forests

MDPI

Article

Omni-Dimensional Dynamic Convolution Meets Bottleneck Transformer: A Novel Improved High Accuracy Forest Fire Smoke Detection Model

Jingjing Qian [1], Ji Lin [1], Di Bai [2,*], Renjie Xu [3] and Haifeng Lin [1,*]

[1] The College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; qianjingjing@njfu.edu.cn (J.Q.)
[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 (D.B.)

Abstract: The frequent occurrence of forest fires in recent years has not only seriously damaged the forests' ecological environments but also threatened the safety of public life and property. Smoke, as the main manifestation of the flame before it is produced, has the advantage of a wide diffusion range that is not easily obscured. Therefore, timely detection of forest fire smoke with better real-time detection for early warnings of forest fires wins valuable time for timely firefighting and also has great significance and applications for the development of forest fire detection systems. However, existing forest fire smoke detection methods still have problems, such as low detection accuracy, slow detection speed, and difficulty detecting smoke from small targets. In order to solve the aforementioned problems and further achieve higher accuracy in detection, this paper proposes an improved, new, high-accuracy forest fire detection model, the OBDS. Firstly, to address the problem of insufficient extraction of effective features of forest fire smoke in complex forest environments, this paper introduces the SimAM attention mechanism, which makes the model pay more attention to the feature information of forest fire smoke and suppresses the interference of non-targeted background information. Moreover, this paper introduces Omni-Dimensional Dynamic Convolution instead of static convolution and adaptively and dynamically adjusts the weights of the convolution kernel, which enables the network to better extract the key features of forest fire smoke of different shapes and sizes. In addition, to address the problem that traditional convolutional neural networks are not capable of capturing global forest fire smoke feature information, this paper introduces the Bottleneck Transformer Net (BoTNet) to fully extract global feature information and local feature information of forest fire smoke images while improving the accuracy of small target forest fire target detection of smoke, effectively reducing the model's computation, and improving the detection speed of model forest fire smoke. Finally, this paper introduces the decoupling head to further improve the detection accuracy of forest fire smoke and speed up the convergence of the model. Our experimental results show that the model OBDS for forest fire smoke detection proposed in this paper is significantly better than the mainstream model, with a computational complexity of 21.5 GFLOPs (giga floating-point operations per second), an improvement of 4.31% compared with the YOLOv5 (YOLO, you only look once) model mAP@0.5, reaching 92.10%, and an FPS (frames per second) of 54, which is conducive to the realization of early warning of forest fires.

Keywords: forest fire smoke detection; YOLOv5; ODConv; BoTNet; SimAM

Citation: 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. https://doi.org/10.3390/f14040838

Academic Editor: Víctor Resco de Dios

Received: 30 March 2023
Revised: 13 April 2023
Accepted: 16 April 2023
Published: 19 April 2023

1. Introduction

The global warming trend has become increasingly pronounced in recent years, and the resulting climate drought and El Niño phenomenon have led to frequent forest fires around the world, killing at least 1000 people worldwide each year. In 2018, severe forest fires in California, USA, killed 66 people and left more than 700 people missing. In 2020,

forest fires in Australia claimed the lives of more than 400 people [1]. Forest fires are a serious threat to human lives and property; they even have a hugely destructive effect on the natural ecosystem of forests and are unpredictable and difficult to rescue [2,3]. Therefore, the prevention of forest fires has always been a top priority in the construction of public systems in various countries. When forest fires occur, smoke often appears earlier than open fires, and the pre-smoke characteristics are more obvious [4–6]. If wildfire smoke can be detected in a timely and accurate manner, it is not only of great significance for forest fire early warning and fighting but can also minimize the loss of life and property.

Manual cruising and observation towers are the oldest and most common means of forest fire monitoring. Manual cruising is susceptible to the influence of climate, communication, traffic, and other factors, such as a large workload that leads to inefficiency, situations when real-time monitoring cannot be achieved, etc. Observation towers [7] have a limited surveillance range, certain dead spot surveillance, and high maintenance costs at a later stage. Satellite monitoring [8] of forest fires still plays an important role, although its monitoring range is wide, the number of satellites is small, the spatial resolution of satellite data is not high, it is not easy to receive information under varying influences of weather, cloud cover, and orbital cycles, and lastly, satellite monitoring may be unable to achieve real-time forest fire monitoring. Aerial forest fire monitoring [9] is also a more effective means of forest fire monitoring, which mainly relies on aircraft or drones for cruising. However, the forest monitoring area is very vast, and its operation costs are very expensive. Traditional early smoke detection of forest fires is mostly achieved by smoke and temperature-sensitive or composite smoke and temperature-sensitive sensors [10,11], which mainly detect smoke particles and rapidly rising ambient temperature to achieve detection. The alarm is triggered only when the smoke concentration or ambient temperature reaches a certain level. However, in real forest environments, smoke spreads quickly, and hardware sensors are characterized by spatial and temporal limitations as well as post-maintenance difficulties. Therefore, it is clear that sensors cannot meet the need for real-time monitoring and early detection and prevention of forest fires in a large and complex environment such as a forest.

As computer technology became more sophisticated, traditional methods of fire smoke detection based on manual feature extraction started to emerge. This method mostly relies on monitoring installed at lookouts to obtain forest fire videos or images, and then features are manually extracted and identifiers are designed. Hidenori et al. [12] applied texture features of smoke to train a Support Vector Machine (SVM) for forest fire smoke recognition. The accuracy of feature extraction and the base number of training samples of this method determines the recognition effect of SVM, but this method stores a large amount of data, and the computational speed is slow. Zhao et al. [13] combined spatio-temporal features and dynamic textures to achieve recognition of smoke through videos. However, smoke is diffuse, so its recorded spatio-temporal energy features can overlap, leading to information redundancy, and additionally, this method cannot completely exclude the interference of foggy weather. Fileonenko et al. [14] performed smoke recognition based on the color and appearance features of the smoke regions of surveillance videos. Taking advantage of the fixed camera, smoke regions were extracted by calculating the roughness of the pixel points at the edges of the smoke and identified using background subtraction, but this method is too sensitive to noise and makes it difficult to achieve accurate and fast smoke detection. Tao et al. [15] performed automatic smoke detection based on a Hidden Markov Model (HMM). In order to perform multi-feature fusion with the dynamic features of the smoke regions, the video color changes of continuous frame sequences were chunked, and each block of a continuous region was classified using Markov models, but this approach still suffers from a complex environment. Traditional video- or image-based methods for forest fire smoke detection have achieved some success but still have some problems. The feature extraction of this method needs to rely on professional knowledge for feature selection, and there is a possibility that the feature design is unreasonable; moreover, it is easily influenced by factors such as lighting and background, leading to poor detection and recognition.

In addition, the robustness of this method is usually poor in the face of the complex and variable forest environment.

After the rapid development of deep learning technologies, many academics have applied them to forest fire smoke detection scenarios. Forest fire smoke detection methods based on deep learning can automatically detect and learn features with more abstract algorithms, learning those features faster with more accurate detection and greater robustness for complex natural forest environments. Zhang et al. [16] expanded the dataset by synthesizing forest fire smoke, and Faster R-CNN was used for detection, which eliminated the complex process of traditional manual feature extraction but increased the data processing cost. Qiang et al. [17] proposed a dual-stream combined forest fire smoke detection algorithm that suppresses background interference by a Tensor Robust Principal Component Analysis (TRPCA) motion detection algorithm and extracts temporal and spatial features of smoke using VGG16 (VGG, visual geometric group network) and a Bidirectional Long Short-Term Memory (BLSTM) network, achieving an accuracy of 90.6%. In regards to the foreground and background factors, long sequences still do not transmit feature information about the starting point of the sequence well. Filonenko et al. [18] used several typical convolutional classification networks, such as AlexNet [19], VGG-16, VGG-19 [20], Resnet [21], and Xception to classify smoke images using Yuan's [22] library of four smoke images for training and analyzing. When comparing the effectiveness of these model networks in recognizing smoke on these data, Xception obtained the best detection. Pan et al. [23] used ShuffleNet, Weakly Supervised Fine Segmentation (WSFS), and Faster R-CNN to predict the presence of fire smoke, but fire smoke was too complex and required extremely high hardware equipment due to excessive training memory usage of the model. Li et al. [24] proposed an Adaptive Depthwise Convolution (ADC) module to adaptively adjust the weights of the convolutional layers to obtain the features of forest fire smoke and obtained an accuracy of 87.26% with an FPS of 43.

In summary, existing forest fire smoke detection algorithms based on deep learning still suffer from low detection accuracy, difficulty in accurately detecting early small target smoke, detection speeds that do not meet daily needs, and high false detection and leakage rates in complex forest environments. Therefore, we propose a high-accuracy forest fire smoke detection model, the OBDS, that is based on YOLOv5, and a series of optimization improvements to address the aforementioned problems. First, to address the problem of sparse forest fire smoke data samples and a single seasonal background, this paper collects publicly available forest fire smoke images on the Internet by crawlers and processes them using an improved data enhancement method with a seasonal style transformation using CycleGAN. Finally, it builds a set containing 25,420 forest fire smoke images and 5000 cloud and fog forest images without smoke for the forest fire smoke dataset. Second, to address the problem of insufficient forest fire smoke feature extraction, this paper introduces an Omni-Dimensional Dynamic Convolution into the backbone, which adaptively decides the weight of each convolution kernel according to the input, overcoming the limitation of static convolution. Additionally, the complementary attention of the convolution kernel is learned in four dimensions, which significantly enhances the capture of contextual smoke information and makes the model better acquire the feature information of forest fire smoke. In addition, this paper introduces the Bottleneck Transformer Net (BoTNet), which combines Convolutional Neural Networks (CNNs) with transformers, enabling the model to more fully extract the global feature information of forest fire smoke images while reducing the computational effort of the model and improving the detection accuracy of small target forest fires. The SimAM attention mechanism is inserted into the neck to enhance the extraction capability of network smoke features without increasing the network parameters while improving the detection accuracy of small target smoke. Finally, we replace the original detection head of YOLOv5 with the Decoupled Head to solve the problem that the original detection head of YOLOv5 has a challenge involving a classification and regression clash, which not only speeds up the convergence of the network but also effectively improves the detection accuracy of the forest fire smoke model.

2. Forest Fire Smoke Dataset

Due to the complexity and variability of the forest environment and the difficulty of sample acquisition, there is no standard dataset for forest smoke identification. Therefore, in this paper, we use a crawler tool to acquire a large number of videos and images of forest fire smoke throughout the web. Considering the small number of forest fire smoke images in a winter context, we have stylistically transformed the dataset by CycleGAN and enriched the dataset in this paper with improved data enhancement methods. The original smoke dataset in this article consists of 15,210 images, including 12,710 forest fire smoke images and 2500 smoke-free cloud and fog forest images. After offline seasonal style transformation using CycleGAN, a total of 25,420 forest fire smoke images were obtained, as well as 5000 smoke-free cloud and fog forest images. The training set, validation set, and test set are divided according to the ratio of 8:1:1. Figure 1 shows some images of the dataset.

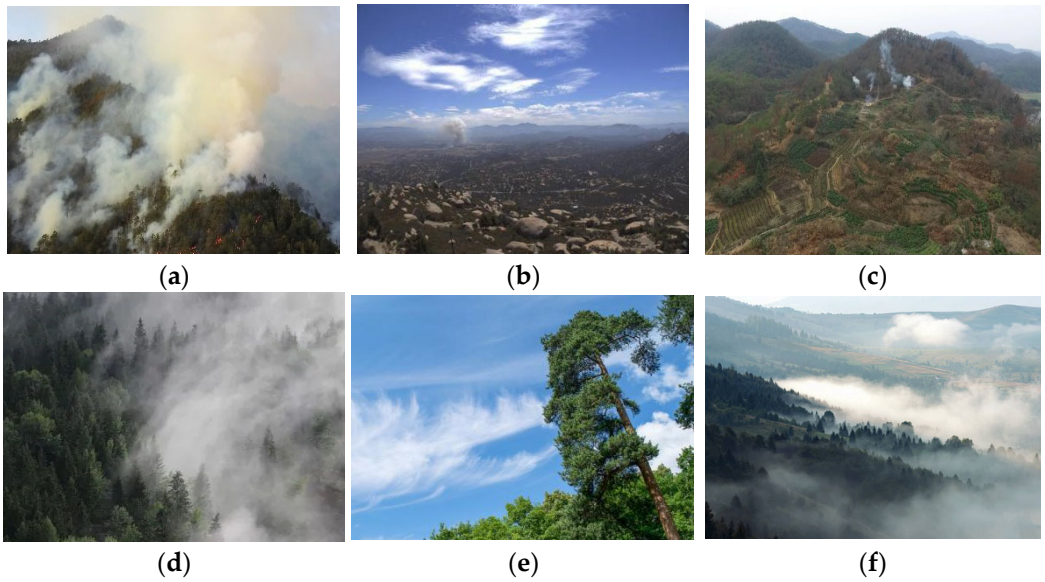


Figure 1. Images of some of the datasets. (**a**–**c**) are images of forest fire smoke; (**d**–**f**) are images of forests without smoke but containing haze, cloud, and fog disturbances.

2.1. Dataset Expansion

In target detection tasks, a large amount of data is very important to improve the generalization ability and robustness of a model. CycleGAN [25] can be used for data augmentation by generating more training data by converting images from one domain to images in another domain, thus improving the performance of the model. In order to expand the forest fire smoke dataset and improve the accuracy of smoke recognition, this paper uses CycleGAN to generate new forest fire smoke images, which expands the smoke samples, and uses an improved mosaic data enhancement method to further improve the generalization ability of the recognition network.

In practical applications, the morphology and color of forest fire smoke may change with the season and weather. If the model is only trained on data from one season, it may not be well adapted to other seasons. The number of publicly available forest fire smoke data is small, and there are fewer forest fire smoke images in winter, which is also the season of frequent forest fires. Therefore, this paper uses CycleGAN to migrate the forest fire smoke images in a seasonal style to enrich the number of forest fire smoke images in the winter so that the model can learn more scenes and changes.

CycleGAN is essentially two GANs that are mirror images of each other, forming a cyclic network, and its network structure is shown in Figure 2.

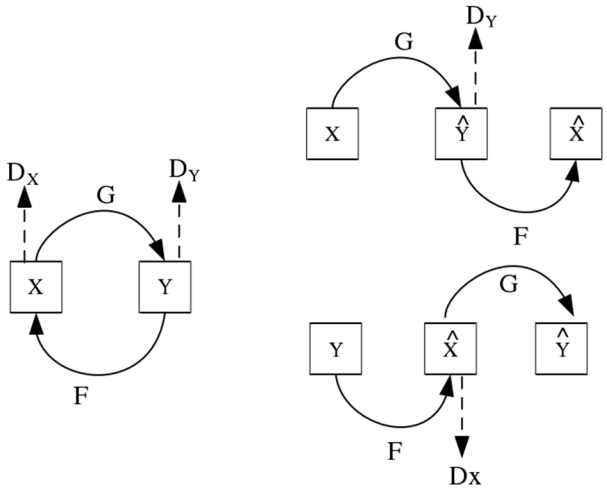


Figure 2. Network structure of CycleGAN.

Where X denotes the input real sample, Y denotes the generated adversarial sample, G and F are the generative networks, and D_X, D_Y is the adversarial network. The X domain can map the image in the X domain to the Y domain by mapping G, which we denote as G(x), and the corresponding discriminator of the generator G is denoted as such. Based on the assumed generators and discriminators, a GAN loss expression, as in (1), can be constructed.

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim P_{data}(y)}[logD_Y(y)] + E_{x \sim P_{data}(x)}[logD_Y(G(x))] \quad (1)$$

This is similar to the traditional GAN generator and discriminator, but the generator G will map all X to the same one in the Y domain. CycleGAN proposes the idea of pairwise mapping in order to solve this problem, i.e., a mapping F is used to reverse the elements in the Y domain to the X. We denote these elements as F(y) and the discriminator corresponding to the generator F as D_X, by defining the loop consistency loss to evaluate the difference between F(y) and the real data x. The requirement is satisfied $F(G(x)) \approx x$ as a way to avoid the problem of mapping to the same picture. Similarly, the two mappings (generators) F and G must also be satisfied $G(F(x)) \approx y$ when migrating images from the Y domain to the X domain.

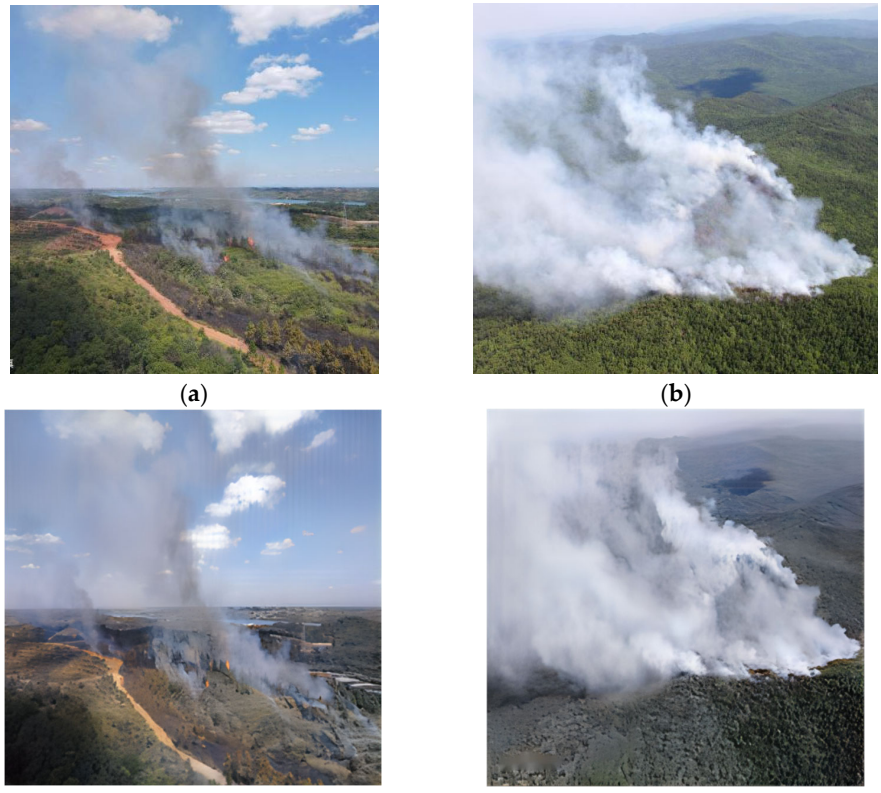
As shown in Figure 3, by using CycleGAN, we can convert the background of forest fire smoke images from summer to winter, which can solve the problem of fewer smoke images in the winter to some extent and expand the dataset so that the model can better learn the background and target features of smoke in various seasons, thus improving the detection of smoke.

2.2. Improved Mosaic Data Enhancement

Mosaic data enhancement is a very important type of data enhancement in YOLOv5. This method loads one image, then additionally selects three images arbitrarily from the dataset, first by random cropping, then stitching them clockwise on one image, and finally scaling them to the set input size and passing them into the model as new samples. This enriches the background of the target to be detected and increases the number of small and medium-sized targets per batch per training, achieving a balance between targets of different scales [26].

Because of the limited number of publicly available images of forest fire smoke, data augmentation of the available data is required to improve the generalization capability of the model. Moreover, because the original mosaic data augmentation method is randomly cropped, there will be a high probability of cropping the forest fire smoke target so that the sample input to the model is only the background; in addition, the original images themselves are inconsistent in scale, which will make the stitched images have more black

and white borders, which will lead to the model training a large amount of useless feature information and affect the speed of model convergence, as shown in Figure 4a. Therefore, this paper improves the mosaic data enhancement method. When loading images, eight additional images are selected for stitching, cropped to the smallest rectangular area that encloses the images, and then random rotation, translation, scaling, and perspective transformation operations are added to the images. As shown in Figure 4b, the improved mosaic data enhancement method reduces more blank boundaries than before the improvement, which reduces useless information, accelerates model convergence, and improves training efficiency; moreover, a large number of different small targets will be formed, which greatly enriches the forest fire smoke data set and significantly improves the performance of the model in detecting small-sized forest fire smoke targets at long distances.



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

(**c**) (**d**)

Figure 3. (**a**,**b**) shows the forest fire smoke images before the seasonal change; (**c**,**d**) shows the forest fire smoke images after the seasonal change.

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

Figure 4. (**a**) shows the mosaic enhancement of the data before the improvement; (**b**) shows the mosaic enhancement of the data after the improvement.

3. The OBDS: An Improved Forest Fire Smoke Detection Model

3.1. The Overall Framework of OBDS

The structure of the forest fire smoke detection model we proposed is shown in Figure 5, and we have made a series of improvements based on the YOLOv5 algorithm and labeled them as "our" in the structure diagram.

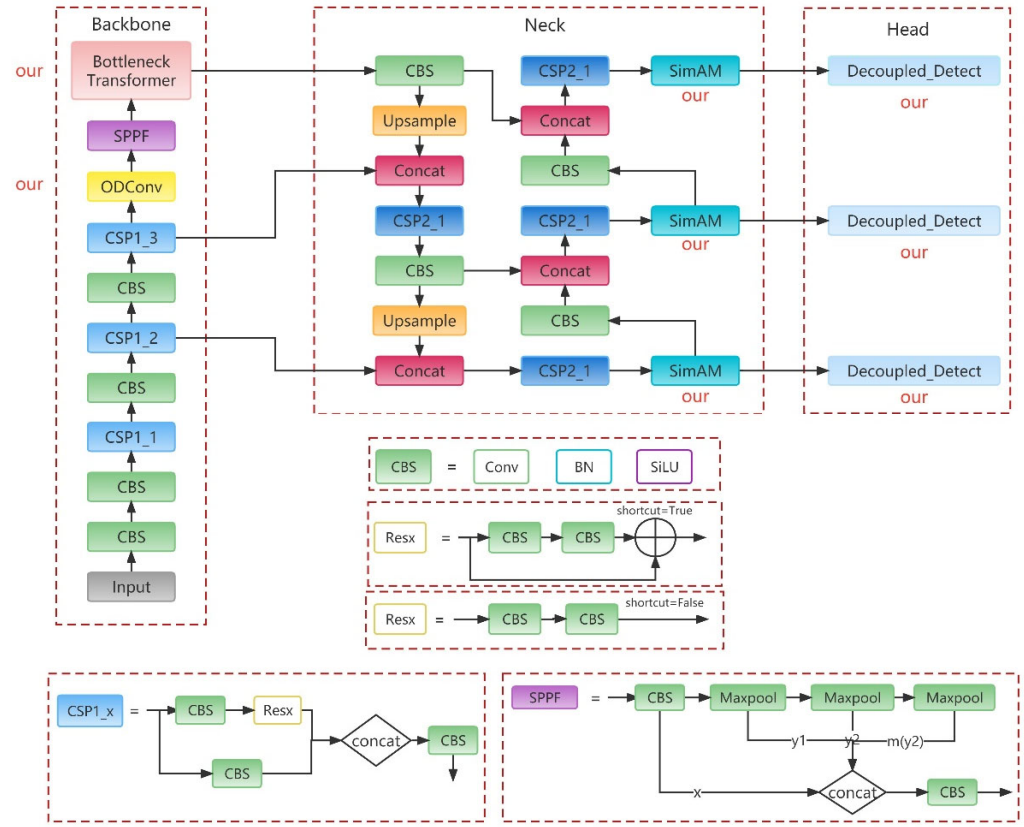


Figure 5. Structure of the forest fire smoke detection model, the OBDS.

First, we insert the Omni-Dimensional Dynamic Convolution (ODConv) into the backbone module of YOLOv5. The ODConv utilizes a novel multidimensional attention mechanism and parallel strategy to learn the complementary attention of the convolution kernel in four dimensions and then adaptively decides the weight of each convolution kernel according to the input, overcoming the limitations of static convolution. The introduction of Omni-Dimensional Dynamic Convolution significantly enhances the ability to capture contextual smoke information, allowing the model to better capture the characteristic information of forest fire smoke.

Second, we insert the Bottleneck Transformer module into the backbone module of YOLOv5. Combining CNNs and transformers overcomes the problem of weak global information extraction in CNNs and solves the problem of a large computation caused by the introduction of self-attention in transformers. It enables the model to fully extract the global feature information of the forest fire smoke images, reduces the computational effort of the model, and improves the detection accuracy of small target forest fire smoke.

In addition, we integrate the SimAM attention mechanism in the neck, which considers both spatial and channel dimensions and can directly derive three weights for the feature map without adding additional parameters, making the network model focus more on the forest fire smoke and suppressing the interference of complex forest background irrelevant information while solving the problem of missed detection of small target forest fire smoke.

Finally, we replace the detection head of YOLOv5 with the Decoupled Head to separate the classification and localization processes, which solves the problem when the

classification and regression tasks conflict in the prediction process due to different concerns, accelerates the convergence speed of the network model, and improves the model detection accuracy at the same time.

3.2. The Basic Framework of the Smoke Detection OBDS Model

Redmon et al. [27–29] proposed a target detection algorithm based on regression ideas, called YOLO. Compared to the R-CNN [30] series, YOLO significantly accelerates the model's execution while essentially maintaining the detection accuracy and satisfying the demand for real-time detection. YOLOv5 is generally faster than other algorithms and has a smaller model weight file, which is especially suitable for small target detection tasks. Therefore, we use YOLOv5 as the detection model. Figure 6 shows the overall structure of it.

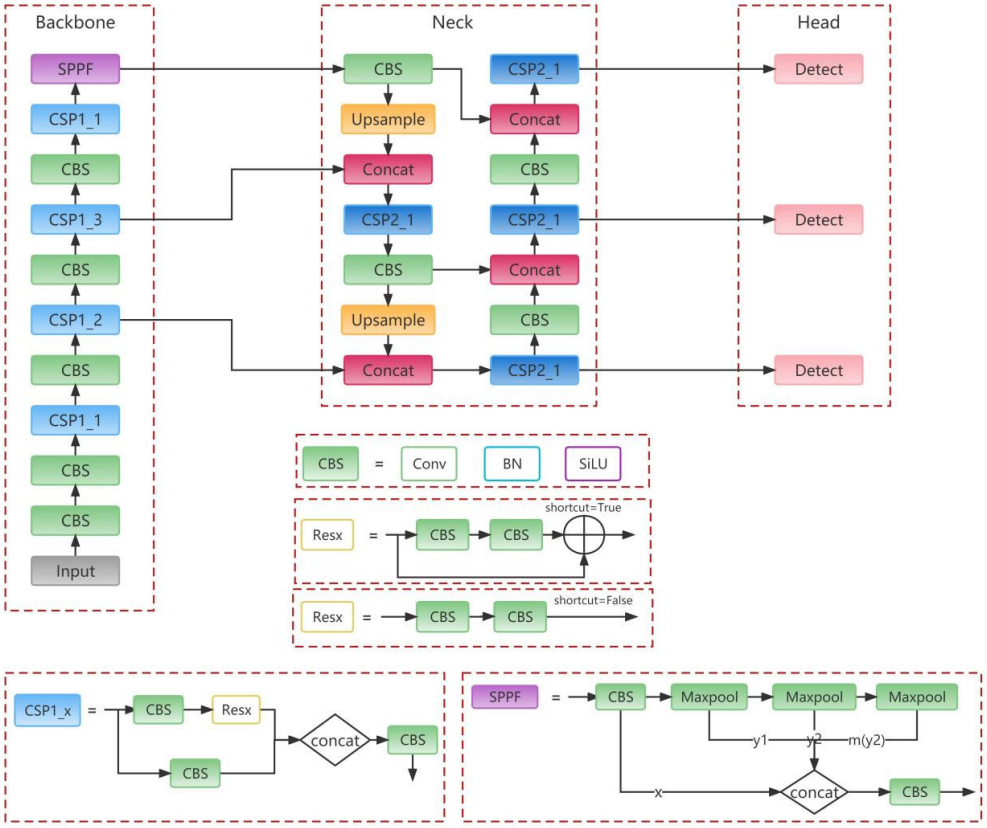


Figure 6. YOLOv5 network structure.

The core components of the backbone network are the convolutional unit CBS (Convolutional Layer + Batch Normalization + SiLU), the Spatial Pyramid Pooling Fast (SPPF) module, and the Bottleneck Cross Stage Partial (CSP) module. The SPPF module performs multiple pooling operations on the feature maps and then splices four different sizes of feature maps. Compared with the spatial pyramid structure of SPP [31], it can fuse high-level semantic features more effectively. Bottleneck CSP borrows the idea of CSP [32] for extracting depth features from images. The detection layer consists of two parts: the Feature Pyramid Network (FPN) and the Pixel Aggregation Network (PAN). The FPN transmits high-dimensional feature information by top-down upsampling, and the PAN is a bottom-up pyramid structure. They are used in combination to improve the feature fusion effect. In the prediction layer, YOLOv5 uses GloU_ Loss as the loss function of the bounding box, which effectively solves the problem that the prediction box does not overlap with the real box and accelerates the prediction box regression. Finally, a weighted non-maximum suppression is used to enhance the recognition of multiple targets and

obtain the best prediction results. Therefore, YOLOv5 is selected as the main framework of the forest smoke fire detection model in this paper, and a series of effective improvements are made based on it.

3.3. Omni-Dimensional Dynamic Convolution

Convolution is one of the basic building blocks of CNNs. In traditional static convolution, the convolution kernel parameters are determined by training, and the same convolution kernel is used to do the same operation on all input images. In other words, static convolution uses the same network structure and parameters for all input data. In addition, previous static networks often enhance the performance of network models by widening the width, depth, and resolution; in previous attention models, attention is mainly applied to the feature maps, such as weighting on different channels of the feature maps (SE Net) and weighting at different spatial locations of the feature maps (spatial attention). Therefore, the static convolutional feature learning capability is inevitably limited when faced with diverse inputs. Therefore, dynamic convolution [33,34] was developed to solve this problem. Dynamic convolution determines the weight of each convolution kernel based on the input of the convolution layer, and an attention mechanism is used to weigh and sum these convolution kernels to obtain a dynamic convolution kernel suitable for that input. The output of the dynamic convolution can be expressed as Formula (2).

$$\mathrm{Output(y)} = \mathrm{x} * (\alpha_1 \mathrm{W}_1 + \alpha_2 \mathrm{W}_2 + \ldots\ldots + \alpha_\mathrm{n} \mathrm{W}_\mathrm{n}), \quad (2)$$

where y and x denote output features and input features, respectively. $\alpha_i (i = 1, 2, \ldots, \mathrm{n})$ is the attention scalar; n is the number of convolutional kernels; and each convolutional kernel $W_i (i = 1, 2, \ldots, \mathrm{n})$ has the same size as the standard convolutional kernel.

The Omni-Dimensional Dynamic Convolution (ODConv) [35] used in this paper utilizes a multidimensional attention mechanism and a parallel strategy to learn the complementary attention of the convolution kernel at any convolution layer along all four dimensions of the kernel space. The four different attentions are the number of input channels of the convolutional kernel, the perceptual field of the convolutional kernel itself, the number of output channels of the convolutional kernel, and the number of convolutional kernels. These four attentions complement each other and are multiplied with the convolution kernel in the order of position, channel, filter, and kernel, making the convolution operation different from all spatial positions, all input channels, all filters, and all kernels of input x, significantly enhancing the capture of contextual information. The output of ODConv can be expressed in Formula (3).

$$\mathrm{Output(y)} = \mathrm{x} * (\alpha_{\mathrm{w1}} \odot \alpha_{\mathrm{f1}} \odot \alpha_{\mathrm{c1}} \odot \alpha_{\mathrm{s1}} \odot \mathrm{W}_1 + \ldots\ldots + \alpha_{\mathrm{wn}} \odot \alpha_{\mathrm{fn}} \odot \alpha_{\mathrm{cn}} \odot \alpha_{\mathrm{sn}} \odot \mathrm{W}_\mathrm{n}), \quad (3)$$

where y and x denote output features and input features, respectively. α_{wi}, α_{fi}, α_{ci}, $\alpha_{si} (i = 1, 2, \ldots, \mathrm{n})$ denote four attentions of attention: scalar, space dimension, input channel dimension, and output channel dimension of the convolution kernel $\mathrm{W}_i (i = 1, 2, \ldots, \mathrm{n})$, respectively. The $\odot$ denotes the multiplication operations in the kernel space along different dimensions. Figure 7 shows the computation process of the ODConv.

The proposed forest smoke fire detection model adopts the ODConv, which can replace the conventional static convolution with a "plug-and-play" approach, adaptively adjust the convolution kernel according to the different input forest smoke fire image data, and extract the effective features of forest smoke and fire in a targeted manner. This significantly enhances the feature extraction capability of the model, strengthens the learning capability of the network, and greatly improves the recognition accuracy of the model for forest smoke fire targets.

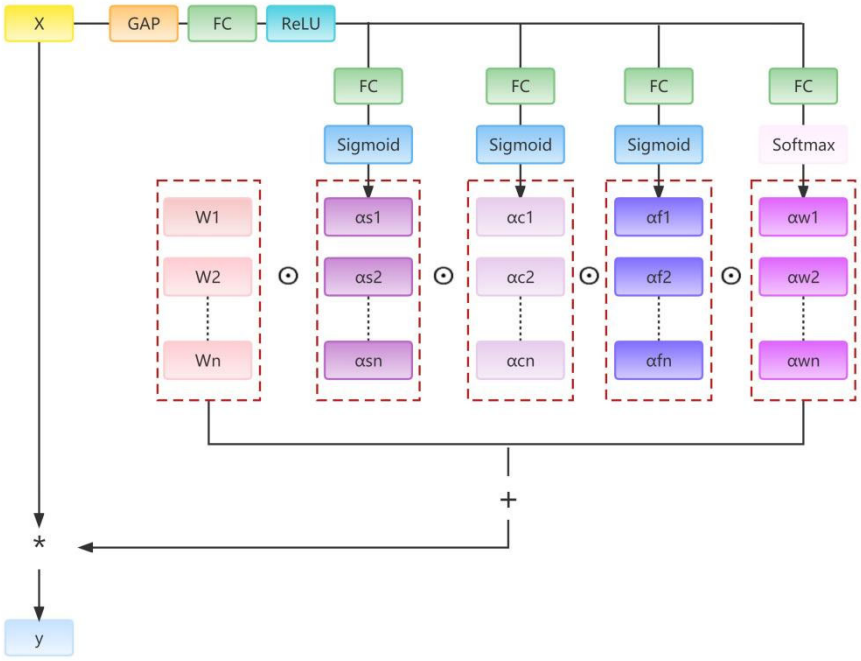


Figure 7. Calculation process of the ODConv.

3.4. Bottleneck Transformer Net (BoTNet)

The advantages of CNNs are parameter sharing and efficient local information aggregation, but for some vision tasks, such as target detection, long-range dependencies are required. In order to integrate global information, CNN-based models need to stack many layers of convolutional layers. As a result, the CNN-based model does not easily obtain global information and is more concerned with the aggregation of local information, so it is easy to lose some features when performing feature extraction, resulting in a decrease in detection accuracy. It is obviously better to use a global approach, such as self-attention, which reduces the number of required network layers and is more powerful and scalable, while the transformer-based [36–39] model has the innate ability to obtain global information and mainly uses the self-attention mechanism to extract intrinsic features. However, the memory and computational effort required by self-attention are quadratic in the spatial dimension, which is particularly problematic when training with high-resolution images.

The BoTNet [40] network is an exploration of the combination of a convolution network and a transformer by researchers from Berkeley and Google. It uses a hybrid approach to simultaneously exploit the feature extraction capability of CNNs and the content and location of self-attention mechanisms in the transformer, achieving a better performance than would be possible with a pure CNN or self-attention mechanism, with an 84.7% accuracy in ImageNet. The combination of a CNN and transformer achieves the effect of complementing each other's strengths and weaknesses. The BoTNet network framework proposes a Bottleneck Transformer to replace the ResNet Bottleneck, i.e., only the global Multi-Head SelfAttention (MHSA) replaces the 3×3 spatial convolution in the last 3 bottleneck blocks of the ResNet framework. Since the introduction of SelfAttention leads to high computation and excessive memory usage, BoTNet adds the self-attention module to the last 3 bottleneck blocks of the ResNet framework. Each bottleneck contains a 3×3 convolution, and MHSA is used to replace this convolution. A 3×3 convolution stride = 2 in the first bottleneck, and the MHSA module does not support stride operation, so BoTNet uses 2×2 mean pooling for downsampling. Figure 8 shows the structure of the BoT (Bottleneck Transformer).

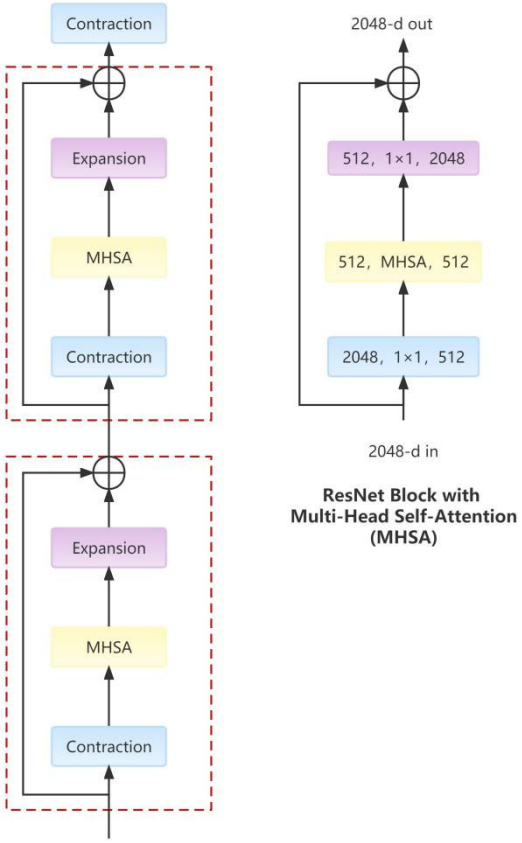


Figure 8. Bottleneck Transformer structure.

At this stage, most deep learning forest fire smoke detection methods are implemented using alternating convolutional layers and pooling layers. Convolutional operations are good at extracting local details, but in high-volume forest fire smoke detection tasks, many convolutional layers often need to be stacked to grasp the global information of forest fire smoke. In contrast, the self-attention in the transformer is good at grasping global information but requires a large amount of data for training. Therefore, this paper proposes to use the BotNet framework, combining a CNN and transformer to complement each other's strengths and weaknesses, to fully extract the global and local feature information of forest fire smoke images while effectively reducing the computational effort of the model and improving the accuracy of small-target forest fire smoke target detection.

3.5. SimAM Attention Mechanism

In order to suppress the non-targeted forest background disturbance information and make the model more focused on the feature information of forest fire smoke, the SimAM attention mechanism is introduced in this paper [41]. Existing attention modules focus on the channel and spatial attention mechanisms, which generate one-dimensional channel or two-dimensional spatial weights from the features and expand the weights of the channel or spatial attention. However, in reality, these two mechanisms should jointly facilitate information selection in the visual process. Therefore, this study introduces the SimAM attention module to assign a unique weight to each neuron. The advantage of the SimAM is that three weights can be directly derived for the feature map without adding additional parameters (i.e., considering spatial and channel dimensions), allowing the network to learn more discriminative neurons so that the network could better extract features.

The SimAM assesses the significance of every neuron based on neuroscientific theories. In neuroscience, information-abundant neurons usually exhibit a different firing pattern than surrounding neurons, and activating neurons usually repress surrounding neurons, i.e., null-field inhibition. In other words, neurons with null domain inhibition

effects should be assigned higher importance. The SimAM attention module evaluates the importance of each neuron based on the idea of null domain inhibition by measuring the linear differentiability between a target neuron and other neurons. Therefore, the following energy function was defined for each neuron, as shown in Formula (4):

$$e_t(w_t, b_t, y, x_i) = \left(y_t - \hat{t}\right)^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x_i})^2, \tag{4}$$

where t and x_i refer to the target neuron and other neurons in a channel of the input feature, respectively. The subscript i represents that the variable being subscripted is the variable in the ith dimension on the spatial dimension, and M represents the total number of neurons on that channel. $\hat{t} = w_t + b_t$ and $\hat{x_i} = w_t x_i + b_t$ are the linear transformed forms of t and x_i, and w_t and b_t are the linear transformed weights and bias terms. All variables in Formula (3) are scalars, and y_t and y_0 are two different values, so Formula (3) reaches a minimum when y_t equals $\hat{t}$ and all $\hat{x_i}$ equals y_0. Minimizing Formula (3) is equivalent to finding linearly branchable branches of the target neuron t and other neurons in the same channel. For simplicity, the SimAM module uses binary labels, i.e., 1 and -1, for y_t and y_0. A regular term is added to Formula (3), and the final energy function is shown in Formula (5).

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_t x_i + b_t))^2 + (1 - (w_t x_i + b_t))^2 + \lambda w_t^2 \tag{5}$$

Formula (5) has analytical solutions on w_t and b_t, which can be solved by Formulas (6) and (7).

$$w_t = -\frac{2(t - \mu_t)}{(t - \mu_t)^2 + 2\sigma_t^2 + 2\lambda}, \tag{6}$$

$$b_t = -\frac{1}{2}(t + \mu_t)w_t\ , \tag{7}$$

where $\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i$ and $\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \mu_t)^2$ are the mean and variance of all neurons in the selected channel except for the target neuron t, respectively. Therefore, the minimum energy can be obtained using Formula (8).

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{8}$$

From the formula of minimum energy, it is clear that the lower the energy of neuron t, the more it is distinguished from the surrounding neurons and the more important it is. Therefore, the importance of a neuron can be calculated by $\frac{1}{e_t^*}$.

Aiming at the characteristics of forest fire smoke in a natural environment, this paper introduces the SimAM attention module to help the model learn the features of forest fires and smoke in images, suppress non-target background interference information, and pay more attention to the feature information of forest fire smoke, thus improving the detection accuracy of forest fire smoke.

3.6. Decoupled Head

There is always a fundamental contradiction in target detection, which is the either/or conflict between classification and regression processing. The Yolo Head classification and regression tasks of the original YOLO family of algorithms are accomplished by a 1×1 convolution. However, classification is more concerned with the texture content of each sample, while regression is more concerned with the edge features of the object's image; therefore, the two have different concerns, which can cause conflicts in the prediction process and lead to model performance degradation. The Decoupled Head is proposed in YOLOX [42], and it appears to solve the aforementioned problems. The Decoupled Head architecture essentially separates the classification and localization processes, and it can

be applied to both one-stage and two-stage target detection algorithms. However, before the advent of YOLOX, the previous YOLO family of algorithms as well as other target detection algorithms had been using a coupling head at the prediction side.

The structure of the Decoupled Head network is shown in Figure 9. For the input feature layer, the Decoupled Head will downscale it using 1×1 convolution and then perform the object classification and target frame coordinate regression tasks using two parallel channels, respectively. To reduce the complexity of the Decoupled Head and increase the model convergence speed, two 3×3 convolutions are used for each channel. The processing results in 3 output values: Cls, Reg, and Obj, where Cls is the kind corresponding to the target frame, Reg is the location information of the target frame, and Obj is whether each feature point contains an object. The 3 output values are fused to obtain the final prediction information.

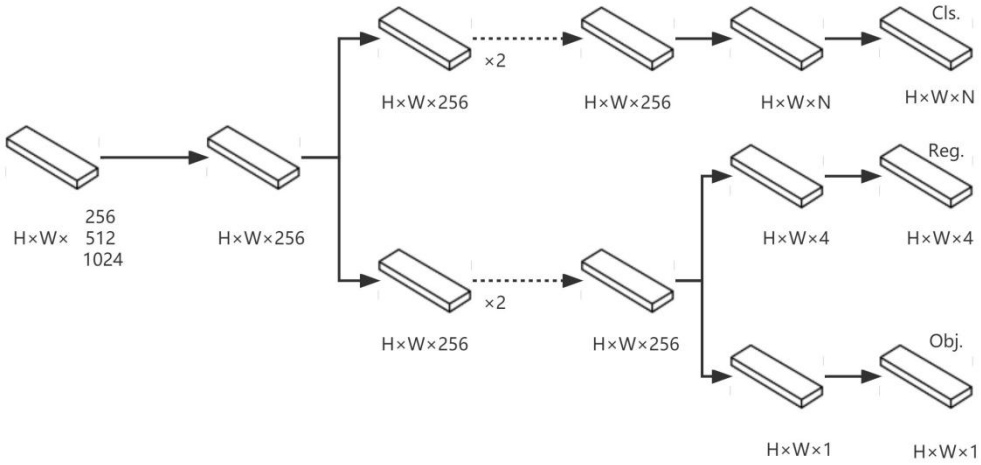


Figure 9. Network structure of the Decoupled Head.

Therefore, this paper uses the Decoupled Head to decouple the classification and regression tasks. The use of the Decoupled Head can not only bring improvements in the accuracy of forest fire smoke detection but also improve the speed of convergence of the forest fire smoke detection model.

3.7. Training

To accomplish forest fire smoke detection, a large amount of labeled data to train deep learning models is required, but in practice, such data is difficult to obtain. Considering the similarity between forest fire smoke and the smoke of other scenes, this paper introduces a migration learning strategy to improve the generalization ability and robustness of the model. USTC-Smoke is a publicly available smoke dataset from the University of Science and Technology of China (http://smoke.ustc.edu.cn/datasets.htm (accessed on 22 September 2022)), which contains smoke from several scenes of smoke. The migration learning process is shown in Figure 10. Firstly, the training is performed based on this dataset, and then the trained model parameter weights are migrated to the new model, which is retrained using the forest fire smoke dataset established in this paper. Table 1 shows the experimental environment of the model OBDS, and Table 2 shows the training parameters. In addition, since forest fire smoke detection is a target detection task that requires processing a large amount of sparse data, an optimizer that is insensitive to sparse gradients is required. Therefore, this paper compares the performance of the model using the Adam and SGD optimization algorithms during the training process.

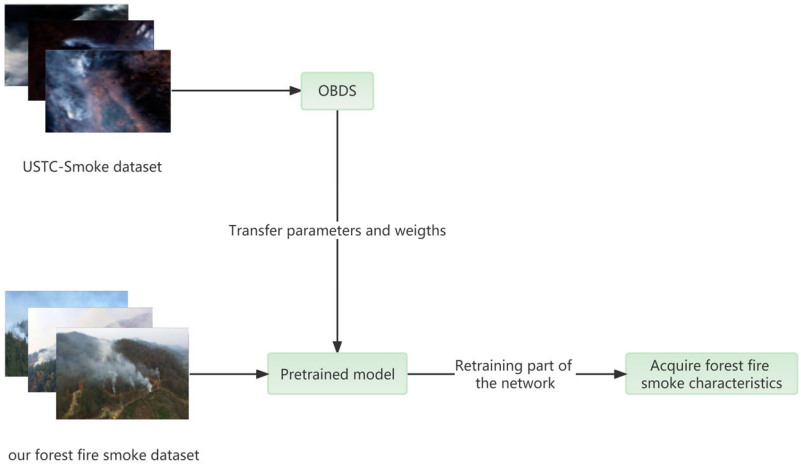


Figure 10. Transfer learning flowchart.

Table 1. Experimental environments.

Experimental Environments	Details
Program Language	Python 3.7
Deep learning framework	Pytorch 1.12.1
GPU Type	NVIDIA GeForce RTX 3060
Operating System	Windows 10

Table 2. Training parameters for our model.

Training Parameters	Details
Epochs	200
Batch-Size	8
Initial Learning Rate	0.01
img-size	640 × 640
Optimization algorithm	SGD and Adam

3.8. Evaluation Metrics

To evaluate the performance of the model for forest fire smoke detection, the evaluation metrics in this paper use Precision, Recall, mAP, speed index FPS, and time. TP refers to the test image and predicted image as both being forest fire smoke images; FP refers to the test as being non-smoke images; FN refers to the test as being smoke images; and prediction is the non-smoke image.

Precision is the ratio of the number of correctly predicted smoke borders to all predicted smoke borders, calculated as follows, and this metric reflects the accuracy of the model for forest fire smoke prediction.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall is the ratio of the number of correctly predicted smoke bounding boxes to the number of all true smoke bounding boxes, and this metric reflects the proportion of smoke samples that are correctly predicted and is calculated as follows:

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

mAP refers to the average accuracy mean, which is the mean value of AP. mAP is the average accuracy rate and is obtained by integrating the Precision-Recall curve with the following formula:

$$AP = \int_0^1 (Precision) d(Recall) \,, \tag{11}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{12}$$

FLOPs (floating point operations) are important metrics for evaluating the computation of a model, which can side-by-side reflect the time spent on model training and testing; they are usually used to measure the time complexity of a model. Although the processing time of the model is affected by hardware conditions, this paper is based on the same hardware device and can reflect the number of operations and computational complexity of the model very objectively. Ignoring a small number of operations in the pooling layer, batch normalization layer, and activation layer, if the bias term is considered, the FLOPs of the convolutional layer can be derived as follows:

$$\text{FLOPs} = \sum_{i=1}^{L_1} \left(C_i \times k_i^2 + 1 \right) \times C_{i+1} \times H_{i+1} \times W_{i+1}, \tag{13}$$

where the number of layers of the convolutional layer is L_1, C_i, and C_{i+1} that denote the number of input and output channels of the i-th convolutional layer, respectively. H_{i+1} and W_{i+1} are the length and width of the output feature map in the i-th convolutional layer; $H_{i+1} \times W_{i+1}$ denotes the size of this output feature map. k_i denotes the size of the convolutional kernel in the i-th layer. The FLOPs operation for the fully connected layer is shown in Formula (14):

$$\text{FLOPs} = \sum_{i=1}^{L_2} (C_i + 1) \times C_{i+1}, \tag{14}$$

where the number of layers of the fully connected layer is L_1 and the weight matrix of the connections is $C_i \times C_{i+1}$. Therefore, the value of the FLOPs after the samples are input into the network is approximately equal to the sum of Formulas (13) and (14).

FPS (frames per second) reflects the number of smoke images that can be detected by the network model per second. f_n indicates the number of detected images, and T represents the detection time used.

$$FPS = \frac{f_n}{T} \tag{15}$$

4. Results

4.1. Training Results

First, we designed two sets of comparison experiments for the choice of optimization algorithm, comparing the mAP_0.5 of our OBDS model when using the Adam and SGD optimization algorithms. The experimental results are shown in Figure 11. It can be seen that the mAP@0.5 of the model is higher when using the Adam optimization algorithm than when using the SGD optimization algorithm, so the former algorithm is used for OBDS in the following experiments. Therefore, the Adam optimization algorithm is used in the following experiments.

Figure 12 shows the convergence process of Precision, Recall, and mAP@0.5 during the training of the smoke detection OBDS model, and the following figures reflect the improvements for faster convergence of the network.

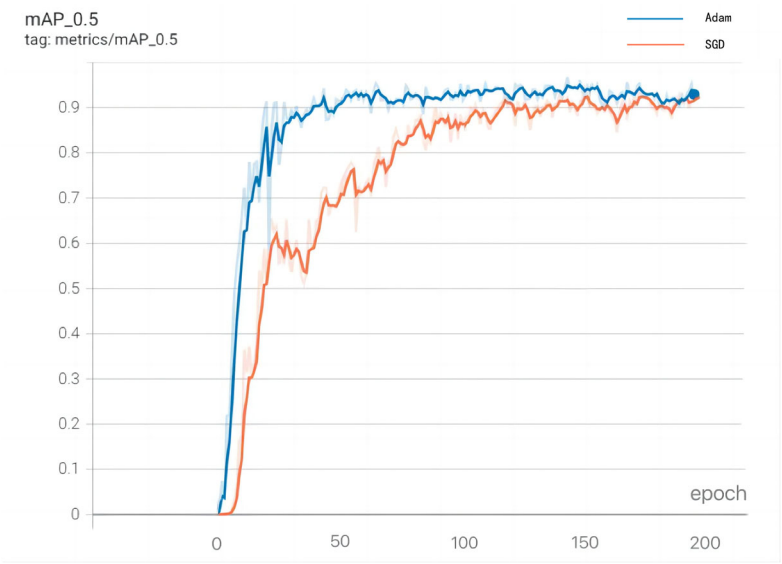


Figure 11. mAP@0.5 of the model when using the Adam and SGD optimization algorithms.

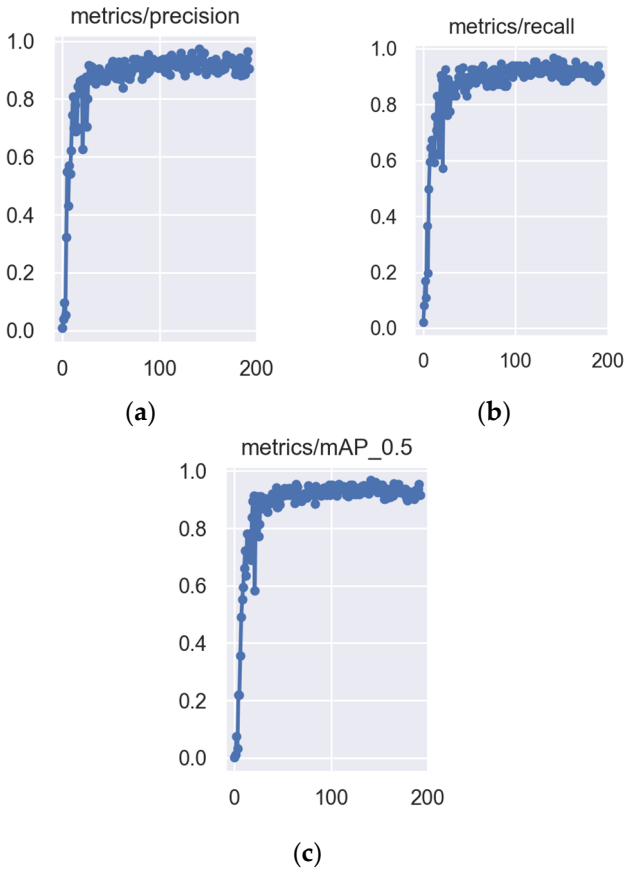


Figure 12. Diagram of the convergence process of evaluation metrics. (**a**) Convergence process diagram for Precision. (**b**) Convergence process diagram for Recall. (**c**) Convergence process diagram for mAP@0.5.

4.2. Ablation Experiments

To verify the effectiveness of each improvement module of the smoke model OBDS, five sets of ablation experiments were designed in this paper, each using the same environment and training parameters. Table 3 shows the experimental results.

Table 3. The results of the ablation experiments.

Model	mAP@0.5 (%)	Precision (%)	Recall (%)	FPS	Time (ms)
YOLOv5	87.79	88.01	87.35	59	16.9
YOLOv5 + ODConv	88.31	87.12	88.12	58	17.2
YOLOv5 + ODConv + BoTNet	90.95	89.56	89.20	56	17.8
YOLOv5 + ODConv + BoTNet + SimAM	91.10	91.02	90.05	55	18.1
YOLOv5 + ODConv + BoTNet + SimAM + Decoupled Head (our model, the OBDS with Adam)	92.10	92.25	91.60	54	18.5

From Table 3, compared with YOLOv5, the "mAP@0.5," Precision, and Recall of the model improved by 3.16, 1.55, and 0.77, respectively, after the introduction of ODConv and BoTNet, indicating that the network's forest fire smoke feature extraction capability was significantly improved. After the introduction of the SimAM attention mechanism and decoupling head, the "mAP@0.5", accuracy, and recall of the YOLOv5 network model are improved by 4.31, 4.24, and 4.25, respectively, compared with those of the YOLOv5 network model, which indicates that the model focuses more on the target area of forest fire smoke and increases the detection accuracy of forest fire smoke. Meanwhile, the FPS of the model reaches 54, i.e., the model can process 54 frames per second, and it takes only 18.5 ms to detect each forest fire smoke image, which can realize the real-time monitoring of forest fire smoke.

In order to further verify the effectiveness of this model OBDS in forest fire smoke detection, the results of comparison experiments between this model OBDS and the mainstream models Fast R-CNN, DETR, Deformable DETR, and Improved YOLO in the field of smoke detection are given in Table 4, and mAP@0.5, FPS, and FLOPs metrics are used to evaluate them. Our model OBDS achieves a computational complexity of 21.5 GFLOPs, a high FPS of 54 frames per second, and the highest mAP@0.5 of 91.2% with a guaranteed lower computational complexity.

Table 4. Results of model comparison experiments.

Model	mAP@0.5 (%)	FPS	GFLOPs
Fast R-CNN [43]	84.21	58	24.5
DETR [44]	87.32	13	26.2
Deformable DETR [45]	89.15	10	39.6
Improved YOLO [46]	91.86	52	20.1
OBDS (our model)	92.10	54	21.5

4.3. Forest Smoke Detection Performance Analysis

Through ablation experiments, we verified that each module of the improved model contributes to the detection of forest fire smoke, and the improved model has a great improvement in the detection effect compared with the original model. OBDS can better extract global and local features of smoke images, effectively suppress the interference of background information in the complex forest environment, and effectively solve the problem of insufficient extraction of forest fire smoke features and low detection accuracy. Meanwhile, the forest fire smoke detection model OBDS can meet the needs of real-time forest fire detection and has good detection of small target forest fire smoke, which can realize early warnings of forest fires.

The detection of forest fire smoke in different backgrounds is shown in Figure 13. As previously shown in Figure 8a, the model OBDS can detect forest fire smoke more accurately. In contrast, the detection results of the unimproved YOLOv5 model are less accurate, and in addition, the original YOLOv5 model misses three smoke targets (Figure 13b).

Furthermore, the detection confidence of OBDS is higher than that of the original YOLOv5 model (Figure 13c,d).

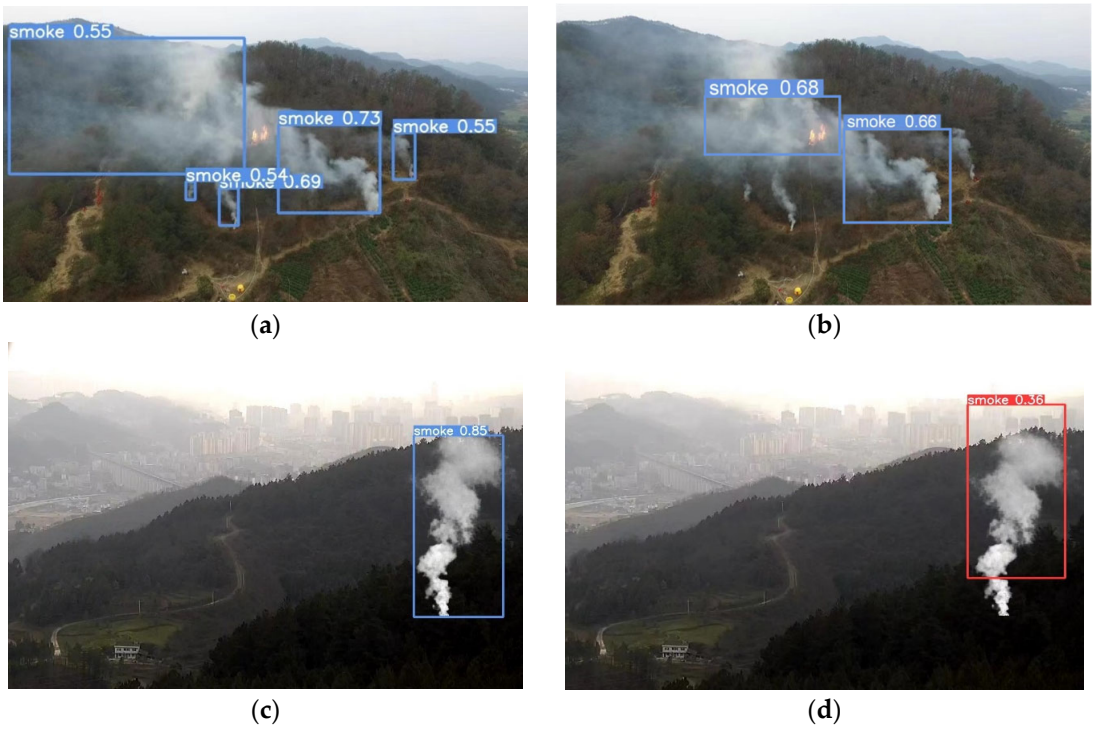


Figure 13. Forest fire smoke detection effect. (**a**,**c**) are the smoke detection results of the present model OBDS; (**b**,**d**) are the smoke detection results of the unimproved YOLOv5.

As shown in Figure 14, the smoke detection confidence of this model is significantly higher than that of the unimproved YOLOv5 model for small target forest fire smoke at a long distance.

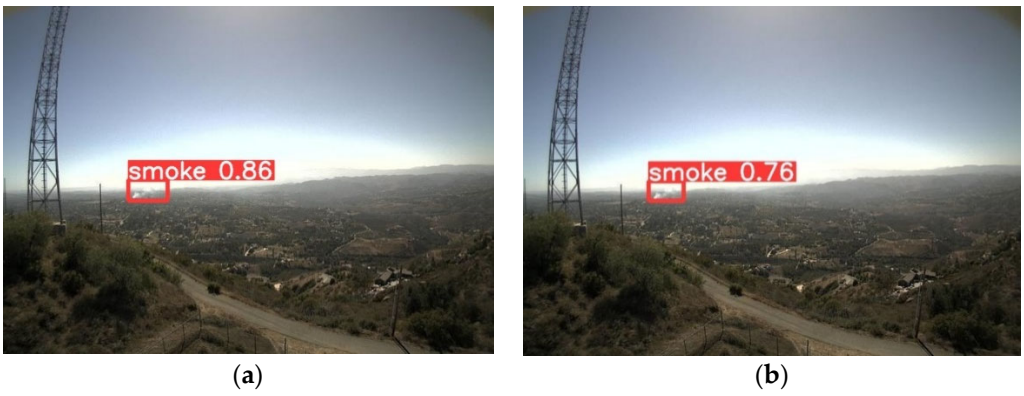
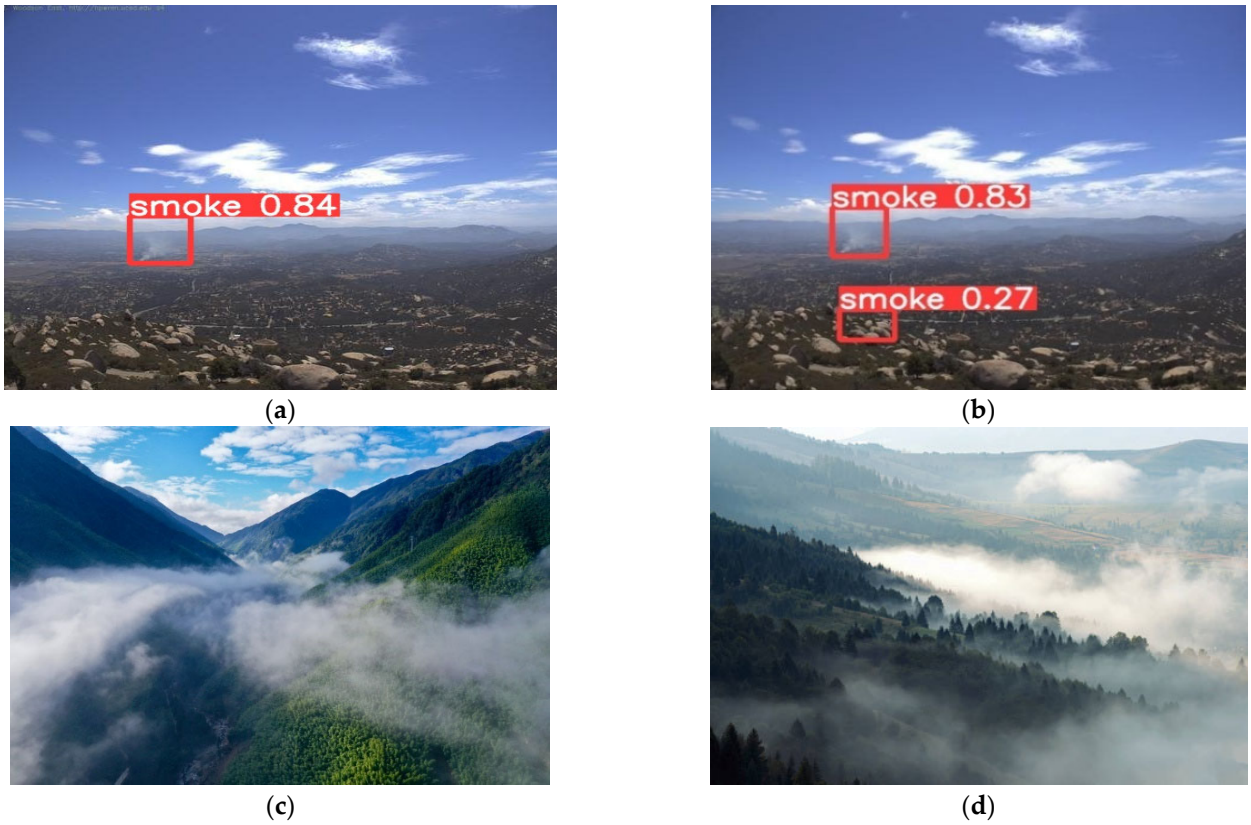


Figure 14. Smoke detection effect of a small forest fire. (**a**) is the smoke detection result of this model of OBDS; (**b**) is the smoke detection result of the unimproved YOLOv5 model.

The experimental results of the complex background interference false detection cases are shown in Figure 15. When the background appears similar to the smoke with clouds, fog, and other interferences, this model does not show false detection cases. Figure 15a,b, are images containing fire smoke, and Figure 15c,d, are images of the forest without fire smoke but with cloud disturbance. Smoke detection of our OBDS model can adapt to various interferences of complex backgrounds without misdetection, whereas the original YOLOv5 model has false detection (Figure 15b).



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

(**c**) (**d**)

Figure 15. Detection effect of the forest fire smoke model against interference ability. (**a**,**c**) are the smoke detection results of the model OBDS; (**b**,**d**) are the smoke detection results of the unimproved YOLOv5 model.

Considering that forest fire smoke monitoring is a task that requires real-time monitoring, the model in this paper detects forest fire smoke videos, and the detection results are shown in Figure 16. The model in this paper can meet the demand for real-time detection, which helps to detect forest fires in time and grasp the smoke's spread.

Figure 16. Forest fire smoke video detection results.

5. Discussion

Forest resources are vital to human beings and, at the same time, play an immense role in maintaining the integrity of the ecological community and overall environment. However, forest fires have occurred repeatedly in recent years, causing great damage to countries around the world and also threatening human lives. If fires can be detected at an early stage of occurrence, they can be extinguished with minimum cost and avoid causing more damage. The forest environment is very complex, especially in densely vegetated forest areas. Forest fires are not easily detected at the early stage of their occurrence, and

smoke is the main feature of the pre-fire period. Using this feature for forest fire smoke detection is crucial for early warning and control of fires.

Traditional forest fire smoke recognition algorithms usually extract specific features from the suspected smoke area and then classify the features for recognition. However, forest fire smoke, as a random turbulent motion phenomenon, is affected by the weather and topography of the forest area, with large differences in appearance and a certain degree of transparency. Additionally, manually selected features often have insufficient discriminatory power and low detection accuracy. In recent years, many scholars have spoken about the application of deep learning techniques to forest fire smoke detection, but due to the limitations of the weak CNN global feature information extraction capability, the existing algorithms still find it difficult to meet the detection needs of high accuracy, low false detection, low leakage, and high real-time performance. In addition, there is no standard dataset for smoke detection because it is difficult to obtain smoke image samples of forest fires. Therefore, this paper uses crawlers to obtain public forest fire smoke images on the web. Considering the problem of a single seasonal background in the dataset, this paper uses CycleGAN to seasonally transform the smoke images, improves the traditional mosaic enhancement algorithm to enrich the dataset, and finally builds a set containing 25,420 forest fire smoke images and 5000 cloud and fog images without smoke. A forest fire smoke dataset containing 25,420 forest fire smoke images and 5000 cloud and fog images without smoke was finally built.

Since traditional static convolution can only operate on a fixed-size convolution kernel, it cannot accommodate input data of different sizes and shapes. In addition, static convolution can only convolve on a single dimension; it cannot capture the dependencies between multiple dimensions. These drawbacks limit the performance and effectiveness of static convolution when dealing with multiple input data. Therefore, this paper introduces Omni-Dimensional Dynamic Convolution to effectively solve the problem of insufficient extraction of static convolutional smoke features by learning the complementary attention of convolutional kernels from four dimensions and adaptively deciding the weight of each convolutional kernel according to the model's input, which significantly enhances the ability to capture contextual smoke information and makes the model better. The model can significantly enhance the capture of contextual smoke information, allowing the model to better capture the feature information of forest fire smoke.

Conventional CNNs can only perform information extraction and feature learning at a local scale and thus perform poorly when processing large and high-resolution images. In addition, traditional CNNs need to reduce the size of the feature map by multiple convolution and pooling operations when detecting targets, which makes it difficult for CNNs to detect small targets. For example, YOLOv5, YOLOv7, and YOLOv8 are all pure convolutional structures. Although YOLOv7 and YOLOv8 are improvements to YOLOv5, YOLOv7, and YOLOv8 are designed for universal detection targets. The OBDS model based on the YOLOv5 framework in this article is designed for the characteristics of forest fire smoke and combines the advantages of transformers for information extraction and feature learning on a global scale. We introduce the Bottleneck Transformer Net (BoTNet), which combines a CNN and transformer to complement each other's strengths and weaknesses so that the model can more fully extract global feature information from forest fire smoke images while reducing the model's computation and improving the detection accuracy of small target forest fires. In addition, this paper incorporates the SimAM attention mechanism to enhance the extraction capability of network smoke features without increasing the network parameters, while improving the detection accuracy of small target smoke. Considering that the original detection head of YOLOv5 detects by dividing the image into a grid, which has a low accuracy for small target localization, in addition to the conflict problem of classification and regression. Therefore, this paper optimizes the detection head by using a decoupling head to separate the classification and regression tasks to predict the location and class of targets more accurately, speed up the convergence of the network, and effectively improve the detection accuracy of the forest fire smoke model.

In the forest fire smoke detection scenario, an optimizer that can converge quickly is needed to optimize the model because the shape and color of smoke may change with time, weather, etc. Meanwhile, since fire smoke detection is a target detection task and needs to deal with a large amount of sparse data, an optimizer that is insensitive to sparse gradients is needed. Based on the above considerations, we compared the performances of the Adam and SGD optimization algorithms, and through the graphs, we found that the mAP0.5 of the model is higher and the model converges faster when using the Adam optimization algorithm. We verified the effectiveness of each optimization module by ablation experiments, and our model OBDS improved by 4.31% compared with the benchmark model mAP0.5, and the FPS reached 54 frames per second, which can meet the real-time detection of forest fire smoke. We compare our OBDS model with the mainstream models (Fast R-CNN, DETR, and Deformable DETR) and the latest models in the field of smoke detection, and the OBDS achieves the highest mAP0.5 of 92.10% with less guaranteed computation (21.5 GFLOPs), which has obvious advantages in forest fire smoke detection.

However, there are still some improvements to be made to this paper. The dataset built in this paper currently contains only daytime and nighttime conditions; therefore, the dataset can be subsequently expanded by collecting smoke images and videos with different weather and ambient colors to achieve better detection results. In the future, we will also be committed to deploying our models on equipment such as UAVs to better support early warning of forest fires.

6. Conclusions

In this paper, a new improved high-precision forest fire smoke detection model is proposed, and a forest fire smoke dataset is collected and established. First, to address the problem of sparse forest fire smoke data samples and a single seasonal background, this paper collects and uses CycleGAN to transform the smoke images with a seasonal style and improves the mosaic data enhancement method to enrich the dataset. Second, to address the problem of insufficient forest fire smoke feature extraction, this paper introduces Omni-Dimensional Dynamic Convolution, which adaptively decides the weight of each convolution kernel according to the input and learns the complementary attention of the convolution kernel from four dimensions, which significantly enhances the ability to capture contextual smoke information and makes the model better acquire the feature information of forest fire smoke. In addition, this paper introduces Bottleneck Transformer Net, which combines a CNN and transformer, enabling the model to fully extract the global feature information of forest fire smoke images while reducing the computational effort of the model and improving the detection accuracy of small target forest fires. The integration of the SimAM attention mechanism makes the network model focus more on the feature extraction of smoke targets while improving the detection accuracy of small target smoke. Finally, this paper replaces the original detection head of YOLOv5 with a decoupling head to speed up the convergence of the network and also effectively improve the detection accuracy of the forest fire smoke model. The experimental results show that the forest fire smoke detection OBDS model proposed in this paper has a high accuracy detection effect on forest fire smoke in various complex background forest environments and has certain advantages compared with the mainstream models, with a computational complexity of 21.5 GFLOPs, mAP@0.5 of 92.10%, and FPS of 54, which can meet the need for real-time detection and is of great value for early warning and fighting forest fires. It has an important application value for early warning and firefighting.

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

Funding: This work was supported by the Key Research and Development Plan of Jiangsu Province (Grant No. BE2021716) and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (2023).

Data Availability Statement: All data generated or presented in this study are available upon request from the corresponding author. Furthermore, the models and code used during the study cannot be shared at this time as the data also form part of an ongoing study.

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

References

1. Wu, Y.D.; Jia, Y.S. Link between climate anomaly and Australia bushfires in 2019−2020. *China Emerg. Rescue* **2020**, *2*, 23–27.
2. Xu, X.; Li, F.; Lin, Z.; Song, X. Holocene fire history in China: Responses to climate change and human activities. *Sci. Total Environ.* **2020**, *753*, 142019. [CrossRef]
3. Yuan, D.; Chang, X.; Huang, P.Y.; Liu, Q.; He, Z. Self-supervised deep correlation tracking. *IEEE Trans. Image Process.* **2020**, *30*, 976–985. [CrossRef] [PubMed]
4. Hu, S.; Zhu, F.; Chang, X.; Liang, X. Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers. *arXiv* **2021**, arXiv:2101.08001.
5. Wang, F.; Zhu, L.; Liang, C.; Li, J.; Chang, X.; Lu, K. Robust optimal graph clustering. *Neurocomputing* **2020**, *378*, 153–165. [CrossRef]
6. Bai, X.; Zhu, L.; Liang, C.; Li, J.; Nie, X.; Chang, X. Multi-view feature selection via nonnegative structured graph learning. *Neurocomputing* **2020**, *387*, 110–122. [CrossRef]
7. Zhang, F.; Zhao, P.; Xu, S.; Wu, Y.; Yang, X.; Zhang, Y. Integrating multiple factors to optimize watchtower deployment for wildfire detection. *Sci. Total Environ.* **2020**, *737*, 139561. [CrossRef]
8. Yao, J.; Raffuse, S.M.; Brauer, M.; Williamson, G.J.; Bowman, D.M.; Johnston, F.H.; Henderson, S.B. Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. *Remote Sens. Environ.* **2018**, *206*, 98–106. [CrossRef]
9. 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.
10. Hefeeda, M.; Bagheri, M. Wireless Sensor Networks for Early Detection of Forest Fires. In Proceedings of the IEEE International Conference on Mobile Adhoc & Sensor Systems, Pisa, Italy, 8–11 October 2007.
11. Fernández-Berni, J.; Carmona-Galán, R.; Martínez-Carmona, J.F.; Rodríguez-Vázquez, Á. Early forest fire detection by vision-enabled wireless sensor networks. *Int. J. Wildland Fire* **2012**, *21*, 938. [CrossRef]
12. Maruta, H.; Nakamura, A.; Kurokawa, F. A new approach for smoke detection with texture analysis and support vector machine. In Proceedings of the International Symposium on Industrial Electronics, Bari, Italy, 4–7 July 2010; pp. 1550–1555.
13. Zhao, Y.; Zhou, Z.; Xu, M. Forest fire smoke video detection using spatiotemporal and dynamic texture features. *J. Electr. Comput. Eng.* **2015**, *2015*, 706187. [CrossRef]
14. Filonenko, A.; Hernández, D.C.; Jo, K.H. Fast smoke detection for video surveillance using CUDA. *IEEE Trans. Ind. Inform.* **2017**, *14*, 725–733. [CrossRef]
15. Tao, H.; Lu, X. Smoke Vehicle detection based on molti-feature fusion and hidden Markov model. *J. Real-Time Image Process.* **2019**, *32*, 1072–1078.
16. Zhang, Q.X.; Lin, G.H.; Zhang, Y.M.; Xu, G.; Wang, J.J. Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images. *Procedia Eng.* **2018**, *211*, 441–446. [CrossRef]
17. Qiang, X.; Zhou, G.; Chen, A.; Zhang, X.; Zhang, W. Forest fire smoke detection under complex backgrounds using TRPCA and TSVB. *Int. J. Wildland Fire* **2021**, *30*, 329–350. [CrossRef]
18. Filonenko, A.; Kunianggoro, L.; Jo, K.H. Comparative study of modern convolutional neural network for smoke detection on image data. In Proceedings of the 2017 10th International Conference on Human System Interactions (HSI), Ulsan, Republic of Korea, 17–19 July 2017; pp. 64–68.
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
20. Simonyan, K.; Zisseman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. Yuan, F.; Shi, J.; Xia, X.; Fang, Y.; Fang, Z.; Mei, T. High-order local ternary patterns with locality preserving projection for smoke detection and image classification. *Inf. Sci.* **2016**, *372*, 225–240. [CrossRef]
23. Pan, J.; Ou, X.; Xu, L. A Collaborative Region Detection and Grading Framework for Forest Fire Smoke using weakly Supervised Fine Segmentation and Lightweight Faster-RCNN. *Forests* **2021**, *12*, 768. [CrossRef]
24. Li, J.; Zhou, G.; Chen, A.; Wang, Y.; Jiang, J.; Hu, Y.; Lu, C. Adaptive linear feature-reuse network for rapid forest fire smoke detection model. *Ecol. Inform.* **2022**, *68*, 101584. [CrossRef]

25. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
26. Zhao, Y.; Gao, F.; Yu, J.; Yu, X.; Yang, Z. Underwater Image Mosaic Algorithm Based on Improved Image Registration. *Appl. Sci.* **2021**, *11*, 5986. [CrossRef]
27. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. *arXiv* **2015**, arXiv:1506.02640.
28. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. *arXiv* **2016**, arXiv:1612.08242.
29. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
30. 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, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
32. Wang, C.Y.; Liao, H.Y.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. *arXiv* **2019**, arXiv:1911.11929.
33. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. *arXiv* **2019**, arXiv:1904.04971.
34. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11030–11039.
35. Li, C.; Zhou, A.; Yao, A. Omni-dimensional dynamic convolution. *arXiv* **2022**, arXiv:2209.07947.
36. Parikh, A.P.; Täckström, O.; Das, D.; Uszkoreit, J. A decomposable attention model for natural language inference. *arXiv* **2016**, arXiv:1606.01933.
37. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st lnternational Conferenceon Neural lnformation Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
39. Lin, J.; Lin, H.; Wang, F. STPM_SAHI: A Small-Target Forest Fire Detection Model Based on Swin Transformer and Slicing Aided Hyper Inference. *Forests* **2022**, *13*, 1603. [CrossRef]
40. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
41. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11863–11874.
42. 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, 18–23 June 2018; pp. 8759–8768.
43. Li, L.; Liu, F.; Ding, Y. Real-time smoke detection with Faster R-CNN. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems, Chongqing, China, 27–29 May 2022; pp. 1–5.
44. Huang, J.; Zhou, J.; Yang, H.; Liu, Y.; Liu, H. A Small-Target Forest Fire Smoke Detection Model Based on Deformable Transformer for End-to-End Object Detection. *Forests* **2023**, *14*, 162. [CrossRef]
45. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
46. Wang, Z.; Wu, L.; Li, T.; Shi, P. A Smoke Detection Model Based on Improved YOLOv5. *Mathematics* **2022**, *10*, 1190. [CrossRef]