Research and Implementation of Forest Fire Detection Algorithm Improvement

Xi Zhou
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 2680620694@qq.com

Changyuan Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: cyw@163.com

***Abstract*—To overcome low efficiency and accuracy of existing forest fire detection algorithms, this paper proposes a network model to enhance the real-time and robustness of detection. This structure is based on the YOLOv5 target detection algorithm and combines the backbone network with The feature extraction module combines the attention module dsCBAM improved by depth-separable convolution, and replaces the loss function CIoU of the original model with a VariFocal loss function that is more suitable for the imbalanced characteristics of positive and negative samples in the forest fire data set. Experiments were conducted on a self-made and public forest fire data set. The accuracy and recall rate of the model can reach 87.1% and 81.6%, which are 7.40% and 3.20% higher than the original model, and the number of images processed per second reaches 64 frames, a growth rate of 8.47%. At the same time, this model was compared horizontally with other improved methods. The accuracy, recall rate and processing speed were all improved in the range of 3% to 10%. The effectiveness of the improved method in this article was verified, and the external perception level of the forest fire scene was deeper.**

Keywords-Fire Target Detection; YOLOv5; CBAM; Depth Separable Convolution; VariFocal Loss

I. INTRODUCTION

Forests are one of the most important ecosystems on Earth. They provide species diversity and support the survival and reproduction of a wide variety of plants, animals and insects. It plays an irreplaceable special role in the balance of the earth's ecosystem, climate regulation, resource protection, and economic development. If humans cherish and protect forests and sustainably use forest resources, forests will give back to humans with long-term ecological and economic value. Fire occurs artificially or naturally in the forest. It will spread uncontrollably and gradually develop into a disaster. It will not only cause immeasurable permanent damage to various resources and properties in the forest but also cause immeasurable permanent damage to humans and other people living in the surrounding area. A huge threat to the life safety of living things.

The type of fire is inseparable from the firefighting strategy. The focus of urban building fires is on controlling the fire and rescuing trapped people. Due to the speed of its spread, the breadth of its scope, and the huge difference between firefighting resource supply and firefighting demand, firefighting in forest scenes focuses on early detection and prevention of fires. Rapid detection of flame signs will be an important measure to prevent forest fires and respond to existing fires. However, because there are many types of fire scenes and the internal conditions are complex, relying solely on fire rescue personnel to screen fire scenes with harsh conditions has many uncertain and limiting factors. Therefore, we can use the camera equipment carried by individual firefighters or obtain information about the fire scene through other channels. The situation is transmitted back to the fire command headquarters for processing, allowing for a further comprehensive and in-depth understanding of the fire situation. Fire scene information perception and interaction are the foundation and premise of firefighting and are directly related to the depth and breadth of digital applications in fire scenes. Once the knowledge and understanding of fire scene information is lost, the fire command department will lose the ability to coordinate and

plan firefighting operations from a high position when a large fire occurs. Therefore, this paper applies digitization to forest fire scenes, utilizing computers to analyze and process forest fire images, thereby reducing the manual analysis workload in firefighting activities. This approach enables timely and efficient detection of fires in the early stages for prompt alarm and response. It also fulfills the requirement of promptly locating and initiating targeted firefighting actions during the development of forest fires. This enhances the external perception of internal conditions at the fire scene. Furthermore, based on the situation at the fire scene, it facilitates task-driven scheduling assistance for individual soldier cooperation, thereby achieving the requirement for organized, efficient, and rational firefighting and disaster relief operations.

II. RELATED WORK

All Fire, generated and spread by humans or nature in forests, can become a disaster [1]. It not only causes incalculable permanent damage to various resources and properties in the forest, but also poses a huge threat to the safety of human and other living beings living in the surrounding areas. This article applies digitalization to forest fire scenes, using computers to analyze and process forest fire scene images, reducing the workload of manual analysis in firefighting behavior, enabling timely and efficient detection in the early stages of fires [2], and providing timely alarms and responses, strengthening the external perception ability of the fire scene for internal conditions, and achieving the requirements of organizing and commanding orderly, efficient, and reasonable firefighting and disaster relief work.

Traditional forest fire detection relies on image processing techniques of classical computer vision to analyze and process the features of flame targets in images, including extracting edge features [3], texture instability and similarity analysis [4], foreground features [5], background modeling [6], and flame color analysis [7]. These classic algorithms have good detection performance, but there are drawbacks such as poor generalization ability [8] and slow detection speed [9]. Based on the characteristics of different fire scenarios, people choose different classic network models and develop various improvement plans for them. Reference [10] improves the GMM algorithm by fusing texture and similarity feature information of different colors in the image, but its learning ability for nonlinear change features is limited. Using depthwise separable convolution and CBAM to form a depthwise separable attention module, a new semantic segmentation network is formed [11], and the fusion multi-scale improved FRCNN [12] method results in slower model processing speed. Building deep neural networks to learn data representation and feature extraction has gradually become a trend in researching fire detection. YOLO has become one of the most optimal object detection algorithms at present. There are improvement options to choose YOLOv3, combined with the CAM [13], or in the detection output module, the improved K-means algorithm optimizes the prior box [14], or adds a variable convolution module [15]. The stability of model accuracy is greatly affected by the environment. On the basis of YOLOv4 network, there are methods such as color enhancement [16], introduction of attention mechanism and residual structure [17], which cannot reduce false positives in certain situations. In YOLOv5, the Neck module was introduced into the weighted bidirectional feature pyramid network [18] to replace the original path aggregation network, transfer learning [19] was adopted to train the model, and the Focal loss function [20] was introduced. The SPP structure was changed by a better performing SPPF structure [21], but the processing speed of the model still cannot meet the timeliness requirements of forest fire detection.

In response to the above issues, this article puts forward an improved fire detection model based on YOLOv5. This model will introduce the attention module dsCBAM, which replaces ordinary convolutions with depthwise separable convolutions, into the backbone network responsible for feature extraction in the YOLOv5 algorithm. This will improve the inference speed of the model and significantly improve its convergence speed. At the same time, the model has advantages in both representation ability in regions of interest and detection robustness in diverse environments.

III. Design of Fire Detection Model

A. Datasets

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

Figure 1. Part of the image data used for training: (a) Fire with poor resolution. (b) Fire in a small area. (c) Fire with flame obstruction. (d) Fire disturbed by smoke.

Since forest fire images cannot be collected and reproduced through experiments, the scene image data are public forest fire scene image data crawled on the Internet, and are collected and simulated by some enterprises and related research institutes to create publicly available data sets. Fire images are used to assist. Taking into account the diversity of real fire situations, images will contain different scenes, weather conditions, and fire intensity. Due to different image data acquisition channels and shooting conditions, image data may encounter several different recognition difficulties as shown in Figure1, such as poor image resolution, small fire range, flame obstruction, smoke interference, etc. The dataset applied to the model in this article is manually selected to filter out images that are not suitable for training in individual extreme cases. Then use labelimg software to label the flame area in VOC data set format using suitable image data.

B. YOLOv5s

YOLO is a classic target detection algorithm known for its efficient real-time detection. In fire detection tasks with high time-efficiency requirements, the YOLO algorithm can meet the need for rapid identification. Currently, the more mature version of YOLO is YOLOv5, and YOLOv5 is divided into s, m, l, and x, according to the complexity of the model. This article will conduct research and discussion based on the lightest YOLOv 5s 6.0 version.

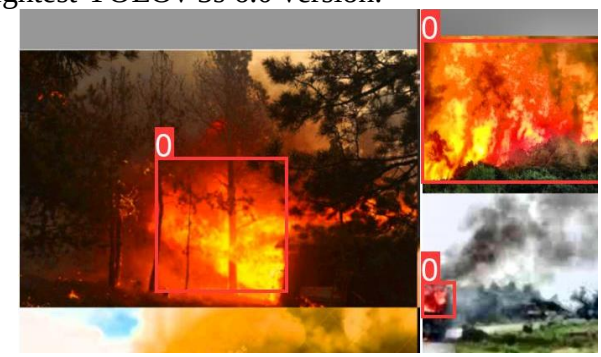


Figure 2. YOLOv 5 input data enhancement method

On the data preprocessing part, YOLOv5s follows the Mosaic method. As shown in Figure 2, it will apply random scaling, random cropping, and random arrangement to the input images for splicing. The processed data is more accurate in detection. Sexuality and discernment abilities are enhanced. Secondly, the difference from 4 version is that the former backbone network apply the single CSP. Two tiny altering structure of CSP are applied in the 5th version. The backbone part adopts CSP1_X and neck part adopts CSP2_X. Normally, the backbone part of YOLOv5s

implements by CBL, C3, and SPPF modules to stack in the Neck part. This can help YOLOv5 better handle targets of different sizes, improve network feature fusion capability, and improve detection performance. Finally, YOLOv5's head output network calculates box and class probability of the target. The head network consists of multiple convolutional layers. It uses threshold filtering and NMS to obtain the final detection result and output the target prediction result. Different from the YOLO algorithm proposed in the previous sequence, YOLOv5 adopts a more lightweight convolution structure, which reduces the amount of calculation and maintains good accuracy.

C. VariFocal Loss

The computing method measures the disparity between the label predicted by the neural network and the expected true label to a certain extent. A good loss function will have a positive impact on the training process and final results of the neural network. YOLOv5 uses a loss function called CIoU (Complete Intersection over Union) to optimize the target detection task. The CIoU loss function takes into account the degree of overlap between the predicted boundary and the real bounding box and optimizes the positioning of the target more accurately. Generally speaking, in practice, the target to be detected in the training image data, that is, the positive sample, only accounts for a small part of the image, especially image data such as a forest fire scene that contains many small flame targets, and most of the area is the background. , constitute the negative samples during training; this will lead to a large number of negative samples in the training data, while the positive samples will account for a relatively small proportion, and the training effect of the model will become worse. Normally, background class negative samples are generally easy-to-separate samples, while target class positive samples are difficult-to-separate samples. As shown in (1), in order to solve the problem of uneven distribution of the two samples, Focal Loss adds weight factors to the samples that α are difficult to separate and those that are easy to separate, increasing the weight of the difficult-to-separate samples and reducing the weight of the easy-to-separate samples, thereby controlling the positive The problem of too large gap between negative samples. Among them, in (1) α Represents balanced weight, $(1-\mathrm{p})^{\gamma}$ is a regulatory factor, γ is an adjustable focusing parameter. Therefore, Focal Loss is suitable for detecting image data of dense targets, and has good effects on data sets with characteristics such as small size, crowding, and occlusion.

$$FL(p,y)=\begin{cases} -\alpha(1-\mathrm{p})^{\gamma}\log(p), y=1 \\ -(1-\mathrm{p})^{\gamma}\log(1-p), others \end{cases} \quad (1)$$

VariFocal Loss is proposed on the basis of Focal Loss, because Focal Loss processes positive and negative samples in a balanced manner, while VariFocal loss only reduces the loss contribution of negative samples without reducing the weight of positive samples in the same way. As shown in (2), the main improvement of VariFocal Loss lies in the introduction of parameter controlled weights for target classification loss, where p is the predicted value of IoU aware classification score (IACS), α and γ is an adjustable scaling factor, and q is a positive sample growth parameter. When it is a negative background sample, $q=0$; When it is the target positive sample, q is equal to the IoU between the generated bbox and the annotation box at that point. In the traditional CIoU loss function, the weights for target classification loss and target localization loss are fixed. VariFocal Loss introduces α and γ parameters to adaptively adjust the loss weights based on the difficulty level of different samples [22]. When the sample is more challenging, larger values of α and γ increase the weight of target classification loss, emphasizing classification accuracy. When the sample is less challenging, smaller values of α and γ decrease the weight of target classification loss, prioritizing localization accuracy and dynamically adjusting the weight of target classification loss. When the sample difficulty is greater, α the value of sum is larger, the weight of the target classification loss increases, and more attention is paid to the accuracy of the classification; when the sample

difficulty is low, the value of sum is small α, γ. The γ weight of the target classification loss decreases, and more attention is paid to the accuracy of the classification Positioning accuracy, dynamically adjust the weight of the target classification loss. By introducing VariFocal Loss, the target detection model can better balance the trade-off between target classification and target positioning, thereby improving the performance of target detection. VariFocal Loss has been applied in some target detection algorithms.

$$VFL(p,y) = \begin{cases} -q\left(q\log(p)+(1-q)\log(1-p)\right), q>0 \\ -\alpha p^{\gamma}\log(1-p), q=0 \end{cases} \quad (2)$$

D. CBAM

In the process of development, the attention mechanism derives various types of attention. According to different classifications of attention, such as multi-scale attention, contextual attention, parallel branch attention, channel attention, spatial attention, etc. Depending on the size of the attention scale, there are several more outstanding models, such as Transformer, SE, CBAM [23], and so on.

CBAM is one of the new lightweight attention modules used to enhance convolutional neural networks. Quantitative attention model convolutional block model. It does not directly calculate the attention map, but separates it, learns attention from channel and spatial respectively, adaptively learns the channel correlation and spatial importance of the input feature map, and simultaneously takes advantage of both to improve the accuracy of feature extraction. Accuracy. Considering the difficulty of fire scene transmission and the portability requirements of individual firefighters, the pixels of fire scene images are generally very low. When it is difficult to apply computer-related algorithms for identification and detection, false alarms and missed detections will occur. The CBAM module, as is shown in Figure 3, inputs the feature map $F,F\in R^{C\times H\times W}$, processes it into a one-bit feature map through the channel attention module $F',F'\in R^{C\times 1\times 1}$, and then uses the spatial attention module to generate a two-dimensional spatial attention map $F'',F''\in R^{1\times H\times W}$.

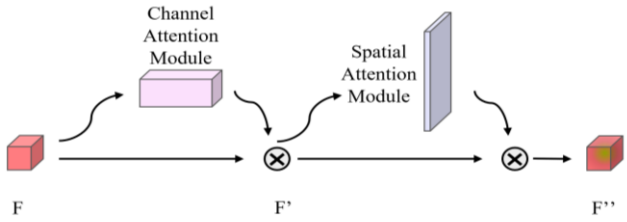


Figure 3. CBAM overall structure

The CAM part , in Figure 4, performs global average pooling and global maximum pooling operations on the input feature layer, and connects the two pooling results using a shared multi-layer perceptron. By performing a weighting operation on the original input feature layer channel-by-channel multiplication, the feature information of different levels of the upper-level output feature map can be extracted. The CAM adopts the global average pooling and global maximum pooling serial structures and combines the results of the two pooling methods to achieve compressed spatial dimensions of the input feature map, with stronger representational power.

The SAM structure, in Figure 5, focuses on which part of the input image information is more significant and is a complement to the CAM in the previous part. To calculate spatial attention, first apply average pooling and maximum pooling along the channel direction of each feature point, stack and aggregate them to generate the channel information of a feature map, and generate two two-bit feature maps. Spatial attention obtains the global maximum feature in the spatial dimension by performing global maximum pooling in the channel dimension and learns the weight of each spatial position through two fully connected layers. In this way, the model can automatically learn the importance of each spatial location, that is, which spatial locations are more important for target positioning.

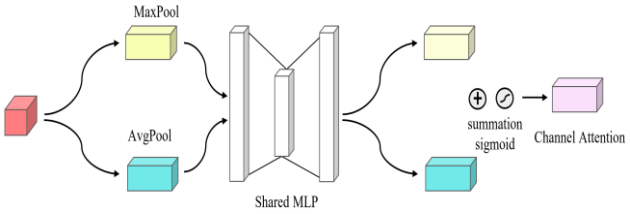


Figure 4. CAM structure

In the forest fire detection task, there may be different key channels for different fire types and backgrounds. Channel attention can help the model adaptively select channel information suitable for the current task, thereby reducing the interference of irrelevant information and improving feature representation. effectiveness. Fires usually appear at specific locations in images, and spatial attention can help the model focus on these important spatial locations and improve target positioning accuracy. By combining channel attention and spatial attention, the CBAM module enables the model to pay more attention to important channel information and spatial position information in the feature extraction stage, thereby enhancing the model's perceptual ability. In the forest fire detection task, CBAM can help the model better understand the correlation and importance of the input feature map, improve the model's detection and positioning capabilities of forest fire targets, and thereby improve the performance and robustness of the forest fire detection system.

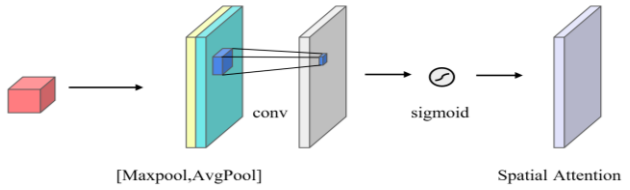


Figure 5. SAM structure

E. Lightweight Convolution

Ordinary convolution is shown in Figure 6. Assume that the number of input channels is M, the size is $D_F \times D_F$, the number of output channels is N, the convolution kernel size is $D_K \times D_K$, and the bias term is ignored b. Then, the amount of calculation required for this convolution operation is

$$Q_c = D_K \times D_K \times M \times N \times D_F \times D_F \quad (3)$$

, the required parameters are shown in (4).

$$P_c = D_K \times D_K \times M \times N \quad (4)$$

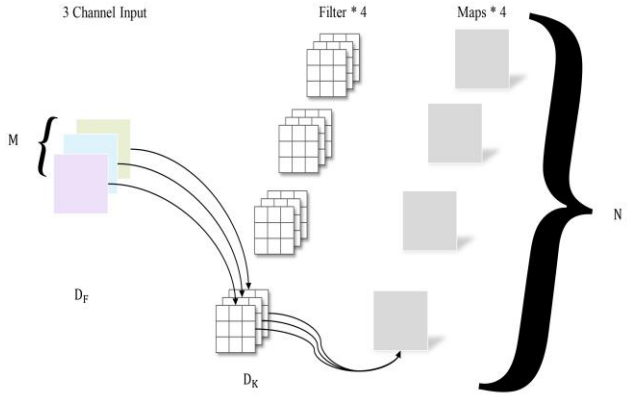
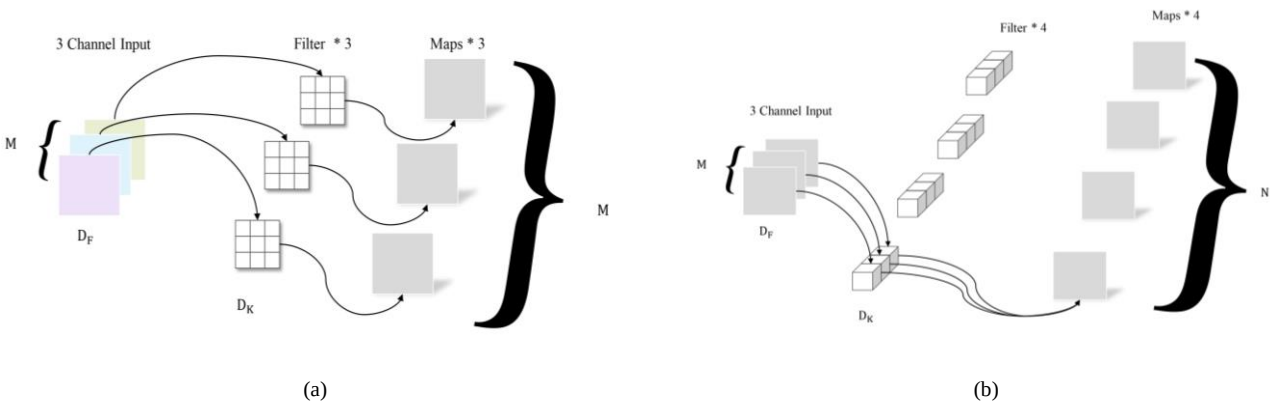


Figure 6. Ordinary convolution

The input feature map of the convolution is divided into g groups, each convolution kernel is also divided into groups accordingly, and the convolution operation is performed in the corresponding group. Each set of convolutions generates one feature map, and a total of g feature maps are generated. The number of groups g is like a control knob. The minimum value is 1, and $g = 1$ the convolution at this time is ordinary convolution; the maximum value is the number of channels of the input feature map C, and $g = C$ the convolution at this time is depth separation convolution, also called channel-by-channel convolution.



(a) (b)

Figure 7. Depth separation convolution: (a) Depth convolution. (b) Pointwise convolution

In other words, depthwise separative convolution is a special form of grouped convolution, where the number of groups is the number of channels of the feature map. That is, each feature map is divided into a group, and convolution is performed within the group. A convolution kernel in the group generates a feature map. This convolutional form is the most efficient form of convolution. Compared with ordinary convolution, multiple feature maps can be generated with the same amount of parameters and calculations, while ordinary convolution can only generate one feature map. Pointwise convolution is just 1×1 an ordinary convolution. Because depth convolution does not integrate inter-channel information, it needs to be used in conjunction with point-by-point convolution. The operation of point-wise convolution is very similar to the conventional convolution operation. The size of its convolution kernel is $1\times1\times M$, M which is the number of channels of the previous layer. Therefore, the convolution operation here will weightedly combine the feature maps of the previous step in the depth direction to generate a new feature map.

Depth separable convolution is equivalent to Figure 7(a). Assuming that this convolution and the ordinary convolution above face the same feature weighting task, the corresponding calculation amount of the depth convolution is

$$Q_{dw} = D_K \times D_K \times M \times D_F \times D_F \tag{5}$$

The parameter quantity is

$$P_{dw} = M \times D_K \times D_K . \tag{6}$$

The corresponding calculation amount of pointwise convolution, as is shown in figure 7(b), is

$$Q_{pw} = 1 \times M \times N \times D_F \times D_F \tag{7}$$

The parameter quantity is

$$P_{pw} = M \times N \times 1 \tag{8}$$

Then the total calculation amount and parameter amount of the combined depth-separable convolution are the sum of the two, and the calculation amount is

$$\begin{aligned} Q_{ds} &= Q_{dw} + Q_{pw} \\ &= D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F, \end{aligned} \tag{9}$$

The parameter quantity is

$$\begin{aligned} P_{ds} &= P_{dw} + P_{pw} \\ &= D_K \times D_K \times M + M \times N \times 1. \end{aligned} \tag{10}$$

Compute the calculation amount and calculation parameters of depthwise separable convolution and ordinary convolution, that is

$$Q_{ds} / Q_c = 1/N + 1/D_k^2 \tag{11}$$

$$P_{ds} / P_c = 1/N + 1/D_k^2 \quad (12)$$

From (11) (12), it can be seen that the calculation amount and parameter amount of the former are $1/N + 1/D_k^2$ times that of the latter. It shows that modified convolution reduces the required parameters and has reference significance in lightweight advanced models. In order to further lightweight the model, the convolution operation in the CBAM module can substitute modified convolution for the initial convolution method, which is referred to as dsCBAM in this article.

F. Improved Forest Fire Detection Algorithm

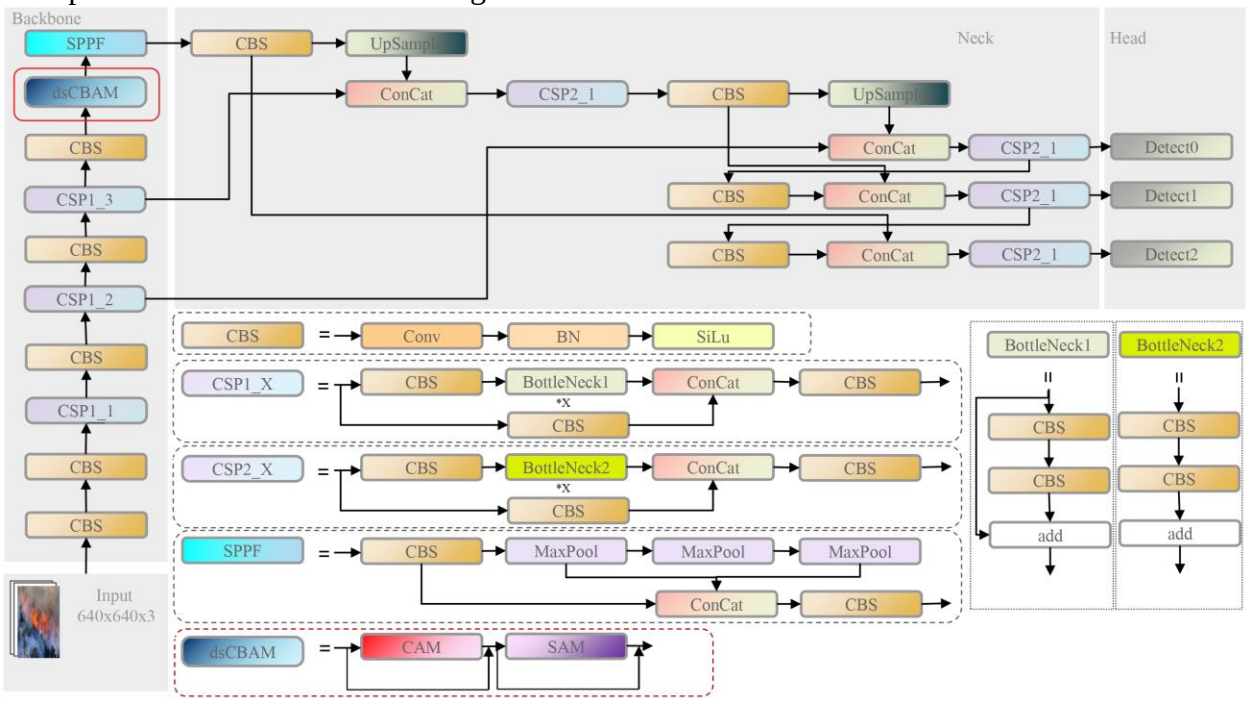


Figure 8. Improved model framework

The forest fire detection framework of this article is shown in Figure 8, which is an improvement based on the YOLOv5s model. In response to the real-time requirements of forest fires, the first step is to improve the lightweight attention model and replace the convolutions in the CBAM module with depth-separable convolutions. Targets usually occupy a small proportion of the screen, which may cause sample imbalance. Replacing the initial loss function with the mentioned earlier function can improve performance for this problem. The third step is to add the lightweight CBAM model to YOLOv5s to enhance the robustness of the forest fire detection system.

The dsCBAM module can adaptively adjust the channel and spatial information through the CAM and SAM, which helps the YOLO network better understand the target structure and contextual relationships in the image and enhance its ability to perceive fire fields. As shown in Figure 9, this article adds the CBAM module to YOLO to replace the last CSP 1_1 module in the original model. Replacing the CSP1_1 module with the CBAM module will enhance the model's ability to perceive targets at different scales, directions, and angles, thereby improving detection accuracy. The introduction of the CBAM module may help reduce noise or redundant information inside the prediction frame, make the target boundary clearer, enhance feature extraction and representation capabilities, and thus help improve detection quality. The CBAM module makes the structure of the backbone network richer and more diverse, making the network more robust to changes and disturbances in the input image, thereby increasing the generalization performance of the model.

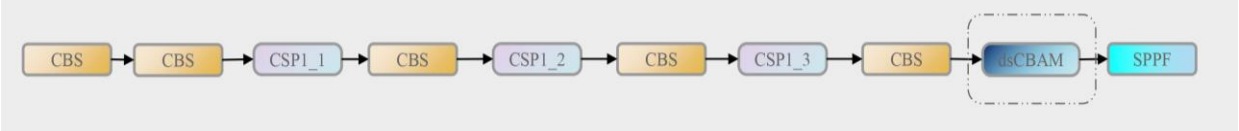


Figure 9. The position of CBAM in YOLOv5s 6.0 version

IV. RESULTS AND DISCUSSION

A. Training

TABLE I. DATASET SETTINGS

Dataset	Training	Test	Validation	Total
Homemade forest fire data set	1442	617	617	2676
Other institutes data set	600	200	200	1000

TABLE II. EXPERIMENTAL SETTINGS

Lab Environment	Detail
programming language	Python3.8.5
operating system	Windows 10
deep learning framework	Pytorch 1.8.0
GPU	4x NVIDIA TITIAN V

The data set needs to be manually screened one by one and the flame targets in it need to be labeled. As shown in Table 1. According to the needs of the experiment, the collected forest fire image data set was divided into a training set, a test set, and a verification set in a ratio of 6:2:2 to carry out experiments on the model.

TABLE III. TRAINING SETTINGS PARAMETERS

Training parameters	Detail
Epochs	100
Batch-size	16
Image-size	6 40 ×640
Initial learning rate	0.01
Optimization algorithm	SGD

The experimental settings of this experiment are shown in Table 2. It shows some parameter settings during the training process of the experiment. In this experiment, the training optimization algorithm uses the default stochastic gradient descent method (SGD). During training, the adaptive moment estimation (Adam) optimization algorithm can be selected according to the actual situation.

B. Model Evaluation

The evaluation index of the public data set Microsoft COCO is recognized as effective and state-of-the-art in the field of object detection. It is used in this article to evaluate the performance of the proposed improved forest fire detection algorithm. The five indicators of P, R, AP, mAP and FPS will be expanded below.

$$P = TP / (TP+FP) \tag{13}$$

P (Precision) refers to the ratio of correctly detected targets to all detection results in (15). Among them, TP (true positive) represents the predicted correct box. The boxes predicted by the model are calculated one by one with the labeled boxes of the image. R (Recall) refers to the proportion of truly detected targets to all real targets in (14).

$$R = TP/(TP+FN) \tag{14}$$

$$AP = \int_0^1 P(r)\,dr \tag{15}$$

AP (Average Precision) essentially describes the performance of the model on a single category. In the multi-category target detection task, each category has an AP value. The metric provide specific numerical values to measure the algorithm's prediction accuracy and target detection capabilities.

$$FPS = 1000 / (pre_{process} + inf + NMS) \tag{16}$$

$pre_{process}$ refers to the preprocessing time for converting the input image into the format required by the algorithm, including image aspect ratio scaling, padding, normalization and other operation times. Value inf refers to the inference time, that is, the forward pass calculation time from inputting the image into the model to the

model output result after preprocessing. NMS It can be understood that post-processing time is mainly the time spent on converting the model output results and other operations. The sum of the three is the total time of image processing. After calculation by formula (16), FPS (Frame Per Second) is obtained. If tested and compared in the same hardware environment, the lightweight effect of the algorithm can be expressed to a certain extent.

C. Ablation Experiment

TableTABLE IV. presents the results of this experiment. The experiments were evaluated separately on the same data set, and eight solutions were compared horizontally, namely (1) original YOLOv5s model; (2) combination of CBAM and YOLOv5s model; (3) combination of SE and YOLOv5s model; (4) Combining ECA with YOLOv5s model; (5) Improving the model combining CBAM with YOLOv5s; (6) Improving the model combining CBAM with Alpha-IoU and YOLOv5s; (7) Improving the model combining CBAM with SIOU and YOLOv5s; (8) Improving CBAM with VariFocal Loss Combined model with YOLOv5s. They show that improvement methods are effective from the four indicators of accuracy P, recall rate R and frames per second (FPS). After adding the CBAM model, the original model's recognition accuracy of flame targets in forest fires has been slightly improved, and the model's ability to perceive flame targets has been further enhanced. By introducing depthwise separable convolution, the speed of data processing of the model is improved, and the degree of lightweight and portability of the model is deepened. Finally, through ablation experiments to compare the three loss functions of Alpha-IoU, SioU, and VariFocal, the loss function proposed in this article was selected as the loss function with the best performance, which verified the importance of suppressing negative samples in improving the performance of the target recognition algorithm. All in all, compared with the traditional YOLO algorithm, the improved model has achieved significant performance improvements in both the difficult detection task of small target detection and the speed of detection.

TABLE IV. COMPARATIVE TEST RESULTS OF THE MODEL

Model	P	R	FPS
YOLOv5s	0.811	0.786	59
YOLOv5s + CBAM	0.814	0.790	60
YOLOv5s + SE	0.810	0.787	5 8
YOLOv5s + ECA	0.812	0.791	5 9
YOLOv5s + dsCBAM	0.812	0.787	62
YOLOv5s + dsCBAM + Alpha-IoU	0.821	0.813	61
YOLOv5s + dsCBAM + SIoU	0.860	0.834	60
YOLOv5s + dsCBAM+ VariFocal （Ours）	0.871	0.816	64

D. Comparision

Figure 10 shows the cpmparative results of the original and the improved. By comparing the initial net and the proposed net to detect four groups of images, the detection results can more intuitively and objectively show that the improved model has better performance. The fire targets in the first set of images can be accurately detected, but the confidence of the improved model is significantly improved. There are three flame targets in the second set of images. The original model can only detect the two larger targets, while the improved model can detect all targets. The flames in the third group of images were blocked to a certain extent by foreground objects and could not be identified by the original model. The improved model accurately identified its location. The proportion of flames in the last set of images is relatively small, and the improved model solves the problem that the original model cannot detect. It can be observed that the improved model does not miss small fire targets and can detect fires more accurately even when the image quality and size are not very high Figures 11 present the robustness experiment, the initial net misclassified forest night lights as flames, while the improved model's attention mechanism enhanced the feature learning of the detection targets, thereby improving the occurrence of false detections.

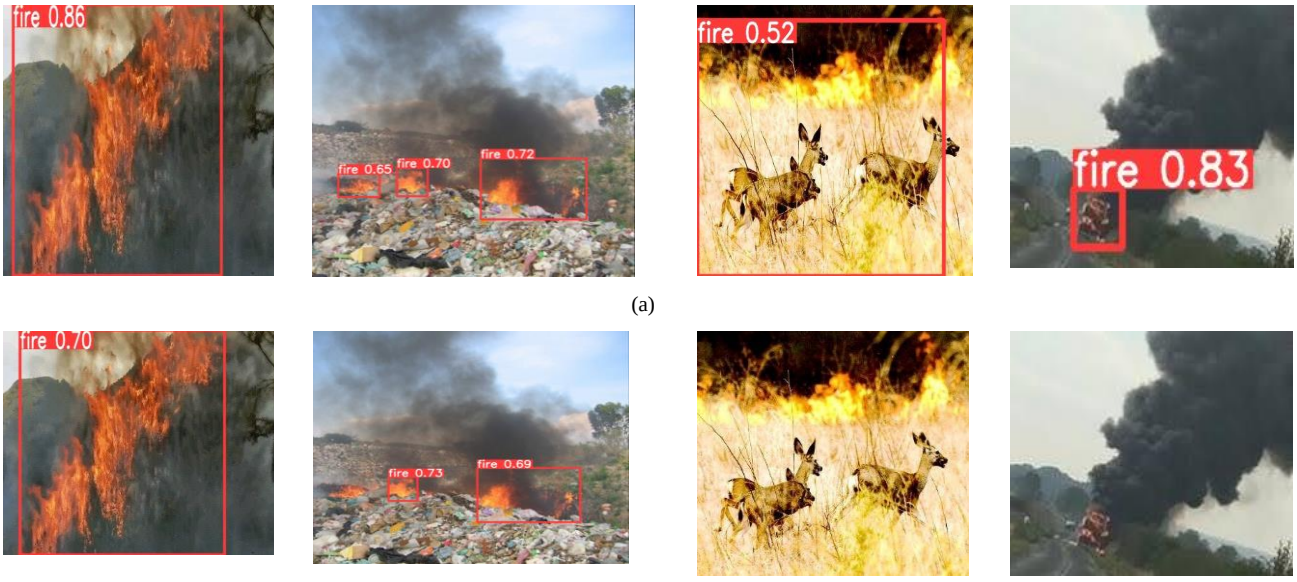
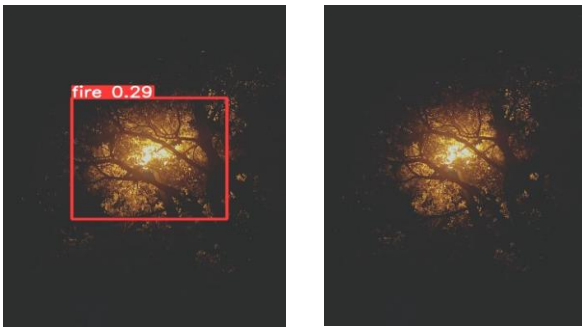


Figure 10. Experimental identification results: (a) Improved model. (b) Original model.

By analyzing and comparing the experimental performance index data in Table 4, we can intuitively observe the robustness of all aspects of the model. Through the comparison of Experiments 1 to 4, we can see that after adding three different attention mechanisms: CBAM, SE, and ECA to the original YOLOv5s model, the results are affected to varying degrees. Among them, As to YOLOv5s combined with SE, the accuracy and FPS of the model have declined, while the recall rate has slightly improved. After the introduction of YOLOv5s in the ECA attention mechanism, the three indicators of P, R have slightly improved, and the model processing speed has almost no change. After the assistance of CBAM, P, R improved more significantly than ECA, and the growth of R was even better, indicating that it can detect more targets and reduce the missed detection problem. The processing speed is also not very high. Significant improvement. From the comparison, CBAM has the best improvement in focusing on small targets and detecting speed and is more in line with the requirements of diverse detection environments. The comparison between Experiment 2 and Experiment 5 directly shows the performance of applying the modified convolution. The experimental accuracy and recall rate result data show that the replacement strategy can shorten the processing time of each image based on ensuring the accuracy of the model, making the lightweight features of the model more prominent. Experiments 5-8 respectively completed the training, verification, and detection tasks of forest fire images by applying the improved CBAM and four different loss functions of the original loss function, Alpha-IoU, SIoU, and VariFocal. Comparing the experimental results of these four different loss functions on the training task, we can observe that the first replacement loss function has a small range of growth in the three indicators measuring model training, but the addition also affects the processing time of the model. Compared with the experimental performance of the SIOU loss and the loss used in Experiments 5 and 6, the training accuracy on the data set has been significantly improved, but this also makes the model pay the price of processing speed. Experiment 8 is the experimental data based on the improvement points proposed in this article. It has good adaptability to changes in input images, lighting conditions, occlusions, etc., and also takes into account the processing speed of the model, so that performance and efficiency are balanced to a certain extent.



(a) (b)

Figure 11. Experimental Experimental results of misdetection of forest street lights at night. (a) Original model. (b) Improved model.

E. Discussion

In this study, we apply CBAM, VariFocal, depthwise separable convolution, and YOLOv5s to the forest fire detection task. By comparing experimental results, we observe that the model achieves significant performance improvements on multiple metrics. First, the CBAM module helps improve the model's attention to key areas, allowing for better detection of features. Secondly, the VariFocal loss function introduces dynamic weight allocation. Furthermore, depthwise separable convolution reduces the computational effort and maintains model performance. Through experiments, we observed that the improved models of CBAM, VariFocal, depthwise separable convolution, and YOLOv5 are highly robust when processing forest fire images in different scenarios, and the model can accurately detect various scales, poses, and density of fire targets, and also has certain adaptability to images under different lighting conditions. Compared with other deep learning-based methods, our method achieves faster inference speed while maintaining high accuracy. Although the improved models of CBAM, VariFocal, depthwise separable convolution, and YOLOv5 achieved good results in the forest fire detection task, there are still some limitations. For example, a model may perform poorly when dealing with low-resolution or blurry images. In addition, the robustness of the model in complex scenarios still needs to be further improved. Future work may need to consider combining multi-modal data, introducing a target tracking module.

V. CONCLUSIONS

This study conducted an in-depth study on the forest fire detection task by applying improved methods of CBAM, VariFocal, depthwise separable convolution, and YOLOv5s, introduced the working principle and working method of the original model, and deeply an alyzed the principles and possible improvements of various improvements. Achievability. For forest fire detection tasks, the assistance of the CBAM structure helps to improve detection model's focus on key areas and the detection accuracy of fire targets. Its mechanism based on channel attention and spatial attention can effectively extract fire features such as flames and smoke in images. The assistance of the advanced loss function is able to overcome imbalance of sample category and present better results. This loss function uses dynamic weight allocation to make the model pay more attention to minority class samples, thereby improving detection accuracy. The application of depthwise separable convolution reduces the computational load of the model while maintaining model performance. This lightweight convolution operation helps improve the running efficiency of the model, making it more suitable for practical fire detection applications. Improvements in YOLOv5 show good performance in forest fire detection. Its fast and accurate target detection capabilities enable the model to monitor forest areas in real-time and detect the occurrence of fires promptly, thus providing the opportunity for rapid response and processing. Experiments are conducted to demonstrate the actual performance of various improvement ideas. Experiments on forest fire detection tasks have proven that the improved method in this article effectively enhances the perception ability of the original YOLO model and achieves good results. The accuracy rate is improved by 0.06 based on the original model, and the number of frames processed per second is 3 frames has been added, which greatly improves the accuracy and efficiency of forest fire detection. The results of this study provide an important reference and foundation for further research and development in the field of forest fire detection. By improving existing models and technologies, we can improve our monitoring and early warning

capabilities for forest fires, thereby reducing the harm of fires to the environment and humans. Future work can explore more deep learning methods and technologies, integrate multi-source data, and enhance the robustness and real-time performance of the algorithm to further advance the development of forest fire detection technology. In summary, the results of this study provide useful exploration for research and practical applications in the field of forest fire detection and demonstrate the potential of CBAM, VariFocal, depthwise separable convolution, and YOLOv5 improved models in fire detection. It is hoped that this article can provide guidance for the prevention and control of forest fires and reduce the harm of fires to the natural environment and human society.

REFERENCES

[1] Liao Shujiang. A preliminary study on the trend of fire spread [J]. Fire Science and Technology, 2012, 31(7): 670-673.

[2] Wang M. Risk Information, Risk Perception and Fire Prevention Behavior [D]. University of Science and Technology of China, 2017.

[3] Lv P T, Li J, Wu L Y et al. Research on automatic edge detection of fire video images [J]. Applied Science. 2003.

[4] YAN Yunyang, GAO Shangbing, GUO Zhibo, et al. Automatic fire detection based on video images [J]. Computer Application Research, 2008, 25(4): 1075-1078.

[5] TAN Yong, XIE Linbai, FENG Hongwei, et al. Image-based flame detection algorithm [J]. Laser & Optoelectronics Progress, 2019, 56(16): 161012.

[6] CUI Bingcheng, CHENG Naiwei, ZHAO Peng. Exploration of smoke image detection method based on matlab [J]. Science and Technology Innovation. 2019, (28).

[7] Gong F, Li C, Gong W, et al. A real-time fire detection method from video with multifeature fusion [J]. Computational intelligence and neuroscience, 2019, 2019.

[8] Li P, Zhao W. Image fire detection algorithms based on convolutional neural networks [J]. Case Studies in Thermal Engineering, 2020, 19: 100625.

[9] Saeed F, Paul A, Karthigaikumar P, et al. Convolutional neural network based early fire detection [J]. Multimedia Tools and Applications, 2020, 79: 9083-9099.

[10] ZHANG Chi, MENG Qinghao, WELL Tao. Video flame detection algorithm based on improved GMM and multi-feature fusion [J]. Laser & Optoelectronics Progress, 2021, 58(4): 0410006.

[11] Jing K, Jia Y, Zhang C, et al. MobileAttentionNet: An Efficient Network for Semantic Segmentation of Forest Fire Images [C]//2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT). IEEE, 2021: 377-380.

[12] Zhang L, Wang M, Ding Y, et al. MS-FRCNN: A Multi-Scale Faster RCNN Model for Small Target Forest Fire Detection [J]. Forests, 2023, 14(3): 616.

[13] Zhang X, Qian K, Jing K, et al. Fire detection based on convolutional neural networks with channel attention [C]//2020 Chinese Automation Congress (CAC). IEEE, 2020: 3080-3085.

[14] ZHAO Yuanyuan, ZHU Jun, XIE Yakun, et al. Improved Yolo-v3 algorithm for real-time flame detection in video images [J]. Journal of Wuhan University (Information Science Edition), 2021, 46(3): 326-334.

[15] DING Hao, WANG Huiqin, WANG Ke. Improved YOLOv3 flame detection algorithm based on dynamic shape feature extraction and enhancement [J]. Laser & Optoelectronics Progress, 2022, 59(24):2410003-2410003-9.

[16] Avazov K, Mukhiddinov M, Makhmudov F, et al. Fire detection method in smart city environments using a deep-learning-based approach [J]. Electronics, 2021, 11(1): 73.

[17] Sun J, Ge H, Zhang Z. AS-YOLO: An improved YOLOv4 based on attention mechanism and SqueezeNet for person detection [C]//2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2021, 5: 1451-1456.

[18] Xue Q, Lin H, Wang F. Fcdm: an improved forest fire classification and detection model based on yolov5 [J]. Forests, 2022, 13(12): 2129.

[19] Xue Z, Lin H, Wang F. A small target forest fire detection model based on YOLOv5 improvement[J]. Forests, 2022, 13(8): 1332.

[20] Yang T, Xu S, Li W, et al. A smoke and flame detection method using an improved yolov5 algorithm [C]//2022 IEEE International Conference on Real-time Computing and Robotics (RCAR). IEEE, 2022: 366-371.

[21] Yang X, Wang Z, He Y, et al. Research on open flame recognition algorithm in construction site based on attention mechanism [C]//2023 15th International Conference on Advanced Computational Intelligence (ICACI). IEEE, 2023: 1-6.

[22] Zhang H, Wang Y, Dayoub F, et al. VariFocalnet: An iou-aware dense object detector [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8514-8523.

[23] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.