An Efficient and Lightweight Detection Model for Forest Smoke Recognition

Xiao Guo 1, Yichao Cao 2 and Tongxin Hu 1, * 

1 School of Forestry, Northeast Forestry University, Harbin 150040, China; gx17719893328@163.com
2 School of Automation, Southeast University, Nanjing 210018, China; caoyichao@seu.edu.cn
* Correspondence: htxhtx@nefu.edu.cn

Abstract: Massive wildfires have become more frequent, seriously threatening the Earth’s ecosystems and human societies. Recognizing smoke from forest fires is critical to extinguishing them at an early stage. However, edge devices have low computational accuracy and suboptimal real-time performance. This limits model inference and deployment. In this paper, we establish a forest smoke database and propose a model for efficient and lightweight forest smoke detection based on YOLOv8. Firstly, to improve the feature fusion capability in forest smoke detection, we fuse a simple yet efficient weighted feature fusion network into the neck of YOLOv8. This also greatly optimizes the number of parameters and computational load of the model. Then, the simple and parametric-free attention mechanism (SimAM) is introduced to address the problem of forest smoke dataset images that may contain complex background and environmental disturbances. The detection accuracy of the model is improved, and no additional parameters are introduced. Finally, we introduce focal modulation to increase the attention to the hard-to-detect smoke and improve the running speed of the model. The experimental results show that the mean average precision of the improved model is 90.1%, which is 3% higher than the original model. The number of parameters and the computational complexity of the model are 7.79 MB and 25.6 GFLOPs (giga floating-point operations per second), respectively, which are 30.07% and 10.49% less than those of the unimproved YOLOv8s. This model is significantly better than other mainstream models in the self-built forest smoke detection dataset, and it also has great potential in practical application scenarios.

Keywords: lightweight model; forest smoke detect; YOLOv8s; SimAM; BiFPN

1. Introduction

Forests provide important ecosystem services, such as climate regulation, the recharge of aquifers, food production, and greenhouse gas sequestration [1]. However, large-scale forest fires have occurred frequently, causing serious damage to the Earth’s ecosystem and human society in recent years. In Australia, Russia, the United States, Canada, and other countries, the burnt area of mega forest fires in recent years has reached tens of thousands or even hundreds of thousands of square kilometers, having serious impacts on the local ecosystem and the personal safety of local residents. According to the statistics of the Fire and Rescue Bureau of China’s Ministry of Emergency Management, 709 forest fires occurred nationwide in 2022, affecting a forest area of about 4689.5 hectares [2]. Wildfires occur primarily in wilderness areas with high levels of flammable materials. In summer and fall, in particular, when the temperature rises and the water content in the forest is low, the likelihood of a fire increases. Additionally, when the wind is strong, fires spread faster. At the same time, forests are more open and oxygen-rich, making it easier for fires to start and spread. Once a fire has spread, it is difficult to control due to changes in terrain, which can cause serious risks to personal safety and economic losses. If a fire occurs on a large scale, it will cause incalculable casualties and economic losses. Therefore, there is an urgent need for the development of a fast and effective fire detection solution.
Early forest fire monitoring methods mainly included ground patrols, lookout monitoring, and aerial patrols. The first two monitoring methods are relatively single-stage approaches, and it is difficult to realize real-time monitoring. The latter method is costly, and it is not easy to realize all-weather monitoring [3]. Traditional smoke detection can be divided into five stages, namely input, preprocessing, feature selection, smoke detection, and output results, mainly focusing on the preprocessing and feature selection stages. Toreyin et al. [4] used the texture features of smoke for detection, calculated the wavelet energy of the suspected motion region using wavelet transform, and judged whether it was smoke or not based on the trend of wavelet energy. Lecun et al. [5] used structural wavelet transform to represent the smoke texture image, calculated the different scales of wavelet transform by using GLCM, and used neural networks to classify the smoke candidate regions. Chen et al. [6] proposed a block-based interframe differencing and LBP-TOP combination method for smoke dynamic characterization. To reduce false alarms, a smoke histogram was constructed to record the most recent classification of smoke candidate regions. In conclusion, while the above studies have introduced various methods to detect texture features of smoke, traditional smoke detection involves complex image processing methods with poor generalization performance and robustness.

With the application and success of deep learning in image processing, more and more research has been conducted to improve the performance of smoke detection with computer vision. Currently, end-to-end object detection methods based on deep learning can be divided into two types. The first group of models involves two-stage object detection based on candidate regions, such as Faster R-CNN [7], R-FCN [8], and Libra R-CNN [9]. The second group involves one-stage regression-based object detection, such as YOLO [10–12], SSD [13], EfficientDet [14], and RetinaNet [15]. Specifically speaking, although two-stage algorithms have high detection accuracy, they are not suitable for real-time detection tasks. Single-stage object detection typically requires only a single forward computation, but two-stage methods may still have some advantages for object tasks that require high accuracy. Zhang et al. [16] constructed a simulated smoke dataset and trained it with their proposed deep convolutional generative adversarial network. They effectively monitored the smoke region and reduced false alarms, but their approach was hardware-demanding and difficult to be widely deployed to meet real-time requirements. Chao et al. [17] proposed a color-guided anchoring strategy and a global information-guided flame detection method for flame detection and combined them with Faster R-CNN for fire detection, which improved the detection speed and overall accuracy. However, Faster R-CNN has a large number of parameters, and its real-time performance of detection is not high. Based on the overall structure of YOLOv5, Zhou et al. [18] used MobileNetv3 as the backbone network and semi-supervised knowledge distillation (SSKD) for training, but the convergence speed and accuracy of the model need to be further improved. Chen et al. [19] proposed an improved multiscale forest fire detection model YOLOv5s-CCAB based on YOLOv5s to address the problem of poor detection accuracy due to the multiscale characteristics and variable morphology of forest fires. Although the accuracy is improved, there is still room for optimizing the operational load of forest fire detection equipment. In conclusion, how to better balance accuracy and speed remains a challenge for forest smoke detection.

In this paper, we propose a lightweight forest smoke detection model based on YOLOv8s. The contributions are as follows:

- We propose an efficient and lightweight forest smoke detection model to successfully detect and localize smoke in the early stages of forest fire spread. Edge devices with limited computing resources and memory can achieve real-time object detection. New ideas and insights are provided for forest fire detection.
- We introduce a BiFPN module in the neck of YOLOv8. This enables the network to aggregate features at different resolutions by incorporating learnable weights to learn the importance of different input features, thus improving the performance of detecting forest smoke.
• SimAM is introduced into the YOLOv8 model to effectively suppress the interference of redundant information on the network by enhancing the attention of the down-sampling unit and the basic unit in the convolutional neural network. To expand the sensory field, SPPF is replaced with focal modulation to learn coarse-grained spatial information and fine-grained feature information to improve the performance of the network.

• The proposed detection model significantly outperforms other methods in the self-built forest smoke detection dataset. Compared with the original YOLOv8s, the mean average precision of forest smoke detection is improved by 3%. The number of parameters and the computational complexity of the model are reduced by 30.07% and 10.49%, respectively.

2. Materials and Methods

2.1. Forest Smoke Detection Based on YOLOv8 Architecture

YOLOv8 [20] was published in January 2023 by Ultralytics, the company that developed YOLOv5. YOLOv8 is designed to be a fast, single-stage object detection algorithm, and is provided in five different scaled versions: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), and YOLOv8x (extra large). It supports image classification, object detection, and instance segmentation tasks and runs on a variety of hardware platforms from the Central Processing Unit (CPU) to the Graphics Processing Unit (GPU). The network structure of YOLOv8 consists of four parts: input, backbone, neck, and head. Input: This mainly consists of mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling. Backbone: The C3 structure of YOLOv5 is replaced by the C2f structure with richer gradient flow, and the number of channels is adjusted differently for different scale models to achieve further lightweightness. At the same time, the SPPF [21] module in Yolov5 is used to fine-tune the model for different scales, which greatly improves the model’s performance. Neck: YOLOv8 combines the idea of PANet and introduces a PAN [22] and FPN [23] architecture in its neck component to enhance the feature fusion capability of the model. This architecture uses top-down upsampling and bottom-up downsampling techniques to fuse strong semantic information and strong localization information, which helps improve the accuracy and robustness of object detection. Head: The original coupled head is replaced by the current mainstream decoupled head structure, which separates the classification and detection heads, and the anchor-based approach is replaced by the anchor-free approach. Since the coupling of the detection head affects the performance of the model, replacing the detection head of YOLO with the decoupled head significantly improves the convergence speed of the model. In addition, loss computation includes two branches: classification and regression. The bounding box regression prediction task employs complete intersection over union (CIoU) and distributed focus loss (DFL) [24]. Meanwhile, the classification task is supported by binary cross-entropy loss (BCE Loss). This design choice helps to improve the recognition accuracy and accelerate the model convergence. The improved YOLOv8s architecture is shown in Figure 1.
2.2. Bidirectional Feature Pyramid Network (BiFPN)

FPN [23] is a top-down approach. It upsamples the coarse location information at the lower level and semantically stronger feature maps at the higher pyramid level to produce higher-resolution features. PAN [22] complements FPN by downsampling from the bottom up so that the top-level features contain image location information. Because FPN uses summation for feature fusion, some detail is lost in the fusion process. Therefore, it may not perform as well as PAN for scenes that require high-precision detection. PAN uses a cascading approach to feature fusion, which can preserve more detail but increases computational complexity.

The bidirectional feature pyramid network (BiFPN) was proposed by Tan et al. in 2020 [14]. Based on the PANet, the BiFPN first removes nodes that are only shallowly sampled without more information. Then, for nodes whose inputs and outputs are located in the same layer, an extra edge is added to fuse more features without increasing the cost. The BiFPN is equivalent to providing different weights to each layer for fusion, allowing the network to pay more attention to the important layers, and it also reduces the node connectivity of some unnecessary layers. In this way, it can better balance the feature information of different scales and improve the detection accuracy of small targets. The BiFPN architecture is shown in Figure 2.

Figure 2. The architecture of the smoke detection network based on improved YOLOv8s.
2.3. Focal Modulation

Focal modulation was presented by Yang et al. [25] in 2022. It mainly incorporates a multilevel feature fusion mechanism in the module to learn both coarse-grained spatial information and fine-grained feature information to improve the performance of the network. Focal modulation can improve the performance of the network according to the object size to increase the focus on hard-to-detect targets, thus improving the detection accuracy of the model.

\[ y_i = M_1(T_1(x_i, X), X), \]
\[ y_i = T_2(M_2(i, X), x_i), \]

Equation (1) is self-attention, where aggregation \( M_1 \) on context \( X \) is performed after calculating the attention score between query and target through interaction process \( T_1 \). In contrast, Equation (2) is focal modulation, where the context features are first aggregated using \( M_2 \) at each location \( i \), and then the query interacts with the aggregated feature based on \( T_2 \) to form \( y_i \). Focal modulation is defined with Equation (3).

\[ y_i = q(x_i) \odot m(i, X), \]

where \( q \) denotes the query mapping function; \( \odot \) indicates the element-wise multiplication; and \( m \) is the context aggregation operation, which consists of two steps: hierarchical semantics in Equation (4) and gated aggregation in Equation (5).

\[ Z^l = f^l_a(Z^{l-1}) \triangleq \text{GeLU}(DWConv(Z^{l-1})) \in R^{H \times W \times C} \]
\[ Z^{out} = \sum_{l=1}^{L+1} C^l \odot Z^l \in R^{H \times W \times C} \]
\[ y_i = q(x_i) \odot h \left( \sum_{l=1}^{L+1} \delta^l_i \times Z^l_i \right) \]

In Equation (4), \( f^l_a \) is the context function of the \( l \)-th layer, generated by deep-wise convolution with kernel size \( k^l \) and the GeLU activation function. Hierarchical semantics extracts context information from local to global ranges through different levels of gran-
In Equation (5), \( G \in \mathbb{R}^{H \times W \times 1} \) is a slice of \( G \) for the level \( l \). Specifically, we use a linear layer to obtain spatial- and level-aware gating weights: \( G = f_G(X) \in \mathbb{R}^{H \times W \times (L+1)} \). Then, we perform a weighted sum through an element-wise multiplication to obtain a single feature map \( Z_{out} \), which has the same size as the input \( X \). Gated aggregation condenses context features at different levels of granularity into a single feature vector, the modulator. Combining the previous interaction and aggregation, the focal modulation formula can be expressed as Equation (6), where \( g_i^l \) and \( Z_i^l \) are the gating values and visual features at position \( i \) of \( G_i^l \) and \( Z_i^l \), respectively. The Focal modulation and its detailed explanation can be found in Figure 3.

![Figure 3](image-url)

**Figure 3.** Detailed explanation of context aggregation (b) in focal modulation (a). The aggregation procedure consists of two steps: hierarchical contextualization to extract contexts from local to global ranges at different levels of granularity and gated aggregation to condense all context features at different granularity levels into the modulator.

### 2.4. Simple and Parameter-Free Attention Module (SimAM)

The images in the forest smoke dataset may contain interference from complex background and environmental factors, such as trees and terrain, which increase the difficulty of object detection. At the same time, as the number of network layers increases, the weight of the interfering information in the feature map also increases, which has some negative impact on the model. Transformers have achieved greater success in natural language processing, image classification, object detection, and image segmentation in recent years. The reason is that self-attention plays a key role and is, therefore, able to support the global interaction of input information. However, its complexity exceeds the number of visual tokens. The high computational complexity of self-attention, especially with high-resolution inputs, may require more computational resources and datasets for training. Therefore, it is not practical to deploy on forest smoke edge computing devices.

In this paper, we introduce the simple and parametric-free attention mechanism (SimAM) [26], a module that can efficiently generate realistic 3D weights to enhance the attention of downsampling and basic units without additional parameters. Compared to other attention mechanisms, the SimAM operates in a more concise manner while allowing the module to maintain its lightweight properties. Therefore, in this work, the SimAM is introduced into the YOLOv8 model to effectively suppress the interference of redundant information in the network and extract essential feature information from the complex background of forest smoke images. This enables the network to pay more attention to
the essential features related to forest smoke images, enhancing the perceptual ability and adaptivity of the model and thus improving the detection accuracy and reducing the network complexity. The calculation formula is shown in Equations (7)–(11).

\[
e_t(w_t, b_t, y, x_t) = (y_t - \tilde{t})^2 + \frac{1}{M-1} \sum_{i=1}^{M-1} (y_0 - \tilde{x}_i)^2
\]  

(7)

where \(\tilde{t} = w_t t + b_t\) and \(\tilde{x}_i = w_i x_i + b_i\) are linear transforms of \(t\) and \(x_i\), and \(x_i\) denote the target neuron and other neurons in a single channel of the input feature, respectively; \(i\) is the index over spatial dimension, and \(M = H \times W\) is the number of neurons on that channel; \(w_t\) and \(b_t\) are the weight and bias of the transform, computed with the following equations:

\[
w_t = -\frac{2}{(t - \bar{t})^2 + 2\hat{\sigma}^2 + 2\lambda} (t - \mu_t)
\]  

(8)

\[b_t = -\frac{1}{2} (t + \mu_t) w_t\]  

(9)

\[e_t^i = \frac{4}{(t - \bar{t})^2 + 2\hat{\sigma}^2 + 2\lambda} (\hat{\sigma}^2 + \lambda)
\]  

(10)

where \(\bar{t} = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i\) and \(\hat{\sigma}^2 = \frac{1}{M-1}\sum_{i=1}^{M-1} (x_i - \bar{t})^2\) denote the mean and variance, respectively, computed for all neurons in the channel except \(t\). Equation (10) shows that the minimal energy \(e_t^i\) and neuron \(t\) are more distinctive from the surrounding neurons and more important for visual processing.

\[
\tilde{X} = \text{sigmoid} \left( \frac{1}{E} \right) \odot X
\]  

(11)

Equation (11) is the augmented feature tensor, where \(E\) groups all \(e_t^i\) across channels and spatial dimensions across all channels and spatial dimensions, while \(\odot\) is a dot product operation. Excessively large values of \(E\) are limited by adding a sigmoid function that does not affect the relative importance of each neuron. The complete 3D weights of SimAM are shown in Figure 4.

Figure 4. SimAM structure diagram. The feature X expansion generates 3D weights, which are normalized with a function. The weights of the target neurons are multiplied by the features of the initial feature map to obtain the final output feature map. The same color indicates that a single scalar is used for each point on that feature map.
3. Results and Analysis

3.1. Datasets and Implementation Details

The Forest Smoke Dataset has a total of 8176 images. The dataset was divided into a training set (6540 images), a validation set (817 images), and a test set (819 images) according to the ratio of 8:1:1. The images covered three scenes of forests, residential areas, and fields. Figure 5 shows some representative sample images on the Forest Smoke Dataset.

![Sample Images](image)

Figure 5. Partial samples from the Forest Smoke Dataset: (a) smoke generated before forest fires; (b) forest fire; (c) smoke from residential areas; (d) field smoke.

In this work, the model experimental environment is shown in Table 1, and the training parameters related to the forest smoke detection model are shown in Table 2. In particular, we selected YOLOv8s as the benchmark for our evaluation among the various versions of YOLOv8.

Table 1. Experimental environment.

<table>
<thead>
<tr>
<th>Test Environment</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming language</td>
<td>Python 3.9</td>
</tr>
<tr>
<td>Operating system</td>
<td>Windows 11</td>
</tr>
<tr>
<td>Deep learning framework</td>
<td>PyTorch 1.10.0</td>
</tr>
<tr>
<td>Run device</td>
<td>NVIDIA GeForce GTX 3090</td>
</tr>
</tbody>
</table>

Table 2. The training parameter of forest fire detection model.

<table>
<thead>
<tr>
<th>Training Parameters</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs</td>
<td>200</td>
</tr>
<tr>
<td>Batch size</td>
<td>32</td>
</tr>
<tr>
<td>Image size (pixels)</td>
<td>640 × 640</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.001</td>
</tr>
<tr>
<td>Optimizer</td>
<td>SGD</td>
</tr>
</tbody>
</table>

3.2. Evaluation Metrics

In the experiments in this study, we used the mean average precision (mAP), precision, and recall to evaluate the performance of forest smoke detection. True positive (TP) was
considered correct only when intersection over union (IOU) ≥ 0.5. The calculations for precision and recall are given with Equations (12) and (13), respectively.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (12)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (13)
\]

where TP denotes the number of correctly detected smoke samples, and FP represents the number of non-smoke samples falsely detected as smoke. The higher the accuracy, the lower the false detection rate. FN indicates the amount of smoke sample leakage.

\[
AP = \frac{\sum_{i \in \{0.5, 0.55, ..., 0.95\}} AP_i}{10} \quad (14)
\]

\[
mAP = \frac{1}{N} \sum_{n=1}^{N} AP(n) \quad (15)
\]

\[
F_1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (16)
\]

The AP is the area under the P–R (precision–recall) curve. It is used to evaluate the performance of object detection for a single class, as shown in Equation (14). mAP is the average of the AP values for all categories, which takes into account the differences between different categories and reflects the model performance more comprehensively. In this study, the accuracy refers to mAP50. The F1 Score is a harmonic average of precision and recall. A higher F1 Score means a better detection effect on forest smoke detection.

### 3.3. Effect of Different Attention Mechanisms

The effect of integrating different attention modules at the end of the YOLOv8s backbone, i.e., after the SPPF, on the model detection performance is shown in Table 3. Adding only the attention mechanism results in a slight increase in the computational cost and the number of parameters of the model, while the accuracy shows a significant change. The GAM [27] attention mechanism takes into account the information of the channel, and the height and width dimensions simultaneously, resulting in a better capture of cross-dimensional interactions. The mean average precision of the model improves by up to 3.1%, but the computational cost increases by 58.71%. On the other hand, the CBAM [28] attention mechanism is able to consider information in both channel and space dimensions but ignores channel–space interactions, thus losing cross-dimensional information. However, there is a certain improvement in model performance. In comparison, the improved SimAM attention mechanism achieves a good trade-off between the different metrics analyzed in this paper, with a 3.3% improvement in the mean average precision and no significant changes in the amount of parameter volume computed.

<table>
<thead>
<tr>
<th>Models</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>mAP50 (%)</th>
<th>Parameter (MB)</th>
<th>GFLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv8s [20]</td>
<td>86.1</td>
<td>82.6</td>
<td>87.1</td>
<td>11.14</td>
<td>28.6</td>
</tr>
<tr>
<td>+GAM [27]</td>
<td>88.3</td>
<td>85.1</td>
<td>90.2</td>
<td>17.68</td>
<td>33.7</td>
</tr>
<tr>
<td>+EMA [29]</td>
<td>86.3</td>
<td>83.9</td>
<td>89.6</td>
<td>11.18</td>
<td>28.9</td>
</tr>
<tr>
<td>+CBAM [28]</td>
<td>85.6</td>
<td>85.0</td>
<td>89.3</td>
<td>11.39</td>
<td>28.7</td>
</tr>
<tr>
<td>+SE [30]</td>
<td>82.4</td>
<td>86.7</td>
<td>89.4</td>
<td>11.16</td>
<td>28.6</td>
</tr>
<tr>
<td>+CoordAttention [31]</td>
<td>86.8</td>
<td>80.8</td>
<td>89.8</td>
<td>11.15</td>
<td>28.4</td>
</tr>
<tr>
<td>+SimAM [26]</td>
<td>86.6</td>
<td>85.3</td>
<td>90.4</td>
<td>11.14</td>
<td>28.6</td>
</tr>
</tbody>
</table>
3.4. Effects of Different Detection Models

To evaluate the performance of the improved model, comparative experiments with mainstream object detection models were conducted using our self-constructed forest smoke dataset. The selected models were two-stage object detection based on candidate regions (Faster R-CNN) and one-stage object detection based on regression (SSD, EfficientDet, YOLOv3-tiny, and YOLOv5). In Table 4, it is obvious that EfficientDet and YOLOv3-tiny achieved the second- and third-best results, with accuracies of 89.37% and 87.9%, and F1 of 0.8 and 0.841, respectively. In comparison, YOLOv5-ShuffleNet [32] and SSD [13] showed lower performance, with accuracies of 75.9% and 76.28%, respectively. In contrast, our improved YOLOv8s model exhibited significant results, with a mean average precision as high as 90.1% and an F1 score of 0.89, which was significantly higher than other object detection models. Additionally, the improved YOLOv8s had stronger generalization performance and robustness.

Table 4. Comparison of the performance of different object detection models.

<table>
<thead>
<tr>
<th>Models</th>
<th>mAP50 (%)</th>
<th>Parameter (MB)</th>
<th>GFLOPs (G)</th>
<th>FPS</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN [7]</td>
<td>82.8</td>
<td>137.1</td>
<td>370.2</td>
<td>56</td>
<td>0.51</td>
</tr>
<tr>
<td>SSD [13]</td>
<td>76.28</td>
<td>26.29</td>
<td>62.75</td>
<td>85</td>
<td>0.7</td>
</tr>
<tr>
<td>EfficientDet [14]</td>
<td>89.37</td>
<td>3.83</td>
<td>4.74</td>
<td>92</td>
<td>0.8</td>
</tr>
<tr>
<td>YOLOv3-tiny [10]</td>
<td>87.9</td>
<td>8.67</td>
<td>12.9</td>
<td>223</td>
<td>0.84</td>
</tr>
<tr>
<td>YOLOv5 [11]</td>
<td>86.7</td>
<td>7.23</td>
<td>16.4</td>
<td>19</td>
<td>0.84</td>
</tr>
<tr>
<td>YOLOv5+ShuffleNet [32]</td>
<td>75.9</td>
<td>3.8</td>
<td>8.0</td>
<td>93</td>
<td>0.74</td>
</tr>
<tr>
<td>Ours</td>
<td>90.1</td>
<td>7.79</td>
<td>25.6</td>
<td>100</td>
<td>0.89</td>
</tr>
</tbody>
</table>

3.5. Ablation Experiment

As shown in Table 5, reconstructing the neck of YOLOv8 using the BiFPN resulted in a 33.84% reduction in parameter count and an 11.89% reduction in computational complexity. In addition, the mAP50 improved by 1.8%. Replacing the SPPF with focal modulation had little effect on model accuracy, parameter count, and computational complexity. However, it is evident that the model’s mean average precision improved by 1.4%, and the detection speed improved by 8 FPS. The introduction of the SimAM attention module resulted in the model parameter counts and computational complexity remaining largely unaltered, and it is worth noting that the mean average precision of the model improved by 3.3%, thus improving the performance of the forest smoke detection model. The combination of the BiFPN and the focal modulation module improved the accuracy of the model while greatly reducing the computation and number of parameters of the network. Overall, on the same dataset, the improved version of YOLOv8 combining the BiFPN, focal modulation, and SimAM modules outperformed the original YOLOv8 model in terms of detection accuracy, computational complexity, and number of parameters. The improved YOLOv8 model had a 3% improvement in the mean average precision, a 30.07% reduction in the number of parameters, and a 10.49% reduction in computational complexity.

Table 5. Comparison results of ablation experiments.

<table>
<thead>
<tr>
<th>BiFPN</th>
<th>Focal Modulation</th>
<th>SimAM</th>
<th>Precision (%)</th>
<th>mAP50 (%)</th>
<th>Parameter (MB)</th>
<th>GFLOPs</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>√</td>
<td></td>
<td></td>
<td>86.1</td>
<td>87.1</td>
<td>11.14</td>
<td>28.6</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>86.6</td>
<td>88.9</td>
<td>7.37</td>
<td>25.2</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td>86.0</td>
<td>88.5</td>
<td>11.54</td>
<td>28.8</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>√</td>
<td>86.6</td>
<td>90.4</td>
<td>11.14</td>
<td>28.6</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>√</td>
<td>86.5</td>
<td>89.3</td>
<td>7.79</td>
<td>25.6</td>
<td>95</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
<td>87.1</td>
<td>90.1</td>
<td>7.79</td>
<td>25.6</td>
<td>100</td>
</tr>
</tbody>
</table>
We selected a few representative images from our test set to better demonstrate the feasibility of the model. The comparison of the detection results between YOLOv8 and our improved forest fire and smoke detection model is shown in Figures 6 and 7. Considering the detection scene, whether it is a residential area or a field, our model is able to identify the target well. In particular, for the small smoke in Figure 6c, the accuracy of the improved model is significantly higher than previous models.

Figure 6. Comparison of smoke detection results before (right) and after (left) improvement: (a,b) and (c,d) show the smoke detection results in residential areas and fields, respectively. The smoke detection results of medium scales and small scales are also shown.

Figure 7. Comparison of the results for the detection of smoke leakage: (a,c) the smoke detection results using our model; (b,d) the smoke detection results using the original YOLOv8s.
Figure 7 compares the effect of the missed detection problem when monitoring ground smoke. The improved YOLOv8s can detect forest smoke more accurately. However, the detection results of the unimproved YOLOv8s are less accurate, missing one smoke target (Figure 7b,d). This may be because the improved YOLOv8s has more robustness and better target detection performance.

3.6. Effects of Different Datasets

In this section, we present the results of a systematic experimental comparison of the EfficientDet, YOLOv3-tiny, YOLOv4, YOLOv5, YOLOv8s, and the improved YOLOv8s models on the self-built forest smoke dataset and the publicly available dataset used by Venancio et al. [33]. As can be seen in Table 6, the improved YOLOv8s performs better in terms of detection accuracy and robustness on the forest smoke dataset. This can be attributed to the fact that the publicly available dataset used by Venancio et al. [33] has a low pixel resolution, which limits its effectiveness in real scenes. In contrast, our self-constructed dataset has a higher pixel resolution and a larger number of diverse scenarios, which allows for a more comprehensive evaluation of the performance of the model in a complex forest environment.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Models</th>
<th>Precision (%)</th>
<th>mAP50(%)</th>
<th>Parameter (MB)</th>
<th>GFLOPs</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>The D-Fire Dataset</td>
<td>EfficientDet [14]</td>
<td>66.3</td>
<td>71</td>
<td>3.79</td>
<td>4.7</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>YOLOv3-tiny [10]</td>
<td>62</td>
<td>68.7</td>
<td>8.6</td>
<td>12.5</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>YOLOv5 [11]</td>
<td>64.8</td>
<td>67.1</td>
<td>7.2</td>
<td>16.2</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>YOLOv4-tiny [33]</td>
<td>60.9</td>
<td>62</td>
<td>5.9</td>
<td>16</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>YOLOv4 [33]</td>
<td>69.8</td>
<td>73</td>
<td>65</td>
<td>142</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>YOLOv8s [20]</td>
<td>64.9</td>
<td>71.9</td>
<td>11.12</td>
<td>28.4</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>The improved YOLOv8s</td>
<td>70.6</td>
<td>74</td>
<td>7.77</td>
<td>25.3</td>
<td>98</td>
</tr>
<tr>
<td>The Forest Smoke Dataset</td>
<td>Ours</td>
<td>87.1</td>
<td>90.1</td>
<td>7.79</td>
<td>25.6</td>
<td>100</td>
</tr>
</tbody>
</table>

4. Discussion

In this study, a self-constructed forest smoke dataset was used as training and test data. However, an actual forest fire scenario has extremely complex features, such as trees, terrain, and other disturbing factors, which makes it more difficult to achieve significant results. In addition, the detection sensitivity of the algorithm will be different for large forest fires and small forest fires. Future studies should continue to expand the dataset to further improve the performance of the model. On the other hand, the proportion of positive and negative samples in the dataset will also affect the detection results. If the proportion of positive samples is high, the model may focus extensively on the positive samples, resulting in poor performance when faced with new negative samples. A balanced ratio of positive and negative samples helps the model learn the object and background more comprehensively, improving model robustness and generalization performance. In real scenarios, the ratio of positive and negative samples must be adjusted according to the specific scenario and the distribution of the dataset.

In the future, we need to further optimize the forest smoke detection algorithm to better adapt to the practical application scenarios of edge computing platforms. Firstly, we will focus on model compression algorithms for neural network models on edge devices. This includes techniques to reduce the memory and computational complexity of the model, such as model pruning [34], knowledge distillation [35], quantization [36], and low-rank decomposition [37]. In addition, the great success of deep learning heavily relies on increasingly large training datasets. Dataset compression can be used to construct a minimal subset of training data from the entire training dataset without significantly affecting the performance of the model [38]. Finally, we will explore automated neural network
architecture design for neural network architecture search. This approach allows for the adaptive generation of optimal network structures based on specific scenario requirements, thus improving the applicability and efficiency of the algorithms. These research directions aim to reduce computational cost and memory requirements and improve the efficiency and utility of object detection. By implementing efficient smoke detection algorithms on edge devices, we can respond to fire threats in a more timely manner and minimize losses.

Compared to the original YOLOv8 model, the proposed improved YOLOv8 model shows significant improvement in terms of accuracy, mAP50 number of parameters, and computational complexity. However, these improvements come at the cost of longer inference times. Experimental results show that the introduction of the FPN structure increases the inference time and reduces the recognition speed of the model. To solve this problem, the focal modulation structure was introduced to replace the original SPPF structure. The performance of the network is improved by learning coarse-grained spatial information and fine-grained feature information. The experimental results show that the introduction of the focal modulation structure can improve the detection speed of the model. In addition, the experimental results in Table 3 show that the combination of attention mechanisms can effectively improve the detection performance. However, it also has the disadvantage of increasing the computational and parameter complexity as well as the inference time, which makes it more demanding for real-time detection tasks. In comparison, the SimAM presented in this paper does not introduce additional parameters and improves the detection accuracy of the model, which is superior to other attention mechanisms.

5. Conclusions

Forest fires are natural disasters with extensive destructive power and rapid spread, posing a great threat to human life, property, and the ecological environment. Therefore, early detection and rapid response to forest fires are crucial.

In this paper, we presented an improved version of the original YOLOv8 model. Firstly, a weighted path aggregation network for multiscale forest smoke detection was fused in the neck to balance the feature information of different scales and optimize the number of parameters and computational load of the model to a great extent. Next, the simple and parametric-free attention mechanism (SimAM) was introduced to address the problem that images in forest smoke datasets may contain complex background and environmental clutter and to improve the perceptual ability and adaptivity of the model. Finally, focal modulation was introduced to increase the focus on hard-to-detect smoke while improving the model’s runtime speed. Compared to the original YOLOv8s, the mean average precision of forest smoke detection improved by 3%, and the number of parameters and computational complexity of the model were reduced by 30.07% and 10.49%, respectively. The proposed detection model significantly outperforms other existing object detection networks in the self-built forest smoke detection dataset and also has advantages in lightweight models. The improved model can be applied to edge devices, such as mobile devices and UAVs, to realize real-time monitoring and early warning and improve the response speed and accuracy of fire events, which is of great significance for early detection and response to forest fires.

Author Contributions: X.G.: conceptualization, methodology, software, writing—original draft preparation. T.H.: funding acquisition, writing—review and editing. Y.C.: writing—review and editing. All authors have read and agreed to the published version of the manuscript.


Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The dataset and code cannot be shared due to specific reasons.
Acknowledgments: We sincerely appreciate Nanjing Enbo Technology Company Ltd. (Nanjing, China) for providing the forest fire smoke dataset for this research. We also greatly appreciate the “Northern Forest Fire Management Key Laboratory” of the State Forestry and Grassland Bureau and the “National Innovation Alliance of Int. J. Wildland Fire Prevention and Control Technology”, China, for supporting this research. We are especially grateful for anonymous constructive comments.

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

References

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.