Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

YOLOv11-based Flame Recognition Algorithm Utilizing a Fusion Dual-stream Attention Mechanism

Rui Gong and **Qiang Li*** 5713855126@163.com and **Jingyu Li**
Information & Communication Branch of State Grid Henan Electric Power Company, Zhengzhou 450001, China
Corresponding author

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Flame change characteristics are affected by ignition source, air pressure, wind direction and other factors, and the traditional method describes the problems of leakage, false alarm and poor real-time performance. Vision-based image detection is one of the important means to solve the above problems. Therefore, a YOLOv11 optimized flame detection algorithm is proposed. First, the feature extraction PCF module is designed to enhance the characterization of different layers of feature maps. Second, the model incorporates the dual-stream mechanism attention mechanism to improve the attention to different scale features. Finally, the model introduces an improved Focal Loss function to optimize the regression accuracy and network robustness in the prediction region. The model is subjected to comparative experiments on the Flame public dataset. The results show that the improved model performs well for flame smoke detection in complex scenarios, reaching 49.6% on mAP50 and 18.9% on mAP75, which is an improvement of 2% and 8% in accuracy compared to the original YOLOv11 model.

Keywords: Deep Learning, Flame Recognition, Feature Extraction, Attention Mechanism

1. Introduction

In recent years, fire accidents have occurred frequently, posing a serious threat to the safety of people's lives and property and the ecological environment. With the rapid development of computer vision, there are three main ideas for real-time flame detection using surveillance images: classification, segmentation and detection.

The classification task accomplishes the judgment of whether an image contains a target or not by building a classifier to map the input image to the corresponding category labels, and the representative models are VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), etc. Segmentation task achieves the target detection method by classifying the image pixel by pixel, representative models are FCN (Long et al., 2015), etc. In the task of detecting fire, detection using classification algorithms is slow and also fails to determine the location of the fire. Segmentation algorithms can accurately determine the area of the fire, but the model is large and difficult to deploy. The detection task combines both localization and classification tasks, using rectangular boxes to mark targets in the original image. Representative models are YOLO series (Redmon and Farhadi, 2018), SSD (Liu et al., 2016). YOLOv4 (Bochkovskiy et al., 2020) network by adding a channel attention module to the prediction header to improve the detection accuracy of smoke. Lightweighting YOLOv3 using a channel pruning method and using it for flame detection. However, most of the studies only detect flame and smoke separately, while both flame and smoke are critical information during fire (Wang et al., 2022).

Based on this problem, this paper proposes an improved YOLOv11 algorithm for flame detection and recognition research (Khanam and Hussain, 2024). First, a data enhancement strategy is introduced to increase the diversity of data; then, a new PCF feature extraction module is designed to replace the C3K2 module, which effectively enhances the characterization ability of the feature maps; then, the CUAM attention mechanism is integrated in the backbone network to improve the model's attention to the features of different scales; and finally, the loss function computation method is optimized to improve the fast convergence speed. The experimental results show that the improved network is capable of fire and smoke recognition in complex scenarios and has certain advantages over other networks in controlled experiments with other methods under the same test platform (Rasheed and Zarkoosh, 2024; Alkhammash, 2025).

2. Preliminaries

YOLOv11 (Rasheed and Zarkoosh, 2024) is capable of providing high-speed and high-accuracy target detection that outperforms most known target detectors, and the network architecture is shown in Figure 1. YOLOv11 optimizes the architecture to achieve improved performance by adding the C3K2 block, the SPFF module, and the C2PSA block. The C3K2 block optimizes the performance of more complex feature extraction by introducing the CSP (Cross Stage Partial) block The C3K2 block optimizes the extraction of more complex features by introducing the CSP (Cross Stage Partial) block, which accomplishes different sized kernel and channel separation strategies. The SPFF (Spatial Pyramid Pooling Fusion) module allows the model to perform better by capturing object attributes at different scales. The C2PSA block combines channel and spatial information to provide a more efficient feature extraction. Together with the Multi-head Attention mechanism to achieve more accurate perception of objects.

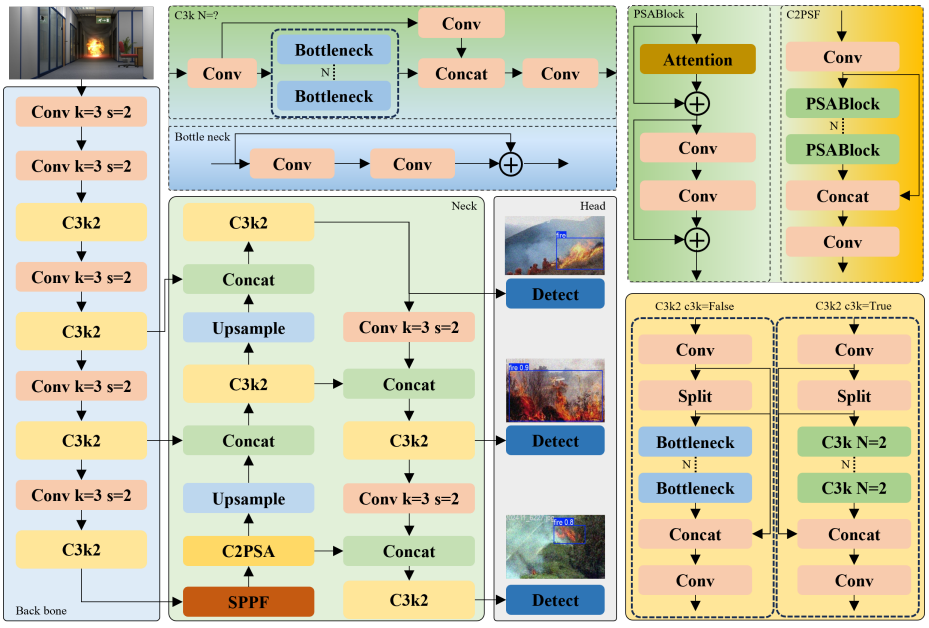


Figure 1: Architecture of YOLOv11.

3. Methods

3.1. PCF Module

Excellent neural networks are based on their network architecture to improve performance. In convolutional neural networks convolutional operations are used as an underlying operation to fuel the development of artificial intelligence. Most of the research on efficient model architectures for scaling efficient layer aggregation networks usually focuses on the number of parameters, the amount of computation and the computational density. Efficient aggregation modules are designed to achieve effective learning on the depth of the network by adjusting the shortest and longest lengths of the gradient paths. However, when using efficient aggregation for channel splicing in large-scale multi-branch stacking, it will cause the number of channels to skyrocket, which in turn reduces the computational efficiency and parameter utilization of the network. To this end, we propose the PCF (Parallel Split-Concatenation Fusion Block) module, which is based on the CSPNet structure, and achieves a better capture of the input tensor by slicing the input tensor in terms of the number of channels, applying multiple classical Bottlenecks in one part to extract the features, and subsequently splicing them with the remaining part. channel correlations, as shown in Figure 2.

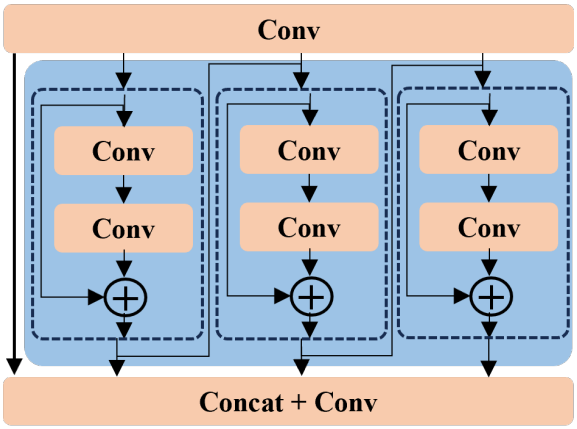


Figure 2: Parallel Split-Concatenation Fusion Block.

3.2. Weighted Hybrid Attention Mechanism

Attention mechanism is an efficient strategy for optimizing the allocation of resources when dealing with large amounts of information in order to cope with the challenge of information overload (Gu et al., 2024). Self-attention mechanism breaks through the effects of the localization constraints of traditional convolution by dynamically modeling the global relationships between input elements. This mechanism accomplishes the memetic representation of features mainly by capturing the advantages of long-range dependencies, dynamic weight assignment, and parallel computing power. The attention mechanism improves the performance of neural networks by capturing contextual information, and the Convolutional Unit Attention Module (CUAM) infers more accurate Channel Attention Module (CAM) through maximum pooling and average pooling, and combines it with the Spatial Attention Module (SAM) and Self-Attention Module (SM) to form a dual-stream self-attention mechanism, as shown in Figure 3.

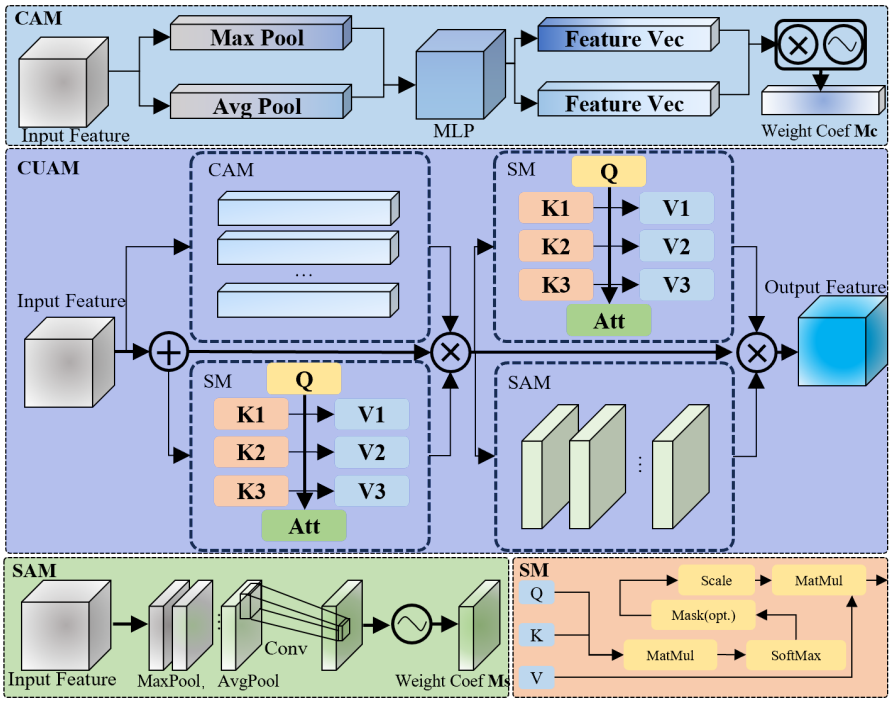


Figure 3: Convolutional Unit Attention Module.

The channel attention module and the spatial attention module are based on local context information and are used to model and weight channel and spatial information, respectively. In addition, the channel attention module generates two-way feature vectors by performing average pooling and maximum pooling on the input feature maps, and then generates two-way feature vectors using a multilayer perceptron (MLP), followed by using the sigmoid activation function to obtain weight coefficients for channel attention M_C. Then, the features required by the spatial attention module are obtained after the weight coefficients are operated by dot product with the input feature map. In addition, each channel combines with the self-attention module to complete the allocation of feature weights using the relationship between the input data, respectively, and compared with the first two mechanisms, the self-attention has the advantage of parallel computation in the computation. The mathematical description of the weight coefficients M_C of channel attention is as follows:

$$M_C(F) = \sigma\left(MLP\left(Avg\left(F\right)\right) + MLP\left(Max\left(F\right)\right)\right) \tag{1}$$

The spatial attention first performs a global pooling operation on the input feature map, followed by stitching the two generated H×W×1 feature maps into H×W×2. After convolution operation, the feature weight coefficients are obtained using the sigmoid activation function M_S. Finally, the weight coefficients and the input feature map are subjected to dot product operation to obtain the output feature map. The weight coefficients M_S of the channel attention are described mathematically as follows:

$$M_S(F) = \sigma\left(f^{7\times7}([Avg\left(F\right), Max\left(F\right)])\right) \tag{2}$$

where F denotes the input feature map. $Avg\left(F\right)$ denotes that the feature map is computed after average pooling. $Max\left(F\right)$ denotes that the features are computed by maximum pooling. $MLP(\cdot)$

denotes that the features are computed by multilayer perceptron. σ denotes that the features are computed by sigmoid activation function. $f^{7\times7}$ denotes a 7×7 convolution operation, which facilitates faster capture of a larger range of contextual information (e.g., the overall contour of a large-size object) without stacking multiple layers of small convolution kernels.

Self-attention mechanism in order to reduce external information dependencies. It allows each element in a sequence to be associated with all other elements, thus capturing the global dependencies within the sequence and more effectively capturing the internal complexity of the feature. Q, K, and V are all obtained by linearly varying the input matrix X represented by a set of words, and for each element in the sequence, the model generates the corresponding query (Q), key (K), and value (V). These vectors are obtained by multiplying the input elements with the weight matrix obtained from training. Q and K are responsible for establishing semantic associations between words, and the distance matrix that expresses the semantic associations is computed by the core algorithm of self-attention. The distance matrix is multiplied by matrix V to get the output matrix of global associations. It is mathematically described as follows:

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d_k})V \tag{3}$$

where d_k denotes the dimension of the Q, K vector; K^T denotes the matrix transpose. $SoftMax(QK^T/\sqrt{d_k})V$ denotes the distance matrix, which establishes the semantic association between words. $\sqrt{d_k}$ is to prevent the value of Q and K from being too large after inner product, softmax is used to calculate the association coefficient of each word corresponding to other words. The distance matrix obtained after softmax. Finally, the distance matrix is multiplied with V to get the output matrix with global associations.

3.3. Loss Function

The loss function gradually reduces the prediction error by calculating the loss value, and the model back-propagation adjusts the parameters. Obtained by calculating the error of target and positive sample prediction, this paper contains three main parts: categorization loss (cls_loss), location loss (iou_loss) and target loss (obj_loss). The loss function depends on the sum of the three parts, and the model uses the cross-entropy loss function as the cls_loss category loss, and adopts the CIoU loss as the iou_loss location loss. the CIoU is defined as:

$$LOSS_{CIoU} = 1 - IOU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha\nu \tag{4}$$

where $\rho^2\left(b, b^{gt}\right)$ represents the Euclidean distance between the centroid of the predicted and true values. represents the diagonal distance between the predicted value and the smallest region of the true value, α is the weight coefficient, and ν is used to measure the consistency of the matrix aspect ratio. α and ν are expressed as follows:

$$\alpha = \frac{\nu}{1 - IOU + \nu} \tag{5}$$

$$\nu = \frac{4}{\pi^2}\left(\tan^{-1}\frac{w^{gt}}{h^{gt}} - \tan^{-1}\frac{w}{h}\right)^2 \tag{6}$$

In addition, the model uses focal Loss as obj_loss, which makes the model focus on difficult to classify samples during training by reducing the weight of easy to classify samples, and the formula

is expressed as follows:

$$LOSS_{CFocal} = \begin{cases} -\alpha(1-y^{'})^{\gamma}\log y^{'}, y=1 \\ -(1-\alpha)^{\gamma}y^{'\gamma}\log(1-y^{'}), y=0 \end{cases} \tag{7}$$

where $y^{'}$ represents the probability of predicting the outcome and y represents the label of the data. focal loss adds γ adjustable factor and α weighting factor. The values of γ and α in the model follow the experimental results in the (Lin et al., 2017), with γ as 2 and α as 0.25.

4. Experiments

4.1. Function Datasets

The publicly available dataset for fire and smoke detection, DFS (Dataset for Fire and Smoke detection), is chosen for this experiment to fulfill the need for simultaneous detection of flame and smoke targets. The dataset covers fire scenarios of multiple scales and also contains interference scenarios with only smoke or similar flames, which can better simulate the real detection environment and is highly challenging (Wu et al., 2023).

The DFS dataset contains a total of 9,462 images, of which more than 8,000 are fire- and smoke-related, and the remaining 1,000 or so contain luminous objects. Each image contains at least one label, and the label categories include FIRE, SMOKE, and OTHER (luminous objects that can be easily misidentified as fire, such as lights, bright-colored objects, etc.). The presence of the "other" category results in an uneven distribution of data, which is preprocessed using data enhancement methods. The experiments in this paper use multiple data enhancement strategies, firstly, 50% of the data is scaled and cropped, then linear superposition is processed with 50% probability, and finally, for the processed data is inputted into the network model proposed in this paper for training (Zhang et al., 2017).

4.2. Results and Analysis

4.2.1. EXPERIMENTAL ENVIRONMENT

In this study, the image size of the dataset was resized to 640×640 pixels. To avoid network overfitting from occurring, pre-processing enhancement operations such as random flipping, rotating, scaling, and cropping are performed on the data. At the time of training, the split of training, validation and test sets is 7:2:1 according to the dataset with reference to YOLOx (Ge et al., 2021) with the following parameters, batch size is set to 16, epochs is set to 300, optimizer is selected as sgd, momentum is set to 0.95, the learning rate is set to 1e-3, and cosine annealing is selected as the way of decreasing the learning rate. The system implementation was carried out on Ubuntu 20.04 system using a graphics card NVIDIA RTX 3060 with 8GB of video memory, vscode as the development environment, and Python 3.9.12, PyTorch 2.0.1 and CUDA11 as the development language.

4.2.2. EVALUATION INDEX

In this paper, Precision (PRE), Recall (RE), Intersection over Union Ratio (IoU), and Mean Accuracy (mAP) are used for quantitative analysis. AP is used to measure the accuracy of target

detection and denotes the area under the precision (PRE) and recall (RE) curves as the mean value. The specific calculations are as follows:

$$RE = \frac{TP}{TP + FN} \tag{8}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i, AP = \int_0^1 PdRE \tag{9}$$

where n is the number of categories. IoU values of 0.5 and 0.75 were taken to evaluate the average accuracy of all categories and individual categories. mAP50 and mAP75 represent the average accuracy of categories with IoU of 0.5 and 0.75, respectively. fAP50, oAP50, and sAP50 represent the accuracy of "fire", "other", and "smoke" with IoU of 0.5. fAP75, oAP75, and sAP75 represent the accuracy of "other", and "smoke" with IoU of 0.75, respectively. and "other" and "smoke" for IoU of 0.5 and 0.75, respectively. fAP75, oAP75, and sAP75 represent IoU of 0.75 for "fire", "other", and "smoke", respectively. "other" and "smoke" with an IoU of 0.75. fAP75, oAP75 and sAP75 represent the "fire", "other" and "smoke" with an IoU of 0.75, respectively.

4.2.3. ANALYSIS OF EXPERIMENTAL RESULTS

The number of parameters of the model in this paper reaches 78.58 M. The experiments show that the average detection accuracy of the model reaches 49.6% when the IoU is 0.5, and the average detection accuracy of the model is 18.9% when the IoU is 0.75. Compared with the YOLOv11 model, the accuracy is improved by 2% and 8%, respectively. The specific data are shown in Table 1.

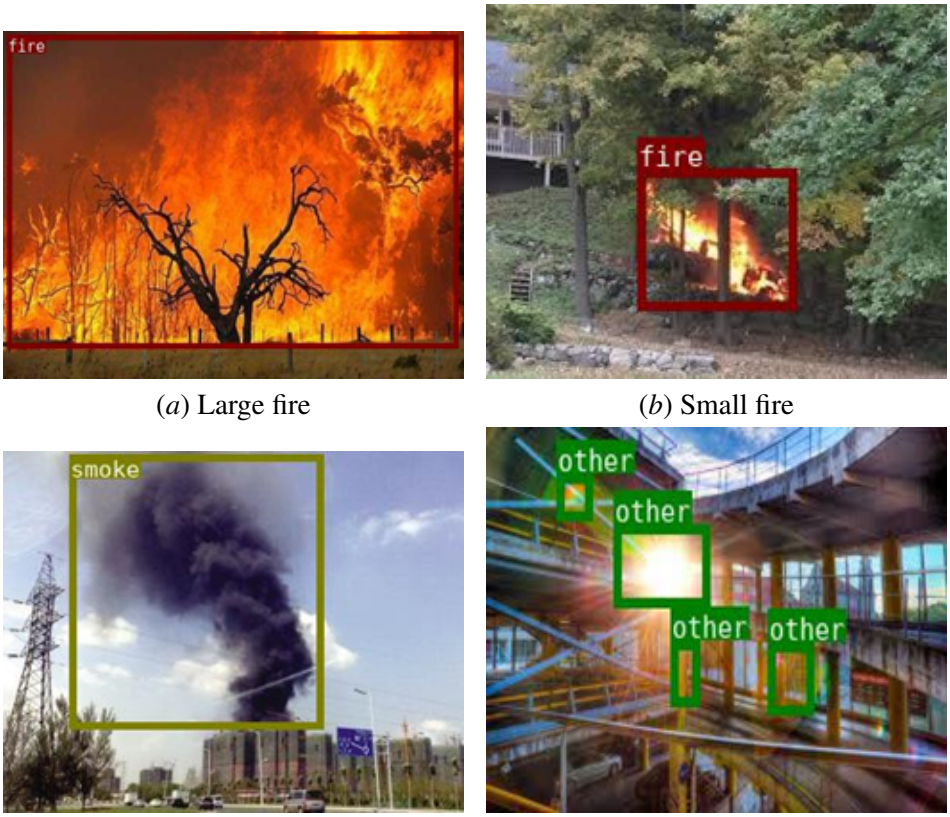
Table 1: Experimental Results.

Index	YOLOv7	YOLOv11	Our	Raising↑
mAP50(%)	46.5	48.3	49.6	2%
mAP75(%)	15.1	17.5	18.9	8%

In order to verify the detection effect of the model more intuitively, this paper selects typical fire pictures containing four common scenarios, such as big fire, small fire, smoke and others. The detection effect is shown in Figure 4. The detection results show that the model is capable of different scales of flames, Fig. 4(a) is a large fire, and the flames account for almost more than 80% or more of the figure, (b) figure has a relatively small percentage of flames, and the model is able to recognize flames at different scales with a high confidence level. In addition, (c) figure contains a large number of smoke targets, smoke on this part of the morphology shows dispersion, the model is still able to accurately identify. The model in (d) correctly recognizes the glowing object as "other". The proposed improved model has some leakage detection, but the possibility of false detection is almost zero, which shows that the model has a strong differentiation ability for flames and other flame-like objects. In conclusion, the improved algorithm can accurately recognize flames and smoke with high confidence in several complex scenes.

4.2.4. COMPARISON EXPERIMENTS

In order to verify the effectiveness of the improved network, we conducted comparison experiments with the current mainstream deep learning target detection algorithms on the public dataset DFS



(*a*) Large fire (*b*) Small fire

(*c*) Smoke (*d*) Other

Figure 4: Qualitative results of this paper's method on four different scene image tasks compared to other detection scenarios.

under the same experimental environment. The comparison results of the three types of detection are shown in Table 2. From the table, it can be concluded that the mAP50 key index, which is the best performance among all evaluation methods, reaches 0.496 for all categories, 0.651 for the "fire" category, 0.286 for the "smoke," "The mAP75 metric requires more accuracy in the target boundary region, and the performance of the improved model is slightly lower, but the performance in the "fire" category is still in the lead, and the overall mAP75 reaches 0.291. In addition, the accuracy of the "other" category is relatively low in both the mAP50 and mAP75 metrics, which is acceptable considering the core objective of flame and smoke detection. In terms of inference speed, the improved algorithm performs better in real-time, although it slightly underperforms the other methods in the mAP75 metric.

Overall, compared to other models, the improved model in this paper provides better recognition on the DFS dataset, and is better able to satisfy the flame and smoke detection task. The average accuracy of the improved model reaches a high level at both IoU of 0.5 and 0.75. Compared to the original YOLOv11 model, the proposed model shows more significant advantages in detection performance.

Table 2: Precision comparison of different methods. (Bold indicates the best)

Method	Backbone	mAP50	mAP75	fAP50	sAP50	oAP50	fAP75	sAP75	oAP75
faster-rcnn	ResNet-50+FPN	0.451	0.156	0.635	0.268	0.452	0.256	**0.063**	0.138
retinanet-fpn	ResNet-50+FPN	0.432	0.146	0.602	0.244	0.412	0.246	0.054	0.154
retinanet-nasfpn	ResNet-50+NASFPN	0.461	0.167	**0.651**	0.257	0.479	**0.286**	0.051	0.156
ssd300	VGG16 Size300	0.433	0.162	0.603	0.245	0.452	0.234	0.048	0.136
ssd512	VGG16 Size512	0.433	0.134	0.603	0.243	0.453	0.236	0.048	0.137
YOLOv3-416	DarkNet-53 Scale:416	0.316	0.112	0.532	0.203	0.347	0.189	0.036	0.097
YOLOv3-608	DarkNet-53 Scale:416	0.396	0.107	0.575	0.218	0.402	0.191	0.024	0.105
YOLOv4	DarkNet-53 Scale:416	0.412	0.108	0.567	0.254	0.414	0.193	0.028	0.099
YOLOv7	DarkNet-53	0.465	0.151	0.645	0.284	**0.458**	0.244	0.052	0.159
YOLOv11	DarkNet-53	**0.483**	**0.175**	0.648	**0.285**	0.455	0.245	0.058	**0.164**
Our	DarkNet-53	**0.496**	**0.189**	**0.651**	**0. 286**	**0.461**	**0.291**	0.051	0.154

5. Conclusion

Flame intelligent recognition is a hot issue in current research, and there are still difficulties in the portrayal as well as the description of the dynamic change features of flame and smoke, which leads to low recognition accuracy and is difficult to meet the practical needs. In order to improve the accuracy of flame recognition and reduce misrecognition, this paper constructs a YOLOv11 flame intelligent recognition model incorporating a dual-stream mechanism. By introducing the PCF module, the characterization ability of the feature map is effectively enhanced; in addition, the CUAM attention mechanism is fused in the backbone network to improve the model's attention to features of different scales; at the same time, the Focal Loss is introduced to improve the computation of the loss function, which improves the accuracy of the target prediction and the robustness of the model. The proposed method is validated on the flame public dataset, and the results show that the improved model is able to detect flame, smoke, and other categories simultaneously in complex scenarios, with the mAP50 reaching 49.6% and the mAP75 reaching 18.9%. Compared with other target detection models, the model has a significant advantage in average accuracy and can realize efficient and accurate flame and smoke recognition.

References

Eman H. Alkhammash. Multi-classification using yolov11 and hybrid yolo11n-mobilenet models: A fire classes case study. *Fire*, 8(1), 2025. doi: 10.3390/fire8010017.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *CoRR*, abs/2107.08430, 2021.

TIANJUN Gu, SUYA Xiong, and Lin XIAO. Diversified generation of theatrical masks based on sasgan. *Journal of Graphics*, 45(1):102–1, 2024. doi: https://doi.org/10.11996/JG.j.2095-302X.2024010102.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *ArXiv*, abs/2410.17725, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, and et al. Ssd: Single shot multibox detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

Areeg Fahad Rasheed and M. Zarkoosh. Yolov11 optimization for efficient resource utilization, 2024.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2022.

Siyuan Wu, Xinrong Zhang, Ruqi Liu, and Binhai Li. A dataset for fire and smoke object detection. *Multimedia Tools and Applications*, 82(5):6707–6726, 2023. doi: 10.1007/s11042-022-13580-x.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017.