

Article

A Smoke Detection Model Based on Improved YOLOv5

Zhong Wang , Lei Wu, Tong Li * and Peibei Shi

School of Computer Science and Technology, Hefei Normal University, Hefei 230601, China; zhongw@ustc.edu.cn (Z.W.); wangz@hfnu.edu.cn (L.W.); pbshi@hfnu.edu.cn (P.S.)

* Correspondence: litong@hfnu.edu.cn

Abstract: Fast and accurate smoke detection is very important for reducing fire damage. Due to the complexity and changeable nature of smoke scenes, existing smoke detection technology has the problems of a low detection rate and a high false negative rate, and the robustness and generalization ability of the algorithms are not high. Therefore, this paper proposes a smoke detection model based on the improved YOLOv5. First, a large number of real smoke and synthetic smoke images were collected to form a dataset. Different loss functions (GIoU, DIoU, CIoU) were used on three different models of YOLOv5 (YOLOv5s, YOLOv5m, YOLOv5l), and YOLOv5m was used as the baseline model. Then, because of the problem of small numbers of smoke training samples, the mosaic enhancement method was used to randomly crop, scale and arrange nine images to form new images. To solve the problem of inaccurate anchor box prior information in YOLOv5, a dynamic anchor box mechanism is proposed. An anchor box was generated for the training dataset through the k-means++ clustering algorithm. The dynamic anchor box module was added to the model, and the size and position of the anchor box were dynamically updated in the network training process. Aiming at the problem of unbalanced feature maps in different scales of YOLOv5, an attention mechanism is proposed to improve the network detection performance by adding channel attention and spatial attention to the original network structure. Compared with the traditional deep learning algorithm, the detection performance of the improved algorithm in this paper was 4.4% higher than the mAP of the baseline model, and the detection speed reached 85 FPS, which is obviously better and can meet engineering application requirements.

Keywords: smoke detection; YOLOv5; dynamic anchor; attention mechanism; loss function

MSC: 68T45



Citation: Wang, Z.; Wu, L.; Li, T.; Shi, P. A Smoke Detection Model Based on Improved YOLOv5. *Mathematics* **2022**, *10*, 1190. <https://doi.org/10.3390/math10071190>

Academic Editor: Cornelio Yáñez Márquez

Received: 12 March 2022

Accepted: 3 April 2022

Published: 6 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Smoke detection is a key link in fire prevention and can play an important role in reducing disaster losses [1,2]. Traditional fire detection technology includes contact-type fire detectors such as temperature and smoke, and these are used in public areas such as shopping malls and subway stations. However, this detection method has shortcomings such as a limited detection range and delayed alarm time. With the development of deep learning technology, video-based fire detection technology has been widely used. As a noncontact fire detection technology, this method has the advantages of fast response speed, a wide detection range, and low hardware cost. Video fire detection technology can be divided into flame detection and smoke detection, according to the detection object. Since smoke appears earlier than flame and is not easy to cover in the initial stage of fire, video-based fire detection technology focuses on detecting smoke.

Although smoke detection technology has been widely used, it is difficult for the accuracy and robustness of existing technology to meet needs for popularization and application in complex smoke scenarios. Due to the small number of fire events and limited real smoke image samples, deep learning-based smoke detection methods are prone to overfitting on limited training samples. In addition, smoke detection scenes are complex and changeable,

especially in indoor and outdoor scenes, and the distribution of topographic features is greatly affected by combustion factors and environmental conditions. The traditional smoke detection algorithm is difficult to apply to all kinds of complex scenes, and the robustness and generalization ability of the algorithm are insufficient.

Aiming at the problems of existing smoke detection technology, this paper proposes a smoke detection model based on the improved YOLOv5 [3]. The steps of the proposed model are as follows:

1. The mosaic enhancement method is used to randomly crop, scale and arrange nine pictures to form a new picture;
2. A dynamic anchor box mechanism is proposed, and anchor boxes are generated for the training dataset through the k-means++ clustering algorithm;
3. A hybrid attention mechanism is proposed by adding channel attention and spatial attention to the original network structure;
4. A large-scale fire dataset is constructed, including 4815 real smoke images and 20,000 synthetic images, and different loss functions are used on multiple models of YOLOv5 for experimental testing.

The remainder of this paper is arranged as follows: Section 2 introduces work related to smoke detection. Section 3 focuses on describing the framework and implementation details of the smoke detection model. Section 4 verifies the performance of the proposed method through experimental tests and, finally, gives a summary and outlook towards future applications.

2. Related Works

Traditional smoke detection technology includes feature extraction and classifier design, and the core research lies in smoke feature research. Common smoke features include color, texture, motion, ability, and other features, as well as feature combinations. Researchers have designed smoke color models through color space to improve the discrimination between smoke and nonsmoke in color space. For example, Miranda et al. [4] counted the pixel value distribution of smoke images based on various color models and analyzed the discrimination of color components of various models through relief features. Yuan et al. [5] used Fisher linear discrimination to build a smoke color model and used the color space conversion matrix to distinguish the color value distribution of smoke and nonsmoke. Distinguishing color features and texture features focuses on the contrast relationship between pixels. Yuan et al. [6] proposed a new smoke texture feature using a high-order local ternary mode by encoding high-order directional derivatives in each pixel and connecting all joint histograms from different orders. Dimitropoulos et al. [7] proposed a new high-order linear dynamic system descriptor based on multidimensional dynamic texture analysis, which analyzes dynamic textures by using the information from various image elements and applies it to smoke detection. In terms of motion features, Zhou et al. [8] used the maximum stable extreme value region detection method to extract the local extreme value region of smoke and detect the propagation motion of potential smoke regions based on a new cumulative region method, which can effectively identify the unique expansion and rising motion of smoke. Lin et al. [9] used volume local binary patterns to extract dynamic texture features and proposed a method to directly extract dynamic texture features based on irregular moving regions. Chen et al. [10] proposed the concept of contrast images for false alarms caused by wavelet minute-minute pure color interferers and judged whether a moving area was smoke by calculating the high-frequency and low-frequency ratio.

Traditional smoke detection technology tries to mine the features of smoke to distinguish other interfering substances and performs smoke detection by manually setting the smoke feature, but the detection rate and false alarm rate have difficulty meeting the application requirements. With the application of deep learning techniques in the field of computer vision, researchers [11,12] have used deep convolutional neural networks for smoke detection, which can learn deeper feature models. Hohberg [13] used firewatch

products to collect a large number of smoke pictures and used CaffeNet and GoogLeNet two-dimensional convolutional neural networks for smoke detection. Three-dimensional convolutional neural networks were used to extract the temporal and spatial information of smoke. Frizzi et al. [14] used a 6-layer convolutional neural network to address the three-classification problem of fire, smoke and nonfire, and used a sliding window on the feature map of the last pooling layer to select suspicious regions, which improved the processing speed. Kim et al. [15] used faster-RCNN for flame detection. Zhang et al. [16] solved the problem of insufficient sample data by inserting real smoke pictures into the forest background and using faster-RCNN to detect wildland forest fire smoke. Filoneko et al. [17,18] used classic convolutional neural networks such as AlexNet, VGG-16, VGG-19, and ResNet to conduct experimental comparisons on four smoke image databases. Sharma et al. [19] used fire images as an imbalanced dataset and used VGG and ResNet networks for fire image classification. Li et al. [20] proposed a 3D, parallel, fully convolutional neural network to segment smoke regions in video sequences, aiming at the problem of false positives in smoke detection. Yin et al. [21] first used deep convolutional motion spatial networks to capture spatial and motion context information and then used temporal pooling layers and recurrent convolutional neural networks to train a smoke model. Yuan et al. [22] proposed a 14-layer convolutional neural network DNCNN for smoke image classification, using a combination of normalization layers and convolutional layers to accelerate the training of the network model. In addition, to further reduce the problem of model overfitting caused by insufficient training samples, they further proposed a new deep multiscale network model. First, a multiscale basic combination layer structure was designed, and then a network model was constructed with multiple basic combination layer structures. Muhammad et al. [23] proposed an energy-saving edge intelligent assisted smoke detection method using deep convolutional neural networks for foggy monitoring environments. Xu et al. [24] proposed a smoke detection method based on a deep saliency network, which uses the circular convolution structure to build a pixel-level saliency detection network, and uses the fused features for saliency reasoning. Li et al. [25] proposed a reconstructed convolutional neural network, which extracts suspicious smoke regions through smoke region proposal, prunes and reconstructs the convolutional neural network model, which can improve the accuracy and efficiency of smoke detection tasks. Liu et al. [26] proposed a two-stage smoke detection method that uses block depth normalization and a convolutional neural network to detect suspicious smoke areas and then uses an SVM classifier to classify suspected areas to reduce false positives. Although smoke detection technology based on deep learning has achieved good results, the field of smoke detection lacks a large sample database, and the smoke scene is complex and changeable. Therefore, this paper proposes a new deep learning algorithm, from the aspects of dataset construction and algorithm improvement.

3. Methodology

3.1. YOLOv5

YOLOv5 is the latest generation target detection network of the YOLO series, which was proposed by Ultralytics in May 2020. It has many characteristics, such as being fast, accurate and lightweight. There are the following four network models in YOLOv5: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Among them, the YOLOv5 network depth and feature map width are the smallest, and the other three models are continuously deepened and widened, based on YOLOv5.

The YOLOv5 network structure includes the following four parts: input, backbone, neck and prediction, as shown in Figure 1.

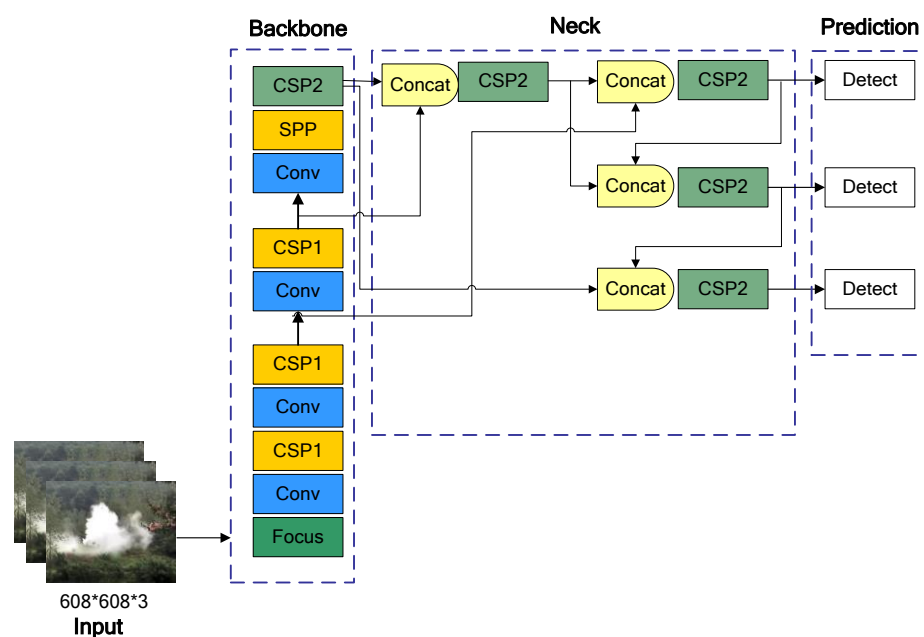


Figure 1. YOLOv5 network structure.

1. The input includes mosaic data enhancement, image size processing and adaptive anchor box calculation. Mosaic data enhancement uses a combination of 4 photos to enrich data diversity, and the use of adaptive anchor box calculation is beneficial to improve the detection speed.
2. The backbone includes focus and CSP [27] modules. The slicing operation in the focus module reduces the number of computations and improves the speed, while realizing downsampling. The CSP module is beneficial for improving the network learning ability and reducing the memory cost.
3. The core of the neck is to use FPN (feature pyramid network) and PAN (path aggregation networks) structures. The FPN and PAN structures realize the fusion and complementation of high-level features and low-level features. The FPN is top-down and uses upsampling to transfer and fuse the new type to obtain the predicted feature map. The PAN adopts a bottom-up feature pyramid. The two complement each other and overcome their respective limitations to strengthen the feature extraction capability of the model.
4. Prediction includes the bounding box loss function and nonmaximum suppression (NMS). YOLOv5 uses GIoU as the loss function, which effectively solves the problem of nonoverlapping bounding boxes. In the target detection prediction result processing stage, the optimal target frame is obtained by using the weighted NMS operation for the screening of numerous target frames.

3.2. Algorithm Optimization

3.2.1. Mosaic-9 Data Augmentation

The traditional YOLOv5 method uses four images, which are randomly cropped and scaled and then randomly arranged and spliced into one image. In this paper, an enhanced version of the mosaic method is used, that is, nine images are randomly combined. First, we take a batch of data from the dataset, randomly take nine images from it each time, cut and stitch at random positions to synthesize new pictures, repeat the batch size times to obtain batch size images with enhanced mosaic data, and then fed them to the neural network for training, as shown in Figure 2.

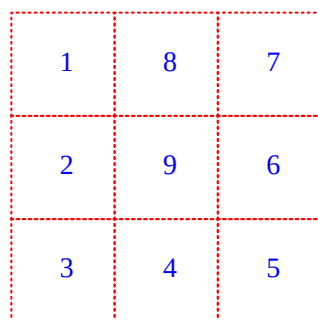


Figure 2. Mosaic-9 data augmentation.

3.2.2. Dynamic Anchor Learning

YOLOv5 uses k-means clustering to generate anchor boxes. Taking the boundary of the training set as the benchmark, three anchor boxes are set under three different sizes of feature maps through the FPN network. The k-means algorithm has some limitations, and the algorithm is affected by the initial setting value and outliers, which leads to the instability of clustering results. Therefore, this paper uses the k-means++ clustering algorithm to cluster the marked target bounding anchor boxes in the smoke dataset many times, to generate a priori boxes of different numbers and sizes to increase the difference between the a priori box and the actual target box as much as possible, and therefore to improve the detection accuracy.

Since the initial anchor box is artificially set and the size of the smoke target in the smoke detection environment is an arbitrary shape, it causes a large loss in the subsequent bounding box regression training and increases the difficulty of model optimization. Therefore, this paper adopts the idea of dynamic anchor feature selection [28], which is proposed based on the ARM module. First, nine anchor boxes are clustered on three different scales by the k-means++ algorithm as the initial anchor box of the model. Then, the ARM module is added to the model to filter out the negative samples of the initial anchor box and the position of the initial anchor box is fine-tuned according to the real value. Then, the ARM module is connected with the YOLOv5 backbone network through bidirectional feature fusion (BFF), the corresponding points on the feature map are fine-tuned, and finally the dynamic refinement anchor frame is obtained as the model training prior frame, as shown in Figure 3. The dynamic anchor frame structure is added as the generation method of the prior frame to better extract the network feature structure, and the dynamic anchor frame has self-learning ability and can obtain parameters suitable for the network through iteration.

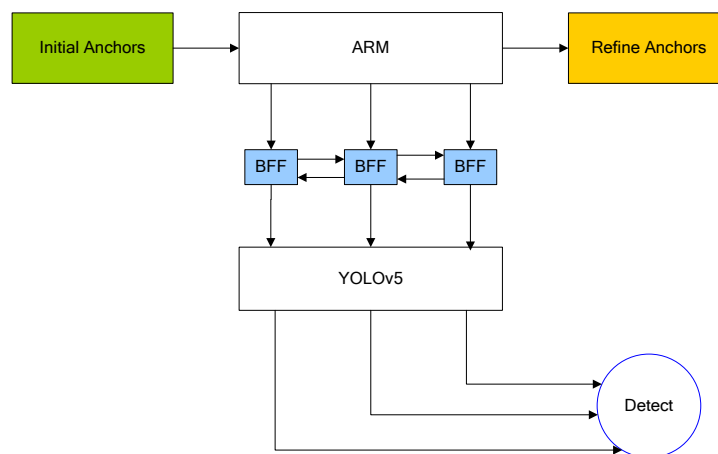


Figure 3. Dynamic anchor learning.

3.2.3. Attention Mechanism

Channel attention: After the image is convolved, the number of channels increases exponentially. Detecting important regions in each channel is a central part of the attention model. After the convolution operation, the image is sent to the channel attention model for weight calculation, different weights are assigned according to the importance of the channel, and the weight is proportional to the importance.

Spatial attention: Not all regions in the sample contribute equally to the detection task; only the part of the region that meets the task requirements needs the algorithm to focus on detection, and the spatial attention model finds such regions in the sample for processing. Dynamic capacity networks use two subnetworks to achieve this task. The low-performance subnetwork is responsible for finding areas that focus on processing, while the high-performance subnetwork is responsible for refining the areas found by the low-performance subnetwork to obtain lower cost and higher accuracy.

By fusing channel attention and spatial attention, a new attention model, the channel-spatial fusion attention model, is proposed. This model reflects the advantages of channel attention and the spatial attention model in the same model and judges the important areas of the image from the two aspects of space and channel. The overall network structure follows the structure of YOLOv5 and uses CSPdarknet53 as the backbone network of the model for training. By adding channel attention after the “Conv+BN+LeakyReLU” module, “CSP×n+A” is added to the model after the convolution operation of the feature maps of three different scales, as shown in Figure 4. Through the mixed use of two kinds of attention, the ability to extract important features of the model is improved to obtain better results.

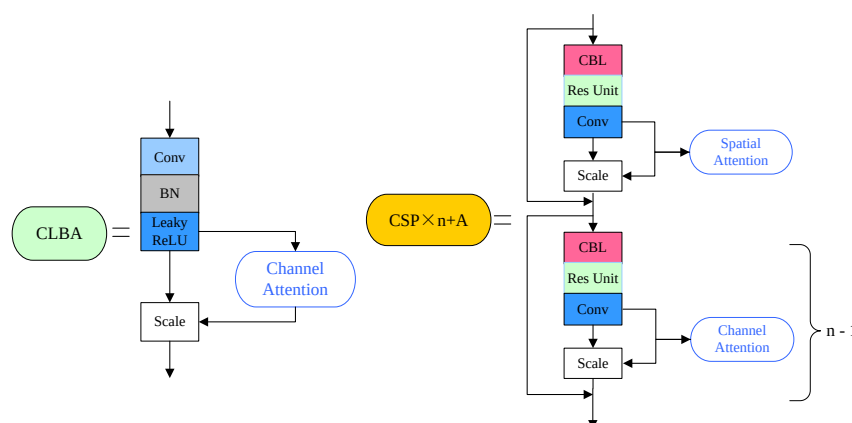


Figure 4. Attention mechanism.

3.3. Boundary Loss Function

The loss function of the target detection task consists of two parts, classification loss and bounding box regression loss. The most commonly used bounding box regression loss is IoU and its improved algorithm. The full name of the IoU algorithm is the intersection and union ratio, which is obtained by calculating the ratio of the intersection and union of the predicted bounding box and the ground-truth bounding box, that is, $\text{IoU}(A, B) = (A \cap B) / (A \cup B)$, where A is the predicted bounding box and B is the ground-truth bounding box. IoU can be used as a distance; then, $\text{Loss}_{\text{IoU}} = 1 - \text{IoU}$. The advantage of IoU is that it can reflect the detection effect of the predicted bounding box and the ground-truth bounding box. In addition, IoU has the characteristics of nonnegativity, scale invariance, identity, symmetry, and triangle inequality. IoU has these two disadvantages:

- When the prediction bounding box and the ground-truth bounding box do not intersect, and $\text{IoU}(A, B) = 0$, the distance between A and B cannot be reflected. At this time, the loss function is not steerable, and IoU loss cannot optimize the situation where the two bounding boxes do not intersect.

- Assuming that the sizes of the prediction bounding box and the ground-truth bounding box are determined, as long as the intersection value of the two boxes is determined and their IoU values are the same, the IoU value cannot reflect how the two boxes intersect.

1. GIoU

$$\text{GIoU}(A, B) = \text{IoU}(A, B) - \frac{|C| - |A \cup B|}{|C|} \quad (1)$$

$$\text{Loss}_{\text{GIoU}} = 1 - \text{GIoU}(A, B) = 1 - \text{IoU}(A, B) + \frac{|C| - |A \cup B|}{|C|} \quad (2)$$

where C is the smallest box containing A and B . GIoU overcomes the disjoint situation of A and B in IoU and has the same scale-invariance properties as IoU. GIoU can be considered the lower bound of IoU, which is less than or equal to IoU. GIoU not only pays attention to the overlapping area but also to other nonoverlapping areas, which can better reflect the degree of overlap between the two. When the two boxes have no overlapping area, GIoU is a number that varies in the range of $[-1, 0]$, and there is a gradient, so optimization can be performed. The optimization direction gradually narrows the distance between the two boxes.

2. DIoU

$$\text{DIoU} = \text{IoU} - \frac{p^2(b, b^{\text{gt}})}{c^2} \quad (3)$$

$$\text{Loss}_{\text{DIoU}} = 1 - \text{DIoU} = 1 - \text{IoU} + \frac{p^2(b, b^{\text{gt}})}{c^2} \quad (4)$$

where b and b^{gt} represent the center points of prediction Box B and ground-truth box B^{gt} , respectively, c represents the square of the diagonal length of the minimum bounding Box C , and p represents the calculation of the Euclidean distance between the two center points. DIoU can directly minimize the distance between two target boxes, so it converges much faster than GIoU. For the case containing two boxes in the horizontal and vertical directions, the DIoU loss can make the regression very fast, while the GIoU loss almost degenerates to the IoU loss.

3. CIoU

$$\text{IoU} = \text{IoU} - \left(\frac{p^2(b, b^{\text{gt}})}{c^2} + \alpha v \right) \quad (5)$$

$$\text{Loss}_{\text{CIoU}} = 1 - \text{IoU} + \frac{p^2(b, b^{\text{gt}})}{c^2} + \alpha v \quad (6)$$

where α is the weight parameter and v is used to measure the similarity of the aspect ratio.

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan \frac{w}{h} \right)^2 \quad (7)$$

$$\alpha = \frac{v}{(1 - \text{IoU}) + v} \quad (8)$$

It can be seen from the definition of α that the loss function will be more inclined to optimize in the direction of increasing overlapping areas, especially when the IoU is zero.

3.4. The Pseudocode of the Proposed Algorithm

We give the pseudocode of the improved method, as shown in Algorithm 1, including parameter initialization, data augmentation, dynamic anchor learning, attention mechanism, and loss function.

Algorithm 1. The proposed algorithm.**Input:** training set**Output:** deep learning model

1. **Initialize:** initialize model parameters, including weights, epochs, batchsize, img_size, etc.;
2. **Data augmentation:** 9 images are randomly combined with Mosaic-9;
3. **Dynamic anchor learning:** The K-means++ algorithm is used to generate the initial anchor, and the ARM module is connected to the backbone network through BF;
4. **Attention mechanism:** CSPdarknet53 is used as the backbone network for training, and a hybrid attention mechanism is added;
5. **Loss Function:** CIoU loss function is used to optimize the regression of boundary box parameters;

4. Experimental Results**4.1. Dataset and Environment**

There is currently no authoritative dataset similar to ImageNet for smoke detection. The datasets used in this paper include real smoke images and synthetic smoke images. The real smoke pictures come from public datasets and network pictures, including 4815 real smoke pictures and 20,000 synthetic smoke pictures. The real smoke pictures come from pictures obtained from sources such as indoor monitoring, outdoor monitoring, field monitoring, field monitoring towers, drone shooting and networks. Since taking smoke pictures is limited by the scene and experimental conditions, it is difficult to simulate fire smoke indoors. Therefore, we generated large-scale smoke synthesis pictures to improve the data size and sample differences of the dataset. We utilized a synthesis system to generate a rich variety of 20,000 synthetic smoke images by extracting the entire rectangular smoke region in the smoke video. Additionally, we collected the same number of nonsmoke images as these smoke images. We divided the dataset containing real smoke images and synthetic smoke images into training and test sets, as shown in Figure 5.



Figure 5. Sample images from the dataset.

The experimental environment of this paper was completed on the Ubuntu18.04 operating system. The GPU model is an NVIDIA GeForce RTX3080Ti, the CPU model is an Intel(R) Xeon CPU E3-1231 v3@3.40 GHz, and the software environment is CUDA10.02 and PyTorch 1.7.

4.2. Model Evaluation

In this paper, precision, recall, average precision (AP), and mean average precision (mAP) are used as model precision evaluation indicators, where AP represents the area under the PR curve, and mAP represents the mean value of AP for each category. The specific formula is as follows:

$$P = \frac{TP}{TP + FP} \quad (9)$$

$$R = \frac{TP}{TP + FN} \quad (10)$$

$$AP = \frac{\sum P}{\text{Num}(\text{objects})} \quad (11)$$

$$mAP = \frac{\sum AP}{\text{Num}(\text{class})} \quad (12)$$

where TP is the number of correct classes predicted to be correct, FN is the number of correct classes predicted to be negative, and FP is the number of negative classes predicted to be correct.

4.3. Experimental Results

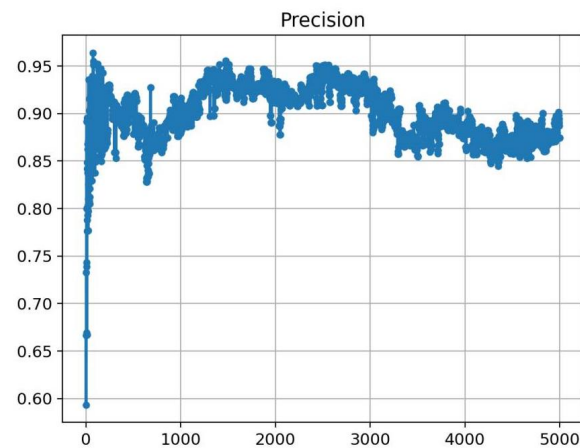
In the network model training phase, the iteration batch size was set to 128, the decay coefficient was 0.0005, the initial learning rate was 0.001, and the total number of iterations was 5000, where iteration batch size represents the number of samples selected for each training, decay represents that the weight is reduced in each iteration to prevent overfitting, learning rate determines the speed at which the parameters reach the optimal value, and the total number of iterations determines the maximum number of model training.

The paper first gives the experimental results of different YOLOv5 models under different loss functions (GIoU, DIoU, CIoU), as shown in Table 1. As seen in Table 1, the detection performance of different models under different loss functions was significantly different. YOLOv5s had the best precision performance on DIoU, and YOLOv5 m had the best recall and mAP performance on CIoU. Overall, YOLOv5 m performed the best, with 2% higher mAP on CIoU and approximately 5% higher mAP on DIoU than other models, while YOLOv5s and YOLOv5 l had similar detection performance on different loss functions. For mAP of different models, CIoU performed the best, followed by GIoU, and DIoU had the worst performance. Although DIoU takes into account the shortcomings of GIoU, the aspect ratio of the bounding box is not considered in the regression process, and its accuracy was lower than that of GIoU. Since CIoU increases the loss of the detection box, the regression accuracy is improved on the basis of DIoU, and the prediction box is closer to the ground truth box. In addition, the detection speeds of YOLOv5s, YOLOv5 m, and YOLOv5 l 132, 102, and 80, respectively.

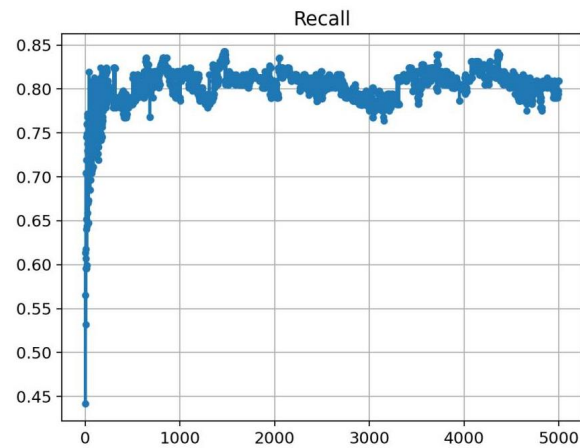
Table 1. Experimental results with different models and loss functions.

Model	Evaluation	GIoU	DIoU	CIoU
YOLOV5s	Precision	0.9621	0.9729	0.9692
YOLOV5m		0.9503	0.9469	0.9499
YOLOV5l		0.946	0.9229	0.942
YOLOV5s	Recall	0.7602	0.6742	0.7079
YOLOV5m		0.7228	0.734	0.7828
YOLOV5l		0.7266	0.6742	0.7302
YOLOV5s	mAP	0.8475	0.7842	0.8503
YOLOV5m		0.8433	0.842	0.874
YOLOV5l		0.8394	0.7933	0.8588

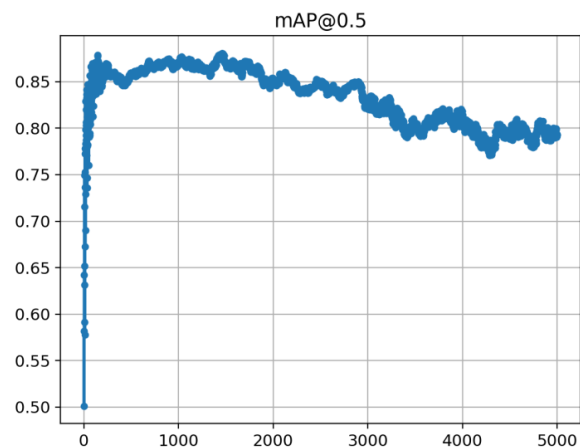
Therefore, the paper uses YOLOv5 m as the benchmark for optimization, which can meet the requirements of real-time detection of engineering applications. CSPDarknet53 is used as the backbone network. The specific training results are shown in Figure 6.



(a)



(b)



(c)

Figure 6. The training results of YOLOv5. (a) Precision. (b) Recall. (c) mAP.

To further verify the robustness of the proposed method, the paper used data augmentation, a dynamic anchor box and an attention mechanism to conduct independent experiments based on YOLOv5 m, and the results are shown in Table 2. Table 2 also gives the traditional two-stage algorithm Faster R-CNN [29], one-stage algorithm SSD [30], and YOLOv3 as comparative experiments and uses mAP and FPS to evaluate and compare each mainstream algorithm. It can be seen in Table 2 that, compared with the traditional

faster-RCNN, SSD, and YOLOv3, the YOLOv5 m benchmark model has obvious advantages, and FPS is superior. After adding data enhancement, a dynamic anchor box and an attention mechanism to the benchmark YOLOv5 m, the optimized model mAP increases by 0.8%, 2.2%, and 2.4%, respectively. Finally, the mAP of our optimized model is 4.4% higher than that of the baseline model, and the detection speed FPS meets the real-time requirements of engineering applications.

Table 2. Experimental results with different models and loss functions.

Model	Backbone	mAP	FPS
Faster R-CNN	Resnet101	83.2%	58
SSD	VGG16	76.8%	67
YOLOv3	Darknet53	82.3%	68
YOLOv5m	CSPDarknet53	87.4%	102
YOLOv5m + Mosaic-9	CSPDarknet53	88.2%	96
YOLOv5m + DA	CSPDarknet53	89.6%	92
YOLOv5m + Attention	CSPDarknet53	89.8%	94
YOLOv5m + Mosaic-9 + DA + Attention	CSPDarknet53	91.8%	85

Figure 7 shows the smoke image detection results of the proposed model in different scenarios, including indoor, outdoor, and wild scenes, and gives the confidence of each sub image detection results. It can be seen from the figure that there are three sub images (4, 8, 12) with a confidence level of 0.8, in which the targets of sub-image 4 and sub-image 12 have a certain similarity with the background, and the target range of sub-image 8 is large and the characteristics of the inner area of the smoke are not obvious. There are two sub-images (9, 11) with a confidence level of 1.0, and their smoke characteristics are more obvious. The confidence levels of the rest of the sub-images are all 0.9, while the sub-image 13 is far away and the target features are not obvious, and the target features of sub-image 15 are greatly affected by the weather. It can be seen in the figure that the model proposed in the paper can correctly detect smoke pictures in different scenarios, and the false alarm rate is low, which also reflects the stability and robustness of the proposed model. It should be noted that the Chinese characters in Figure 7 are not important information to others.



Figure 7. Detection results with the proposed model.

5. Conclusions

Considering the current difficulties in smoke detection, this paper built a large-scale smoke detection dataset with real smoke pictures and synthetic smoke pictures and utilized a large number of different YOLOv5 models (YOLOv5s, YOLOv5 m, YOLOv5 l) and different loss functions (CIoU, GIoU, DIoU). In the experiment, YOLOv5 m was finally used as the benchmark model, nine images were used for data enhancement, the initial anchor frame was generated for the training dataset through the k-means++ clustering algorithm, the ARM module was added to the model, and the ARM module was connected to the YOLOv5 backbone network through BF. By fusing channel attention and spatial attention, a channel-spatial fusion attention model was proposed. The experimental results show that the optimized model is significantly better than the traditional detection model, which is 4.4% higher than the mAP of the benchmark model YOLOv5 m, and the detection speed FPS meets the requirements of engineering applications. By improving the YOLOv5 model, this paper increased the detection ability of the model, improved the robustness of the model, and achieved accurate and real-time smoke detection. The next step will be to deploy the performance improvement meter of the model on embedded devices to realize real-time smoke detection in field environments.

Author Contributions: Conceptualization, Z.W.; methodology, Z.W.; software, L.W.; validation, T.L.; formal analysis, P.S.; investigation, P.S.; resources, T.L.; data curation, T.L.; writing—original draft preparation, T.L.; writing—review and editing, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 61976198), Natural Science Research Key Project for Colleges and University of Anhui Province (No. KJ2019A0726), High-level Scientific Research Foundation for the Introduction of Talent of Hefei Normal University (No. 2020RCJJ44), and the Scientific Research Project by Enterprises of Hefei Normal University (HXXM2022007).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chen, S.; Cao, Y.; Feng, X.; Lu, X. Global2Salient: Self-adaptive feature aggregation for remote sensing smoke detection. *Neurocomputing* **2021**, *466*, 202–220. [\[CrossRef\]](#)
2. Asiri, N.; Bchir, O.; Ben Ismail, M.M.; Zakariah, M.; Alotaibi, Y.A. Image-based smoke detection using feature mapping and discrimination. *Soft Comput.* **2021**, *25*, 3665–3674. [\[CrossRef\]](#)
3. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A Forest Fire Detection System Based on Ensemble Learning. *Forests* **2021**, *12*, 217. [\[CrossRef\]](#)
4. Miranda, G.; Lisboa, A.; Vieira, D.; Queiroz, F.; Nascimento, C. Color feature selection for smoke detection in videos. In Proceedings of the 2014 12th IEEE International Conference on Industrial Informatics (INDIN), Porto Alegre, Brazil, 27–30 July 2014; pp. 31–36.
5. Wei, Y.; Jie, L.; Jun, F.; Yongming, Z. Color model and method for video fire flame and smoke detection using Fisher linear discriminant. *Opt. Eng.* **2013**, *52*, 027205. [\[CrossRef\]](#)
6. Yuan, F.; Shi, J.; Xia, X.; Fang, Y.; Fang, Z.; Mei, T. High-order local ternary patterns with locality preserving projection for smoke detection and image classification. *Inf. Sci.* **2016**, *372*, 225–240. [\[CrossRef\]](#)
7. Dimitropoulos, K.; Barmoutis, P.; Grammalidis, N. Higher Order Linear Dynamical Systems for Smoke Detection in Video Surveillance Applications. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 1143–1154. [\[CrossRef\]](#)
8. Zhou, Z.; Shi, Y.; Gao, Z.; Li, S. Wildfire smoke detection based on local extremal region segmentation and surveillance. *Fire Saf. J.* **2016**, *85*, 50–58. [\[CrossRef\]](#)
9. Lin, G.; Zhang, Y.; Jia, Y.; Xu, G.; Wang, J. Smoke detection in video sequences based on dynamic texture using volume local binary patterns. *KSII Trans. Internet Inf. Syst.* **2017**, *11*, 5522–5536. [\[CrossRef\]](#)
10. Chen, J.; Wang, Y.; Tian, Y.; Huang, T. Wavelet based smoke detection method with RGB Contrast-image and shape constrain. In Proceedings of the 2013 Visual Communications and Image Processing (VCIP), Kuching, Malaysia, 17–20 November 2013; pp. 1–6.

11. Cheoi, K. An Intelligent Fire Leaning and Detection System. *J. Korea Multimed. Soc.* **2015**, *18*, 359–367. [[CrossRef](#)]
12. Kim, Y.J.; Kim, E.G. Image based fire detection using convolutional neural network. *J. Korea Inst. Inf. Commun. Eng.* **2016**, *20*, 1649–1656. [[CrossRef](#)]
13. Hohberg, S.P. Wildfire Smoke Detection Using Convolutional Neural Networks. Master's Thesis, Freie Universitat Berlin, Berlin, Germany, September 2015.
14. Frizzi, S.; Kaabi, R.; Bouchouicha, M.; Ginoux, J.M.; Moreau, E.; Fnaiech, F. Convolutional neural network for video fire and smoke detection. In Proceedings of the IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 24–27 October 2016; pp. 877–882.
15. Kim, Y.J.; Kim, E.G. Fire detection system using faster R-CNN. In Proceedings of the International Conference on Future Information & Communication Engineering, Seoul, Korea, 22–24 May 2017; Volume 9, pp. 261–264.
16. Zhang, Q.-X.; Lin, G.-H.; Zhang, Y.-M.; Xu, G.; Wang, J.-J. Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images. *Procedia Eng.* **2018**, *211*, 441–446. [[CrossRef](#)]
17. Filonenko, A.; Kurnianggoro, L.; Jo, K.H. Comparative study of modern convolutional neural networks for smoke detection on image data. In Proceedings of the 2017 10th International Conference on Human System Interactions (HSI), Ulsan, Korea, 17–19 July 2017; pp. 64–68.
18. Filonenko, A.; Kurnianggoro, L.; Jo, K.H. Smoke detection on video sequences using convolutional and recurrent neural networks. In Proceedings of the International Conference on Computational Collective Intelligence, Rhodes, Greece, 29 September–1 October 2021; Springer: Cham, Switzerland, 2017; pp. 558–566.
19. Sharma, J.; Granmo, O.C.; Goodwin, M.; Fidge, J.T. Deep convolutional neural networks for fire detection in images. In Proceedings of the International Conference on Engineering Applications of Neural Networks, Rhodes Island, Greece, 25–28 September 2015; Springer: Cham, Switzerland, 2017; pp. 183–193.
20. Li, X.; Chen, Z.; Wu, Q.M.J.; Liu, C. 3D Parallel Fully Convolutional Networks for Real-Time Video Wildfire Smoke Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *30*, 89–103. [[CrossRef](#)]
21. Yin, M.; Lang, C.; Li, Z.; Feng, S.; Wang, T. Recurrent convolutional network for video-based smoke detection. *Multimed. Tools Appl.* **2018**, *78*, 237–256. [[CrossRef](#)]
22. Yin, Z.; Wan, B.; Yuan, F.; Xia, X.; Shi, J. A Deep Normalization and Convolutional Neural Network for Image Smoke Detection. *IEEE Access* **2017**, *5*, 18429–18438. [[CrossRef](#)]
23. Muhammad, K.; Khan, S.; Palade, V.; Mehmood, I.; De Albuquerque, V.H.C. Edge Intelligence-Assisted Smoke Detection in Foggy Surveillance Environments. *IEEE Trans. Ind. Inform.* **2019**, *16*, 1067–1075. [[CrossRef](#)]
24. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, Z.; Jia, Y.; Wang, J. Video smoke detection based on deep saliency network. *Fire Saf. J.* **2019**, *105*, 277–285. [[CrossRef](#)]
25. Li, C.; Yang, B.; Ding, H.; Shi, H.; Jiang, X.; Sun, J. Real-time video-based smoke detection with high accuracy and efficiency. *Fire Saf. J.* **2020**, *117*, 103184. [[CrossRef](#)]
26. Liu, T.; Cheng, J.; Yuan, Z.; Hua, H.; Zhao, K. Video smoke detection with block DNCNN and visual change image. *KSII Trans. Internet Inf. Syst.* **2020**, *14*, 3712–3729. [[CrossRef](#)]
27. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
28. Li, S.; Yang, L.; Huang, J.; Hua, X.S.; Zhang, L. Dynamic anchor feature selection for single-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6609–6618.
29. Li, L.; Liu, F.; Ding, Y. Real-time smoke detection with Faster R-CNN. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems, Chongqing, China, 27–29 May 2022; pp. 1–5.
30. Xu, G.; Zhang, Q.; Liu, D.; Lin, G.; Wang, J.; Zhang, Y. Adversarial Adaptation From Synthesis to Reality in Fast Detector for Smoke Detection. *IEEE Access* **2019**, *7*, 29471–29483. [[CrossRef](#)]