WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n2.07
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Innovative YOLOv5s-Based Algorithm for Real-Time Tunnel Fire Detection

Yangyang Zheng [1], Baofeng Ji [1, 2], Liqin Yue [3], Xiaohui Song [4], Yifeng Liu [1]

[1] Henan University of Science and Technology, Luoyang 471023, China

[2] Longmen Laboratory, Luoyang 471000, China

[3] Huanghe University of Science and Technology, Zhengzhou 450000, China

[4] Henan Academy of Sciences, Zhengzhou 450000, China

ABSTRACT

Aiming at the special characteristics of tunnel spatial environment and the problems of confusing smoke and fire in tunnel fires with high real-time detection requirements, an improved algorithm for tunnel smoke and fire detection based on YOLOv5s is proposed. The Focus module is replaced by a convolutional layer, the number of convolutions in B_CSP is reduced, and the spatial pyramid pooling structure SPP is replaced by SPPF, which reduces the number of parameters in the network and improves the detection efficiency of the model. In order to better optimize the anchoring frame, the dynamic, non-monotonically focused Wise-IOU bounding box loss function is used to replace the GIOU, which speeds up the convergence of the network and improves the accuracy of network detection. The CA attention mechanism is incorporated at the end of the Backbone layer to dynamically adjust the importance between channels, which enables the model to better capture the flame features in the image and improve the model performance. The experiments use 2050 tunnel smoke and fire datasets as training samples, and the experimental results show that the accuracy of the improved model on the dataset is increased by 3.2 percentage points, and the network model can achieve 98.9% precision and 95.1% recall as well as 99.2% average precision by testing the network model on the dataset, and the detection speed is increased to 148 FPS. It can meet the real-time prevention and detection of tunnel fire in daytime, nighttime or poor vision, and has good accuracy and robustness.

KEYWORDS

Lightweight networks; Tunnel fire detection; Yolov5; CA attention mechanism

1. INTRODUCTION

With the implementation of major national strategies such as transportation power, new infrastructure construction and western development, China's transportation infrastructure construction has achieved unprecedented development. Since entering the 21st century, China has become the country with the largest number and longest mileage of tunnel construction scale in the world. Statistics show that there are 21999.3km of highway tunnels in China [1], of which more than 70% are long tunnels and extra-long tunnels. Highway tunnels have complex structure, narrow space, strong closure, and few entrances and exits, and fire accidents are the most hazardous type of accidents, with complex causes and uncertainty, which will destroy the internal structure of the tunnel, pose great threats to the personal and property safety of trapped people, and have a large negative impact on the community. Therefore, rapid and accurate identification of tunnel fire is very necessary, and is the focus and difficulty of tunnel fire research, while contributing to the emergency treatment and

prevention of tunnel fire accidents. Traditional fire detection uses fire detectors, mainly including smoke detectors, temperature sensors and so on. Although fire can be detected to a certain extent, there are limitations in special environments such as tunnels. For example, smoke detectors may have delayed alarms due to ventilation conditions inside tunnels; temperature sensors may not be able to respond effectively until the source of fire approaches the detector. In addition, these traditional methods often do not provide information on the exact location and size of the fire and are not effective enough for rapid and effective response and resource allocation.

In recent years, image detection techniques based on deep learning have been developing rapidly: Khan et al. use the lightweight EfficientNet convolutional neural network and DeepLabv3+ network to segment and then classify smoke images, which effectively reduces the false detection rate [2-4]; Lin et al. combine the Faster RCNN and 3DCNN models for smoke detection, which improves the accuracy of target localization and detection accuracy. Lin et al. combined the Faster RCNN and 3DCNN model for smoke detection, which improved the accuracy of target localization and detection precision, but because the generation and classification of candidate frames are carried out in two steps during the detection process, the model is more complex, the detection speed is slow, and the model generalization is poor [5 - 7]; Park et al. integrated Elastic into the Backbone module of the YOLOv3 network, and the Random Forest classifier can identify whether the target is a flame or not, and the method can detect the fire information under the complex scene effectively. This method can effectively detect the fire information in complex scenes, but the model is large and the arithmetic requirement for computational reasoning is high, which can not meet the requirements of real-time airborne equipment inspection [8]; Xie et al. based on the single-stage target detection algorithm use K-means clustering to get the anchor box suitable for smoke, avoiding the interference of extraneous information in the image, and the method improves the detection accuracy under the condition of increasing the number of parameters and decreasing the computation rate [9]; Xu et al. proposed a pixel-level and target-level algorithm to improve the detection accuracy [10]. Xu et al. proposed a pixel-level and target-level fusion of salient target detection algorithm for detecting smoke in open space, which preserves the original feature information of the image to a large extent, but the network itself has insufficient utilization of features, resulting in low average classification performance [10].

Among the YOLO series algorithms, the only ones with good detection accuracy and real-time performance are YOLOv4, YOLOv5 and YOLOX fire detection networks [11-13].YOLOv4 weight file is too large; YOLOX is optimized for YOLOv5, but the network model is more complex and the gain is not necessarily ideal; and the YOLOv5 network has a simple structure, which is easy to be applied to the highway tunnel scenario for the smoke The YOLOv5 network has a simple structure and can be easily applied to the road tunnel scenario to improve the smoke and flame features. In this paper, we adopt the YOLOv5 framework and optimize the feature extraction part of the backbone network to enhance the deep learning of smoke and flame features, and realize the lightweight and improved design of the model. Large-scale datasets are used for model training to meet the network's demand for diverse training data. The experimental results show that the improved YOLOv5s model is robust to morphological changes, interfering objects and feature blurring of smoke and fire targets while meeting the real-time requirements in early tunnel fires, which well meets the requirements in the above industrial scenarios.

2. MATERIALS AND METHODS

2.1. Basic Structure of YOLOv5 Network

YOLOv5 is a single-stage target detection algorithm proposed by Ultralytics, which is further improved on the basis of YOLOv4, with smaller model size and faster inference. The YOLOv5 network structure is divided into Input, Backbone, Neck and Prediction part of the Input, Overall network structure is shown in Figure 1.

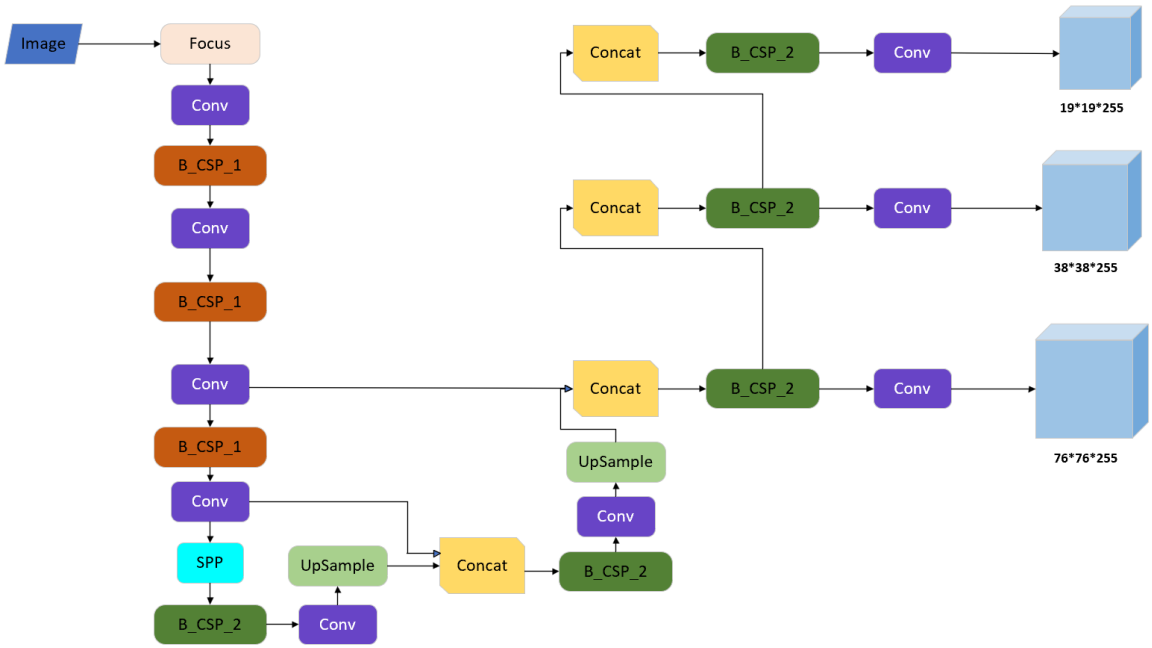

Figure 1. YOLOv5 network architecture

2.1.1. Input

At the input side, the training set images are preprocessed using adaptive image scaling, adaptive anchor boxes, and mosaic data enhancement. At the input side for different datasets, there are anchor box with initial set length and width. In network training, the network outputs a prediction box based on the initial anchor box, and then compares it with the groundtruth box , calculates the gap between the two, and then updates it in reverse to iterate the network parameters.Mosaic data augmentation randomly uses four images, randomly scales them, and then randomly distributes them for splicing, which greatly enriches the detection dataset, especially the random scaling adds a lot of small targets, which makes the network more robust. Mosaic data enhancement method combined with adaptive image scaling, the original image is spliced and then adaptively add the least black edges, and then the image is uniformly scaled to 640x640x3 and sent to the network for training. The results of the image processing at the input side are shown in Figure 2.

Figure 2. Image preprocessing results at the input

2.1.2. Backbone

Backbone is the workhorse of the detection network, including the Focus module, the convolutional layer, the B_CSP module and the SPP module.The Focus structure is unique to YOLOv5, and its function is slicing. The input matrix is changed into 4 low-dimension matrices with the information of the original matrix after the interval sampling operation, and then the 4 matrices are stacked in the same dimension, which makes the new matrix after stacking YOLOv5 designs two kinds of B_CSP structures, CSP1_X is used in Backbone, and the residual structure is adopted. The B_CSP structure is mainly borrowed from the Cross Stage Paritial Network (CSPNet) structure, which solves the problem of gradient in the network. structure, which solves the problem of repeating gradient information in the network by integrating the gradient changes into the feature map, thus reducing the number of parameters in the model, which ensures both the speed and accuracy of inference and also reduces the model size. By adjusting the depth and width of B_CSP and the residual component, four network models, YOLOv5s, YOLOv5m, YOLOv51, and YOLOv5x, are obtained respectively. The SPP module is a spatial pyramid pooling, which is downsampled by four different sizes of maximal pooling layers for multi-scale fusion to increase the sensory field. The Neck part is mainly used for generating feature pyramids, which will enhance the detection of objects at different scales by the model, allowing it to recognize the same object at different sizes and scales.

2.1.3. Neck

The Neck terminal contains two structures, the Feature Pyramid Networks (FPN) and the Path Aggregation Network (PAN). The FPN structure passes features from top down and fuses the information through up-sampling, while the PAN structure is a bottom-up feature pyramid and enhances the network feature fusion capability through down-sampling. sampling to enhance the network feature fusion. The main feature pyramid is generated to enhance the detection of the model for objects with different scaling scales, so that the same object with different sizes and scales can be recognized accurately.CSP_2 is used in Neck, and there is no residual structure.

2.1.4. Head

The Head module is the final part of the detection, which mainly predicts the image features, generates the bounding box, predicts the category to which the target belongs and calculates the loss function. The bounding box contains x, y, w, h, confidence. X ,Y represents the relative value between the center of the bounding box and the convenience of the grid. w, h represents the proportion of the width and height of the bounding box relative to the whole image. Confidence indicates whether the mesh contains objects and how accurately the coordinates of the bounding box are predicted, and is calculated by the formula:

$$\text{Confidence} = \text{P}_\text{r}(\text{Object}) \times \text{IOU}_{\text{pred}}^{\text{truth}} \tag{1}$$

In this case,

$$\text{IOU}_{\text{pred}}^{\text{truth}} = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where: $P_r(Object)$ indicates whether the bounding box contains the object or not, when the bounding box does not contain the object $P_r(Object)$ is 0, and vice versa is 1; IOU_{pred}^{truth} is the intersection and concurrency ratio between the predicted bounding box and the object's real region; A is the area of the predicted bounding box; and B is the area of the target's real region. Three loss functions are computed in the prediction side, namely DIOU_Loss, CIOU_Loss, GIOU_Loss. GIOU_Loss is used as the default loss function and has three different scales of outputs, corresponding to three different target predictions, namely, large, medium, and small [12].

2.2. YOLOv5s Model

The target detection algorithm based on YOLOv5 deep learning contains four versions, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x, of which YOLOv5s has the smallest weight. Considering the weight file size, recognition accuracy and detection speed of the model, YOLOv5s, which has the fastest detection speed and relatively high recognition accuracy, is selected for the study.The detection network of the YOLOv5s architecture consists of three detection layers, which input feature maps with pixel dimensions of 80×80, 40×40, and 20×20, respectively, for the detection of image objects with different sizes. Each detection layer finally outputs pre-selected boxes corresponding to the category, predicted bounding boxes and categories of the target raw image to generate labels to achieve image fire target detection.

3. MODEL IMPROVEMENT

3.1. Feature Extraction Network Improvement

In recent years, CNNs have shown great promise in computer vision tasks and are widely used in various scenarios. However, the application of neural network models on mobile and embedded

devices is still a great challenge due to the limitation of storage space and computational power. In order to reduce the amount of network computation and avoid large weights in the fire detection model, this study uses a Conv convolutional network with 6 convolutional kernels and a step size of 2 to replace the Focus structure in the Backbone, which not only maintains the improved detection speed of the Focus network, but also improves the detection accuracy to a certain extent. In addition, the B_CSP module is improved by removing the convolutional layer in the bridging branch of the original module, and connecting the input feature mapping of the B_CSP module to the output feature mapping of another branch directly in depth, which effectively reduces the number of parameters in the module. In addition, in order to further investigate how to reduce the network computation and lower the model weight, SPPF is used instead of SPP [14] layer. The original SPP uses three maximal pooling, with convolution kernels of 5*5, 9*9, and 13*13, and the features are connected to the input through different maximal pooling. While SPPF can better use the features that are constantly pooled, in SPPF the convolution kernel 5*5 is used instead of the previous large convolution kernel, the feature map after the first maximum pooling is pooled for the second time and the output is connected to the input, the image output after the second pooling is pooled for the third time and the output is connected to the previous output and the very first input, and the structure is shown in Figure 3. The SPPF structure is placed at the end of the backbone network, followed by reducing the number of convolutional networks from 9 to 6 in the second B_CSP of the backbone layer and using residuals in the last B_CSP.

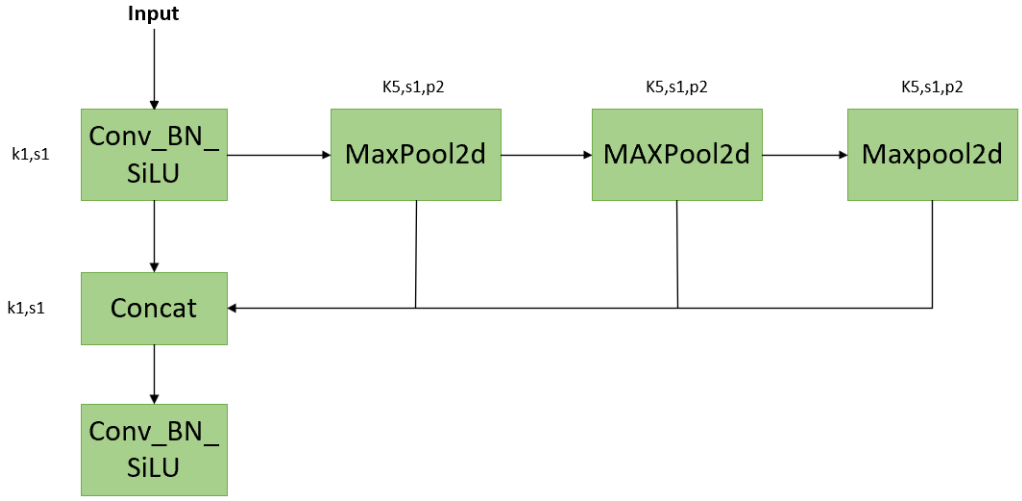

Figure 3. SPPF layer structure

3.2. Optimization of Loss Function

GIOU is used as the loss function of Bounding box in YOLOv5. Due to its validation dependence on IOU, it leads to too slow training convergence and low Bounding box prediction accuracy. If the prediction box and the GT(Ground truth) box appear to be non-intersecting, the gradient will become 0 at this time, the neural network will not be able to optimize, and the relative positions of the prediction box and the real box cannot be distinguished. Considering the above, a new Wise-IOU [16] is used to replace GIOU by utilizing outliers β instead of IOU to evaluate the quality of anchor box with the following equation:

$$\beta = \frac{\bar{L}_{IOU}}{L_{IOU}} \in [0, +\infty) \tag{3}$$

The dynamic nonmonotonic focusing mechanism is introduced with the following equation:

$$L_{\mathrm{WIOUV3}} = rL_{\mathrm{WIOUV1}}, r = \frac{\beta}{\sigma a^{\beta-\sigma}} \tag{4}$$

$$L_{\mathrm{WIOUV1}} = R_{\mathrm{WIOU}} L_{\mathrm{IOU}},$$
$$R_{\mathrm{WIOU}} = \exp\left(\frac{(x-x_{\mathrm{gt}})^2 + (y-y_{\mathrm{gt}})^2}{W_{\mathrm{g}}^2 + H_{\mathrm{g}}^2}\right) \tag{5}$$

Where: L_{WIOUV3} is the nonmonotonic focusing loss function of the prediction box. The parameters in Eq. (5) are schematically shown in Figure 4.

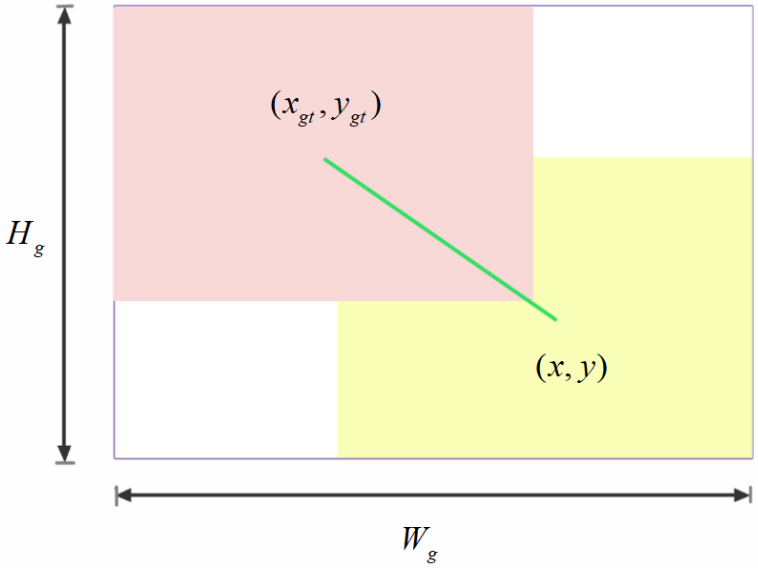

Figure 4. Schematic diagram of Wise-IOU parameters

Where w, h, (x, y) denote the width and height dimensions and center coordinates of the predicted frame, respectively; w_{gt}, h_{gt}, (x_{gt}, y_{gt}) denote the width and height dimensions and center coordinates of the real frame, respectively; W_i, H_i denote the intersection width and height dimensions, respectively; W_g, H_g denote the minimum border width and height dimensions, respectively.

L_{IOU} is dynamic, automatically adjusting the threshold as the model is trained, which makes the criteria for dividing the quality of the anchor box dynamic as well and provides a judicious gradient gain assignment strategy that reduces the competitiveness of high-quality anchor box while minimizing the deleterious gradients generated by low-quality samples. This allows Wise-IOU to dynamically and non-monotonically focus on common samples, improving the model's generalization ability and overall performance.

3.3. CA Attention Mechanism

The attention mechanism (ATM) [15] is a widely used technique in fields such as computer vision and natural language processing, which mimics the human attention mechanism to enable the network to selectively focus on the most important or relevant information in the input. The CA(Coordinate attention) attention mechanism [16] is introduced at the end of the Backbone layer, which is an attention mechanism that combines channel attention and location information, and helps the model to better localize and recognize the target by embedding the location information into the channel attention. Compared with the SE attention module, this module not only acquires the long-range

dependence on the spatial direction, but also enhances the expression of the position information of the features, while increasing the global receptive field of the network. By pooling the input features in the X-axis and Y-axis directions with one-dimensional adaptive averaging, we obtain independent direction-aware features with X-axis and Y-axis information, where one spatial direction captures the long-range dependence and the other retains the accurate position information. The two one-dimensional features are spliced in the W dimension with a convolution and a nonlinear activation function, followed by splitting the features in the channel dimension, and two feature maps with long-range dependencies in specific spatial directions are obtained by the convolution and the Sigmoid activation function, which can be applied to the input feature maps in a complementary way to enhance the target of interest. By feature fusion with the original features, the final feature map with attention weights in the width and height directions is obtained. The experimental comparison shows that it is more effective after placing it in the SPPF layer, which allows for more accurate target detection and recognition and improves the overall performance.

4. EXPERIMENTAL DESIGN AND ANALYSIS OF RESULTS

4.1. Experimental Design

4.1.1. Experimental environment

The experimental platform is Windows 11, with 16.0 GB of RAM and NVIDIA RTX4060 GPU, based on PyTorch2.0, using CUDA11.8 for computational acceleration, and training and validation under the same hyperparameters. The number of training rounds is set to 250, the batch size is 4, and the image size is 640. The specific configurations are shown in table 1.

Table 1. Platform configuration

Project	Configuration
Operating System	Windows 11
GPU	NVIDIA RTX4060
CUDA Version	11.8
Python	3.10
Deep Learning Framework	PyTorch 2.0

4.1.2. Dataset

In order to solve the problems of single scenes, interference of fire-like smoke objects, and difficulty in recognizing small targets in the existing fire and smoke dataset, this experiment uses crawler technology to collect fire videos and images from major websites, extracts the collected videos frame by frame, mixes them with fire and smoke images of various tunnel scenes to get 2050 images of different times and locations, and uses the labeling tool labelImg to label each image to create a more comprehensive dataset. This dataset contains 1518 images of only flame; 532 images contain both flame and smoke, making it more complex to recognize. At the same time, in order to classify fire class smoke scene information, this experiment in the dataset involved joining 10% of the negative samples for reference learning in order to reduce the false detection rate. The dataset is divided into a training set, a validation set, and a test set according to the ratio of 8:1:1, and finally, we get 1640 images in the training set, 205 images in the validation set, and 205 images in the test set.

4.1.3. Evaluation metric

Evaluation metrics such as accuracy (P), recall (R), mean average precision (mAP), and detection speed were used as follows:

Accuracy rate. The proportion of all targets predicted by the model that are predicted correctly, also known as the check accuracy rate, can be expressed as follows:

$$P = \frac{TP}{TP + FP} \tag{6}$$

Where: TP represents the number of IOUs greater than the set threshold, which is calculated only once for the same GT box, and IOU represents the intersection and concurrency ratio of the area of the prediction box and the GT box; FP represents the sum of the number of redundant prediction box with IOUs less than the set threshold or detected for the same real box.

2) Recall. The proportion of all labeled targets that are correctly predicted by the model, also known as the check all rate, can be expressed by the following formula:

$$R = \frac{TP}{TP + FN} \tag{7}$$

Where: FN represents the number of true box that were not detected.

3) Average Precision (AP). The evaluation index for reconciling the contradictory variables of accuracy and recall in target detection, the area enclosed by the curve with recall as the horizontal axis and accuracy as the vertical axis, the calculation formula can be expressed as follows:

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{8}$$

4) Average precision mean. A measure of recognition accuracy, which is the mean value of all categories of AP, and the calculation formula can be expressed as follows:

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{9}$$

Where: AP_i denotes the average precision of a single category, and C denotes the number of all categories.

5) Detection speed. Measures the computing power of the model and refers to the number of images that the model can process per second.

4.2. Experimental Results

4.2.1. Analysis of experimental results

By testing the training model with a test set of 2050 photos, the improved YOLOv5s model recognizes tunnel fireworks with 98.5% precision and 95.1% recall. Some of the test results are shown in Figure 5.

Figure 5. Partial test set identification results

The accuracy and recall are plotted as the Y-axis and X-axis, respectively, and then calculated to obtain the P-R curve corresponding to tunnel fire detection as shown in Figure 6, with a mAP value of 0.992.

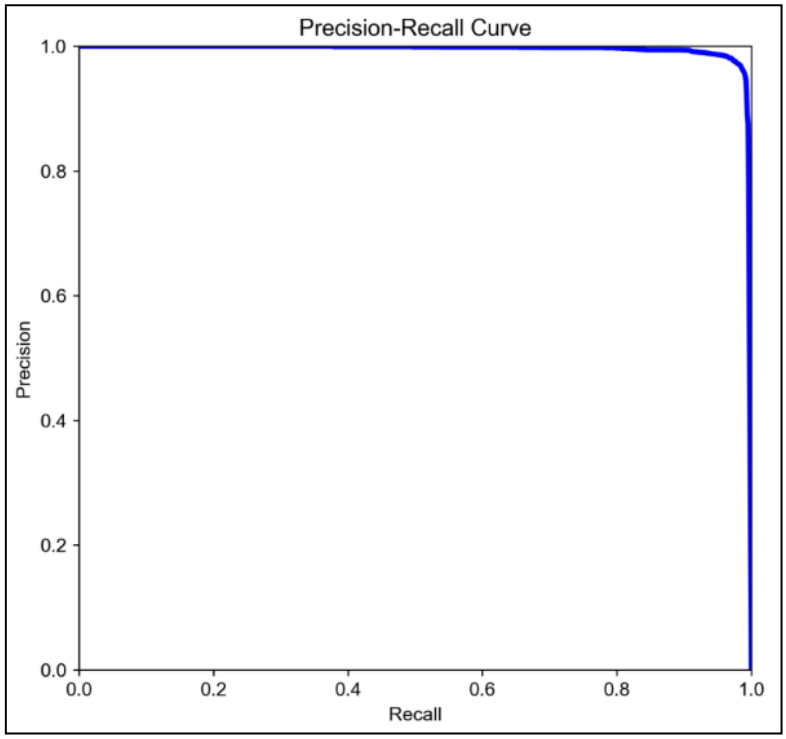

Figure 6. P-R curve

The training results can be viewed using the visual training tool, and the parameter maps generated after training the improved YOLOv5s model are shown in Figure 7.

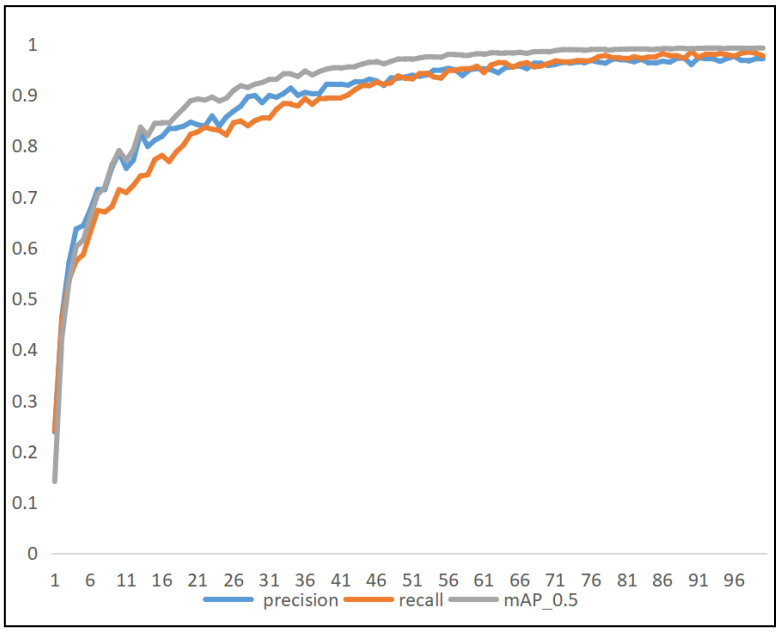

Figure 7. YOLOv5 model parameter map

In order to further verify the superiority of the algorithm in this paper, it is compared with other algorithms in the same dataset and experimental environment for experiments, and the results are shown in Table 2.

Table 2. Performance comparison table of five algorithms

algorithms	Precison	Recall	mAP@50/%	FPS
Faster R-CNN [14]	0.439	0.820	71.9	21.00
YOLOv3 [15]	0.879	0.513	73.0	88.00
YOLOv4 [16]	0.830	0.680	75.0	64.00
YOLOv5	0.97	0.956	96.0	138.0
The algorithms in this paper	0.989	0.951	99.2	148.00

In the same hardware environment and development environment, the improved YOLOv5 model is more complete, and according to Table 2, it can be seen that there are some improvements in accuracy, recall, precision, and detection time, and the FPS reaches 148 frames per second. Comparing with the original model training results, in the comparison of Figure 8(a) and Figure 9(a), it can be clearly seen that the improved algorithm is significantly higher than the original model in terms of detection accuracy. In the comparison of Figure 8(b) and Figure 9(b), it can be seen that in the case of poor field of view and a large amount of smoke occlusion, the improved model fusion CA attention mechanism locks in the target features by complementary fusion of two-direction feature maps with the original feature maps to detect the flames more accurately. In Figure 8(c), the original Yolov5s model has a leakage that cannot reliably detect the target, while the improved algorithm in this paper uses the CIoU loss function, continues to improve the aspect ratio of the prediction box and the GT box, considers the overlap area of the BBOX, the distance from the center point, and the consistency of the BBOX aspect ratio, improves the localization accuracy, reduces the target leakage rate, has stronger discriminative power, and is able to accurately detect the flame target. It can well meet the application scenarios described in this paper and can accurately detect small target flames even under the condition of smoke interference in such a dim and narrow environment as the tunnel, which provides reliable technical support for tunnel safety monitoring.

(a) (b) (c)

Figure 8. Recognition results of the model before improvement

(a) (b) (c)

Figure 9. Recognition results of the improved model

5. CONCLUSION

In order to improve the shortcomings of the existing flame detection algorithms in tunnel scenarios, this paper investigates the algorithm based on YOLOv5s, which has high real-time requirements for detection, and proposes an improved algorithm for detecting smoke and fire in tunnels based on YOLOv5s. The flame feature extraction network is improved by replacing the Focus structure with a convolutional layer and using the SPPF instead of the SPP layer for its lightweight design to reduce the computational pressure. In the loss function of YOLOv5s, the bounding box loss function is changed to a dynamic non-monotonic focusing mechanism, WIoU (wise intersection over union), which improves the bounding box regression performance of the detection model. In order to improve the detection effect of the detection algorithm on the flame target, the CA attention mechanism is fused at the very end of the Backbone layer, so that the attention mechanism sees the feature maps of the whole Backbone part with global vision, which improves the ability to extract the features of the flame target, and improves the model's ability to generalize the model in the actual scene. After the above improvements, the average accuracy of the model reaches 0.992, which lays the foundation for the accurate recognition of tunnel fire. The process of tunnel fire occurrence is actually the process of accumulation of danger, this paper utilizes small target flame and smoke detection for identification, comprehensively measures the danger of fire, nips the fire in the bud, and serves as an efficient early warning.

ACKNOWLEDGEMENTS

This work was supported by the SRTP project of Henan University of Science and Technology (2023124).

REFERENCE

[1] 2020 statistical bulletin on the development of the Transportation industry [J]. Finance & Accounting for Transport, 2021(6): 92-97.

[2] Khan S, Muhammad K, Hussain T, et al. Deepsmoke: Deep learning model for smoke detection and segmentation in outdoor environments [J]. Expert Systems with Applications, 2021, 182: 115125.

[3] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.

[4] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[5] Pan J, Ou X, Xu L. A collaborative region detection and grading framework for forest fire smoke using weakly supervised fine segmentation and lightweight faster-RCNN [J]. Forests, 2021, 12(6): 768.

[6] Lin G, Zhang Y, Xu G, et al. Smoke detection on video sequences using 3D convolutional neural networks [J]. Fire Technology, 2019, 55: 1827-1847.

[7] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in neural information processing systems, 2015, 28.

[8] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks [C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[9] Park M J, Ko B C. Two-step real-time night-time fire detection in an urban environment using Static ELASTIC-YOLOv3 and Temporal Fire-Tube [J]. Sensors, 2020, 20(8): 2202.

[10] Wang Y, Hua C, Ding W, et al. Real-time detection of flame and smoke using an improved YOLOv4 network [J]. Signal, Image and Video Processing, 2022, 16(4): 1109-1116.

[11] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection [J]. arXiv preprint arXiv:2004.10934, 2020.

[12] Jocher G, Chaurasia A, Stoken A, et al. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations [J]. Zenodo, 2022.

[13] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021 [J]. arXiv preprint arXiv:2107.08430, 2021.

[14] Tang H, Liang S, Yao D, et al. A visual defect detection for optics lens based on the YOLOv5-C3CA-SPPF network model [J]. Optics Express, 2023, 31(2): 2628-2643.

[15] NIU Z Y, ZHONG G Q, YU H.Areview on the attention mechanism of deep learning [J]. Neurocomputing, 2021, 452: 48-62.

[16] Cai M, Xu Q, Chen C, et al. Formation control with lane preference for connected and automated vehicles in multi-lane scenarios [J]. Transportation research part C: emerging technologies, 2022, 136: 103513.

[17] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.