A Fire and Smoke Detection Model Based on YOLOv8 Improvement

Pengcheng Gao*
School of Cyber Security, Gansu University of Political Science and Law, Lanzhou 730070, Gansu, China

***Abstract*—The warning of fire and smoke provides security for people's lives and properties. The utilization of deep learning for fire and smoke warning has been an active area of research, especially the use of target detection algorithms has achieved significant results. For improving the fire and smoke detection performance of model in different scenarios, a high-precision and lightweight improvement based on the model of You Only Look Once (YOLO), is developed. It utilizes partial convolutions to reduce the complexity of model, and add an attention block to acquire the cross-space learning capability. In addition, the neck network is redesigned to realize bidirectional feature fusion. Experiments show that it has significantly improved the results for all metrics in the public Fire-Smoke dataset, and the size of the model has also been widely reduced. Comparisons with other popular target detection models under the same conditions indicate that the improved model has the best performance as well. In order to have a more visual comparison with the detectability of the original model, the heatmap experiments are also established, which also demonstrate that it is characterized by less leakage rate and more focused attention.**

***Keywords*—Fire and smoke detection; deep learning; computer vision; YOLO**

I. Introduction

The warning of disaster is a broad field that many researchers have devoted themselves to study in recent years. There are many categories of disasters, including floods and fires, which must be monitored at an early stage, so that precautionary measures can be taken. It is very necessary to detect and monitor disasters including floods, typhoons and fires at an early stage and take relevant preventive measures. Among these disasters, fire is one of the most hazardous, which often inflicts a serious threat to people's property and lives, and also causes huge losses to public facilities and ecological resources [1].

In many cases, the fire detection is still based on the traditional smoke sensor and temperature sensor [2], [3], [4]. When the value detected by the sensor exceeds a certain threshold, the alarm and fire extinguishment system will be activated [5]. This method is more effective in some relatively small indoor environments, but in some scenes, like a factory and forest, which are relatively open and easy to cause the rapid spread of fire, this method is often unable to quickly detect the occurrence of fire, and it is difficult to accurately provide the fire location information. Therefore, how to improve the ability of fire detection, as well as accurately and rapidly detect the fire have become the focus and direction of current research in this area. With the popularity of video surveillance and the iteration of image processing, it becomes a mainstream of current research to detect the occurrence of fire by learning the characteristics of flame and smoke through processing image sets. This research is mainly divided into three categories: target classification models, target segmentation models and target detection models [6].

Since target classification models can only determine whether flame and smoke are present in the image, and target segmentation models need to build a large number of pixel-level labelled datasets for training, both types of models have certain limitations when performing such tasks. Target detection models have the functions of classifying and locating the target to be detected, which can quickly detect whether a fire occurs or not, and also accurately select the target through the anchor box, so the target detection model is more suitable for dealing with this type of task, which is also a future research direction.

Existing models have two shortcomings in flame and smoke detection, which are worthy of continuous improvement. Firstly, at the time of initial fire, the measurement of object is little and the feature is not distinct enough, thus making it more difficult to be detected. Secondly, the current target detection models for the flame and smoke are generally too complicated to be applied to the equipment with different performance, resulting in insufficient practicability. The main reason for this problem is that the modules used in the model improvement scheme proposed by the researchers, as well as the improvements to the model structure, significantly increase the complexity of inference, resulting in slower model computation [6]. For the existing issue, the objective of paper is to achieve a lightweight and high-precision object detection model by improving an existing model. The significance of this study is to make the improved model more practical, which can be easily deployed on various terminal devices for detection tasks in different scenarios, so that the model has the ability to detect and locate the flame and smoke targets more quickly and accurately in order to reduce the losses caused by disasters.

The work established in this paper is based on the improvement of YOLOv8n. Under the premise of improving the precision, we compressed the magnitude of model by reducing the parameters required for the operation. As a result, the developed model has the characteristics of both high precision and light weight. Three major innovations are shown below:

1) For the purpose of decreasing the size of model, A new block C2f-faster is constructed by replacing the Bottleneck Block in the original YOLOv8n with FasterNet Block.

2) By utilizing the Efficient Multi-scale Attention (EMA) block into the network, it is more conducive to fuse the contextual information at different scales, and make the neural network extract the feature from the input better.

3) By redesigning the Neck layer of yolov8n, the Bidirectional Feature Fusion (BiFF) is realized to improve the detectability.

The rest part involves five sections: Section II presents the related researches. Section III illustrates the structure of the YOLOv8n model and its advantages over previous versions, and then introduces the three improvements based on this model. Section IV mainly presents the dataset and setting used in the experiments. Section V demonstrates the effects of different improvement methods through ablation experiments, and the results of comparative experiment with YOLOv3t, YOLOv4t, YOLOv5n, YOLOv6n, YOLOv7t are shown in this Section. Moreover, comparisons of detection and heatmap are also finished. Section VI analyzes the experimental results and summarizes the whole work.

II. Related Works

The algorithm based on object classification models determines whether the input image contains fire or smoke category information and outputs the corresponding label. Based on the VGG16 model, He et al. [7] introduced an attention block and FPN feature fusion block to obtain an improved classification effect of smoke and smoke-like targets. However, the usage scenario of the improvement has some restrictions. Besides, the situation of smoke cannot be identified. RYU et al. [8] used Harris corner detector and HSV channel to pre-process the flame, and then captured features from Inceptionv3 model to improve the accuracy, but the pre-processing took a long time. Nguyen et al. [9] developed a method which combines the CNN and Bi-LSTM to extract spatial domain and temporal domain features of flame simultaneously. However, the large number of fully connected layers in the network made the computation heavy, and made it difficult to deploy.

Compared with the target classification model that can only judge whether there are flame and smoke in the image, the target segmentation model can get the shape, size and other details from the loaded pictures, and then judge the spread trend of fire. U-net, proposed by Ronneberger et al. [10], is a model which is applied extensively in the image segmentation field. It constructs a network similar to the letter U through the encoder and decoder structure, and utilizes this structure to make the output which is extracted by the encoder part fuse in the decoder part to get multi-scale features. Inspired by the complete convolutional network (FCN), Yuan [11] proposed a target segmentation model with good performance in the segmentation of fuzzy smoke images. Frizzi et al. [12] established a network structure based on VGG16 to detect and locate flame and smoke, and outperformed U-net and Yuan-net in different indicators. The algorithm based on the target segmentation model can provide more detailed fire information, but the size of the model is usually large. Besides, a large number of pixel-level labeled datasets are used in the training of this type of model, which will undoubtedly consume a lot of time.

The algorithm based on object detection model can classify and locate multiple flame and smoke targets by different anchors in the input image. Park et al. [13] integrated the ELASTIC block [14] into the backbone of YOLOv3 to detect candidate regions, generated a bag-of-features (BoF) histogram for the target region, and then passed the BoF into the random forest classifier to detect the target. It is difficult to deploy the model to embedded devices because of its high requirement of graphics operation. Xue et al. [15] mainly added a 160*160 head into the YOLOv5 model to obtain a better capability when detecting small targets and utilized the CBAM [16] that includes broader identities to improve the perception of model. From the experimental results, the value of mAP is improved, but the value of Frame Per Second (FPS) is decreased.

III. Improved Methodology

YOLOv8 is a one-stage target detection algorithm released by Ultralytics in January 2023 based on YOLOv5 [17]. This version can be used in performing image classification, target detection, target tracking and other tasks. The entire network is composed of three components: the Backbone extracts feature maps from the loaded picture; the Neck aggregates the features of different layers and passes it to the predicting part; and the Head makes predictions about the target and its location information. Compared with the previous version of YOLO algorithm, YOLOv8 demonstrates better detection performance on the COCO dataset. Moreover, YOLOv8 provides different models according to the size, such as n, s, m, l, and x. The model becomes larger in turn, which is controlled by depth, width, and max channels. The model chosen for improved is the smallest of the above, YOLOv8n, which suits better with the objective of this work.

The constitution of YOLOv8 is displayed in Fig. 1. To realize further lightweight, the C3 block in the former version is updated by the C2f block in YOLOv8 [18]. In the Neck layer, the former convolutional module in the up-sampling layer in YOLOv5 is deleted, and the output from different layers are straightly loaded into the up-sampling stage [19]. Decoupled Head is adopted in the Head part, which captures the position of target and category information separately and aggregates them after learning in different paths of network. Compared with the Coupled head in YOLOv5, it can efficiently enhance the model's performance to generalize and increase its robustness [20]. Unlike the Anchor-Base used in the previous YOLO series to predict the position and size of the Anchor, YOLOv8 uses the Anchor-Free detection method, which means it does not need to preset the Anchor, thus reducing the time-consuming and required arithmetic power [17].

The flame and smoke detection task is often limited by device resources. In order to be applied to as many different scenarios as possible, a lightweight and low latency model is a basic condition for it to be deployed on different devices. On this basis, realizing high accuracy as much as possible is also an improvement direction for the model. A new model called YOLOv8n-EBF is improved and proposed.

Fig. 2 shows the main structure of YOLOv8n-EBF. As mentioned before, YOLOv8n-EBF mainly makes three

improvements on the original network. Firstly, a new FasterNet Block consisting of partially convolution is used to change the Bottleneck in the C2f to constitute a new module called C2f-faster. There are total seven C2f-faster modules used in the network, which can effectively decrease the magnitude of the network and further affect the computing speed. Secondly, the EMA block is used to strengthen the extraction of target features. Finally, a modified Neck layer structure is utilized for fusing the output feature maps of four C2f-faster modules in the Backbone across space. The extracted multi-scale features are loaded into the network to obtain the detecting improvement. The three followed subsections specify the details of each modified module.

A. The Improved C2f Module

The C2f references the design idea of Efficient Layer Aggregation Networks [21] to obtain richer gradient information by branching more gradient streams in parallel, which in turn results in higher accuracy and lower latency.

The convolutional kernels and operations are widely used in deep learning networks, and the process often require a large amount of computational support. For alleviating the issue of slow inference process generated by convolutional operation in the model, Chen [22] proposed a new partial convolution, called PConv. It replaces the regular form of convolution by utilizing one PConv of c_p channels and one 1×1 convolution of $c - c_p$ channels to combine a hammer-like structure, as shown in Fig. 3(a) and Fig. 3(b). Compared with one regular $k * k * c$ kernel convolution, shown in Fig. 3(c), the participants in the improved convolution module is reduced from $k^2 \cdot c$ to $k^2 \cdot c_p + (c - c_p)$, which not only achieves a similar effect but also greatly reduces the amount of computation when it is used for calculation. Based on the partial convolution, they constructed a new network module, FasterNet Block, shown in Fig. 4 below, which is used to extract features. It contains one 3×3 PConv layer and two 1×1 regular Convolution layers, which has a similar structure and function with Bottleneck block. Therefore, it is utilized to propose an improved module called C2f-faster, shown in Fig. 5 below.

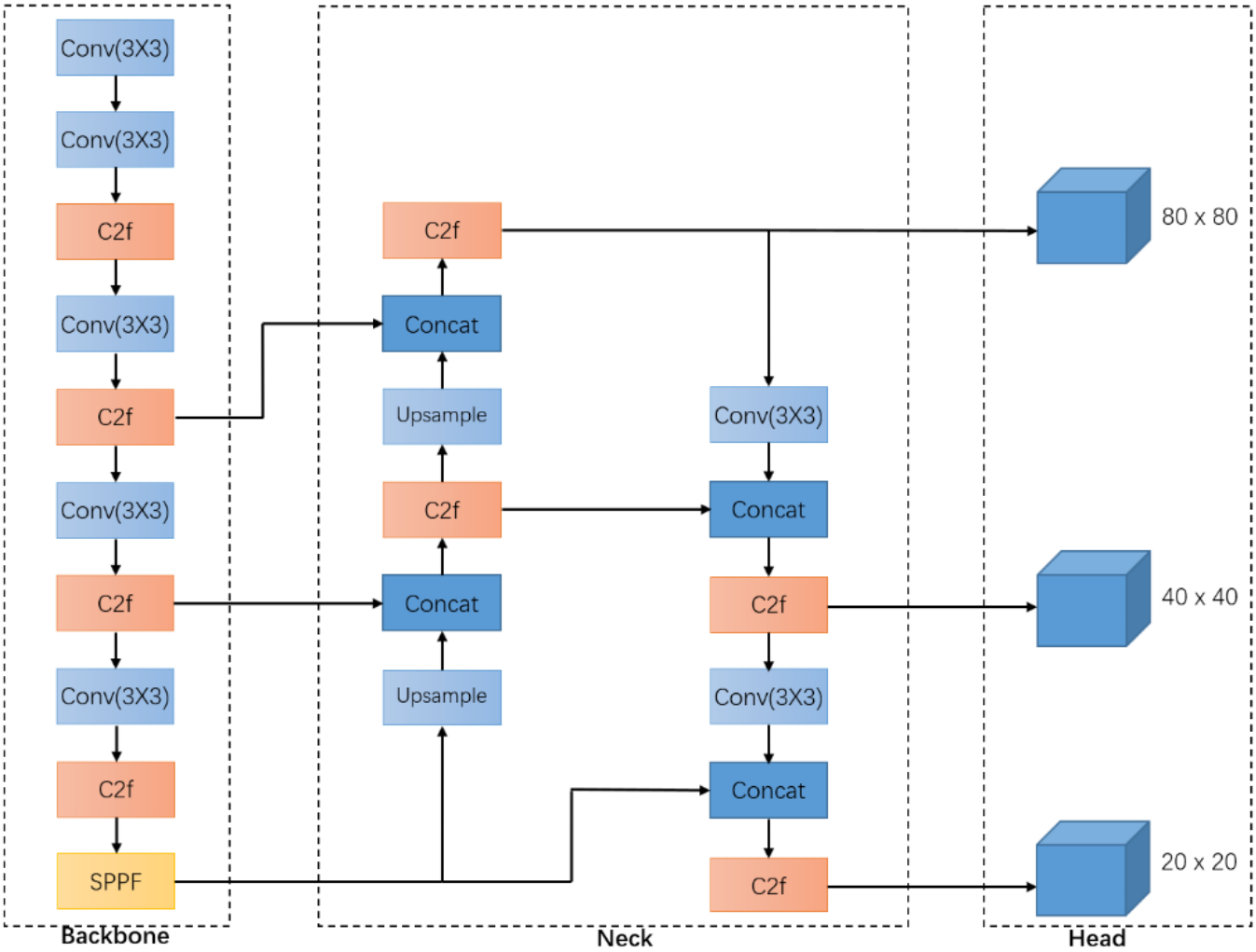

Fig. 1. The structure of YOLOv8.

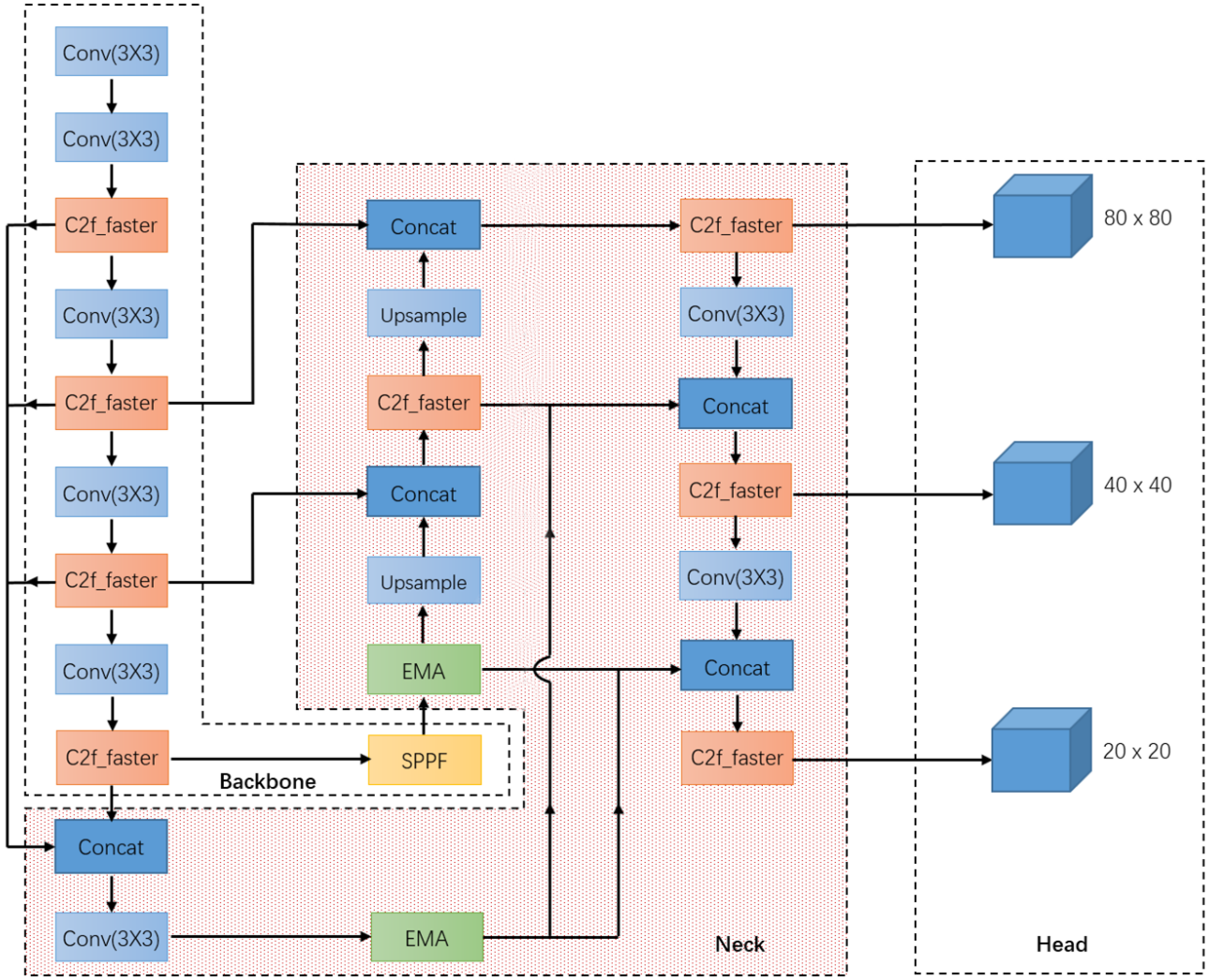

Fig. 2. The structure of YOLOv8n-EBF.

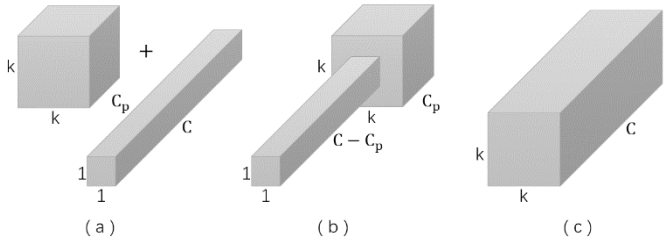

Fig. 3. (a) Structures of convolutional variants; (b) A hammer-like structure which is constituted by one PConv and one 1*1 Conv; (c) One regular k*k*C kernel Conv.

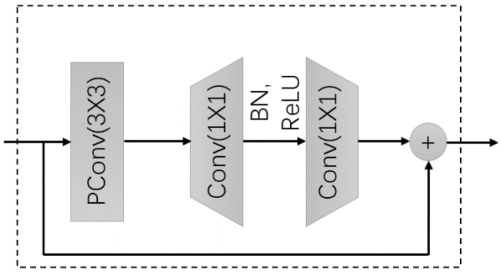

Fig. 4. The structure of FasterNet block.

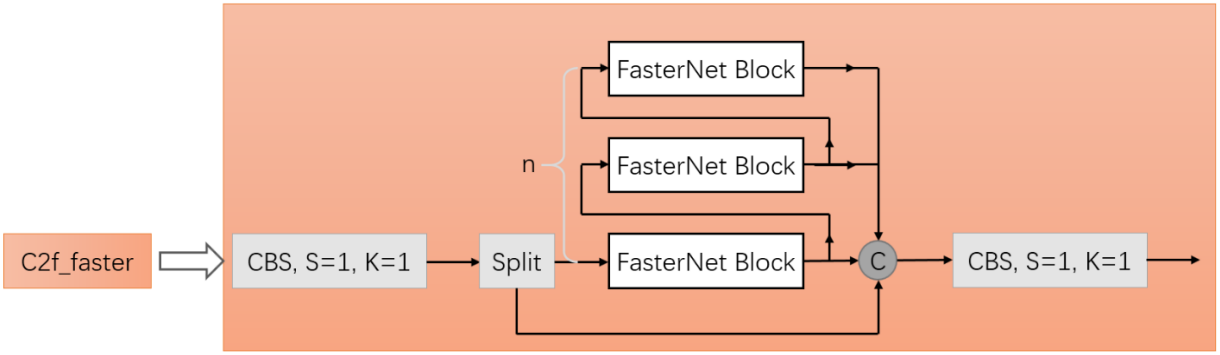

Fig. 5. An improved C2f formed by replacing bottleneck blocks with FasterNet blocks.

B. *Efficient Multi-Scale Attention (EMA) Module*

By invoking the attention module can capture the important image information, allowing the model to focus on detecting the key areas and obtaining the significant features of the target, which plays an important character in all kinds of computer vision tasks [23]. In this paper, EMA module [24] is utilized into the improved model to enhance the detection capability. Fig. 6 reveals the principle of EMA module. For the input feature map $X \in R^{C\times H\times W}$, EMA divides the channel dimension into G sub-features, $X=[X_0,X_i,...,X_{G-1}]$, $X_i \in R^{C\times H\times W}$, and makes $G \ll C$, which enable the model to obtain different semantic features. This module captures the weights of grouped features during two parallel paths which contains one 1×1 convolution path and one 3×3 convolution path. The parallel substructure reduces the depth of the networks, and avoids the dimensionality reduction by merging some of the channels at the same time, maintaining the features of each channel. Similar to the Coordinate Attention [25], a global average pooling operation is added for encoding operations in the X and Y directions of the channel in the 1×1 branch, and these two encoded features are concatenated and convolved with a 1×1 kernel convolution. The output is then decomposed into two vectors and fitted using a Sigmoid nonlinear activation function. Finally, the cross-channel interaction is achieved by multiplying the aggregated channel attention, which efficiently captures the inter-channel dependencies and preserves the spatial information in the channel.

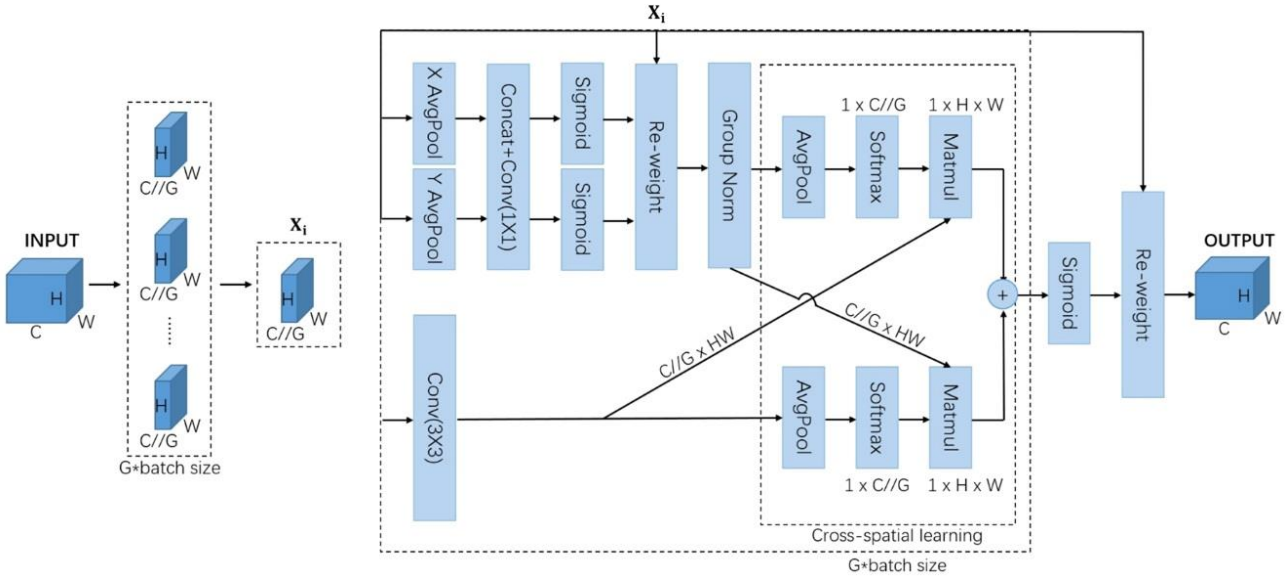

Fig. 6. The structure of EMA module.

In another branch, one 3×3 convolutional kernel is added for capturing multi-scale features and constitutes with the 1×1 branch for aggregating the cross-space information. The main approach is to encode the outputs of the 1×1 branch and the 3×3 branch by a global average pooling operation and convert them to a $1\times C//G$ dimensional shape after passing through a normalization function and a reshape operation, and then multiply it with the feature vector $C//G\times H*W$ of the other branch after dimensionality reduction, as shown in the formula below:

$$R=R_1^{1\times C//G}\times R_2^{C//G\times H*W} \quad (1)$$

The output R that fuses contextual information from different branches enables the neural network to produce a better attention for the feature map. Moreover, it is multiplied with the original input after a Sigmoid activation function and a dimensional transformation to obtain the final output feature

map. Since the size of EMA's input and output are same, which makes it convenient to directly add into the YOLOv8n network.

C. Redesigned Neck Layer

The feature fusion of different scales is a significant approach to improve image processing. To obtain richer image feature information, an improved structure for YOLOv8n network with Bidirectional Feature Fusion (BIFF) is proposed, as shown in Fig. 7. To entirely utilize the important semantic information in the high-dimensional feature maps as well as the target feature information contained in the medium- and low-dimensional feature maps, we aggregated the feature maps of four different layers in the backbone and then fused them with others in the neck part.

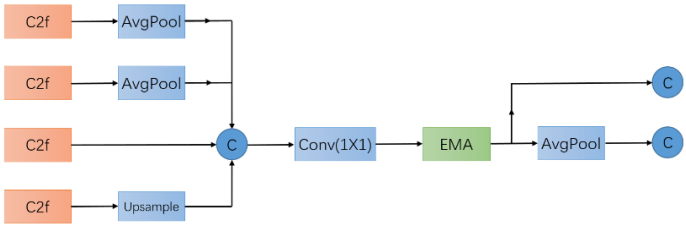

Fig. 7. The process of BiFF network.

According to the structure diagram of the original YOLOv8n, it can be known that there are four C2f modules in the backbone, which can generate four different scales of feature maps, i.e., 160×160, 80×80, 40×40, and 20×20. In the redesigned neck network, the low-dimensional feature maps of 160×160 and 80×80 are chosen to be reduced to the 40×40 size by the average pooling operation, and the high-dimensional feature map of 20×20 was scaled up to the size of 40×40 by up-sampling operation. The low-dimensional feature maps tend to contain more spatial information due to smaller receptive fields, while the high-dimensional feature maps with larger receptive fields tend to contain more semantic information [26]. The reason for choosing to scale these three feature maps to the size of 40×40 is that this size of feature map can contain the information in both the low- and high-dimensional feature maps, and will not cause the loss of information due to being too large or small. These four feature maps are concatenated in series and then downscaled by a point wise convolution and fed into the EMA module. As mentioned in the subsection above, The EMA module mainly works on slicing the feature information of C channels into G groups and performs feature extraction on different parallel paths, and finally generates the feature maps that incorporate multi-scale information. In one branch, it is combined with another 40×40 map in the original Neck network, and in another branch, it is scaled to 20×20 for combining with the same size feature map in the network by average pooling operation. At this point, an improved Neck network for cross-space feature fusion induced by the backbone layer is constructed.

The network has three main advantages as followed:

1) This network is combined with the original Neck network to realize two-way feature fusion, which strengthens the expression of features, and thus improves the performance of the detector;

2) It mainly consists of parallel structure, which is faster in computation;

3) It mainly utilizes four existing feature maps. The subsequent experimental part shows that this improvement only increases a few parameters.

IV. Environment and Dataset

A. Experimental Environment and Evaluation Criterion

This work is established in the following environment: the CPU is an 8-core Xeon Gold 5218R; the memory capacity is 32GB; the graphics card is a Tesla V100-SXM2 with 16GB of memory. The version of Python is 3.8.8, Pytorch is 1.8.0, CUDA is 11.7, and YOLOv8n is ultralytics 8.0.147. The models in the experiments did not use pre-trained weights, and the main hyperparameter values are shown in Table I below.

In the experiment, the Precision, Recall, Average Precision, mAP@.5, Parameters and GFLops are chosen as the evaluation criterion. The criteria for sample classification are shown in Table II.

TABLE I. Description of the Main Hyperparameters

Hyperparameter	Value
Lr	0.01
Lrf	0.01
Momentum	0.937
Weight_decay	0.0005
Batch-size	16
workers	8
Epochs	200

TABLE II. Criteria for Sample Classification

Classification	Explanation
TN	Predicting the correct quantity of negative samples
FN	Predicting the incorrect quantity of negative samples
TP	Predicting the correct quantity of positive samples
FP	Predicting the incorrect quantity of positive samples

1) P (Precision), the scale of positive samples predicted correctly to samples predicted as positive, is calculated as:

$$P=\frac{TP}{TP+FP} \quad (2)$$

2) R (Recall), the scale of positive samples predicted correctly to all true positive samples, is calculated as:

$$R=\frac{TP}{TP+FN} \quad (3)$$

3) AP (Average Precision), which reflects the average prediction ability for a single target category. The higher the value of AP, the better the detectability of the model in this category. The calculation formula is:

$$AP = \int_0^1 P(R)\, dR \quad (4)$$

4) mAP, which reflects the average predictive ability of the model for all categories, i.e., averaging the AP values for all categories, is calculated as follows:

$$mAP = \frac{1}{n} \sum_{i=1}^{n} (AP)_i \quad (5)$$

where, n means all predicted categories, and $(AP)_i$ means the average precision of the ith category. mAP@.5 is used as an evaluation criterion in the experiment, which means that when the overlap between the predicted box and the GT box is greater than 0.5, i.e., IoU>0.5, the prediction is judged to be correct, and the relevant values are calculated using this as a benchmark.

5) Parameters and GFLops which reflect the model size and computational complexity, are used to measure the ease with which a model can be deployed in end devices.

B. Dataset

A high-quality dataset allows the model to extract features more efficiently during training. We selected a public dataset Fire-Smoke, which contains 3961 photos. The labels of the dataset are categorized into Fire, Smoke, compared to the single-label dataset, this dataset enables the model to detect both fires that can be directly observed and fires that are obscured by objects by detecting smoke.

Training and validation sets are split 9:1. Fig. 8 displays some representative pictures. The scenes cover indoor scenes such as living rooms, bedrooms, offices, and hallways, as well as outdoor scenes such as factories, forests, streets, and buildings. Besides, it contains pictures at different distances from close view to distant view, it contains pictures with only flames, pictures with only smoke, and pictures with both flames and smoke.

Overall, the selected dataset contains a rich collection of scenarios covering enough features of flames and smoke to make the trained model generalizable and applicable to detection work in different environments.

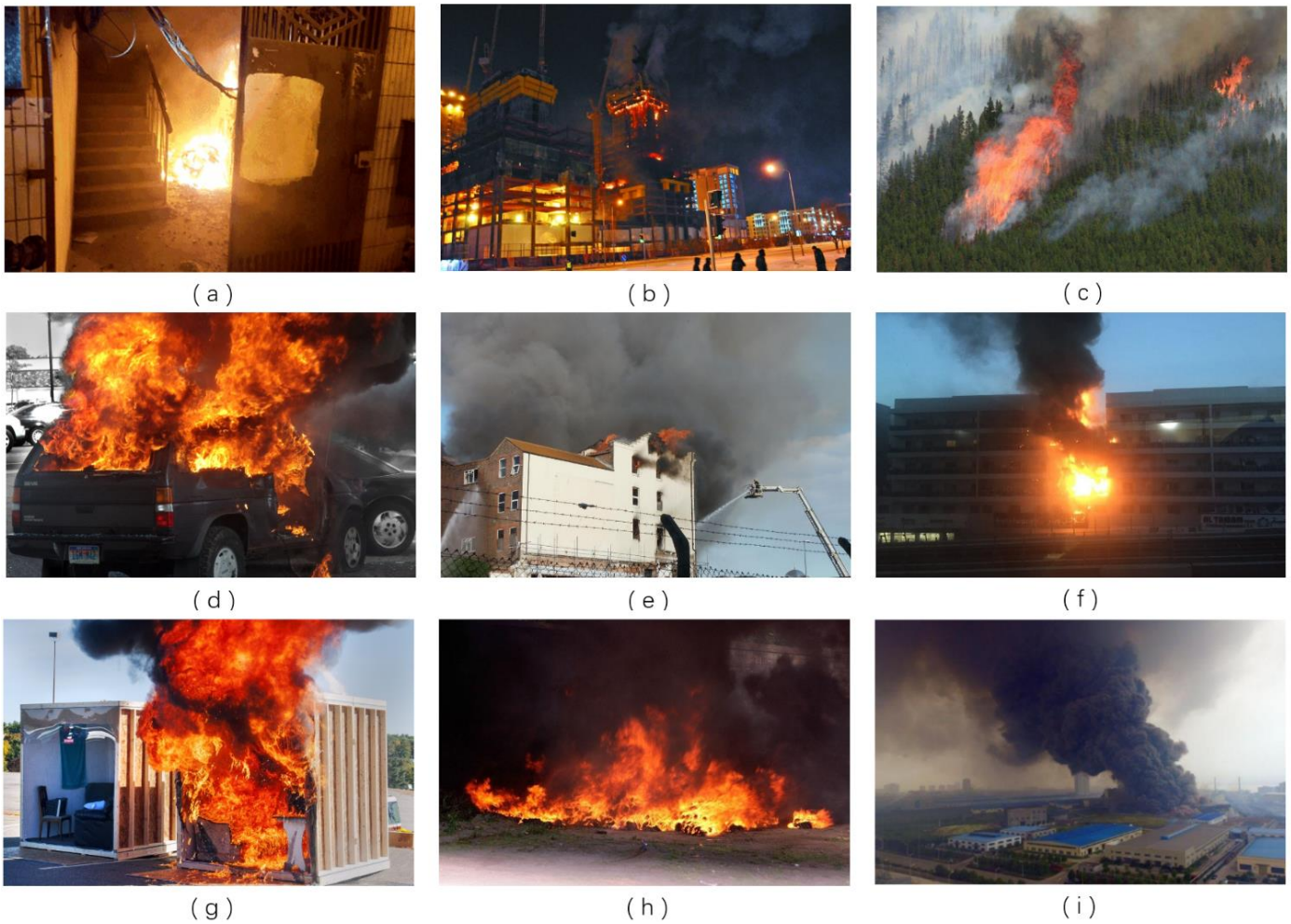

Fig. 8. Representative fire and smoke images selected from the dataset: (a) Fire in a corridor, (b) Fire in a building, (c) Fire in a forest, (d) Fire in close-up, (e) Fires in mid-range, (f) Fires in far-range, (g) Images dominated by flames, (h) Images with both flames and smoke, (i) Images dominated by smoke.

V. Results and Analysis

A. Ablation Experiment

To verify the effects of different methods proposed in this work on the original network, four sets of ablation experiments were carried out for YOLOv8n.

The first experiment used C2f-faster modules to replace all the C2f modules in the network. The second experiment added an EMA module after the SPPF module. The third experiment used BiFF to form a new Neck network. In the end, the fourth experiment used the three improvement methods mentioned above to form the complete network YOLOv8n-EBF. All the experiments were established on the same environment. The results are listed in the Table III.

From the ablation experiments, the Precision, Recall, and mAP@.5 of the redesigned network are improved by 4.7%, 1.9%, and 3.1%, respectively, compared to YOLOv8n, while the parameters decrease by 19.7%, and GFLops decrease by 18.3%. Replacing C2f with C2f-faster efficiently reduces parameters, and increases Precision as well as mAP@.5 by 3.7% and 1.5%, respectively, but Recall has a slight decrease. The addition of the EMA module increases the network with almost no parameters and GFLops, and enables the neural network to generate better attention for the feature maps by fusing contextual information at different scales, resulting in a certain improvement in the overall detection ability. A new neck network was constructed by adding a bottom-to-top path to enable bi-directional feature fusion with the network. With only a 3.7% growth in parameters, Precision increases by 1.0%, Recall increases by 1.5%, and mAP@.5 increases by 1.3%, indicating that the improved neck network can indeed have positive effects. In summary, compared to the YOLOv8n, the overall performance of YOLOv8n-EBF model is improved with a large reduction in complexity. These improvements result in a lighter model with higher accuracy at the same time.

B. Comparative Experiment

In order to further verify the difference in performance between the YOLOv8n-EBF and other models on the flame and smoke detection, this paper conducts comparative experiments. Five classical small-sized models in the field of target detection, i.e., YOLOv3-tiny, YOLOv4-tiny, YOLOv5n, YOLOv6n, YOLOv7-tiny, are selected. The performances of each model after training are displayed in Table IV.

YOLOv3-tiny has the largest number of Parameters and GFLops among different versions of YOLO above. It has more than twelve million Parameters, which is five times more than the improved YOLOv8n-EBF, and 19.0 GFLops, which is 2.8 times more than the latter. In terms of model size, YOLOv8n-EBF is 4.8MB, only 10.4% of YOLOv4t, which is the smallest among all models and can be easily deployed in different devices. In terms of detection ability, YOLOv4-tiny has the worst performance in this experiment, with a value of mAP@.5 of only 43.1%, and YOLOv8n-EBF has an improvement of 74.2% for this parameter. The only other models with a mAP@.5 above 70% are YOLOv5n, YOLOv6n, and YOLOv8n, and their performance is relatively similar, with results close to 71.9%. Compared to YOLOv8n-EBF, the latter has a mAP@.5 of 75.0%, which is the highest of all models. In addition to this, the other parameters of YOLOv8n-EBF are at the highest level compared to other models.

C. Comparison of Detection Effects

At the end of training, the obtained weight parameter model is used to detect the target samples and mark the location of the detected objects. The results are shown in the Fig. 9 below, with the original image, the detected image of YOLOv8n, and the detected image of the improved model in the left-middle-right of each row, respectively.

TABLE III. The Results of Ablation Experiments

Model	Parameter	GFLops	P/%	R/%	mAP@.5/%
YOLOv8n	3011238	8.2	73.1	63.4	71.9
YOLOv8n-C2f-faster	2306038	6.4	76.8	63.2	73.4
YOLOv8n-EMA	3011252	8.2	74.4	64.2	72.6
YOLOv8n-BiFF	3122100	8.5	74.1	64.9	73.2
YOLOv8n-EBF	2416914	6.7	77.8	65.3	75.0

TABLE IV. The Results of Comparative Experiments

Model	Parameter	GFLops	Size/MB	P/%	R/%	mAP@.5/%
YOLOv3t	12133156	19.0	23.2	67.8	61.0	66.5
YOLOv4t	6056606	16.4	46.3	30.4	69.9	43.1
YOLOv5n	2508854	7.2	5.0	73.7	63.4	71.6
YOLOv6n	4238342	11.9	8.3	75.7	62.8	71.5
YOLOv7t	6017694	13.2	11.7	67.9	67.2	69.6
YOLO8n	3011238	8.2	6.0	73.1	63.4	71.9
YOLOv8n-EBF	2416914	6.7	4.8	77.8	65.3	75.0

1) In the detection comparison of Fig. 9 (a) with a bright light and unobstructed situation, both the original YOLOv8n and YOLOv8n-EBF are able to detect the smoke in the picture, but the anchor box of the former model locates inaccurate compared to the improved network, as shown in Fig. 9(b) and Fig. 9(c).

2) In the detection comparison of Fig. 9(d) with a low light and obstructed situation, the original model is capable of detecting the fire in the picture, but the inaccuracy range of the anchor still exists. As shown in Fig. 9(e) and Fig. 9(f), a larger portion of the selected box for smoke is a building rather than a target to be detected, while the improved model is more accurate obviously.

3) In the detection comparison of Fig. 9(g) with a high contrast, the YOLOv8n-EBF detects all four targets which is shown in Fig. 9(i), but the result of original YOLOv8n in Fig. 9(h) only detects three large-sized targets but not the smallest flame in the picture, which appeared to be a missing detection.

4) In the detection comparison of Fig. 9(j) with a low contrast, the original YOLOv8n model also occurs a similar result, which detects one smoke target and two flame targets, but not the small flame located in the center of the picture, as shown in Fig. 9(k). Moreover, when framing the flames on the left side, it appears more obvious that the anchor box cannot cover the target, i.e., the framing is inaccurate. However, it can be observed from Fig. 9(l) that YOLOv8n-EBF performs significantly better, detecting all targets and being able to accurately localize them.

Overall, the improved model has a better detectability for different sizes, and can accurately recognize the target in the presence of environmental interference and object occlusion.

Fig. 9. Comparison of detection effects of original image, YOLOv8n and YOLOv8n-EBF.

D. Comparison of Heatmap Effects

In order to have a more intuitive understanding of the focused region of the model on the image and to make the decision-making process of the network better interpretable, Grad-CAM [27] is applied to generate heatmaps in this paper. In order to better compare with for synthesis, we used the same images as above for the experiments. The same images chosen in the previous section are used for comparisons. The settings, especially the layer, are consistent in the experiment, and the results are shown in Fig. 10. The left-center-right of each row shows the original image, the heatmap of YOLOv8n, and the heatmap of YOLOv8n-EBF, respectively.

1) Comparing the heatmaps in Fig. 10(a), (b) and (c), the focus area of YOLOv8n is more inclined to the right side of the image, and it is larger and more distributed in the whole heatmap. Compared with the improved network, which focuses on the region of the target to be detected, the latter is more concentrated, which obviously has a better detection effect.

2) When comparing the heatmap effect of Fig. 10(d), (e) and (f), the focus area of YOLOv8n also appears to be more scattered, focusing on parts of the image not related to the flame, such as the building in the upper left corner and the extinguished vehicle in the lower right corner. However, the improved model focuses precisely on the flame region.

Fig. 10. Comparison of heatmap effects of original image, YOLOv8n and YOLOv8n-EBF.

1) In the comparison of heatmaps in Fig. 10(g), (h) and (i), YOLOv8n only focuses on two flame targets below the image, and only one flame target is highlighted, i.e. the red area in the picture, while the color covered another flame is lighter, which indicates that the level of attention is not high enough, in addition, this model does not focus on the flame target above the image. Multiple flame targets are better considered in the improved model, not only highlighting the two flames below the image, but also focusing on the target above the image.

2) In the comparison of heatmaps in Fig. 10(j), (k) and (l), the highlighted area of the original model is in the upper right, which can be found from the original that this area is not smoke, but a brighter background. YOLOv8n-EBF focuses on a more scattered area than other situations, but it can be seen that the highlighted areas are still the part of flame and smoke.

From the four sets of heatmap comparisons above, we can more intuitively see that YOLOv8n-EBF developed with more focused attention is able to locate the aim more accurately.

VI. Conclusion

In this paper, three improvements are made to the YOLOv8n model and all experiments are performed on a public dataset. First, ablation experiments are performed to show that each method contributes to the promotion of model's performance. Subsequently, comparison experiments with six different models are conducted to demonstrate that the algorithm not only has better detection capabilities but has a lightweight characteristic at the same time. Finally, the paper conducts detection comparison experiments as well as heat map comparison experiments to provide a more straightforward comparison with the original network. The conclusion of the established work are as follows:

1) The dataset used in this experiment contains abundant pictures of flame and smoke, which makes the model can effectively detect both of them and has a good generalization to apply to detecting tasks in different environments. The ability to detect smoke makes the model capable of detecting obstructed combustible and early fire, reducing the leakage problem caused by single-target detection.

2) The improved model involves the EMA blocks and a developed neck network to improve feature fusion in different dimensions. In the comparison experiments of detection and heatmap, this model shows a higher sensitivity and more focused attention to targets of different scales, which enables the model to locate the target more accurately and reduces the leakage rate.

3) By replacing the Bottleneck in the original C2f module with a new FasterNet block composed of partial convolution to form the new module called C2f-faster, the complexity is effectively reduced. The parameters of YOLOv8n -EBF are about 2.4 million, the GFLops is about 6.7, and the size of the model is only 4.8MB. Therefore, it is convenient to be deployed in various terminals.

4) The improved model achieves 77.8% precision, 65.3% recall and 75.0% mAP@.5. The network has improved Precision, Recall and mAP@.5 by 4.7%, 1.9% and 3.1%, respectively, compared to YOLOv8n, with a reduction of 19.7% in parameters and 18.3% in GFLops. According to the experiments, it can be observed that the complexity of YOLOv8n-EBF has greatly decreased compared to YOLOv8n, while all the indicators measuring the detection performance have been significantly improved. It is superior to the former in terms of performance and complexity optimization, which further confirms the effectiveness of the improvement.

References

[1] F. Cui, “Deployment and integration of smart sensors with IoT devices detecting fire disasters in huge forest environment,” Computer Communications, vol. 150, pp. 818-827, 2020.

[2] J. Zhang, W. Li, N. Han, and J. Kan. “Forest fire detection system based on a ZigBee wireless sensor network,” Frontiers of Forestry in China, vol. 3, pp. 369–374, 2008.

[3] M. F. Othman, and K. Shazali, “Wireless sensor network applications: A study in environment monitoring system,”Procedia Engineering, vol. 41, pp. 1204-1210, 2012.

[4] K. B. Shaban, A. Kadri, and E. Rezk, “Urban air pollution monitoring system with forecasting models,” IEEE Sensors Journal, vol. 16, no. 8, pp. 2598-2606, 2016.

[5] M. Kumar, P. K. Singh, M. K. Maurya, and A. Shivhare, “A survey on event detection approaches for sensor based IoT,” Internet of Things, vol. 22, p. 100720, 2023.

[6] Y. Zhu, Y. Si, and Z. Li,“Overview of smoke and fire detection algorithms based on deep learning,” Computer Engineering and Applications, vol. 58, no. 23, pp. 1-11, 2023.

[7] L. He, X. Gong, S. Zhang, L. Wang, and F. Li, “Efficient attention based deep fusion CNN for smoke detection in fog environment,” Neurocomputing, vol. 434, pp. 224-238, 2021.

[8] J. Ryu and D. Kwak, “Flame detection using appearance-based pre-processing and convolutional neural network,” Applied Sciences, vol. 11, no. 11, p. 5138, 2021.

[9] M. D. Nguyen, H. N. Vu, D. C. Pham, B. Choi, and S. Ro, “Multistage real-time fire detection using convolutional neural networks and long short-term memory networks,” IEEE Access, vol. 9, pp. 146667-146679, 2021.

[10] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Springer International Publishing, 2015, pp. 234-241.

[11] F. Yuan et al., “Deep smoke segmentation,” Neurocomputing, vol. 357, pp. 248-260, 2019.

[12] S. Frizzi, M. Bouchouicha, J. M. Ginoux, E. Moreau, and M. Sayadi, “Convolutional neural network for smoke and fire semantic segmentation,” IET Image Processing, vol. 15, no. 3, pp. 634-647, 2021.

[13] M. J. Park and B. C. Ko, “Two-step real-time night-time fire detection in an urban environment using Static ELASTIC-YOLOv3 and Temporal Fire-Tube,” Sensors, vol. 20, no. 8, p. 2202, 2020.

[14] H. Wang, A. Kembhavi, A. Farhadi, A. L. Yuille, and M. Rastegari, “Elastic: Improving cnns with dynamic scaling policies,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 2258-2267.

[15] Z. Xue, H. Lin, and F. Wang, “A small target forest fire detection model based on YOLOv5 improvement,” Forests, vol. 13, no. 8, p. 1332, 2022.

[16] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.

[17] F. M. Talaat and H. ZainEldin, "An improved fire detection approach based on YOLO-v8 for smart cities." Neural Computing and Applications, vol. 35, no. 28, pp. 20939-20954, 2023.

[18] T. Wu and Y. Dong, "YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition," Applied Sciences, vol. 13, no. 24, p. 12977, 2023.

[19] X. Wang, H. Gao, Z. Jia, and Z. Li, "BL-YOLOv8: An Improved Road Defect Detection Model Based on YOLOv8," Sensors, vol. 23, no. 20, p. 8361, 2023.

[20] E. Soylu and T. Soylu, "A performance comparison of YOLOv8 models for traffic sign detection in the Robotaxi-full scale autonomous vehicle competition," Multimedia Tools and Applications, pp. 1-31, 2023.

[21] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2023, pp. 7464-7475.

[22] J. Chen et al., "Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2023, pp. 12021-12031.

[23] J. Park J, S. Woo S, J. Y. Lee, and I. S. Kweon, "A simple and light-weight attention module for convolutional neural networks," International journal of computer vision, vol. 128, no. 4, pp. 783-798, 2020.

[24] D. Ouyang et al., "Efficient Multi-Scale Attention Module with Cross-Spatial Learning," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1-5.

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

[26] J. Wang et al., "Deep high-resolution representation learning for visual recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 10, pp. 3349-3364, 2020.

[27] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, IEEE, 2017, pp. 618-626.