fire

MDPI

Article

CNTCB-YOLOv7: An Effective Forest Fire Detection Model Based on ConvNeXtV2 and CBAM

Yiqing Xu [1], Jiaming Li [2], Long Zhang [2], Hongying Liu [3] and Fuquan Zhang [2,*]

[1] School of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China
[2] College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China
[3] School of Computer and Artificial Intelligence, Nanjing University of Science and Technology ZiJin College, Nanjing 210024, China
* Correspondence: zfq@njfu.edu.cn

Abstract: In the context of large-scale fire areas and complex forest environments, the task of identifying the subtle features and aspects of fire can pose a significant challenge for the deep learning model. As a result, to enhance the model's ability to represent features and its precision in detection, this study initially introduces ConvNeXtV2 and Conv2Former to the You Only Look Once version 7 (YOLOv7) algorithm, separately, and then compares the results with the original YOLOv7 algorithm through experiments. After comprehensive comparison, the proposed ConvNeXtV2-YOLOv7 based on ConvNeXtV2 exhibits a superior performance in detecting forest fires. Additionally, in order to further focus the network on the crucial information in the task of detecting forest fires and minimize irrelevant background interference, the efficient layer aggregation network (ELAN) structure in the backbone network is enhanced by adding four attention mechanisms: the normalization-based attention module (NAM), simple attention mechanism (SimAM), global attention mechanism (GAM), and convolutional block attention module (CBAM). The experimental results, which demonstrate the suitability of ELAN combined with the CBAM module for forest fire detection, lead to the proposal of a new method for forest fire detection called CNTCB-YOLOv7. The CNTCB-YOLOv7 algorithm outperforms the YOLOv7 algorithm, with an increase in accuracy of 2.39%, recall rate of 0.73%, and average precision (AP) of 1.14%.

Keywords: forest fire recognition; ConvNeXtV2; YOLOv7; CBAM; ELAN-CBAM; CNTCB-YOLOv7

Citation: Xu, Y.; Li, J.; Zhang, L.; Liu, H.; Zhang, F. CNTCB-YOLOv7: An Effective Forest Fire Detection Model Based on ConvNeXtV2 and CBAM. *Fire* **2024**, *7*, 54. https://doi.org/10.3390/fire7020054

Academic Editor: Grant Williamson

Received: 8 December 2023
Revised: 8 February 2024
Accepted: 9 February 2024
Published: 12 February 2024

1. Introduction

Forests are a vital component of the Earth's ecosystem, providing rich biodiversity and habitats for numerous plants and animals. Their presence contributes to maintaining ecological balance, promoting species interactions, and ensuring the stability of ecosystems [1,2]. However, forest fires devastate habitats and biodiversity. They are categorized by location into ground, surface, and crown fires, differing in behavior and impact. Their size, measured by the area burned or heat release rate (HRR), evolving from growth to decay phases, is influenced by environmental conditions and management. They can swiftly engulf vegetation and trees, leaving many wildlife species without their homes. Additionally, the significant carbon emissions released by forest fires exacerbate global climate change [3,4]. This climate change, in turn, further increases the risk of forest fires, creating a vicious cycle.

The early detection of forest fires allows for prompt action and emergency response. This helps in quickly controlling the fire and reducing the damage and loss caused by the fire [5–8]. At present, there are many ways to detect forest fires. Observation towers are a common way to see if forest fires are happening [9,10]. With the development of satellite remote sensing technology, people begin to observe forest fires by satellite [11,12]. The

deployment of sensors to detect forest fires is also one of the common ways that real-time forest environment detection can quickly discover the fire situation. Deep learning models can perform real-time processing and analysis, enabling rapid detection and response to forest fires [13–15]. This is crucial for emergency rescue and fire control, as it reduces the response time and helps minimize the damage caused by fires to some extent [16–19]. As a deep-learning-based object detection algorithm, YOLOv7 offers a high detection accuracy and inference speed. It is currently widely applied in the field of forest fire detection.

In order to address the limitations of traditional methods and reduce false alarms and complexity, Yar et al. proposed an improved YOLOv5s model that integrates a Stem module in the backbone of YOLOv5, replaces the larger kernel with a smaller kernel in the neck, and adds a P6 module in the head. Their model outperforms 12 other detection models and contributes a medium-scale annotated fire dataset for future research [20]. Al-Smadi et al. proposed a new framework that reduces the sensitivity of various YOLO detection models. Different yolo models, such as YOLOv5 and YOLOv7, are compared with Fast R-CNN (Region-based Convolutional Neural Network) and Faster R-CNN in detection performance and speed. The results show that the proposed method achieves significantly better results than the most advanced target detection algorithms while maintaining a satisfactory level of performance under challenging environmental conditions [21]. Zhou et al., based on the overall structure of YOLOv5 and MobileNetV3 as the backbone network, used semi-supervised knowledge extraction (SSLD) for training, which improved the convergence speed and accuracy of the model [22]. Dilli et al. used the target detection library YOLO model based on DL to carry out early wildfire detection on UAV thermal images, and used the significance graph integrated with thermal images to solve the shortcomings of using thermal images. The proposed approach is considered capable of providing technical support for night monitoring to reduce the catastrophic loss of forest resources and human and animal life in the early stages of wild forest fires [23]. Zhang et al. proposed a multi-scale convergent coordinated pyramid network with mixed attention and fast Robust NMS (MMFNet) for the rapid detection of forest fire smoke [24]. Jin et al. designed an enveloping self-focusing mechanism to solve the problem of identifying bad fire sources, focusing on the characteristics of the channel and spatial direction, and collecting contextual information as accurately as possible. In addition, a new feature extraction module is constructed to improve the detection efficiency while preserving the feature information [25]. In summary, these studies share a common focus on improving fire detection performance using various modifications and enhancements to YOLO-based models. They explore different techniques, such as integrating new modules, comparing YOLO models with other detection algorithms, utilizing semi-supervised learning, and incorporating attention mechanisms. Despite their promising results, these studies may still face challenges in addressing specific issues, such as sensitivity to environmental conditions, identification of bad fire sources, and efficient feature extraction. Further research and development are needed to optimize these models and address their limitations.

With the continuous advancement of the YOLO series algorithms, YOLOv7 has emerged as a remarkable innovation, offering improved accuracy and faster processing speeds compared to its predecessor, YOLOv5. The application of YOLOv7 in forest fire detection holds great potential for enhancing the effectiveness of such detection efforts. However, the task of detecting forest fires poses certain challenges, particularly in scenarios where the fire area is extensive and the forest background is complex. In such cases, the model may struggle to capture the intricate details and distinguishing features of the fire.

To address these challenges and bolster the applicability of the YOLOv7 algorithm in forest fire detection, this research focuses on augmenting the model's capabilities by incorporating ConvNextV2 and ConvFormer networks. ConvNeXtV2 integrates self-supervised learning techniques along with Fully Convolutional Masked AutoEncoder (FCMAE) and Global Response Normalization (GRN) layers, enhancing the model's performance in various recognition tasks. Conv2Former employs a simple convolutional modulation layer instead of the self-attention mechanism, and compared with residual modules, the

convolutional modulation operation in Conv2Former can also adapt to the content of the input. Moreover, to enhance the model's ability to discern crucial information amidst complex forest backgrounds, an attention mechanism is introduced through the ELAN-CBAM module, building upon the ELAN structure. This culmination of efforts gives rise to the CNTCB-YOLOv7 algorithm for forest fire detection.

Compared with the standard YOLOv7 algorithm, the CNTCB-YOLOv7 algorithm places greater emphasis on global information, effectively reducing false detection and elevating both the detection accuracy and AP. Leveraging these improvements, the research contributes to the study of forest fire behavior and the identification of key characteristics that aid in the understanding and prediction of forest fire propagation. This, in turn, facilitates more proactive and targeted firefighting strategies, ultimately leading to improved forest fire management and mitigation efforts. In addition, real-time monitoring and analysis of forest fire situations can collect a large amount of fire data, which is helpful for studying the spread patterns and characteristics of forest fires under different environments and conditions, such as the rate of fire spread. The improved model performance can also support the establishment of more accurate forest fire risk prediction models, enhancing the ability for early warning and forecasting. In conclusion, the proposed method in this study provides technical support for in-depth research on forest fire behavior and forest fire management.

2. Materials and Methods

2.1. Hyperparameter Settings and Dataset

2.1.1. Hyperparameter Settings

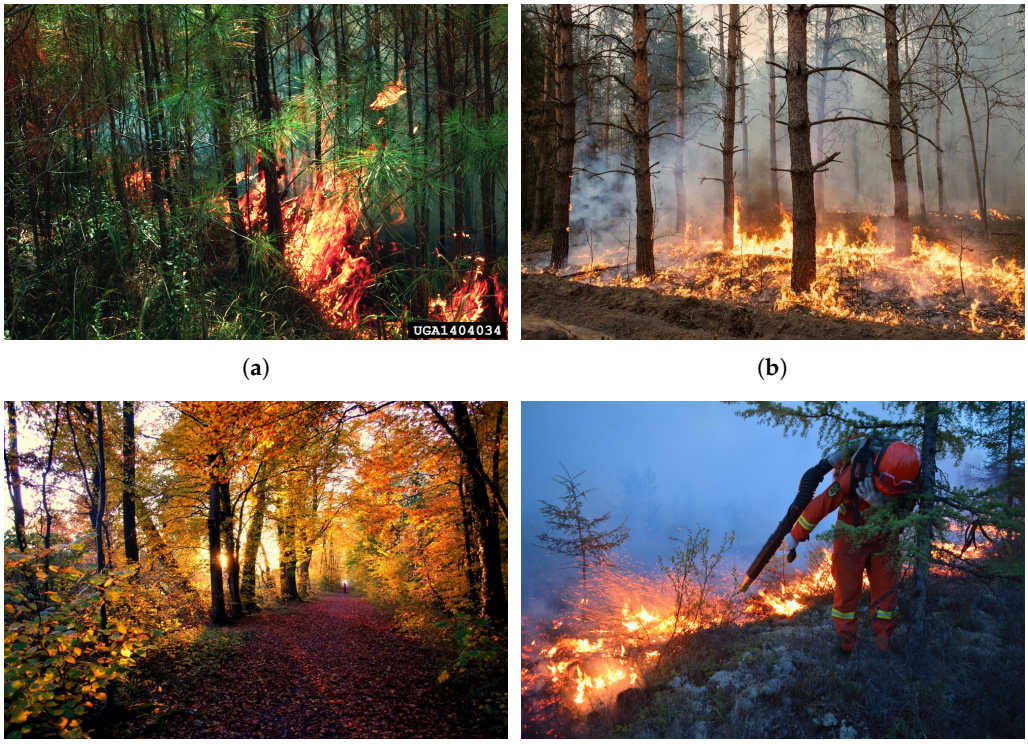
The hyperparameter settings in the experiments include the image size, epochs, batch size, initial learning rate (Lr0), and optimizer. Image size determines the input size of the model, usually measured in pixels, set to 640 × 640 pixels in our experiments. Epochs determine the number of iterations the model goes through the entire dataset during the training process, with 200 epochs set for this study. Batch size refers to the number of samples used to update the model weights each time, and here it is set to 8. Initial learning rate (Lr0) determines the initial learning speed of the model, which is set at 0.01 in our case. The optimizer determines the optimization method used by the model to find local optimal solutions, and in this study, stochastic gradient descent (SGD) is used as the optimization method.

The aforementioned settings, which contribute to enhancing the training process and the performance of the models, are derived from experimental trials and empirical assessments. The optimal configurations of these hyperparameters are influenced by the characteristics of the datasets and the architecture of the models. It is crucial to adjust these values when conducting different experiments to ensure the best possible outcomes.

2.1.2. Dataset

In order to obtain forest fire images required by model training, we employed various data collection methods. Firstly, we downloaded traditional forest fire images and non-forest fire images using web crawler technology. Secondly, we extracted a series of frames from downloaded forest fire videos to serve as additional forest fire images. Moreover, we utilized publicly available fire datasets, such as the BoWFireDataset [26]. The combined use of these data sources contributes to enhancing the quality and effectiveness of the model training. A total of 2590 images were obtained. Among the collected images, 2058 images were positive sample images with forest fire, while the remaining 532 images were negative sample images without forest fire. To ensure the compatibility of the input data with our model's requirements, all images were uniformly resized to a resolution of 640 × 640 pixels. This standardization is crucial for maintaining consistency across the dataset and facilitating efficient processing by the models employed in our study. Furthermore, considering the specific context of detecting large-scale fires within complex forest environments, certain images underwent cropping, aimed at enhancing the proportional representation of fire

within these images. Finally, the prepared forest fire dataset was divided into the training set and verification set according to the ratio of 8:2. Figure 1 shows some fire and non-fire images included in the dataset.



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

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

Figure 1. Dateset of forest fires: (**a**,**b**,**d**) fire images; and (**c**) non-fire image.

2.1.3. Model Performance Evaluation Index

In this paper, the task of forest fire detection is classified as a binary problem, that is, it is judged as fire or non-fire. For the forest fire category, fire is a positive sample and non-fire is a negative sample. In the binary classification problem of forest fire, the following four situations usually occur in the data sample, which are True Positive (TP), the result predicted by the model is positive sample, and the actual number of samples is positive sample, that is, the fire is predicted by the model, and the real picture is also fire, respectively. If the example is True Negative (TN), the result predicted by the model is negative samples, and the actual number of samples is negative samples; that is, it is predicted by the model as non-fire, and the real picture is also non-fire. In the case of False Positive (FP), the predicted result of the model is positive samples, but it is actually the number of samples of negative samples, that is, the number of samples that are misjudged as fire without fire. False Negative example (FN): The result predicted by the model is negative samples, but it is actually the number of samples of positive samples; that is, the number of samples that misjudge the fire as non-fire [27].

Precision is the proportion of true positive samples out of all the samples predicted as positive by the model. The calculation method is shown as Equation (1) [28].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall is the proportion of true positive samples that are accurately predicted as positive by the model, out of all the true positive samples. The calculation method is shown as Equation (2) [29].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Average precision (AP) is a metric that measures the average precision. It is obtained by calculating the area under the Precision–Recall (P-R) curve generated by plotting precision (P) on the x-axis and recall (R) on the y-axis. The calculation formula for AP is shown as Equation (3) [30].

$$\mathrm{AP} = \int_0^1 \mathrm{P(R)dR} \tag{3}$$

When calculating AP, the average precision values for different classes are weighted and averaged to obtain the mean average precision (mAP) [31]. The calculation formula is shown as Equation (4), where n represents the total number of classes and AP_i represents the AP value for the i-th class.

$$\mathrm{mAP} = \frac{\sum_{i=1}^{n} \mathrm{AP}_i}{n} \tag{4}$$

mAP is commonly used to evaluate object detection algorithms. In this paper, we focus on forest fire detection, a single class, so we use the AP metric with a 50% Intersection Over Union (IOU) threshold, referred to as AP50 [32].

2.2. YOLOv7 Algorithm Structure

YOLOv7 is an object detection model known for its high accuracy, ease of training, and deployment capabilities [33]. It has a faster network speed compared with the YOLOv5 model and achieves better results on the MS COCO (Microsoft Common Objects in Context) dataset. The overall network structure of YOLOv7 is shown in Figure 2, which shares similarities with the network structure of YOLOv5, with the main difference being the internal network modules. At the input end, YOLOv7 uses the same Mosaic data augmentation method as YOLOv5, as well as adaptive anchor box calculation and adaptive image scaling. The main backbone network of YOLOv7 incorporates the Extended Efficient Layer Aggregation Networks (E-ELAN) and Max Pooling (MP) modules, merging the model's Neck and Head layers into a unified Head layer. In Figure 2, MP1 and MP2 are two separate MP modules used in the YOLOv7 backbone network.

As shown in the diagram, the CBS (Convolution-BatchNorm-Silu) module consists of three components: a convolutional layer, a Batch Normalization layer, and a Silu (Sigmoid Linear Unit) activation function. The ELAN (Effective Layer Aggregation Network) module is an effective hierarchical aggregation network that employs a feature fusion technique to enhance the model's feature extraction capability and obtain stronger feature representations. The ELAN module has two main branches. One branch adjusts the number of channels using a 1×1 convolutional kernel, while the other branch adjusts the number of channels with a 1×1 convolutional kernel and then performs feature extraction using four consecutive 1×1 convolutional kernels. The outputs of the four branches are then concatenated to obtain the final output. The Efficient Layer Aggregation Networks-Higher (ELAN-H) module is similar to the ELAN module in structure, but it differs in the number of selected output features to be concatenated in the second branch, which is higher.

The MP module also consists of two main branches, as shown in Figure 3, and its main purpose is to perform downsampling operations on the feature maps. One branch uses max pooling followed by a 1×1 convolutional layer with a stride of 1 to adjust the number of channels. The other branch adjusts the number of channels with a 1×1 convolutional layer and then performs downsampling using a 3×3 convolutional layer with a stride of 2. The outputs of the two branches are concatenated to obtain the final downsampling output.

The SPPCSPC module, as a component of the YOLOv7 structure, effectively extracts image features and improves the detection accuracy of the model. The SPPCSPC module consists of two parts: SPP (Spatial Pyramid Pooling) and CSP (Cross Stage Partial). The SPP part is primarily responsible for performing feature pooling at different scales to extract features of varying sizes. The CSP part aims to reduce the number of parameters and further enhance the feature extraction capabilities. The structure of the SPPCSPC module is shown in Figure 4, featuring multiple branches of max pooling. Each max pooling branch operates at a different scale. The pooling operations at different scales have different receptive fields,

allowing the model to better handle objects of varying sizes and ensuring the effectiveness of the detection process.

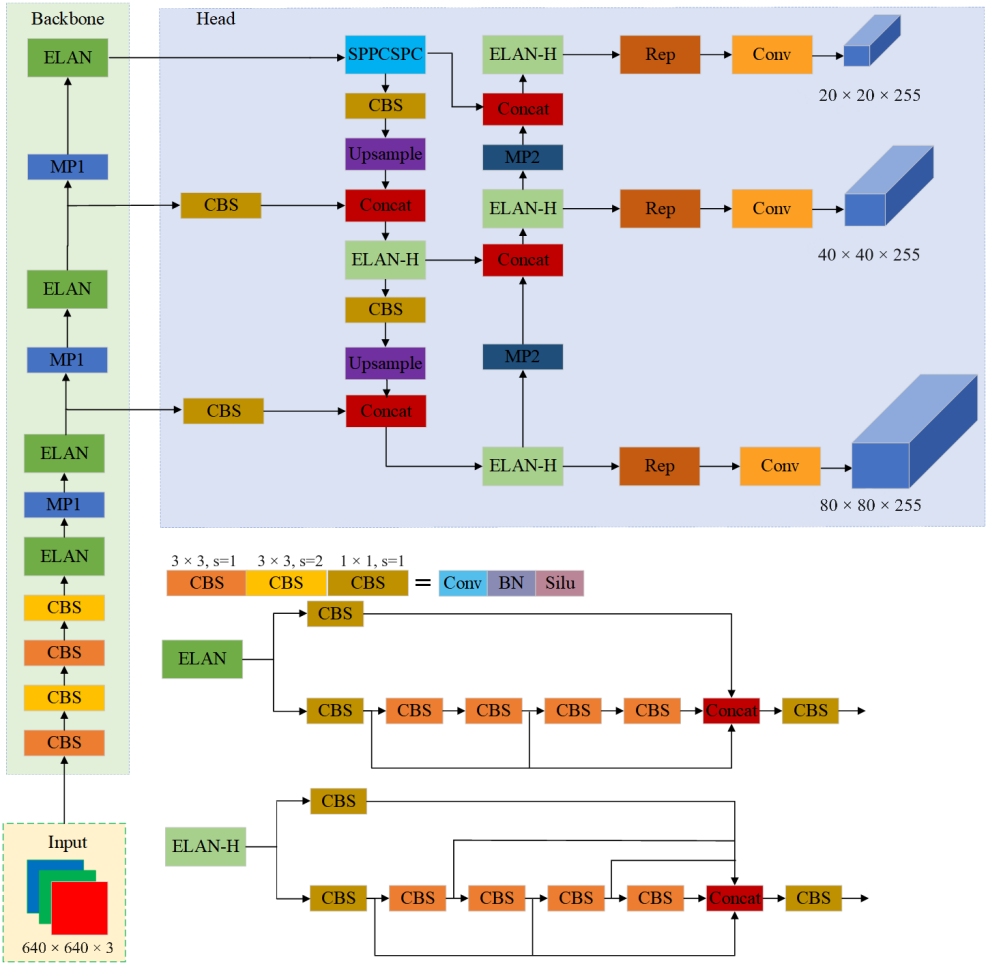


Figure 2. YOLOv7 Model Architecture.

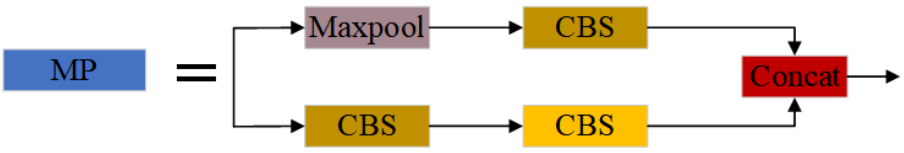


Figure 3. MP module.

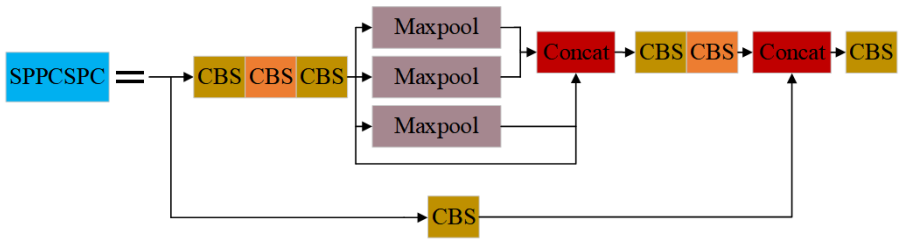


Figure 4. SPPCSPC module.

2.3. Improving the Network Used by the YOLO7 Algorithm

2.3.1. ConvNeXtV2

The ConvNeXt model was proposed by leveraging the network structure of the Swin Transformer and using the ResNet-50 architecture as a base, as described in [34]. The performance of the ConvNeXt model on COCO detection and ADE20K surpasses

that of the Swin Transformer. The authors of the ConvNeXt model trained the ResNet-50 network model using ViTs' strategy, which yielded better results compared with the original. Building upon this baseline, a series of experiments were conducted. The ConvNeXt model applies downsampling to the feature maps using a 4×4 convolutional kernel with the same stride as the Swin Transformer system, resulting in a slight improvement in accuracy. Various sizes of convolutional kernels were experimented with in the ConvNeXt model, and the results indicate that the 7×7 kernel achieved the best performance and highest accuracy.

Modern convolutional neural networks, such as ConvNeXt, have demonstrated an advanced performance in various scenarios, thanks to continuous improvements in representation learning frameworks and architectures. Researchers attempted to combine ConvNeXt with self-supervised learning techniques like masked autoencoders (MAE), but the resulting performance was unsatisfactory. Therefore, it was proposed to add a FCMAE and GRN layer to the structure of the previous version of ConvNeXt (ConvNeXt V1) to enhance feature competition between channels. This new model, which incorporates both self-supervised learning techniques and architectural improvements, is referred to as ConvNeXt V2 [35]. The block structures of the ConvNeXt V1 and ConvNeXt V2 are shown in Figure 5. Compared with ConvNeXt V1, the GRN layer was added after the MLP (Multi-Layer Perceptron) layer and the redundant LayerScale was dropped in ConvNeXt V2.

FCMAE is a novel self-supervised learning framework that comprises a ConvNeXt encoder based on sparse convolution and a lightweight ConvNeXt block decoder. This framework is capable of efficiently processing masked inputs and reduces the computational cost of pre-training. GRN is a normalization technique used in convolutional neural networks to enhance contrast and selectivity between channels, with the main goal of improving the model's performance in recognition tasks. GRN consists of three steps: global feature aggregation, feature normalization, and feature calibration. The ConvNeXt V2 part in Figure 5 illustrates the structure of adding GRN to the ConvNeXt block. In the ConvNeXt V2 model, the adoption of fully convolutional masked autoencoder and global response normalization techniques further enhances the model's performance in various recognition tasks.

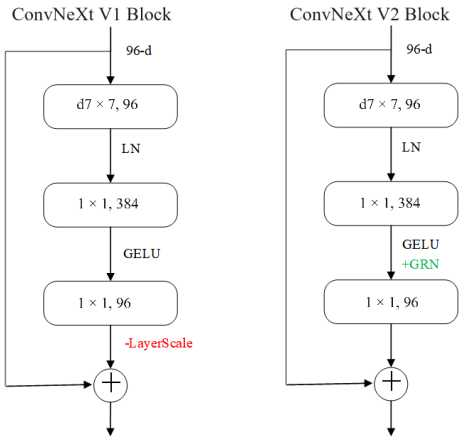


Figure 5. Block structures of ConvNeXt V1 and ConvNeXt V2. In ConvNeXt V2, the GRN layer (in green) was added after the dimension-expansion MLP layer and the LayerScale (in red) was dropped.

2.3.2. Conv2Former

The self-attention mechanism in transformers can model global pairwise dependencies and provide a more efficient way of encoding spatial information. However, when processing high-resolution images, self-attention can be computationally expensive. ConvNext, by borrowing the design and training approach from transformers, achieves a better performance than some common transformers. To date, how to effectively construct more powerful models using convolutions remains a hot research topic.

In Conv2Former, when processing high-resolution input images, a simple convolutional modulation layer is used instead of self-attention, which can save memory consumption compared with self-attention. Moreover, compared with residual modules, the convolutional modulation operation in Conv2Former can also adapt to the content of the input [36]. As shown in Figure 6, on the left is the self-attention operation, where the output of each pixel is obtained by taking the weighted sum of all the positions. Similarly, this process can be simulated by the convolutional modulation operation on the right side of the figure, which calculates the output of a large kernel convolution and performs a Hadamard product with the value representation. The results show that using convolution to obtain the weight matrix can also achieve good results.

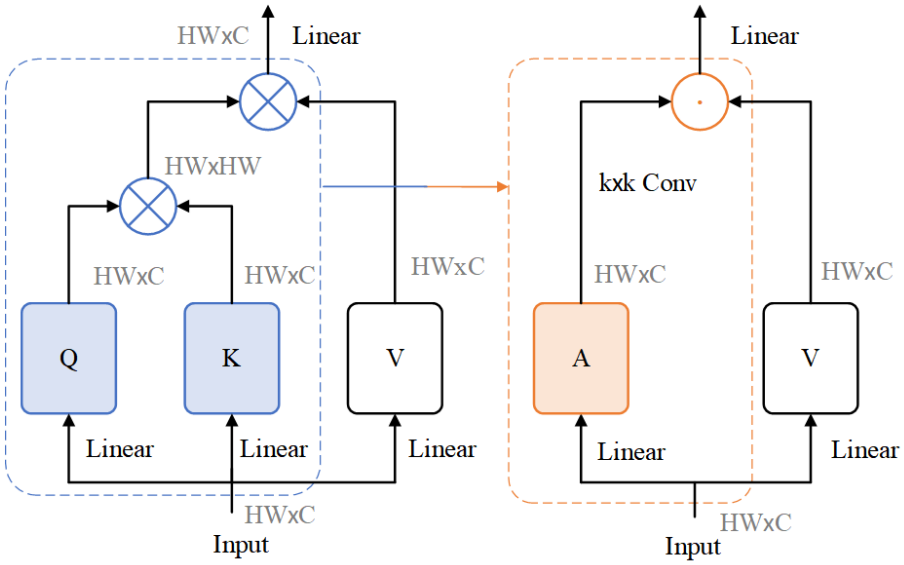


Figure 6. Self attention mechanism and convolutional modulation operation.

2.4. Improved Strategy for YOLOv7

In order to improve feature extraction and information fusion for forest fire detection in larger and more complex scenarios, and to enhance the detection accuracy of the YOLOv7 algorithm, this study modifies the backbone network and the Head layer of the YOLOv7 algorithm. Specifically, high-performance ConvNeXtV2, Transformer-style Conv2Former, introduced in the previous chapter, are used to replace the first and last ELAN modules in the backbone network, as well as all ELAN-H modules in the head layer. As a result, multiple improved versions of the YOLOv7 algorithm are obtained, namely ConNeXtV2-YOLOv7 and ConvFormer-YOLOv7. As the overall network architecture is similar, this study only presents the network structure of ConNeXtV2-YOLOv7, as shown in the Figure 7.

2.4.1. Backbone and Head Improvement

To enhance the performance of convolutional neural network(CNN) models, a common approach is to introduce attention mechanisms. Attention mechanisms can suppress irrelevant noise information and allow CNN models to focus more on useful information, thereby improving the model's expressive power to handle different visual tasks. Additionally, attention mechanisms can select and compress feature maps, suppressing non-essential information and reducing the dimensionality of feature maps, thereby reducing computational complexity. Attention mechanisms can improve the model's robustness to factors such as occlusion and noise, making the model more robust. Furthermore, the introduction of attention mechanisms provides interpretability and visualizability, making the model's outputs more intuitive and understandable. To further improve the performance of the YOLOv7 algorithm and make the network pay more attention to important information

in the current forest fire detection task, attention mechanisms are introduced to aggregate local information of feature maps. Specifically, improvements are made to the remaining ELAN module in the backbone network, as indicated by the "Attention" label in Figure 8. Four types of attention mechanisms are experimented with individually.

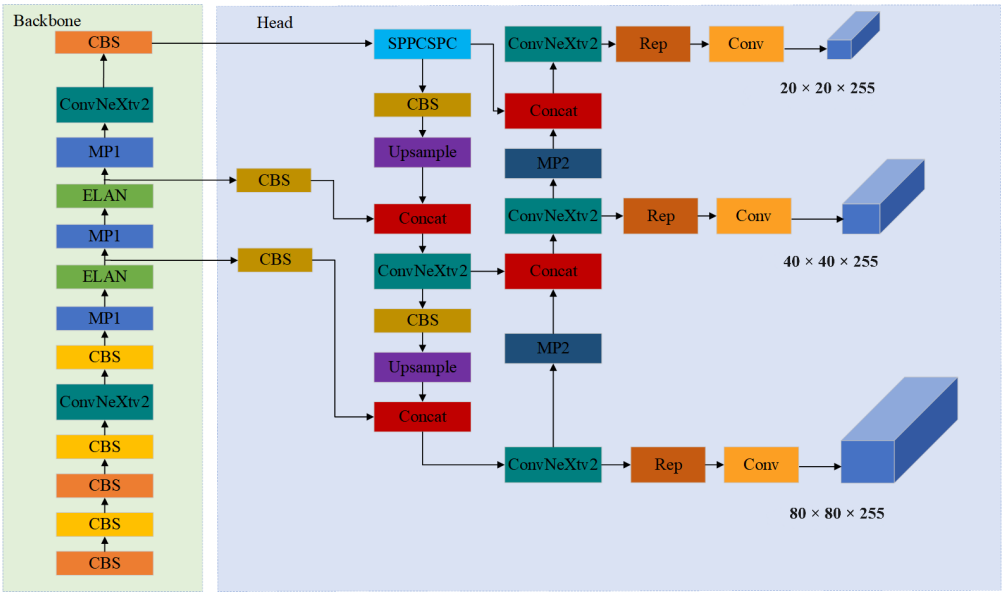


Figure 7. The network structure of ConNeXtV2-YOLOv7.

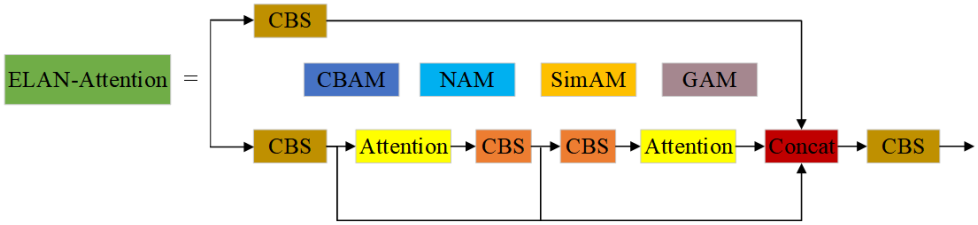


Figure 8. The structure of introducing the attention mechanism for ELAN.

2.4.2. ELAN Structures That Introduce Attention Mechanisms

Normality-based Attention Module (NAM), as a lightweight and efficient attention module based on normalization , is often used in image classification and target detection tasks in deep learning. NAM proposed an attentional calculation method that can be weighted for input feature graphs [37]. The importance of weights is expressed by the normalized scaling factor, so as to suppress irrelevant channels and pixel information in images. In this way, differences between input values can be better distinguished, allowing the network to focus more on the features that are most useful for the task at hand.

Attention mechanisms are commonly used in various computer vision tasks to improve model performance and have received widespread attention. However, the importance of preserving both channel and spatial information for enhancing cross-dimensional interactions is often overlooked. Therefore, a Global Attention Mechanism (GAM) is proposed, which aims to improve the performance of deep neural networks by reducing information redundancy and amplifying global interaction representations [38]. GAM draws inspiration from the sequential channel attention mechanism of CBAM and redesigns the sub-modules.To maintain cross-dimensional information, the channel attention sub-module in GAM uses a 3D arrangement and employs a multi-layer perceptron to amplify spatial dependencies across dimensions. Additionally, in the spatial attention sub-module, the concentration of spatial information is achieved through the use of two convolutional layers. Experimental results on image classification tasks such as CIFAR-100 and ImageNet-1k demonstrate that this attention mechanism exhibits an excellent performance in models like ResNet and MobileNet.

The Convolutional Block Attention Module (CBAM) is an attention mechanism that combines both spatial and channel attention to aggregate the local information of feature maps. The channel attention module and spatial attention module are two independent submodules of CBAM, allowing the network to focus more strongly on important information and perform weighted attention on both spatial and channel dimensions, achieving a plug-and-play effect [39]. When a feature map is input to CBAM, it first goes through the channel attention module. In the channel attention module, the feature map undergoes two parallel operations: max pooling and average pooling, which compress the feature map into two one-dimensional feature vectors. These vectors are then passed through a shared fully connected layer, and the results are added together. Finally, the sigmoid activation function is applied to obtain the channel attention features.The channel attention features are multiplied with the input features to obtain the input features for the spatial attention module. This process also includes max pooling and average pooling operations. After pooling, the results are concatenated based on channels and passed through a convolutional layer to adjust the channels to 1. The sigmoid activation function is applied to obtain the spatial attention feature map.The spatial attention feature map is multiplied with the input of this module to obtain the final generated feature map.

Currently, attention modules usually suffer from two problems. First, they can only refine features along the channel or spatial dimension, thus limiting the flexibility of learning their attention weights across channels as well as spatial variations. In addition, their structures such as pooling need to be composed of a complex set of elements. Therefore, based on neuroscience theory, the SimAM module is proposed for solving these problems. After considering the spatial and channel dimensions, the 3D weights are inferred from the current neurons, and then the neurons are refined, allowing the network to learn more discriminative neurons.SimAM, as a conceptually simple but effective attention module , is able to infer feature maps in a layer without adding parameters to the original network compared with common spatial as well as channel attention module 3D weights [40]. In addition, an optimized energy function is proposed so as to derive the importance of each neuron. On the CIFAR-10 and CIFAR-100 datasets, the SimAM module has a better performance in terms of accuracy compared with common attention modules such as SE and ECA.

3. Results

3.1. Comparison of Multiple Model Results

In this section, we consider applying various structures to the YOLOv7 algorithm and compare the performance of the models to find a better model for forest fire detection. Table 1 shows a comparison of the experimental results for different models. Conv2Former has a better performance than traditional CNN-based models [36], and in order to further improve the model's performance, the Conv2Former-YOLOv7 algorithm was proposed. However, according to the results, applying Conv2Former-YOLOv7 to forest fire detection did not achieve the expected performance in terms of accuracy, recall rate, and AP. ConvNeXtV2 can enhance channel-wise feature competition and has shown a superior performance in various visual tasks by using fully convolutional mask autoencoders and global response normalization techniques. Therefore, the ConvNeXtV2-YOLOv7 algorithm was proposed. The experimental results showed that compared with the YOLOv7 algorithm, the ConvNeXtV2-YOLOv7 algorithm achieved an accuracy of 85.81% and an increase of 2.02%. It also improved the recall rate by 0.59% and the AP by 0.61%. After comprehensive comparison of overall performance, the ConvNeXtV2-YOLOv7 algorithm is more suitable for forest fire detection.

3.2. An Experimental Comparison of Attentional Mechanisms

In order to further enhance the model's generalization ability, suppress irrelevant features and pixel information in forest fire images, and better distinguish the differences between input values, the network should pay more attention to the most useful features for the current task. Therefore, in this section of the experiment, based on the performance of

the ConNeXtV2-YOLOv7 algorithm, attempts were made to embed NAM, SimAM, GAM, and CBAM modules in the ELAN structure of its backbone network. In order to verify the feasibility of the method, comparative experiments were conducted between the attention-mechanism-integrated ConvNeXtV2-YOLOv7 algorithm and YOLOv7 and ConvNeXtV2-YOLOv7 algorithms. As shown in Table 2, it can be observed that compared with the ConNeXtV2-YOLOv7 algorithm, the introduction of SimAM and GAM attention modules did not effectively improve the model's performance; instead, there was a slight decline. However, by introducing the NAM and CBAM attention modules, there was a certain improvement in accuracy. Specifically, the introduction of the NAM attention module led to a decrease of 3.81% in the recall rate, while the introduction of the CBAM attention module showed a slight improvement. Additionally, both models showed varying degrees of improvement in AP, with the introduction of the CBAM attention module showing a more significant improvement. In terms of parameter count, the introduction of the GAM attention module increased the parameter count to 50.1 million (M), while the algorithms incorporating the NAM, SimAM, and CBAM modules all showed a decrease in parameter count compared with the YOLOv7 algorithm, and the parameter count was relatively close.

Table 1. Comparison of experimental results of different models.

Model	P, %	R, %	AP, %
YOLOv7	83.79	81.12	87.22
Conv2Former-YOLOv7	83.17	80.43	87.22
ConvNeXtV2-YOLOv7	85.81	81.71	87.83

Table 2. Experimental comparison of adding different attention mechanisms.

Model	P, %	R, %	AP, %	Parameter, M
YOLOv7	83.79	81.12	87.22	37.2
ConNeXtV2-YOLOv7	85.81	81.71	87.83	34.48
ConNeXtV2-YOLOv7 + NAM	86.03	77.9	88.07	33.71
ConNeXtV2-YOLOv7 + SimAM	83.75	81.46	87.67	33.71
ConNeXtV2-YOLOv7 + GAM	84.82	79.92	87.05	50.1
ConNeXtV2-YOLOv7 + CBAM	86.18	81.85	88.36	33.73

Through comprehensive comparison, the performance of the ConNeXtV2-YOLOv7 algorithm was improved by incorporating the CBAM attention mechanism, leading to the proposal of the CNTCB-YOLOv7 forest fire detection method. Building upon the ConNeXtV2-YOLOv7 algorithm, embedding the CBAM module in the ELAN structure of the backbone network effectively improved the model's accuracy. Figure 9 illustrates the structure of ELAN-CBAM, which enhances global interactions while preserving channel and spatial information, thereby improving the performance and detection effectiveness of the network model.

Furthermore, as shown in Table 3, compared with the YOLOv7 algorithm, the CNTCB-YOLOv7 algorithm achieved an accuracy of 86.18%, an improvement of 2.39%. The recall rate and AP were also improved by 0.73% and 1.14%, respectively. Additionally, in terms of model lightweightness, the CNTCB-YOLOv7 algorithm only requires 33.73 M, a reduction of 3.47 M compared with the YOLOv7 algorithm. This reduction in computational resource usage helps improve the inference speed of the model.

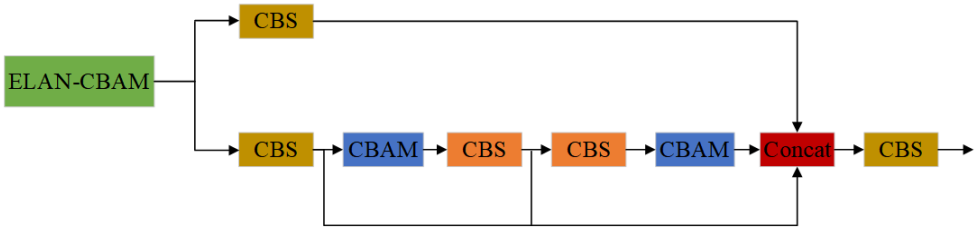


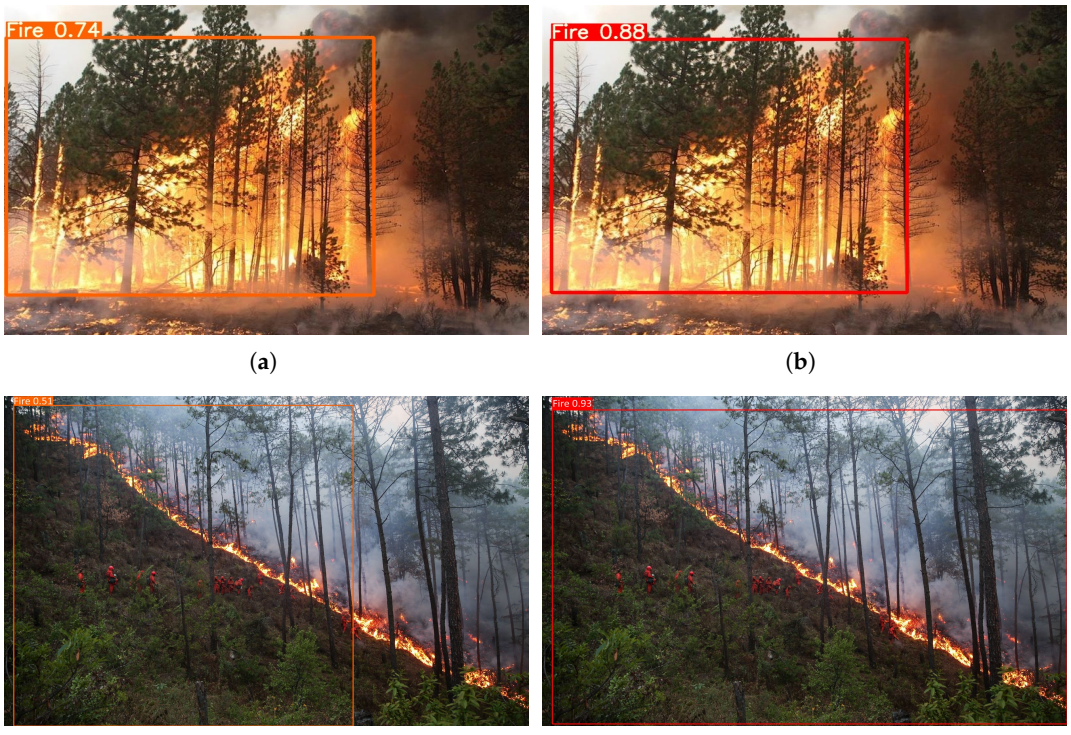
Figure 9. ELAN-CBAM structure.

Table 3. Experimental results of YOLOv7 and CNTCB-YOLOv7.

Model	P, %	R, %	AP, %	Parameter, M
YOLOv7	83.79	81.12	87.22	37.2
CNTCB-YOLOv7	86.18	81.85	88.36	33.73

In addition, the YOLOv7 algorithm and the proposed CNTCB-YOLOv7 algorithm were tested on a dataset of test images, and partial test results are shown in Figures 10 and 11.

Figure 10 shows the test results for large-scale forest fires. In Figure 10a,c, we can see the test results of the YOLOv7 algorithm, while Figure 10b,d show the test results of the CNTCB-YOLOv7 algorithm. From Figure 10a,b, it can be observed that when the forest fire image represents a large-scale crown fire, the CNTCB-YOLOv7 algorithm performs better in terms of detection compared with the YOLOv7 algorithm. Furthermore, from Figure 10c,d, it can be seen that when the forest fire image represents a large-scale surface fire, the YOLOv7 algorithm only detects a portion of the fire area in the image, while the CNTCB-YOLOv7 algorithm is able to detect all fire areas in the image. The CNTCB-YOLOv7 algorithm pays more attention to the global information of forest fires compared with the YOLOv7 algorithm, resulting in a better detection performance.

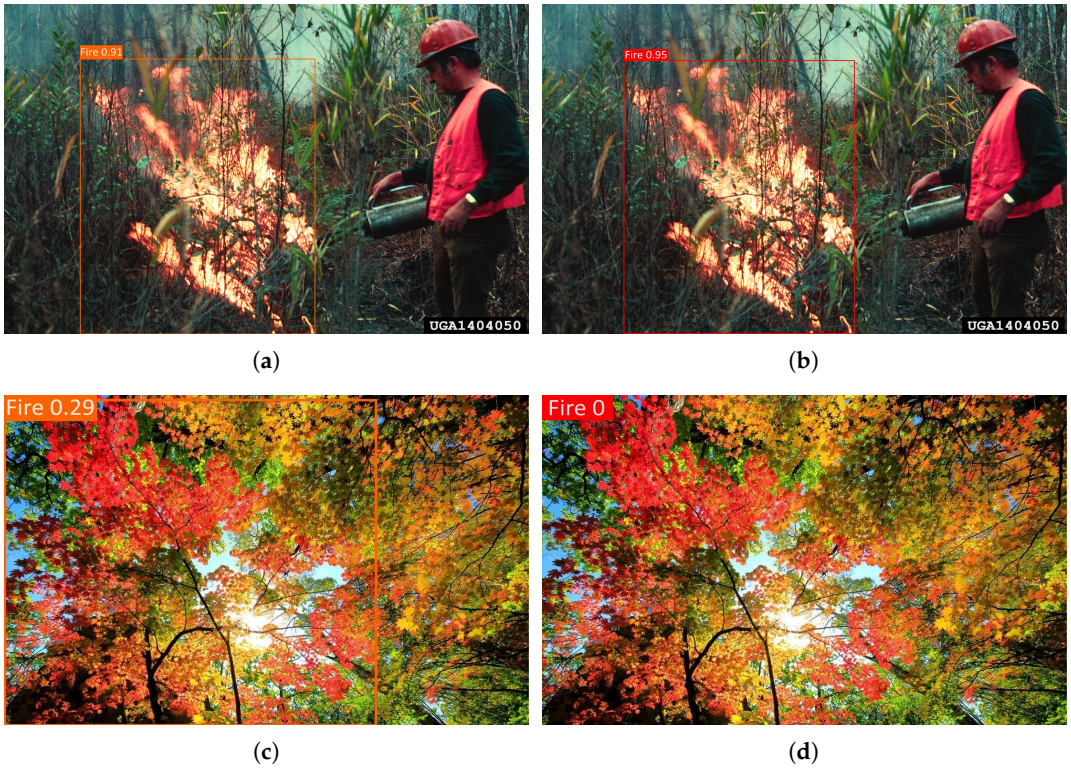


(**a**) (**b**)

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

Figure 10. Test image results with a large range of forest fires: (**a**,**c**) YOLOv7 algorithm; and (**b**,**d**) CNTCB-YOLOv7 algorithm.

Figure 11 show the test results for complex forest backgrounds. In Figure 11a,c, we can see the test results of the YOLOv7 algorithm, while Figure 11b,d show the test results of the CNTCB-YOLOv7 algorithm. From Figure 11a,b, it can be observed that when there is background interference similar to the color of the fire in the image, both the YOLOv7 and CNTCB-YOLOv7 algorithms did not produce false detections. Additionally, the detection performance of the CNTCB-YOLOv7 algorithm is superior to that of the YOLOv7 algorithm. Furthermore, as shown in Figure 11c,d, when there are images with colors and textures similar to the fire, the YOLOv7 algorithm might have result in false detections, while the CNTCB-YOLOv7 algorithm did not lead to such occurrences.



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

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

Figure 11. Test image results with a complex forest background: (**a**,**c**) YOLOv7 algorithm; and (**b**,**d**) CNTCB-YOLOv7 algorithm.

4. Discussion

In order to improve the feature representation capability and detection accuracy of the model, and to make the network pay more attention to the most useful features for the current task, further enhancing the model's generalization ability, we made corresponding improvements to the YOLOv7 algorithm. The experimental results show that the proposed CNTCB-YOLOv7 algorithm surpassed the YOLOv7 algorithm in terms of precision, recall, and mean average precision, and it had a lower parameter count and faster inference speed. We first introduced the ConvNeXtV2 and Conv2Former network structures, replacing parts of the ELAN modules in the YOLOv7 algorithm to enhance the model's feature representation ability and detection accuracy. Comparative experiments revealed that the ConvNeXtV2-YOLOv7 algorithm was more suited for forest fire detection tasks, thus it was chosen as the base model for further improvements. On this foundation, we introduced an attention mechanism by embedding the CBAM module within the backbone network's ELAN structure, achieving the aggregation of local information in feature maps. This enabled the network to focus more on critical information in forest fire detection tasks, leading to the development of the CNTCB-YOLOv7 algorithm. The introduction of the CBAM module significantly improved the model performance, while reducing the parameter count, which is advantageous for enhancing the inference speed.

This methodology offers technical support for in-depth research on forest fire behavior and management. Early high-precision detection can shorten response times, helping to quickly controlling the spread of fires and mitigate losses. Real-time monitoring and analysis of forest fires can collect extensive fire data, aiding in the study of fire spread patterns and characteristics under different environments and conditions, such as the rate of fire spread. The improved model performance also supports the development of more accurate forest fire risk prediction models, enhancing early warning and forecasting capabilities.

The CNTCB-YOLOv7 algorithm, characterized by a superior detection accuracy, can contribute significantly to an enhanced comprehension of fire behavior. This, in turn, facilitates the implementation of more proactive firefighting strategies, thereby bolstering the overall management and mitigation of forest fires.

However, there may be potential limitations in the model's generalization ability to different environments and conditions, which could be addressed in future work by exploring more diverse datasets and incorporating additional attention mechanisms. While our model demonstrates performance improvements over YOLOv7, it lacks comparative analysis with other prevalent models like Faster R-CNN, SSD, or RetinaNet.

In current study, the evaluation of our model primarily relied on metrics such as Precision, Recall, AP, and mAP. These metrics were selected due to their direct relevance to the performance goals of our classification task, especially in the context of our uniquely self-collected dataset. However, our analysis lacks crucial statistical measures that are essential for understanding the variability and reliability of the model's performance across different scenarios. This absence might limit the depth of our findings in terms of statistical consistency and reliability.

In future work, we plan to enhance our model through further refinement by incorporating additional attention mechanisms, conducting comprehensive comparisons with other prevalent models, and expanding our evaluation criteria. This expansion includes introducing standard deviations and confidence intervals in our analysis to provide a more comprehensive statistical understanding of our model's performance. Additionally, we aim to test and validate our model on a broader range of datasets. This expansion will not only enhance the generalizability of our findings, but also allow us to assess the model's performance under different scenarios and conditions. In addition, we will explore the applicability of our model in diverse domains by considering additional data types, such as LiDAR (Light detection and ranging) data [41,42]. This broader range of datasets will enable us to thoroughly test and validate the robustness and versatility of our model across various fields.

Furthermore, future research could focus on optimizing the model's inference speed and reducing computational resource utilization, making it more suitable for real-time monitoring and analysis of forest fires. Furthermore, we plan to investigate how the enhanced model could support more accurate forest fire risk prediction models, thereby aiding in forest fire management and mitigation efforts.

5. Conclusions

This article presents a forest fire detection model based on ConvNeXtV2 and CBAM, named CNTCB-YOLOv7, designed to enhance the feature extraction and information fusion capabilities of the YOLOv7 algorithm to address challenges in large-scale fire areas and complex forest backgrounds. Firstly, we introduced networks such as ConvNeXtV2 and Conv2Former into the structure of YOLOv7 to find the best-performing network model. Then, we improved the ELAN structure in the backbone network using attention mechanisms and proposed the ELAN-CBAM structure. Based on the comparison of experimental results, we proposed a CNTCB-YOLOv7 Forest fire detection method. Compared with the YOLOv7 algorithm, the CNTCB-YOLOv7 algorithm achieved a 2.39% improvement in accuracy, and the recall rate and AP were improved by 0.73% and 1.14%, respectively. Additionally, the parameter count of CNTCB-YOLOv7 decreased by 3.47 M compared with

the YOLOv7 algorithm, reducing the utilization of computational resources and helping to improve the model's inference speed.

Our future work includes refining the model with additional attention mechanisms, conducting thorough comparisons, and expanding the evaluation criteria. We aim to validate its performance on diverse datasets, optimize inference speed for real-time monitoring of forest fires, and explore its potential to support risk prediction models for better forest management.

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

Funding: This work was supported in part by the Start-up Fund for New Talented Researchers of Nanjing Vocational University of Industry Technology (Grant No. YK22-05-01) and the Open Foundation of Industrial Software Engineering Technology Research and Development Center of Jiangsu Education Department (201050621ZK007) and the Humanities and Social Science Fund of Ministry of Education for the project "Method of Link Prediction in Signed Networks" (Grant No. 17YJAZH071).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Tiemann, A.; Ring, I. Towards ecosystem service assessment: Developing biophysical indicators for forest ecosystem services. *Ecol. Indic.* **2022**, *137*, 108704. [CrossRef]
2. Seidl, R.; Turner, M.G. Post-disturbance reorganization of forest ecosystems in a changing world. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2202190119. [CrossRef]
3. Sorge, S.; Mann, C.; Schleyer, C.; Loft, L.; Spacek, M.; Hernández-Morcillo, M.; Kluvankova, T. Understanding dynamics of forest ecosystem services governance: A socio-ecological-technical-analytical framework. *Ecosyst. Serv.* **2022**, *55*, 101427. [CrossRef]
4. O'Connor, A.; Audretsch, D. Regional entrepreneurial ecosystems: Learning from forest ecosystems. *Small Bus. Econ.* **2023**, *60*, 1051–1079. [CrossRef]
5. Chowdary, V.; Gupta, M.K. Automatic forest fire detection and monitoring techniques: A survey. In *Intelligent Communication, Control and Devices: Proceedings of ICICCD 2017*; Springer: Singapore, 2018; pp. 1111–1117.
6. Bu, F.; Gharajeh, M.S. Intelligent and vision-based fire detection systems: A survey. *Image Vis. Comput.* **2019**, *91*, 103803. [CrossRef]
7. Dhall, A.; Dhasade, A.; Nalwade, A.; Mohan Raj, V.K.; Kulkarni, V. A survey on systematic approaches in managing forest fires. *Appl. Geogr.* **2020**, *121*, 102266. [CrossRef]
8. 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**, *4*, 838. [CrossRef]
9. Amiri, T.; Banj Shafiei, A.; Erfanian, M.; Hosseinzadeh, O.; Beygi Heidarlou, H. Using forest fire experts' opinions and GIS/remote sensing techniques in locating forest fire lookout towers. *Appl. Geomat.* **2022**, *15*, 45–59. [CrossRef]
10. Kucuk, O.; Topaloglu, O.; Altunel, A.O.; Cetin, M. Visibility analysis of fire lookout towers in the Boyabat State Forest Enterprise in Turkey. *Environ. Monit. Assess.* **2017**, *189*, 329. [CrossRef] [PubMed]
11. Wang, Z.; Yang, P.; Liang, H.; Zheng, C.; Yin, J.; Tian, Y.; Cui, W. Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery. *Remote Sens.* **2022**, *14*, 45. [CrossRef]
12. Kang, Y.; Jang, E.; Im, J.; Kwon, C. A deep learning model using geostationary satellite data for forest fire detection with reduced detection latency. *GISci. Remote Sens.* **2022**, *59*, 2019–2035. [CrossRef]
13. Chowdary, V.; Deogharia, D.; Sowrabh, S.; Dubey, S. Forest fire detection system using barrier coverage in wireless sensor networks. *Mater. Today Proc.* **2022**, *64*, 1322–1327. [CrossRef]
14. Peng, Y.; Wang, Y. Real-time forest smoke detection using hand-designed features and deep learning. *Comput. Electron. Agric.* **2019**, *167*, 105029. [CrossRef]
15. Lin, J.; Lin, H.; Wang, F. A Semi-Supervised Method for Real-Time Forest Fire Detection Algorithm Based on Adaptively Spatial Feature Fusion. *Forests* **2023**, *2*, 361. [CrossRef]

16. Dong, M.; Sun, M.; Song, D.; Huang, L.; Yang, J.; Joo, Y.H. Real-time detection of wind power abnormal data based on semi-supervised learning Robust Random Cut Forest. *Energy* **2022**, *257*, 124761. [CrossRef]
17. Seydi, S.T.; Saeidi, V.; Kalantar, B.; Ueda, N.; Halin, A.A. Fire-Net: A deep learning framework for active forest fire detection. *J. Sens.* **2022**, *2022*, 8044390. [CrossRef]
18. Vipin, V. Image processing based forest fire detection. *Int. J. Emerg. Technol. Adv. Eng.* **2012**, *2*, 87–95.
19. Chen, G.; Zhou, H.; Li, Z.; Gao, Y.; Bai, D.; Xu, R.; Lin, H. Multi-Scale Forest Fire Recognition Model Based on Improved YOLOv5s. *Forests* **2023**, *2*, 315. [CrossRef]
20. Yar, H.; Khan, Z.A.; Ullah, F.U.M.; Ullah, W.; Baik, S.W. A modified YOLOv5 architecture for efficient fire detection in smart cities. *Expert Syst. Appl.* **2023**, *231*, 120465. [CrossRef]
21. Al-Smadi, Y.; Alauthman, M.; Al-Qerem, A.; Aldweesh, A.; Quaddoura, R.; Aburub, F.; Mansour, K.; Alhmiedat, T. Early Wildfire Smoke Detection Using Different YOLO Models. *Machines* **2023**, *11*, 246. [CrossRef]
22. Zhou, M.; Wu, L.; Liu, S.; Li, J. UAV forest fire detection based on lightweight YOLOv5 model. *Multimed. Tools Appl.* **2023**, 1–12. [CrossRef]
23. Dilli, B.; Suguna, M. Early Thermal Forest Fire Detection using UAV and Saliency map. In Proceedings of the 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 14–16 December 2022; pp. 1523–1528.
24. Zhang, L.; Lu, C.; Xu, H.; Chen, A.; Li, L.; Zhou, G. MMFNet: Forest Fire Smoke Detection Using Multiscale Convergence Coordinated Pyramid Network with Mixed Attention and Fast-robust NMS. *IEEE Internet Things J.* **2023**, *10*, 18168–18180. [CrossRef]
25. Jin, C.; Zheng, A.; Wu, Z.; Tong, C. Real-time fire smoke detection method combining a self-attention mechanism and radial multi-scale feature connection. *Sensors* **2023**, *23*, 3358. [CrossRef]
26. Chino, D.Y.; Avalhais, L.P.; Rodrigues, J.F.; Traina, A.J. Bowfire: Detection of fire in still images by integrating pixel color and texture analysis. In Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Brazil, 26–29 August 2015; pp. 95–102.
27. Yang, S.; Wang, Y.; Wang, P.; Mu, J.; Jiao, S.; Zhao, X.; Wang, Z.; Wang, K.; Zhu, Y. Automatic Identification of Landslides Based on Deep Learning. *Appl. Sci.* **2022**, *12*, 8153. [CrossRef]
28. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
29. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2022**, arXiv:2010.16061.
30. Henderson, P.; Ferrari, V. End-to-end training of object class detectors for mean average precision. In Proceedings of the 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 198–213.
31. Xue, Q.; Lin, H.; Wang, F. FCDM: An Improved Forest Fire Classification and Detection Model Based on YOLOv5. *Forests* **2022**, *13*, 2129. [CrossRef]
32. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
33. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
34. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
35. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. *arXiv* **2023**, arXiv:2301.00808.
36. Hou, Q.; Lu, C.Z.; Cheng, M.M.; Feng, J. Conv2Former: A Simple Transformer-Style ConvNet for Visual Recognition. *arXiv* **2022**, arXiv:2211.11943.
37. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. *arXiv* **2021**, arXiv:2111.12419.
38. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
39. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
40. 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, Virtual, 18–24 July 2021; pp. 11863–11874.
41. Xue, X.; Jin, S.; An, F.; Zhang, H.; Fan, J.; Eichhorn, M.P.; Jin, C; Chen, B; Jiang, L; Yun, T. Shortwave radiation calculation for forest plots using airborne LiDAR data and computer graphics. *Plant Phenom.* **2022**, *2022*, 9856739. [CrossRef]
42. Jiang, K.; Chen, L.; Wang, X.; An, F.; Zhang, H.; Yun, T. Simulation on Different Patterns of Mobile Laser Scanning with Extended Application on Solar Beam Illumination for Forest Plot. *Forests* **2022**, *13*, 2139. [CrossRef]