FSFYOLO: A Lightweight Model for Forest Smoke and Fire Detection

Yinglai HUANG, Jing LIU, Liusong YANG*
College of Computer and Control Engineering, Northeast Forestry University, Harbin, Heilongjiang 150040, China

***Abstract*—The detection and identification of forest smoke and fire are critical for forest fire prevention efforts. However, current forest smoke and fire target detection algorithms confront obstacles such as high memory usage, computational costs, and deployment difficulty. Regarding these key issues, this paper presents FSFYOLO, a lightweight forest smoke and fire detection model based on the YOLOv8s model. To efficiently extract key features from forest smoke and fire images while reducing computational redundancy, the lightweight network EfficientViT is used as the backbone network. A lightweight detection head, Partial Convolutional Head (PCHead), is designed using the shared parameters idea to greatly minimize the amount of parameters and computations by leveraging shared convolutional layers and branched processing, thus achieving the lightweight design of the model. In the neck network, a lightweight feature extraction module, C2f-FL, is built to more fully extract local features and surrounding contextual information to widen the receptive field. Additionally, a Coordinate Attention (CA) mechanism is integrated into both the backbone and neck networks to capture cross-channel information, directional awareness, as well as position-sensitive information, improving the model's capacity to precisely pinpoint fire and smoke in forests. The experimental outcomes results on our self-constructed forest smoke and fire dataset demonstrate that FSFYOLO reduces the number of parameters and computation by 47.6% and 60.9%, respectively, compared to the original model, while improving precision, recall, and mAP50 by 1.3%, 1.0%, and 1.0%, respectively. This demonstrates that FSFYOLO strikes a good compromise between model lightweighting and detection accuracy.**

***Keywords*—Forest smoke and fire; target detection; lightweight; YOLOv8; EfficientViT**

I. Introduction

Forests are one of Earth's most valuable natural resources. They not only provide essential materials and minerals for production, but also play a critical role in maintaining ecological balance, preventing and mitigating drought, and conserving water resources [1], [2], [3]. However, forest fires often go undetected until they have spread across vast areas, making them difficult or even impossible to control and extinguish [4]. Such fires can cause irreversible and devastating damage to the environment, including contributing to global warming, soil erosion, the extinction of rare species of flora and fauna, and impairing the forest's ability to self-regulate [5], [6]. Moreover, these fires pose significant risks to human life, infrastructure, and property [7]. Thus, quickly detecting forest fires and accurately identifying smoke areas is crucial for enabling firefighting personnel to take timely action, controlling the spread of the fire, which helps reduce the damage to ecosystems, infrastructure, and loss of life caused by forest fires [8].

The detection methods for forest fires are divided into smoke detection and flame detection [9]. Smoke, as an early indicator of fire, appears sooner, covers a larger volume, spreads faster, and is more easily detected by the naked eye [10], making it a critical clue for early fire detection. Flames, on the other hand, are essential for accurately pinpointing the fire's location [11], with color and varying shapes serving as key visual features that provide valuable information for firefighting efforts [12]. Therefore, integrating both smoke and fire detection significantly enhances the accuracy of forest fire monitoring, helping to protect forest resources and mitigate damage [13].

As a result of the quick development of computer vision technology, digital image processing techniques have been extensively used to identify forest fires. For the purpose of detection, early digital image processing techniques mainly extract the color, shape, and texture properties of smoke and flames. Unfortunately, manual feature extraction is heavily depended upon by these methods, and susceptibility to subjective human factors, as well as environmental complexities such as weather and lighting conditions, often leads to unsatisfactory detection performance [14]. Recently, the advances in deep learning have opened up new approaches to identifying forest fires. Deep learning models greatly improve the accuracy and robustness of fire detection models by providing benefits in terms of accuracy, detection speed, deployment flexibility, and adaptability to various fire characteristics [15], [16].

Despite promising progress in forest smoke and fire detection, several challenges remain unresolved. A key issue is how to achieve high detection accuracy, particularly in forest fire scenarios where the background is complex, interference is high, and the morphology of smoke and flames is highly variable. Furthermore, designing lightweight models for resource-constrained devices, such as edge and mobile devices, remains a critical research challenge. Thus, with the goal of addressing these concerns, this study proposes a lightweight forest smoke and fire detection model (FSFYOLO) built on YOLOv8s, which aims to reduce the computational load and parameter count through a lightweight design, enhancing detection accuracy. This makes it more feasible to deploy on resource-constrained devices, such as edge and mobile devices, allowing for rapid and accurate detection of smoke and fire in the early stages of a forest fire. This facilitates the issuance of timely warnings, reduces the time for rescue operations, and

minimizes the severe harm and losses caused by the spread of fires.

The main contributions of this paper are as follows: First, EfficientViT, a lightweight network, is employed as the backbone for YOLOv8s. Second, a new lightweight detection head, PCHead, is designed using the concept of shared parameters. Third, to fully extract local features and contextual information, the neck network is using the lightweight feature extraction module C2f-FL. Finally, a coordinate attention mechanism is introduced to capture direction-aware and position-sensitive information from forest smoke and fire images.

The remaining significant sections of this document are listed below: Section II provides a review of related research. Section III describes the improved model in this study. Section IV summarizes the dataset, experimental setup, parameters, and assessment measures that were employed during the studies. Section V performs pertinent experiments and discusses the findings. Finally, Section VI summarizes the entire effort of this study.

II. Related Works

With its strong feature extraction and pattern recognition capabilities, deep learning can automatically extract important information about forest fires from vast amounts of photos and videos. Therefore, deep learning-based recognition techniques have been used extensively in forest smoke and fire detection missions due to their notable benefits in forest smoke and fire recognition in recent years.

This research in [17] aimed to detect early forest fire smoke by improving the deformable DETR model. This approach improves detection capabilities for little or unobtrusive smoke by incorporating modules like Dense Pyramid Pooling. An iterative bounding box combination technique is described for producing more exact bounding boxes. In addition, a forest fire smoke dataset was created to validate the capability of the improved network. However, the improved model still has a larger number of parameters.

In the study [18], based on SqueezeNet, an efficient lightweight forest fire detection network was proposed. The model integrates Attention Gate (AG) units into the skip connections to enhance key features and suppress irrelevant information. Standard convolutions are replaced with depthwise convolutions, and a channel shuffle operation is introduced to optimize feature transmission. Although the model achieves good segmentation accuracy for forest fires, it may have limitations in broader fire detection tasks.

This paper in [19] described a methodology for detecting forest fires automatically that combines the Atom Search Optimizer (ASO) and deep transfer learning. The ResNet50 model is utilized to generate feature vectors, and the ASO is used to optimize the ResNet model's hyperparameters. A quasi-recurrent neural network model is used for fire categorization, with promising recognition and detection results.

The authors in [20] improved the YOLOv5 model to classify and detect forest fires. By incorporating the Weighted Bi-directional Feature Pyramid Network (BiFPN) and the Convolutional Block Attention Module (CBAM), the model enhances its ability to recognize various types of fires in complex backgrounds. The bounding box loss function adopts SIoU loss and introduces directionality to accelerate model convergence, effectively detecting different types of forest fires.

The research in [21] introduced a multi-task learning model for forest fire detection, which includes detection, classification, and segmentation tasks. The model includes a diagonal random origin swapping data augmentation approach that significantly enhances detection performance for small fire targets. When compared to single-task models, the upgraded model reduces missed and incorrect detections and has better feature extraction capabilities.

This paper in [22] improved the YOLOv8 model by adding a large-object detection head and introducing an Efficient Multi-Scale Attention (EMA) mechanism to reduce background noise and improve the identification of smoke targets and large-scale fires. The proposed path aggregation network bag structure further improves accuracy in detecting fires and smoke with uneven feature distributions and variable shapes. The improved model achieves higher detection accuracy.

III. Improved Methodology

A. The Forest Smoke and Fire YOLO Model

YOLOv8 is a high-performance object detection algorithm, available in five versions: n, s, m, l, and x, ranging from small to large. These versions have the same network structure, with differences only in network depth and width. YOLOv8s offers notable advantages, including strong feature extraction capabilities, high accuracy, compact size, and ease of deployment. Therefore, this study uses YOLOv8s as the baseline network and proposes a lightweight forest smoke and fire detection model, named FSFYOLO (Forest Smoke and Fire YOLO). Fig. 1 shows the FSFYOLO network structure.

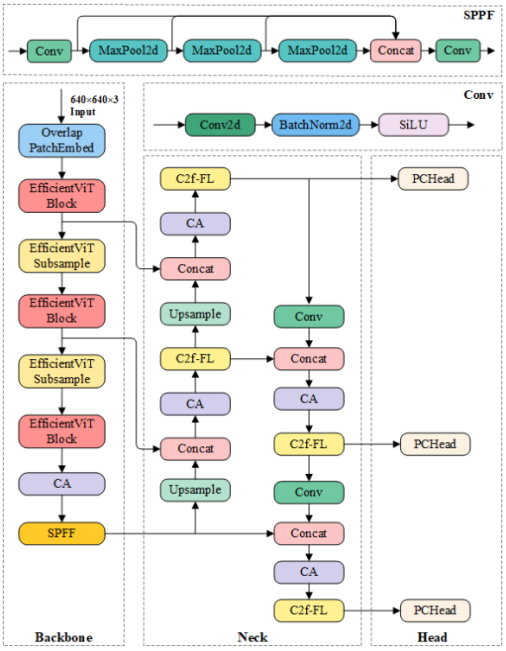

Fig. 1. Architecture of FSFYOLO.

The FSFYOLO network enhances the ability to accurately capture smoke and fire features in images while reducing computational redundancy by using the lightweight EfficientViT as the backbone network. Partial convolution is introduced, and a lightweight detection head, PCHead, is designed by sharing convolutional layers and branching processing, which productively minimizes the count of parameters and computational cost. In the neck network, to fully extract local features of smoke and fire as well as the surrounding contextual information, the LCFE block is proposed. This block is combined with the FasterNet Block and then integrated into the C2f module to form the lightweight feature extraction module C2f-FL, which expands the receptive field and lowers computational complexity. The coordinate attention, which extracts location sensitivity, directional awareness, and cross-channel information from fire and smoke images, is included into the backbone and neck networks. This mechanism filters out superfluous features in forest smoke and fire images, suppressing the impact of unrelated background information.

B. *EfficientViT*

The YOLOv8 backbone network, composed of multiple convolutional and pooling layers, results in high computational and storage costs. Furthermore, it struggles to accurately capture both local and global features of smoke and fire when processing cross-scale information. To address these issues, this study adopts EfficientViT as the backbone network of YOLOv8s. EfficientViT [23] is a high-speed vision transformer model that strikes a balance between speed and accuracy by optimizing memory efficiency and reducing attention computation redundancy. The EfficientViT network consists of overlapping patch embedding layers, EfficientViT blocks, and EfficientViT subsample layers, as shown in Fig. 2.

The input feature map first passes through the overlapping patch embedding layer, which divides the input into 16×16 patches and transforms them into vector tokens of a specified dimension, enabling better learning of the underlying features of the feature map.

The EfficientViT block is the core module of the EfficientViT network, with each block consisting of a sandwich layout formed by 2N FeedForward Network (FFN) layers, a token interaction layer, and Cascaded Group Attention (CGA). The token interaction layer, built with depthwise convolution (DWConv), is placed before the FFN layers to better capture local features in the image, thereby enhancing the model's overall performance. Unlike the conventional Multi-head Self-Attention Mechanism (MHSA), the CGA mechanism first divides the heads before generating Q, K, and V, and adds each head's output to the following head's input, thus providing each head with different features, improving the diversity of the attention maps. The outputs of all heads are spliced together and then passed through a linear layer to get the final output. Additionally, comparable to group convolution, this method lowers computational complexity and parameter count by lowering the Q, K, and V layers' input and output channels by a factor of 1/G, where G is the number of groups.

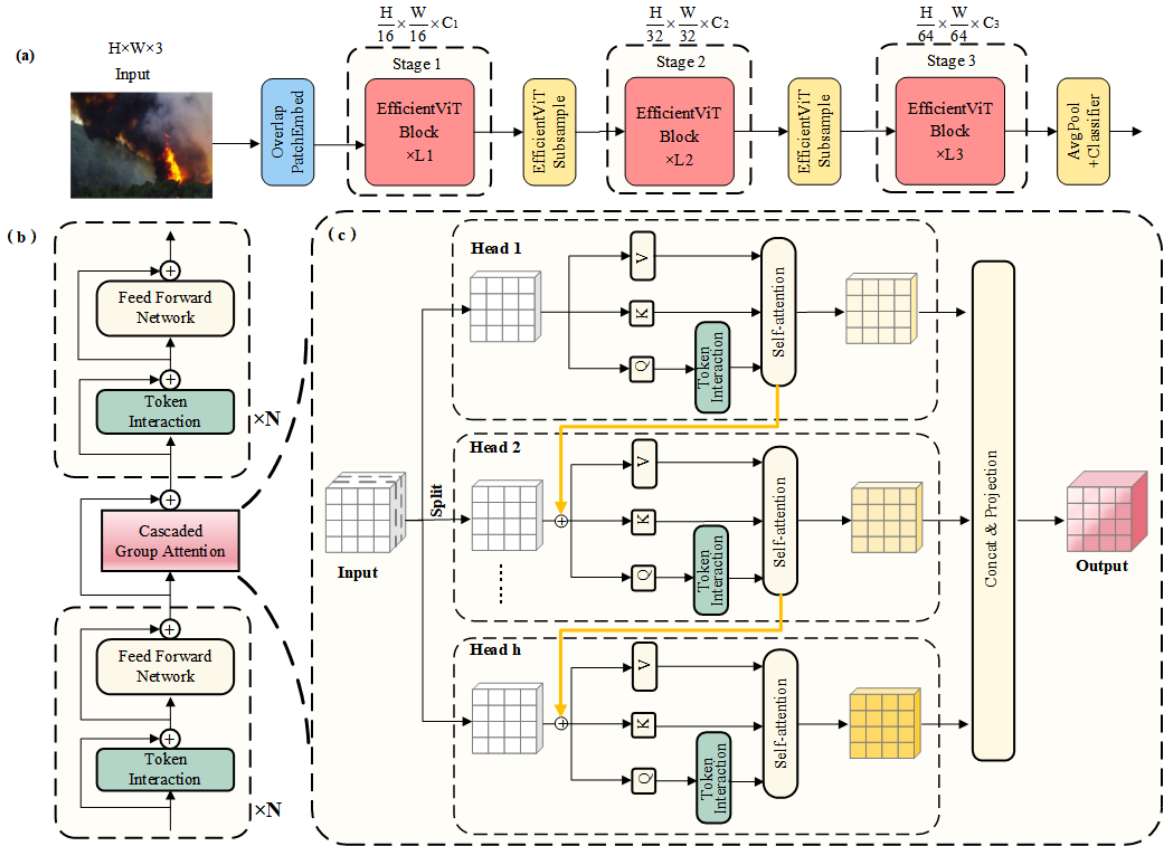

Fig. 2. (a) Architecture of EfficientViT; (b) Structure of sandwich layout block; (c) Structure of cascaded group attention.

CGA is expressed by the formula:

$$\tilde{X}_{ij} = Attn(X_{ij}W_{ij}^{Q}, X_{ij}W_{ij}^{K}, X_{ij}W_{ij}^{V}) \quad (1)$$

$$\tilde{X}_{i+1} = Concat[\tilde{X}_{ij}]_{j=1:h}W_{i}^{P} \quad (2)$$

$$X_{ij}^{'} = X_{ij} + \tilde{X}_{i(j-1)}, 1 < j \le h \quad (3)$$

In Eq. (1) and Eq. (2), $\tilde{X}_{ij}$ represents the output of X_{ij} after being processed by the self-attention mechanism in the j-th head. X_{ij} refers to the j-th slice of the input feature map X_i, i.e. $X_i=[X_{i1},X_{i2},...,X_{ih}]$, , where $1<j\le h$ and h represent the total number of attention heads. $W_{ij}^{Q},W_{ij}^{K},W_{ij}^{V}$ denotes the weight matrix, and W_i^P represents the linear layer.

In Eq. (3), $X_{ij}^{'}$ represents the sum of X_{ij}, the j-th slice of X_i and the output $\tilde{X}_{i(j-1)}$ from the (j-1)-th head, as obtained through Eq. (1) and Eq. (2). At this point, $X_{ij}^{'}$ replaces X_{ij} as the j-th head's original input feature map.

The EfficientViT subsampling layer downscales the feature map. Unlike traditional Transformer models, EfficientViT uses an inverted residual block in the subsample block instead of a self-attention layer, reducing potential information loss during downsampling.

C. The Lightweight Design of the Detection Head

Both branches of the YOLOv8 detecting head start with two 3×3 convolution modules, then a Conv2d module. Finally, they calculate the Cls and Bbox losses individually. Fig. 3 shows the specific structure of the YOLOv8 detection head.

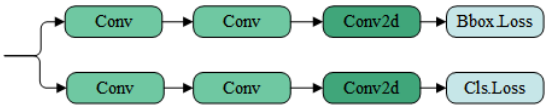

Fig. 3. YOLOv8 detection head structure.

To detect smoke and fire targets at different scales, the YOLOv8 detection head requires more convolution operations to process multi-scale feature maps, which increases the depth and number of parameters in the network. We adopt Partial Convolution (PConv) from the FasterNet Block [24] to modify the lightweight design of the YOLOv8s detection head to address these issues. Fig. 4 displays the FasterNet Block and PConv network structure diagram.

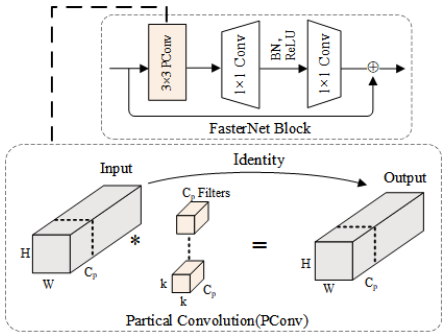

Fig. 4. Network structure of FasterBlock and PConv.

Comprising three layers, the FasterNet Block is composed of PConv layer, 1×1 convolutional layer, and 1×1 2D convolutional layer. PConv selectively applies regular convolution to specific input channels to extract spatial features; the remaining input channels remain unaltered and are directly translated to the output channels, resulting in significant computational redundancy reduction.

The new detection head first shares a PConv layer and a 1×1 convolutional layer, and then branches into two paths. Each path computes the Bbox loss and Cls loss, respectively, after passing through a Conv2d module. This new detection head is called PCHead (Partial Convolutional Head), and its structure is shown in Fig. 5.

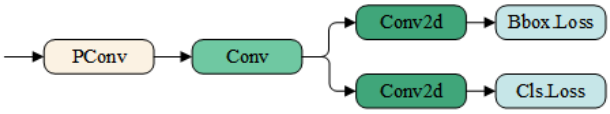

Fig. 5. PCHead structure.

D. C2f-FL Lightweight Feature Extraction Module

1) Local and Contextual Feature Extraction (LCFE) block: Traditional convolution operations confront issues such as information loss and a limited receptive field when capturing the diverse features of forest smoke and fires in various conditions. This hinders the model's ability to effectively extract local forest smoke and fire features and the corresponding surrounding contextual information, which in turn affects the extraction of fire-related features and restricts the expansion of the receptive field. In order to overcome this limitation, we propose a Local and Contextual Feature Extraction (LCFE) block, as illustrated in Fig. 6. The LCFE block is designed to efficiently capture local forest smoke and fire features while also extracting related contextual information, broadening the model's receptive field, and improving the network's capacity to identify fire features.

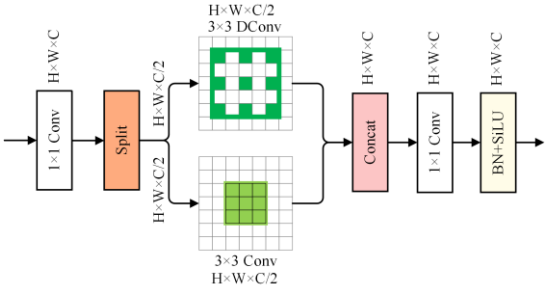

Fig. 6. LCFE block structure.

The LCFE block integrates information from several channels of the input feature map using a 1×1 convolution, thus enabling information interaction within the receptive field. The feature map is then partitioned evenly along the channel dimension into two sub-feature maps with an equal number of channels. One sub-feature map uses 3×3 conventional convolution to extract local features from eight neighboring vectors, while the other uses 3×3 dilated convolution to capture contextual information as well as widen the receptive field. Subsequently, concatenation of the extracted characteristics

occurs along the channel dimension; the local features and contextual information are combined using a 1×1 convolution. The fused features are then normalized and nonlinearized using batch normalization (BN) and the SiLU activation function, yielding the final output. Through this design method, the LCFE block is able to fully capture the intricate characteristics of smoke and fires, improving the model's recognition capabilities.

$$f_1, f_2 = Split(F_n(X)) \tag{4}$$

$$f = W(F_n(Concat(F_n(f_1), F_m^d(f_2)))) \tag{5}$$

In this context, X represents the input feature map, F_n denotes an n × n convolution operation, and F_m^d refers to a dilated convolution with a dilation rate d and a kernel size of m × m. f_1 and f_2 are the output feature maps, uniformly divided along the channel dimension. $W(\cdot)$ denotes the Batch Normalization (BN) and SiLU activation procedures; $Concat(\cdot)$ denotes concatenation along the channel dimension; $Split(\cdot)$ denotes the operation of splitting the feature map along the channel dimension. Ultimately, after being processed by the LCFE block, the output feature map f is obtained.

2) C2f-FL module: In YOLOv8, the C2f module makes use of a bottleneck structure made up of many convolutional layers, which requires repetitive dimensionality reduction and channel expansion of the input feature map. This approach can cause information loss while also increasing the model's computational complexity and parameter count. To tackle these issues, this study introduces the FasterNet Block, which reduces the model's parameter load and computational complexity while achieving efficient spatial feature extraction. Additionally, the LCFE block is incorporated into the forward propagation of the FasterNet Block, forming the FL block. This FL block serves as the bottleneck module within the C2f module of the neck network, resulting in the new lightweight feature extraction module, C2f-FL. This improvement reduces the parameters and computation required for smoke and flame feature extraction, effectively enhancing both target feature extraction and overall model efficiency. The structure of the FL block and C2f-FL is shown in Fig. 7.

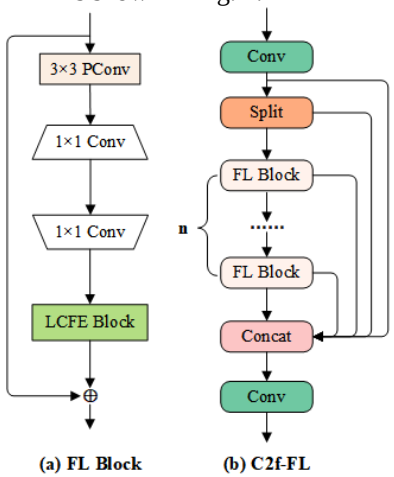

Fig. 7. Structure of FL block and C2f-FL module.

E. Coordinate Attention (CA)

Given the complexity of background textures and the abundance of irrelevant information in forest smoke and fire images, existing attention mechanisms often focus only on channel dependencies, neglecting the importance of spatial information. This results in significant redundancy in the spatial dimension of the extracted feature maps. To address this issue, a coordinate attention (CA) [25] mechanism is integrated into the network: before the backbone network's SPPF module and after the neck network's feature fusion module. This approach enhances the extraction of both channel and spatial information, effectively filtering out redundant features in forest smoke and fire images. Fig. 8 depicts the structure of the coordinate attention mechanism.

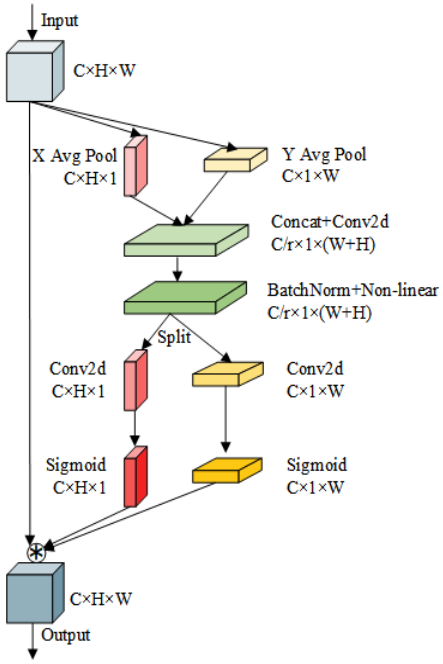

Fig. 8. Structure of the coordinate attention mechanism.

Two-dimensional global pooling is broken down by CA into two one-dimensional global poolings in separate directions. The input feature map is aggregated separately along the vertical and horizontal directions, resulting in two independent feature maps that are direction-aware and position-sensitive. This approach enables the capture of long-range dependencies along different spatial dimensions while avoiding the loss of spatial details, thus accurately preserving positional information from the original image. Subsequently, these direction-specific feature maps undergo operations such as stacking and normalization to encode attention maps. Finally, these attention maps are applied to the input feature maps in a complementary manner through elemental multiplication.

IV. Dataset and Experimental Setup

A. Dataset

One of the challenges in this study is the absence of a publicly available, unified forest fire dataset. To address this, a custom dataset containing instances of smoke and fire was created. The images in the dataset come from two sources: the

first involves collecting forest smoke and fire images and videos from the Internet, extracting one frame every 5 frames of the videos to create still images; the second source includes partial images from the FLAME dataset [26] published by Northern Arizona University. The dataset contains a total of 3,895 images, all manually labeled using LabelImg. It is separated into two sets: training and validation, in a 4:1 ratio, with 3,116 images for training and 779 for validation. Fig. 9 depicts sample photos from the collection.

Fig. 9. Partial experimental data.

B. Experimental Environment

The system used for the experiments in this study runs on Windows 11 with 16GB of RAM. The hardware includes a 13th Gen Intel(R) Core(TM) i7-13700H CPU and an NVIDIA GeForce RTX 4060 Laptop GPU. Pycharm was used as the software environment, and the PyTorch deep learning framework, based on Python, was utilized. The Python version used was 3.10.

C. Hyperparameter Settings

In this work, all models' hyperparameters are kept consistent during training and validation. The experimental parameters utilized for network training are listed in Table I:

TABLE I. EXPERIMENTAL HYPERPARAMETERS

Parameter name	Configuration
Size of the input image	640×640
Optimizer	SGD
Batchsize	16
Epochs	150
Momentum	0.937
Lr0	0.01
Weight decay	0.0005

D. Evaluation Criterion

Precision (P), Recall (R), Mean Average Precision (mAP), Giga Floating Point Operations per Second (GFLOPs), number of parameters (Params), and model weight file size (in MB) were used as assessment measures to analyze the network performance in this study.

Precision refers to the fraction of true positive samples among those predicted as positive by the model. It is calculated using the following formula:

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

Recall evaluates the proportion of true positive samples that are correctly predicted among all actual positive samples. Its calculation is as follows:

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

where TP, FP and FN denote the number of true, false positive and false negative cases, respectively.

AP represents the average precision for a single target category. It is calculated using the following formula:

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

mAP is the average of AP values across all categories. It is calculated using the following formula:

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

The Intersection over Union (IoU) represents the ratio of the intersection area to the union area between the predicted bounding box and the ground truth bounding box. By setting different IoU thresholds, corresponding mAP values can be obtained. In this study, mAP at IoU = 0.5 (mAP50) is adopted as the evaluation metric to assess the model's localization and classification capabilities for detected objects, providing a comprehensive evaluation of its overall detection performance.

GFLOPs is an indication for determining a model's computational complexity; a lower GFLOPs value indicates reduced computational cost. The parameter count measures the size of the model, and fewer parameters help accelerate the training process. The number of bytes in the file created during training is referred to as the weight file size for the model, with smaller weight files facilitating deployment and operation on resource-constrained devices.

V. RESULTS AND DISCUSSION

A. Ablation Experiments

To verify the effectiveness of the proposed detection model FSFYOLO in smoke and fire detection tasks, four groups of ablation experiments were conducted. Throughout these tests, the same dataset and hyperparameters were used to train each model, with only the modules under evaluation being changed. Table II presents the experimental results obtained.

The ablation experiment outcomes show that incorporating EfficientViT as the backbone network for YOLOv8s leads to reductions of 24.6% in the number of parameters, 28.2% in GFLOPs, and 22.0% in the size of the weight file compared to the original network. This proves that EfficientViT effectively reduces the model's complexity and computational burden. Additionally, by using the PCHead as the detection head, which shares convolutional layers and employs branch-based processing, the parameters and computational load are further reduced on the YOLOv8s-EfficientViT foundation, greatly decreasing model complexity. In addition, the network is further lightweighted by using the lightweight feature

extraction module C2f-FL, resulting in reductions of 16.9% in parameters, 17.9% in GFLOPs, and 15.3% in weight file size compared to YOLOv8s-EfficientViT-PCHead. The recall and mAP50 of the model compare favorably to the original network in terms of performance, suggesting that adding the C2f-FL module can recognize target items more thoroughly and lower the miss detection rate. Lastly, by adding the CA mechanism, the model's precision increases by 1%, reducing false positives in smoke and fire detection, as it better captures directional and positional information from the images. Importantly, the inclusion of CA does not significantly increase the parameter count or computation load, demonstrating that performance improvements are achieved without a notable increase in model size. The experimental results confirm that the FSFYOLO model proposed achieves superior performance in forest smoke and fire image recognition tasks.

TABLE II. THE RESULTS OF ABLATION EXPERIMENTS

YOLOv8s	EfficientViT	PCHead	C2f-FL	CA	P	R	mAP50	Params/10^6	GFLOPs	Weight File Size/MB
+					0.904	0.873	0.923	11.13	28.4	21.4
+	+				0.896	0.872	0.922	8.39	20.4	16.7
+	+	+			0.906	0.868	0.926	6.86	13.4	13.7
+	+	+	+		0.907	0.886	0.931	5.70	11.0	11.6
+	+	+	+	+	0.917	0.883	0.933	5.83	11.1	11.8

B. Comparative Experiments

To further validate the effectiveness of the improved network in detecting forest smoke and fire, a comparative analysis was conducted under the same experimental settings, dataset, and training strategies, using other mainstream object detection models. These models include SSD [27], YOLOX [28], YOLOv5, YOLOv6 [29], YOLOv7 [30], RT-DETR [31], the improved YOLOv5s model by Yang et al. [32], and the improved YOLOv8s model by Kong et al. [33]. Performance indicators such as precision (P), recall (R), mAP50, number of parameters, GFLOPs, and model weight file size were used as evaluation criteria. Table III presents the experimental results obtained.

The experimental results indicate that the SSD, RT-DETR, and YOLOv6s detection algorithms have relatively large parameter counts and computational costs. These factors result in larger model weight files and impose higher demands on hardware resources. In contrast, the parameter count of the FSFYOLO model is approximately one-fourth that of SSD, one-fifth that of RT-DETR, and one-third that of YOLOv6s. Therefore, SSD, RT-DETR, and YOLOv6s are deemed unsuitable for lightweight, real-time detection of forest smoke and fire. The YOLOv5s, YOLOv7, and FSFYOLO models include some identical feature extraction modules, such as the Conv module, and share similar network architectures. As a result, they exhibit only minor differences in parameter count, computational complexity, and model weight size. However, FSFYOLO achieves significant advantages by optimizing the C2f feature extraction module and introducing the CA mechanism to enhance feature extraction. Consequently, FSFYOLO outperforms YOLOv5s and YOLOv7 in terms of accuracy, recall, and mAP50, demonstrating a clear and distinct advantage. Although YOLOXs has a computational cost close to that of FSFYOLO, its parameter count is 34.8% higher, and its weight file is 2.9 times larger, with lower precision, recall, and mAP50 compared to FSFYOLO. Yang et al. [32] proposed an improved YOLOv5s model by adding C3Ghost and Ghost modules, resulting in parameter counts and GFLOPs of 3.78×10^6 and 8.3, respectively. However, due to the Ghost module only performing standard convolution operations on half of the spatial features, the model exhibits limitations in capturing detailed features for complex forest fire smoke detection tasks. Consequently, its precision, recall, and mAP50 do not surpass FSFYOLO. Similarly, Kong et al. [33] introduced Efficient Multi-Scale Attention (EMA) and GSConv to optimize YOLOv8s, reducing its parameter count, GFLOPs, and weight file size, but all performance metrics remain lower than those of FSFYOLO. Compared to the baseline YOLOv8s, the improved network outperforms the original in all key metrics. Based on a comprehensive analysis of all evaluation metrics, FSFYOLO offers significantly enhanced detection capabilities and is better suited for forest fire detection compared to other models.

C. Visual Analysis

To better further confirm the FSFYOLO network performance in forest smoke and fire detection tasks, we conducted a visual analysis utilizing the Gradient-weighted Class Activation Mapping (Grad-CAM) [34] approach. Grad-CAM heatmaps show that deeper colors indicate more attention to the respective locations, while lighter colors suggest less attention. We chose a few typical photos from the experimental validation set to compare the detection performance of YOLOv8s with FSFYOLO under various fire conditions. Fig. 10 displays the detection results, with (a) and (b) containing both fire and smoke, (c) containing only smoke, and (d) containing only fire.

It is evident from the heatmaps that YOLOv8s and FSFYOLO are equally adept at identifying and pinpointing target locations that are characterized by smoke and fire. But the FSFYOLO model has a higher confidence level, focuses more accurately on the regions of fire and smoke in the image, and focuses on regions very close to the actual smoke and fire shapes. Specifically, YOLOv8s may be affected by complex background interference present in forest fire images, leading to a dispersed focus on target objects. In contrast, FSFYOLO, through its lightweight design and optimized feature extraction modules, not only suppresses background noise but also enhances the precise capture of target features, improving accuracy. This again demonstrates the effectiveness of FSFYOLO in the task of forest smoke and fire detection.

TABLE III. THE RESULTS OF COMPARATIVE EXPERIMENTS

Model	P	R	mAP50	Params/10^6	GFLOPs	Weight File Size/MB
SSD	0.900	0.667	0.848	23.75	136.8	91.1
YOLOXs	0.905	0.864	0.908	8.94	13.4	34.3
YOLOv5s	0.899	0.879	0.920	7.03	15.8	13.7
YOLOv6s	0.898	0.884	0.917	18.50	45.3	38.7
YOLOv7	0.904	0.870	0.916	6.02	13.2	11.7
RT-DETR	0.890	0.837	0.901	32.81	108.0	63.0
Yang et al. [32]	0.898	0.874	0.917	3.78	8.3	7.7
Kong et al. [33]	0.897	0.879	0.922	8.56	20.9	16.6
YOLOv8s	0.904	0.873	0.923	11.13	28.4	21.4
FSFYOLO	0.917	0.883	0.933	5.83	11.1	11.8

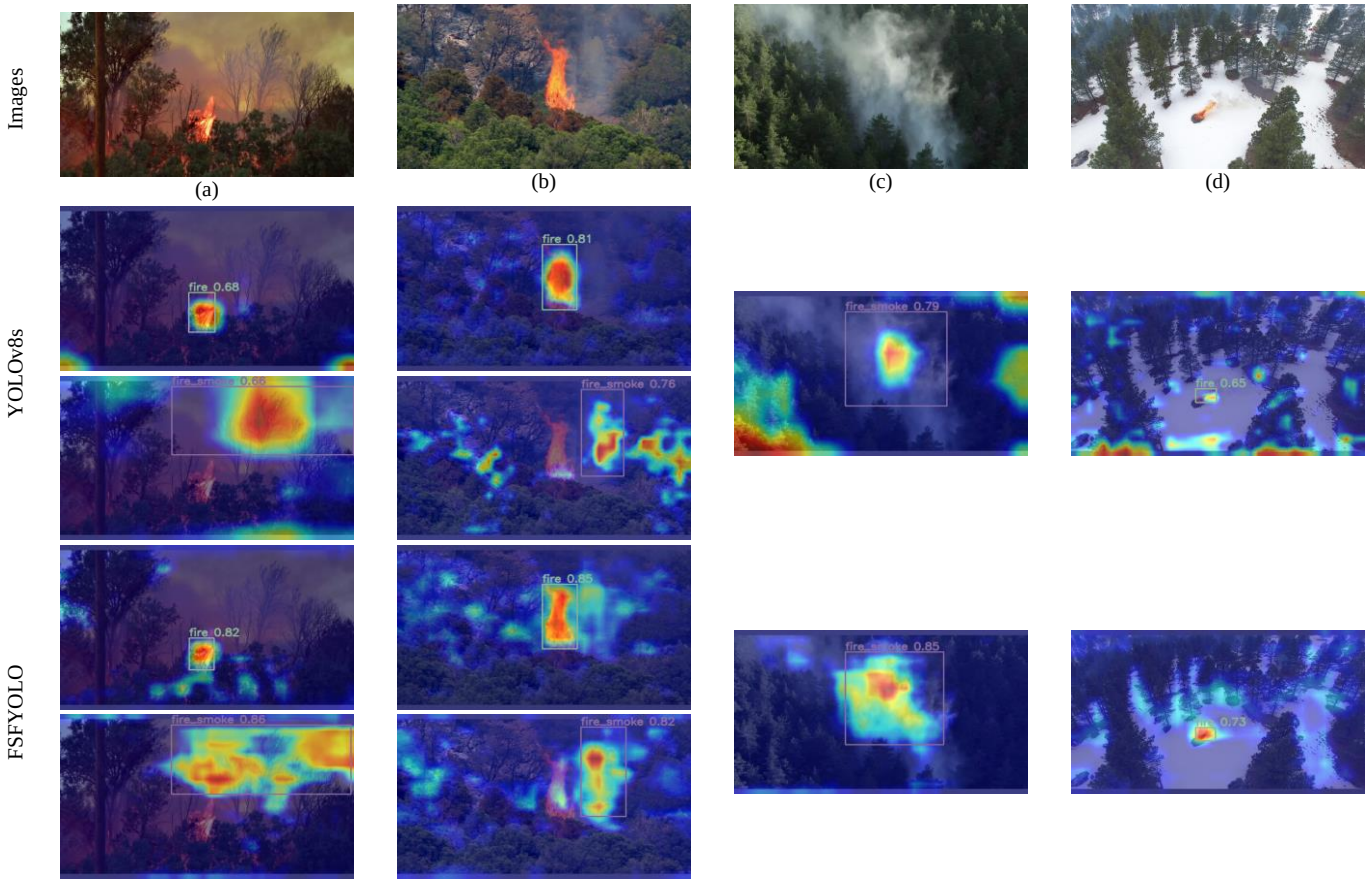

Fig. 10. Schematic visualization of YOLOv8s and FSFYOLO detection results.

D. Generalization Verification

In this experiment, the proposed FSFYOLO model was evaluated for generalization performance using the open FM-VOC Dataset18644 [35], which contains instances of smoke and fire. The dataset contains 16,844 photos depicting fires in a variety of contexts, including building fires, grassland fires, indoor fires, forest fires, and road fires. Experiments were conducted to compare the YOLOv8s and FSFYOLO models on the FM-VOC Dataset18644. The experimental conditions, parameter settings, and evaluation metrics were consistent with those used in other experiments in this study. The results are shown in Table IV.

As Table IV illustrates, compared to the baseline YOLOv8s model, the FSFYOLO model achieved improvements of 1.3%, 2.3%, and 1.3% in precision, recall, and mAP50, respectively, on the FM-VOC Dataset18644. Additionally, the FSFYOLO model demonstrated reductions of 47.6%, 60.9%, and 44.9% in parameter count, GFLOPs, and model weight size, respectively. The experimental results confirm that the FSFYOLO model achieves a good balance between detection performance and lightweight design, effectively detecting and recognizing fires

across different scenarios. Therefore, the FSFYOLO model exhibits excellent generalization performance, which will broaden the application of target detection in forest fire scenarios.

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS ON THE FM-VOC DATASET18644

Evaluation Metrics	YOLOv8s	FSFYOLO
P	0.917	0.930
R	0.872	0.895
mAP50	0.939	0.952
Params/10^6	11.13	5.83
GFLOPs	28.4	11.1
Weight File Size/MB	21.4	11.8

VI. CONCLUSION

To accomplish precise and timely forest fire detection, we present FSFYOLO, a lightweight forest smoke and fire detection network based on YOLOv8s. To begin, EfficientViT, a lightweight transformer network, serves as the backbone network to improve network feature extraction capability. Second, a lightweight detection head, PCHead, is designed using the shared parameters idea, which decreases the model's complexity while preserving detection performance. Third, the lightweight feature extraction module C2f-FL is introduced to effectively capture local features of forest smoke and fire, as well as relevant surrounding contextual information, which achieves the dual enhancement of model computation efficiency and feature extraction capability. Finally, a coordinate attention mechanism is integrated to extract both channel and spatial location information, filtering out superfluous features in forest fire images. Experimental validation shows that the FSFYOLO network achieves higher accuracy than other networks, with significantly reduced parameter count and computational cost, satisfying real-time needs for forest smoke and fire detection. Additionally, the FSFYOLO network is easily deployable on resource-constrained devices, providing an effective method for forest fire detection. However, this study still has some limitations. For example, the forest fire dataset used in the experiments is relatively small, and the range of scene coverage is insufficient. Moreover, although the FSFYOLO network decreases the number of parameters and computing costs while increasing accuracy, there is still room for optimization in the model structure and performance. In future work, we will focus on expanding the forest fire dataset to cover more complex scenarios and exploring more efficient, parameter-reduced methods to further enhance the performance and precision of forest fire detection models.

REFERENCES

[1] A. Khan, B. Hassan, S. Khan, R. Ahmed and A. Abuassba, "DeepFire: A novel dataset and deep transfer learning benchmark for forest fire detection," Mob Inf Syst, vol. 2022, no.1, p. 5358359, 2022.

[2] S.D. Wang, L.L. Miao and G.X. Peng, "An improved algorithm for forest fire detection using HJ data," Procedia Environ Sci, vol. 13, pp. 140-150, 2012.

[3] M.S. Anggreainy, B. Kurniawan and F.I. Kurniadi, "Reduced false alarm for forest fires detection and monitoring using fuzzy logic algorithm," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 7, pp. 535-541 2022.

[4] A.A. Alkhatib, "A review on forest fire detection techniques," Int J Distrib Sens Netw, vol. 10, no. 3, p. 597368, 2014.

[5] S.T. Seydi, V. Saeidi, B. Kalantar, N. Ueda and A.A. Halin, "Fire - Net: A Deep Learning Framework for Active Forest Fire Detection," J Sensors, vol. 2022, no. 1, p. 8044390, 2022.

[6] M. Yandouzi, M. Grari, I. Idrissi, M. Boukabous, O. Mous-saoui, M. Azizi, K. Ghoumid and A.K. Elmiad, "Forest fires detection using deep transfer learning," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 8, pp. 268-275, 2022.

[7] Q. Jiao, M. Fan, J. Tao, W. Wang, D. Liu and P. Wang, "Forest fire patterns and lightning-caused forest fire detection in Heilongjiang Province of China using satellite data," Fire, vol. 6, no. 4, p. 166, 2023.

[8] Y. Peng and Y. Wang, "Real-time forest smoke detection using hand-designed features and deep learning," Comput Electron Agric, vol. 167, p. 105029, 2019.

[9] J. Zhan, Y. Hu, G. Zhou, Y. Wang, W. Cai and L. Li, "A high-precision forest fire smoke detection approach based on ARGNet," Comput Electron Agric, vol, 196, p. 106874, 2022.

[10] X. Zheng, F. Chen, L. Lou, P. Cheng and Y. Huang, "Real-time detection of full-scale forest fire smoke based on deep convolution neural network," Remote Sensing, vol. 14, no. 3, p. 536, 2022.

[11] P.E.N.G. Bo, "Research on classification of forest fire risk based on GIS technology in Xichang City, Sichuan Province," Journal of Sichuan Forestry Science and Technology, vol. 42, no. 5, pp. 53-57, 2021.

[12] C. Emmy Prema, S.S. Vinsley and S. Suresh, "Efficient flame detection based on static and dynamic texture analysis in forest fire detection," Fire technology, vol. 54, pp. 255-288, 2018.

[13] A. Gaur, A. Singh, A. Kumar, A. Kumar and K. Kapoor, "Video flame and smoke based fire detection algorithms: A literature review," Fire technology, vol. 56, pp. 1943-1980, 2020.

[14] J. Lin, H. Lin and F. Wang, "A semi-supervised method for real-time forest fire detection algorithm based on adaptively spatial feature fusion," Forests, vol. 14, no. 2, p. 361, 2023.

[15] J. Lin, H. Lin and F. Wang, "STPM_SAHI: A Small-Target forest fire detection model based on Swin Transformer and Slicing Aided Hyper inference," Forests, vol. 13, no.10, p. 1603, 2022.

[16] Z. Guan, X. Miao, Y. Mu, Q. Sun, Q. Ye and D. Gao, "Forest fire segmentation from Aerial Imagery data Using an improved instance segmentation model," Remote Sensing, vol. 14, no. 13, p. 3159, 2022.

[17] J. Huang, J. Zhou, H. Yang, Y. Liu and H. Liu, "A small-target forest fire smoke detection model based on deformable transformer for end-to-end object detection," Forests, vol. 14, no. 1, p. 162, 2023.

[18] J. Zhang, H. Zhu, P. Wang and X. Ling, "ATT squeeze U-Net: A lightweight network for forest fire detection and recognition," IEEE Access, vol. 9, pp. 10858-10870, 2021.

[19] K. Alice, A. Thillaivanan, G.R.K. Rao, S. Rajalakshmi, K. Singh and R. Rastogi, "Automated forest fire detection using atom search optimizer with deep transfer learning model," ICAAIC, pp. 222-227, IEEE, 2023.

[20] Q. Xue, H. Lin and F. Wang, "Fcdm: an improved forest fire classification and detection model based on yolov5," Forests, vol. 13, no. 12, p. 2129, 2022.

[21] K. Lu,, J. Huang, J. Li, J. Zhou, X. Chen and Y. Liu, "MTL-FFDET: A multi-task learning-based model for forest fire detection," Forests, vol. 13, no. 9, p. 1448, 2022.

[22] T. Zhang, F. Wang, W. Wang, Q. Zhao, W. Ning and H. Wu, "Research on fire smoke detection algorithm based on improved YOLOv8," IEEE Access, vol. 14, pp. 117354-117362, 2024.

[23] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in Proceedings of the IEEE/ CVF conference on computer vision and pattern recognition, pp. 14420-14430, 2023.

[24] J. Chen, S.H. Kao, H. He, W. Zhuo, S. Wen, C.H. Lee and S.H.G. Chan, "Run, don't walk: chasing higher FLOPS for faster neural networks," in

Proceedings of the IEEE/ CVF conference on computer vision and pattern recognition, pp. 12021-12031, 2023.

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

[26] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P.Z. Fulé and E. Blasch, "Aerial imagery pile burn detection using deep learning: The FLAME dataset," Computer Networks, vol. 193, p. 108001, 2021.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu and A.C. Berg, "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21-37, Springer International Publishing, 2016.

[28] G. Zheng, L. Songtao, W. Feng, L. Zeming and S. Jian, "YOLOX: Exceeding YOLO series in 2021," arXiv preprint arXiv: 2107.08430, 2021.

[29] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie and Y. Li, "YOLOv6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv: 2209.02976, 2022.

[30] 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, pp. 7464-7475, 2023.

[31] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu and J. Chen, "Detrs beat yolos on real-time object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965-16974, 2024.

[32] J. Yang, W. Zhu, T. Sun, X. Ren and F. Liu, "Lightweight forest smoke and fire detection algorithm based on improved YOLOv5," PLoS one, vol. 18, no. 9, p. e0291359, 2023.

[33] D. Kong, Y. Li and M. Duan, "Fire and smoke real-time detection algorithm for coal mines based on improved YOLOv8s," Plos one, vol. 19, no. 4, p. e0300502, 2024.

[34] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.

[35] X. Geng, Y. Su, X. Cao, H. Li and L. Liu, "YOLOFM: an improved fire and smoke object detection algorithm based on YOLOv5n," Scientific Reports, vol. 14, no. 1, p. 4543, 2024.