = \frac{x_1^{gt} + x_2^{gt}}{2}, y_c^{gt} = \frac{y_1^{gt} + y_2^{gt}}{2}, x_c^{pred} = \frac{x_1^{pred} + x_2^{pred}}{2}, y_c^{pred} = \frac{y_1^{pred} + y_2^{pred}}{2} \tag{20}$$

$$w_{gt} = x_2^{gt} - x_1^{gt}, h_{gt} = y_2^{gt} - y_1^{gt}, w_{pred} = x_2^{pred} - x_1^{pred}, h_{pred} = y_2^{pred} - y_1^{pred} \tag{21}$$

Here, $|C|$ represents the minimum bounding box area covering the predicted box and the ground truth box, (x_c^{pred}, y_c^{pred}) and (x_c^{gt}, y_c^{gt}) are the center coordinates, and w_{pred}, h_{pred} and w_{gt}, h_{gt} are the width and height of the two boxes, respectively.

2.7. FFYOLO

This paper presents a forest fire detection method based on YOLOv8 shown in Figure 11, named FFYOLO, which incorporates CPDA attention module in the model's Backbone section and replaces the detection head with MCDH, so as to enhance the model's feature extraction capability and classification accuracy. Compared to the original detection head, MCDH reduces 60% of the parameters.

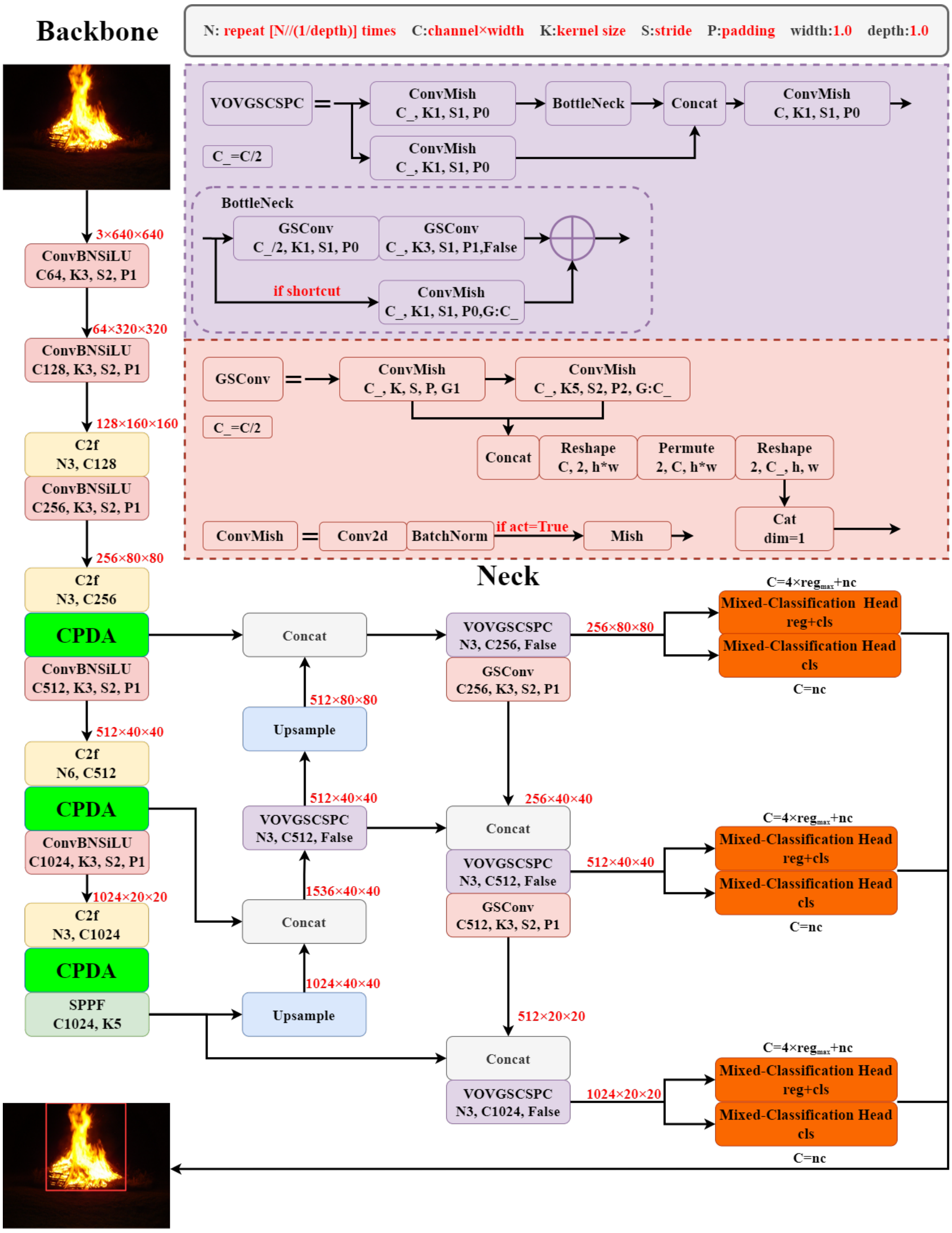


Figure 11. Model structure of FFYOLO.

We also redesigned the structure of the backbone. Compared to the original model, we reduced the repetition times of the second C2f module from 6 to 3 and replaced the

SConv in the Neck section with lightweight GSConv. Additionally, the C2f module in the Neck section is replaced by the VOVGSCSPC module, which combines VOVnet [31], and GS convolution to reduce the model's complexity. This modification engendered a 25.4% reduction in parameters and a 30.6% decrement in FLOPs, while preserving the model's learning capability with the original design.

In the initial stages of feature extraction within the Backbone section, all connections must be preserved to ensure the integrity of information flow to the Neck section. The use of GSConv at initial stage can impede the flow of information and lead to excessive computational complexity. In the Neck section, the majority of information is transferred to the channel dimension, eliminating the need for further information compression. Therefore, it is more appropriate to use GSConv at this stage. The comparison of parameters for the FFYOLO and YOLOv8 modules in Neck section is presented in Table 1.

Table 1. Comparative Analysis of Neck Architectures in FFYOLO and Original YOLOv8.

Neck with GSconv (FFYOLO)			Original Neck (YOLOv8)		
Layer Name	**In/Out Channel**	**Params**	**Layer Name**	**In/Out Channel**	**Params**
VoVGSCSPC(L15)	768/256	307,296	C2f(L12)	768/256	591,360
VoVGSCSPC(L18)	384/128	77,872	C2f(L15)	384/128	148,224
GSConv(L19)	128/128	75,584	Conv(L16)	128/128	147,712
VoVGSCSPC(L21)	384/256	208,992	C2f(L18)	384/256	493,056
GSConv(L22)	256/256	298,624	Conv(L19)	256/256	590,336
VoVGSCSPC(L24)	768/512	827,584	C2f(L21)	768/512	1,969,152

3. Experiments and Analysis

3.1. Dataset

The dataset used in the experiment consists of the D-Fires [32] dataset and data collected from various sources on the Internet. The D-Fires dataset is specifically designed for machine learning and object detection algorithms related to fire and smoke. To ensure the quality of the dataset, a large portion of images with resolutions lower than 384 × 384 were excluded, resulting in a final set of 10,099 images. These images are then annotated using the labeling annotation tool.

The dataset covers a wide range of forest fire scenarios. Additionally, it includes over 500 images of backgrounds that do not contain any fire or smoke, serving as negative samples to enhance the model's ability to distinguish non-forest fire scenes. The dataset is divided into training, testing, and validation sets in a ratio of 7:2:1. A portion of images are exhibited in Figure 12. Detailed information about the dataset is listed in Table 2.

Table 2. Details of Dataset.

Dataset	Number of Images	Number of Targets	Number of Smoke	Number of Fires
Training	7069	17,082	11,941	5141
Validation	2019	4835	3354	1481
Testing	1011	2839	1964	875

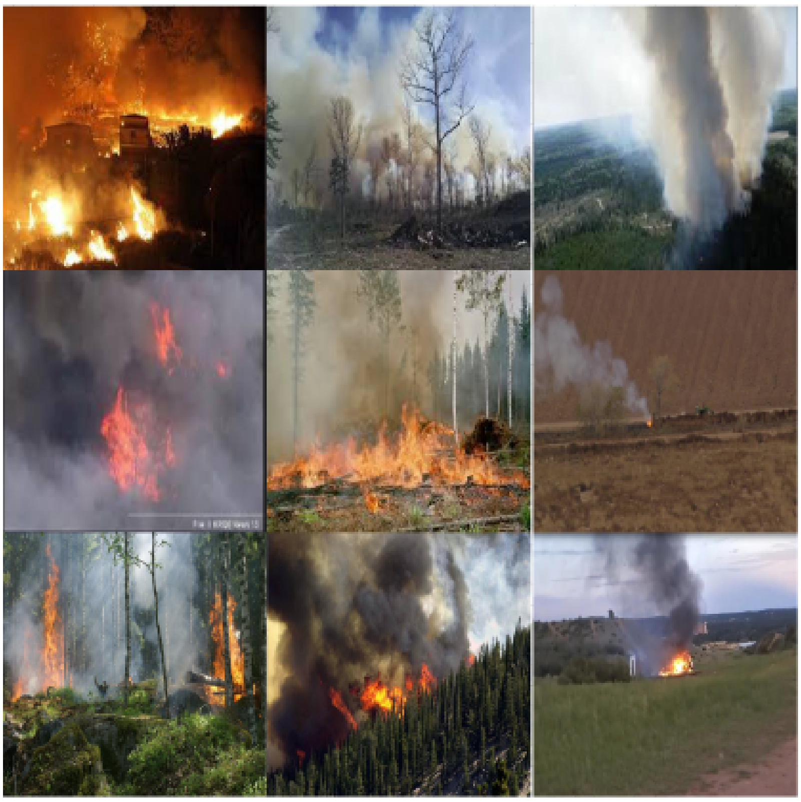


Figure 12. Examples of the experimental data.

3.2. Experimental Environment

This paper employed the PyTorch framework (version 1.13.1) and Python (version 3.8) for model development. All models were trained on an RTX A6000 GPU under the Linux Ubuntu 22.4 operating system. The experimental setting is listed in Table 3.

Table 3. Experimental Setting.

Input Image Size	Epochs	Optimizer	Learning Rate Scheduling	SGD Momentum	Batch Size	Weight Decay
640 × 640	300	SGD	Linear decay (0.01:0.0001)	0.937	32	0.0005

The training regimen incorporated multi-scale input size adjustment, wherein the input image size varied randomly (±50%, step = 32) for each epoch.

3.3. Model Evaluation

The model's performance was gauged by prevalent metrics in object detection: $\mathrm{mAP}_{0.5}$, AP, Parameters, and FLOPs. Herein, $\mathrm{mAP}_{0.5}$ signifies the mean average precision at an IoU threshold of 0.5 on the test set, and AP signifies the average precision of a certain class at an IoU threshold of 0.5 on the test set. The Parameters and FLOPs metrics evaluate the model's complexity in spatial and temporal dimensions, respectively. The frames per second (FPS) metric delineates the processing speed of the model, indicating the number of images processed per second at an input image size of 640 × 640.

The calculation equations are as follows:

$$P = \frac{TP}{TP + FP} \tag{22}$$

$$R = \frac{TP}{TP + FN} \tag{23}$$

$$AP = \int_0^1 P(R)dR \tag{24}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N}\int_{0}^{1} P_i(R)dR \tag{25}$$

$$FPS = \frac{1}{T} \tag{26}$$

where N represents the number of classes; P and R are precision and recall, respectively; TP, FP, and FN correspond to True Positive, False Positive, and False Negative; and T refers to the time required to detect a single image.

3.4. Detect Performance and Analysis

3.4.1. Effectiveness of CPDA

We introduced the SE, CBAM, and CA attention mechanisms into YOLOv8 to compare their performance with CPDA. The comparison results are shown in Figure 13, which shows the performance of incorporating attention modules into the model backbone. It can be seen that the detection accuracy of the model with CPDA attention mechanism is significantly improved.

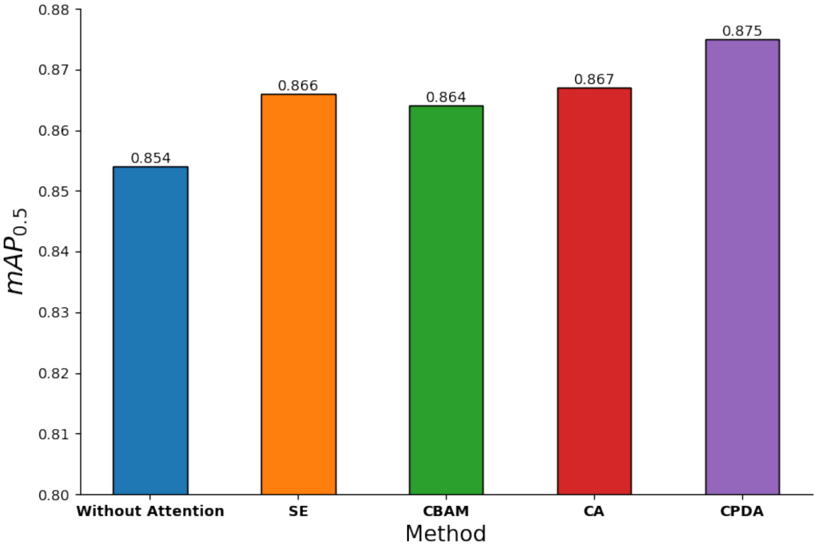


Figure 13. Performance of different attention mechanism.

3.4.2. Effectiveness of MCDH

Relevant experiments that are shown in Figure 14 were conducted on the setting of α in the Formula (1) of Section 2.3. When α is relatively small, the weight of the main classification branch decreases, resulting in a decrease in mAP$_{0.5}$ compared to the original model. Conversely, when α approaches 1, mAP$_{0.5}$ is close to the result of the original model.

3.4.3. Effectiveness of MPD-IoU

Comparison of the model performance with four different IoU loss functions is shown in Figure 15. Compared to the original model with CIoU, YOLOv8 achieved a 1.0% improvement in mAP$_{0.5}$ when MPDIoU was adopted.

3.4.4. Ablation Experiments

The experimental outcomes are encapsulated in Table 4, illuminating the efficiency of the proposed model in forest fire detection tasks.

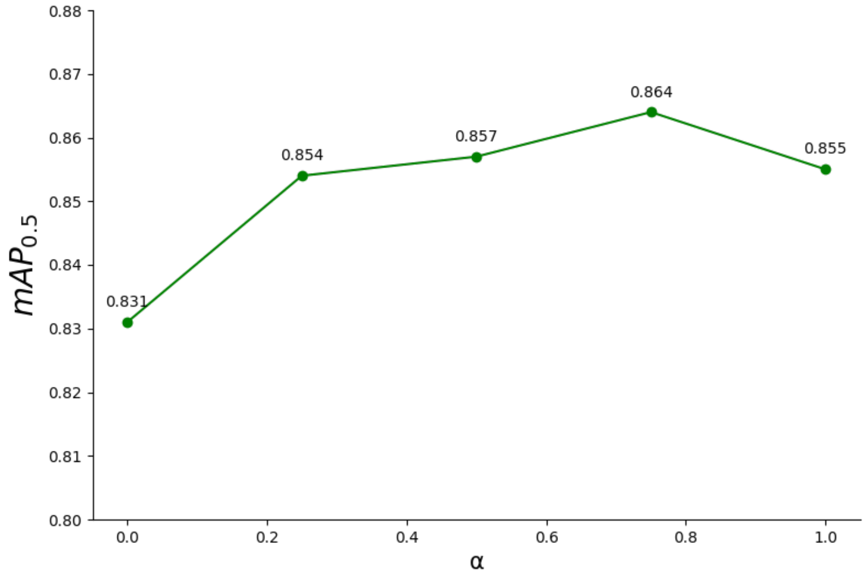


Figure 14. Results of different weight coefficients in MCDH.

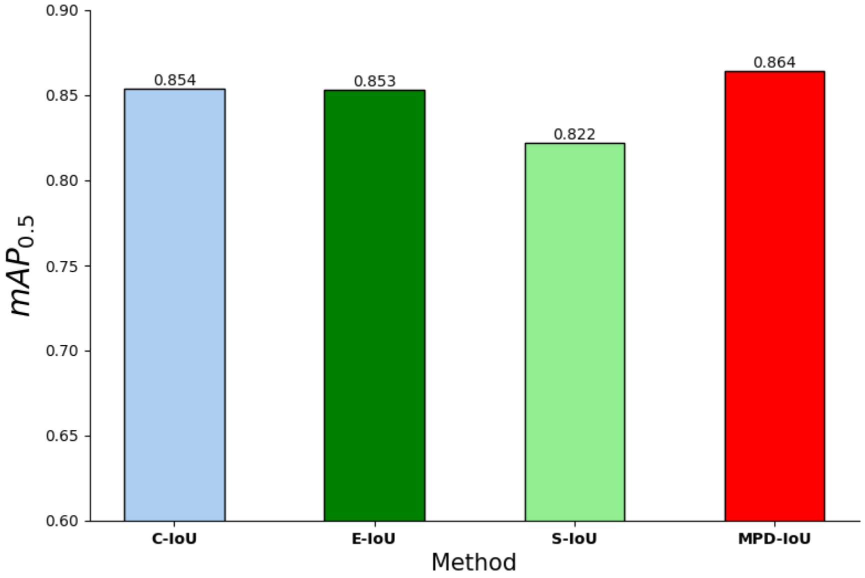


Figure 15. $\text{mAP}_{0.5}$ results of four IoU methods.

Table 4. Ablation experiment results.

Baseline	IoU	CPDA	MCDH	GSConv	Params (M)	FLOPs (G)	FPS	AP_{smoke}	AP_{fire}	$\text{mAP}_{0.5}$
YOLOv5	CIoU	-	-	-	7,015,519	15.8	175	86.0%	84.4%	85.2%
YOLOv8	CIoU	-	-	-	11,166,560	28.8	172	88.0%	82.8%	85.4%
YOLOv8	MPDIoU	-	-	-	11,166,560	28.8	172	89.0%	83.7%	86.4%
YOLOv8	CIoU	✓	-	-	12,235,750	31.2	166	88.8%	86.2%	87.5%
YOLOv8	CIoU	-	✓	-	9,470,310	22.1	169	89.2%	83.5%	86.4%
YOLOv8	CIoU	-	-	✓	8,992,470	24.6	178	88.5%	84.5%	86.5%
YOLOv8 *	CIoU	-	-	-	11,166,560	28.8	171	88.6%	84.7%	86.6%
YOLOv8	CIoU	✓	✓	-	10,569,718	24.6	158	88.6%	86.6%	87.6%
YOLOv8	CIoU	-	✓	✓	7,326,422	18.0	178	89.1%	83.3%	86.2%
YOLOv8	CIoU	✓	-	✓	10,091,878	27.1	164	88.9%	86.5%	87.7%
FFYOLO	MPDIoU	✓	✓	✓	8,343,638	19.5	188	89.1%	87.6%	88.3%
FFYOLO *	MPDIoU	✓	✓	✓	8,343,638	19.5	188	89.5%	88.1%	88.8%

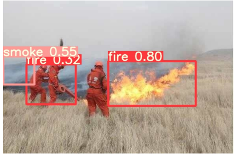
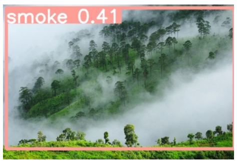
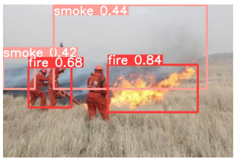
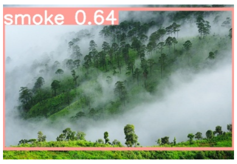
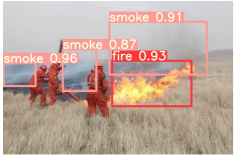
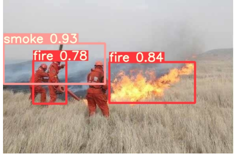
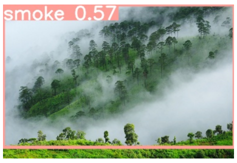
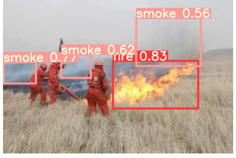
The Baseline that marked with * means model is trained with knowledge distillation strategy.

3.4.5. Model Comparison and Visualization

Figure 16 illustrates the distribution of various models concerning average precision and inference time, with proximity to the top-left corner indicating superior performance. The final model, as depicted, achieves a commendable balance of high precision and rapid inference speed, showcasing its efficacy.

Table 5 illustrates the performance of different models in various scenarios. In (a), YOLOv5, YOLOv8, and RetinaNet misclassified firefighters as fire, while FFYOLO can distinguish them well. (b) demonstrates the detection performance at different distances from the target. YOLOv5s, YOLOv8s, and RetinaNet existed with varying degrees of missed detections, and FFYOLO can successfully detect the majority of targets. In (c), due to the presence of fog near the ground, most models incorrectly identified it as smoke. FFYOLO possessed excellent recognition ability in this scenario. Additionally, Faster-RCNN utilized Resnet50 as the Backbone and also achieved a good performance, but at the cost of a 97.7 MB weight file, while FFYOLO's weight file size was only 17.0 MB.

Table 5. Visual display of detection results.

Model	Detect Results		
YOLOv5			
YOLOv8			
FasterRCNN			
RetinaNet			
FFYOLO			
	(**a**)	(**b**)	(**c**)

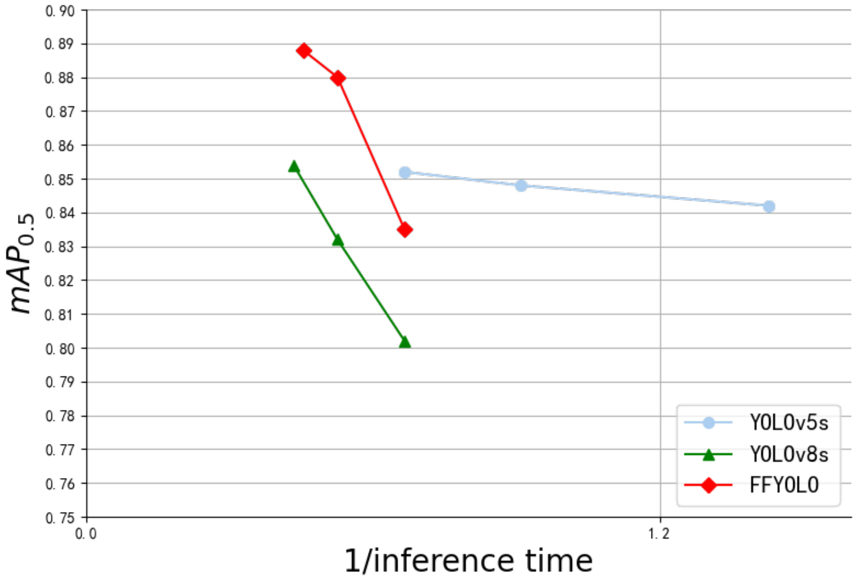


Figure 16. Comparative analysis of inference time and mean average precision.

4. Conclusions

The complexity of a forest environment and the varying environmental factors bring challenges to forest fire detection, which often lead to poor detection and false alarms.

In recent years, with the development of technology and hardware, deep learning algorithms have gradually become the mainstream of forest fire detection. Compared to traditional machine learning algorithms, the forest fire detection algorithm based on deep learning technology has the advantages of high accuracy and good generalization. However, most of them have difficulty in meeting the real-time detection requirements due to their high complexity or excessive parameters.

This paper proposes the FFYOLO model for forest fire detection, aiming to address these challenges. The CPDA attention mechanism is a specially designed CPDA attention mechanism to enhance the feature extraction capabilities of fire and smoke. Additionally, we replace the original detection head with MCDH and introduce GSConv to reduce parameters and complexity while maintaining accuracy. Finally, the MPDIoU and knowledge distillation training strategy is introduced to reduce false and missed detection rates, minimizing the risk of overfitting. In the experiments of this paper, 10,099 forest fire images were partitioned into training, testing, and validation sets in a ratio of 7:2:1. FFYOLO achieved an $mAP_{0.5}$ of 88.8%, and FPS improved by 9.3%; thus, the effectiveness of our improvements is validated.

Therefore, compared with the original YOLOv8 model, FFYOLO shows higher accuracy and efficiency. On one hand, FFYOLO has less parameters and computational complexity, which makes it easier to deploy on low-power devices. On the other hand, FFYOLO is more robust and can detect the forest fire in most complex scenarios.

Forest fire detection is inherently a process of multi-source data fusion prediction. The method proposed in this paper focuses on detecting forest fire using RGB images. In real-world scenarios, the information provided by images is usually limited. Challenges arise when the scene is too bright, the fire is concealed under trees, or the target is far away, making the features of fire and smoke less discernible in the image and consequently making it difficult for the model to detect fire and smoke. Sensors also play a crucial role in forest fire detection, such as infrared sensors, multispectral sensors, and hyperspectral sensors [33], which provide richer information that cannot be captured by images alone. In the future, we will focus on how to combine image data with corresponding sensor data in forest fire detection. This integration of image and sensor information can lead to the development of a more robust forest fire detection model that utilizes multi-source data, thereby improving the accuracy of forest fire detection.

Author Contributions: Conceptualization, B.Y. and Y.Z.; methodology, B.Y. and Y.Z.; software, Y.Z.; validation, B.Y., Y.Z. and Z.L.; formal analysis, B.Y. and Y.Z.; investigation, Z.L. and T.L.; resources, Y.Z.; data curation, Y.Z.; writing–original draft preparation, B.Y. and Y.Z.; writing–review and editing, B.Y., Y.Z., Z.L. and T.L.; visualization, Y.Z.; supervision, B.Y.; project administration, B.Y. and Z.L.; funding acquisition, B.Y. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under grant agreement No. 61972357, by the National innovation and Entrepreneurship training program for college students under grant agreement No. 202311057037, and by the Industry-University-Research innovation fund of Chinese colleges under grant agreement No. 2022IT009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Kanwal, R.; Rafaqat, W.; Iqbal, M.; Song, W. Data-Driven Approaches for Wildfire Mapping and Prediction Assessment Using a Convolutional Neural Network (CNN). *Remote Sens.* **2023**, *15*, 5099. [CrossRef]
2. Kinaneva, D.; Hristov, G.; Raychev, J.; Zahariev, P. Application of artificial intelligence in UAV platforms for early forest fire detection. In Proceedings of the 2019 27th National Conference with International Participation (TELECOM), Sofia, Bulgaria, 30–31 October 2019; pp. 50–53.
3. Xu, R.; Yu, P.; Abramson, M.J.; Johnston, F.H.; Samet, J.M.; Bell, M.L.; Haines, A.; Ebi, K.L.; Li, S.; Guo, Y. Wildfires, global climate change, and human health. *N. Engl. J. Med.* **2020**, *383*, 2173–2181. [CrossRef] [PubMed]
4. Johnston, L.M.; Wang, X.; Erni, S.; Taylor, S.W.; McFayden, C.B.; Oliver, J.A.; Stockdale, C.; Christianson, A.; Boulanger, Y.; Gauthier, S.; et al. Wildland fire risk research in Canada. *Environ. Rev.* **2020**, *28*, 164–186. [CrossRef]
5. Yang, X.; Tang, L.; Wang, H.; He, X. Early detection of forest fire based on unmaned aerial vehicle platform. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–4.
6. Sah, S.; Prakash, S.; Meena, S. Forest Fire Detection using Convolutional Neural Network Model. In Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Tumkur, Karnataka, India, 7–9 April 2023; pp. 1–5.
7. Chen, T.H.; Wu, P.H.; Chiou, Y.C. An early fire-detection method based on image processing. In Proceedings of the 2004 International Conference on Image Processing, ICIP'04, Singapore, 24–27 October 2004; Volume 3, pp. 1707–1710.
8. Ding, X.; Gao, J. A new intelligent fire color space approach for forest fire detection. *J. Intell. Fuzzy Syst.* **2022**, *42*, 5265–5281. [CrossRef]
9. Celik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [CrossRef]
10. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Proceedings, Part I 14, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
13. Barmpoutis, P.; Dimitropoulos, K.; Kaza, K.; Grammalidis, N. Fire detection from images using faster R-CNN and multidimensional texture analysis. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8301–8305.
14. Li, P.; Zhao, W. Image fire detection algorithms based on convolutional neural networks. *Case Stud. Therm. Eng.* **2020**, *19*, 100625. [CrossRef]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
16. Liu, H.; Hu, H.; Zhou, F.; Yuan, H. Forest flame detection in unmanned aerial vehicle imagery based on YOLOv5. *Fire* **2023**, *6*, 279. [CrossRef]
17. Qian, J.; Lin, J.; Bai, D.; Xu, R.; Lin, H. Omni-Dimensional Dynamic Convolution Meets Bottleneck Transformer: A Novel Improved High Accuracy Forest Fire Smoke Detection Model. *Forests* **2023**, *14*, 838. [CrossRef]
18. Li, J.; Xu, R.; Liu, Y. An Improved Forest Fire and Smoke Detection Model Based on YOLOv5. *Forests* **2023**, *14*, 833. [CrossRef]
19. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 3490–3499.

20. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
21. 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, 18–23 June 2018; pp. 7132–7141.
22. Woo, S.; Park, J.; Lee, J.Y.; 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.
23. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
24. Xiao, Z.; Wan, F.; Lei, G.; Xiong, Y.; Xu, L.; Ye, Z.; Liu, W.; Zhou, W.; Xu, C. FL-YOLOv7: A Lightweight Small Object Detection Algorithm in Forest Fire Detection. *Forests* **2023**, *14*, 1812. [CrossRef]
25. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *arXiv* **2022**, arXiv:2211.12905.
26. 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.
27. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
28. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp.4694–4703.
29. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
30. Siliang, M.; Yong, X. MPDIoU: A loss for efficient and accurate bounding box regression. *arXiv* **2023**, arXiv:2307.07662.
31. Lee, Y.; Hwang, J.W.; Lee, S.; Bae, Y.; Park, J. An energy and GPU-computation efficient backbone network for real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 752–760.
32. de Venancio, P.V.A.; Lisboa, A.C.; Barbosa, A.V. An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices. *Neural Comput. Appl.* **2022**, *34*, 15349–15368. [CrossRef]
33. Varotsos, C.A.; Krapivin, V.F.; Mkrtchyan, F.A. A new passive microwave tool for operational forest fires detection: A case study of Siberia in 2019. *Remote Sens.* **2020**, *12*, 835. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.