fire

MDPI

Article

Anchor-Free Smoke and Flame Recognition Algorithm with Multi-Loss

Gang Li [1], Peng Chen [1,*], Chuanyun Xu [2,*], Chengjie Sun [1] and Yingli Ma [1]

[1] School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China; ligang@cqut.edu.cn (G.L.); sunchengjie@stu.cqut.edu.cn (C.S.); myl@stu.cqut.edu.cn (Y.M.)
[2] School of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
* Correspondence: peng@2020.cqut.edu.cn (P.C.); xcy@cqnu.edu.cn (C.X.)

Abstract: Fire perception based on machine vision is essential for improving social safety. Object recognition based on deep learning has become the mainstream smoke and flame recognition method. However, the existing anchor-based smoke and flame recognition algorithms are not accurate enough for localization due to the irregular shapes, unclear contours, and large-scale changes in smoke and flames. For this problem, we propose a new anchor-free smoke and flame recognition algorithm, which improves the object detection network in two dimensions. First, we propose a channel attention path aggregation network (CAPAN), which forces the network to focus on the channel features with foreground information. Second, we propose a multi-loss function. The classification loss, the regression loss, the distribution focal loss (DFL), and the loss for the centerness branch are fused to enable the network to learn a more accurate distribution for the locations of the bounding boxes. Our method attains a promising performance compared with the state-of-the-art object detectors; the recognition accuracy improves by 5% for the mAP, 8.3% for the flame AP50, and 2.1% for the smoke AP50 compared with the baseline model. Overall, the algorithm proposed in this paper significantly improves the accuracy of the object detection network in the smoke and flame recognition scenario and can provide real-time fire recognition.

Keywords: smoke and flame recognition; anchor-free; path aggregation network; multi-loss

Citation: Li, G.; Chen, P.; Xu, C.; Sun, C.; Ma, Y. Anchor-Free Smoke and Flame Recognition Algorithm with Multi-Loss. *Fire* **2023**, *6*, 225. https://doi.org/10.3390/fire6060225

Academic Editor: Grant Williamson

Received: 3 May 2023
Revised: 2 June 2023
Accepted: 2 June 2023
Published: 4 June 2023

1. Introduction

Fire is a devastating natural disaster that often causes enormous social and ecological damage and severe economic losses. Early fire recognition, warning, and autonomous response can effectively reduce fire damage [1]. In typical buildings, physical signal-based detectors, such as smoke detectors, heat-release infrared flame detectors, and ultraviolet flame detectors, are widely used for fire alarms. However, these conventional physical sensors require proximity to the source of fire; thus, they cannot be used in environments such as large-space buildings and open spaces (e.g., forests, construction sites, ports, and grasslands) and cannot provide accurate fire details such as the fire location, size, and extent of the burning. Therefore, computer vision-based smoke and flame recognition are essential tasks for modern surveillance systems.

Fire recognition based on computer vision includes smoke and flame recognition. Smoke recognition is mainly based on features such as the shape and color of the smoke and processing methods, include wavelet transform, neural network, and fuzzy algorithm. Flame recognition is mainly based on features such as the flame color, shape, and dynamic characteristics, and the processing methods include the neural network, Support Vector Machine, Markov model, and expert system. Binti et al. [2] performed fire recognition based on RGB and YCbCr features in 2015. Li et al. [3] 2017 proposed a fire recognition algorithm based on flame color, dynamics, and flicker characteristics. Wang et al. [4] extracted a variety of features for forest fire recognition in 2019, including color, texture, area, and shape features. All of the above researchers constructed feature extractors to

improve the accuracy of smoke and flame recognition, but these manually designed feature extractors are highly redundant, which makes it difficult to accurately recognize fire.

In recent years, several deep learning-based methods for smoke and flame classification and detection have been proposed to improve the accuracy and efficiency of smoke and flame recognition. Muhammad et al. [5] used a transfer learning method to improve performance by using MobileNet as a backbone network and fine-tuning the fully connected layer on a small flame dataset. Sharma et al. [6] combined a pretrained Visual Geometry Group-16 (VGG-16) model and Residual Network-50 (ResNet-50) to develop a fire recognition system. With the increase in object detection methods, fire recognition requires not only determining whether a fire has occurred but also locating and extracting the exact area of the fire. Several different object detection methods have been proposed for the fire recognition task. Wu et al. [7] proposed a new architecture based on the You Only Look Once (YOLO) algorithm that utilizes the newly added smoke class and flame area changes to minimize false recognition. Xie et al. [8] proposed a video fire recognition method that uses deep static features extracted by Convolutional Neural Networks (CNNs) and dynamic features based on motion flicker extracted by background subtraction and flicker detection to improve accuracy. The work in Ref. [9] developed an image-based fire detection algorithm based on the YOLO-V3 network to detect smoke and flames. Despite their success, most previous studies focused on model optimization, and the complex situations in real environments need to be considered. We observe the following three problems in existing practices: (1) existing algorithms mainly recognize fire through flames, but smoke is more important for fire recognition; (2) anchor boxes with fixed shapes and sizes cannot meet the recognition of smoke and flames with variable shapes and sizes; (3) the algorithms are not accurate enough to locate smoke and flames with uncertain and ambiguous boundaries. To address the shortcomings of the current fire recognition algorithms, and borrowing from the anchor-free network architecture in the field of object detection, this paper proposes an anchor-free smoke and flame recognition algorithm.

The main contributions are summarized as follows:

- Our algorithm employs a pixel-by-pixel approach to directly predict the bounding box locations and corresponding class of the objects, resulting in faster training and testing as well as a lower training memory footprint.
- Incorporating the channel attention mechanism and new connections into the multi-scale feature fusion network makes the network focus on the channel features with foreground information and improves the accuracy of the algorithm.
- We use a multi-loss fusion method to provide more accurate and informative bounding box locations of smoke and flame objects by modeling the flexible distribution for bounding boxes.

2. Related Work

2.1. Smoke and Flame Recognition Algorithms

Existing vision-based smoke and flame recognition algorithms are divided into image processing-based and deep learning-based methods. Zhang et al. [10] used an algorithm combining Fast Fourier Transform (FFT) and wavelet to analyze video flame contours. Jiang et al. [11] used the improved Canny edge detector to detect the fire region. Kosmas et al. [12] implemented a fire recognition system modeling fire behavior by employing various spatiotemporal features using linear dynamical systems and a bag-of-systems approach. Toulouse et al. [13] developed a method that focused on detecting the geometric features of flames, such as the location and length, and classified the fire images' pixels based on the flame color and presence of smoke. The work in [14,15] developed vision-based fire detection models to improve the detection of fire in buildings. In recent years, deep learning methods have been widely and effectively applied in different ways in smoke and flame recognition research. The work in Ref. [16] reviewed the state-of-the-art applications of Intelligent fire detection in building and construction. Ba et al. [17] proposed a smoke recognition model incorporating spatial and channel attention mechanisms in

CNN to enhance the feature representation for scene classification. Wu et al. [18] presented a novel intelligent fire detection approach through video cameras for preventing fire hazards from losing control in chemical factories and other high-fire-risk industries. Park et al. [19] proposed ELASTIC-YOLOv3, a fire recognition method for urban environments that can quickly recognize a fire at night in urban areas by reflecting its nighttime characteristics. To analyze fire emergency scenes, Sharma et al. [20] proposed a method that uses a deep learning image segmentation network to recognize objects based on their build material and vulnerability. Muhammad et al. [21] integrated the principal component analysis as a preprocessing module with the improved YOLO-V3 to boost the network predictions for smoke in the wild. The work in Ref. [22] presented an attention-based CNN model for the detection of fire and used the gradient-weighted class activation mapping method to visualize and locate the fire in images. However, the existing smoke and flame detection network is not accurate enough due to the irregular shapes, unclear contours, and large-scale changes in smoke and flames. Table 1 summarizes the recent developments in the field of flame and smoke recognition algorithms. In this paper, we follow the deep learning-based object detector design, and we show it is possible to achieve higher accuracy for smoke and flame recognition with optimized network architectures.

Table 1. Recent developments and comparison of different flame and smoke recognition algorithms.

Basic Methodology	Dataset	Improvement Direction	Ref.
Image Processing	Flame	Combining wavelet and FFT	[10]
	Flame	Improving Canny Edge Detector	[11]
	Flame	Combining spatio-temporal flame modeling and dynamic texture analysis	[12]
	Flame/smoke	Combining traditional image features and machine learning methods	[13]
Deep Learning	Flame/smoke	Improving CNN model structure	[14]
	Flame/smoke	Video fire detection model using indoor closed-circuit television surveillance	[15]
	Smoke	Improving CNN model structure	[17]
	Flame	Improving CNN model structure	[18]
	Flame	Improving CNN model structure	[19]
	Flame/smoke	Multitask learning	[20]
	Smoke	Improving CNN model structure	[21]
	Flame	Improving CNN model structure	[22]

2.2. Anchor-Free Object Detectors

Researchers have proposed anchor-free object detection algorithms to overcome the shortcomings of existing anchor-based object detection algorithms, which remove the predefined anchor boxes and directly predict the object's bounding boxes and classes. The object detection process of YOLO-V1 [23] is a regression problem that can directly extract features from the input image to predict the bounding boxes and class probabilities. DenseBox [24] is considered to be the earliest anchor-free method. DenseBox takes each pixel as a centroid, predicts a bounding box for each pixel, and then filters the bounding box using Non-Maximum Suppression (NMS). The detection accuracy of the algorithm on small objects, such as face detection, is significantly higher than that of the anchor-based algorithms. CornerNet [25] is a single-stage anchor-free object detector that detects pairs of corners of a bounding box and groups them to form the final detected bounding box. CenterNet [26], inspired by CornerNet, uses the object's centroid as the detection center, and can detect objects by adding their boundaries and size information. FCOS [27] employs a centerness branch to mitigate the issue of excessive and improper bounding boxes, which utilizes multiple binary classifiers for object classification. Compared with the anchor-based object detection network, the anchor-free object detection network is more conducive to the recognition of smoke and flames without fixed shapes and sizes. The proposed network

architecture in this paper adopts a comparable structure to FCOS but incorporates a more robust backbone network.

2.3. Multiscale Feature Fusion

One of the important research elements of the object detection task is how to effectively represent and process the multiscale features output from the backbone. The feature pyramid network (FPN) [28] is the earliest multiscale feature fusion network that proposes a top-down fusion path to fuse multiscale features. The Path Aggregation Network (PANet) [29] adds an extra bottom-up path on top of the FPN to achieve a higher performance in multiscale feature fusion. Tan et al. [30] proposed a weighted bidirectional feature pyramid network (BiFPN) for simple and efficient multiscale feature fusion. BiFPN introduces learnable weights to learn the importance of different input features while iteratively applying top-down and bottom-up multiscale feature fusion. PANet achieves a good balance between speed and accuracy in these multiscale feature fusion networks. In this paper, we aim to optimize multiscale feature fusion with the attention mechanism.

3. Methods

3.1. Anchor-Free Smoke and Flame Recognition Network Architecture

Existing smoke and flame recognition algorithms use anchor-based object detection methods to recognize smoke and flames. These methodologies rely heavily on the design of anchor boxes, thereby limiting their capacity to account for the shapes and sizes of smoke and flames, especially when their scale is still insignificant. The smoke and flame recognition algorithm based on the anchor-free network architecture no longer uses anchor boxes, directly predicting the locations and the class of the target box containing the pixel based on each pixel. Since the anchor-free network architecture generates only one predicted box for each pixel, the architecture is simpler, and the computational effort is significantly reduced compared with that of the anchor-based network architecture. At the same time, the hardware requirements of the model are also reduced to facilitate deployment. The network architecture proposed in this paper adopts a comparable structure to FCOS, which contains a backbone network, a multiscale feature fusion network, and five detection heads; the overall architecture is shown in Figure 1. EfficientNet [31] is used as the backbone network to perform basic feature extraction after image preprocessing. A multiscale feature fusion network is introduced after the backbone network to further exploit the multiscale information of the image. The feature fusion network utilizes the information of a total of five feature layers, using a top-down and bottom-up fusion structure, which can utilize both the semantic features of the higher layers and the image information of the lower layers and is beneficial for the recognition of smoke and flames at different scales. The different detection heads of the network are responsible for the recognition of smoke and flames at different scales. The features of the smoke and flames are further extracted to predict the classes and locations of the bounding boxes.

Let $F_i \in R^{H \times W \times C}$ be the feature map at layer i of the network, and the ground-truth bounding boxes of each input image are defined as:

$$B_i = ({x_0}^{(i)}, {y_0}^{(i)}, {x_1}^{(i)}, {y_1}^{(i)}, c^{(i)}) \quad (1)$$

where $({x_0}^{(i)}, {y_0}^{(i)})$ and $({x_1}^{(i)}, {y_1}^{(i)})$ are the coordinates of the top-left and bottom-right corners of the bounding box, respectively, and $c^{(i)}$ is the class the object in the bounding box belongs to. For each location (x, y) on the feature map F_i, it can be mapped onto the input image as $\left(\frac{s}{2} + xs, \frac{s}{2} + ys\right)$ (s is the feature map downsampling step), which is near the center of the respective field region of the location (x, y). Unlike the anchor-based object detectors, the anchor-free object detector directly takes location (x, y) as a training sample instead of an anchor box.

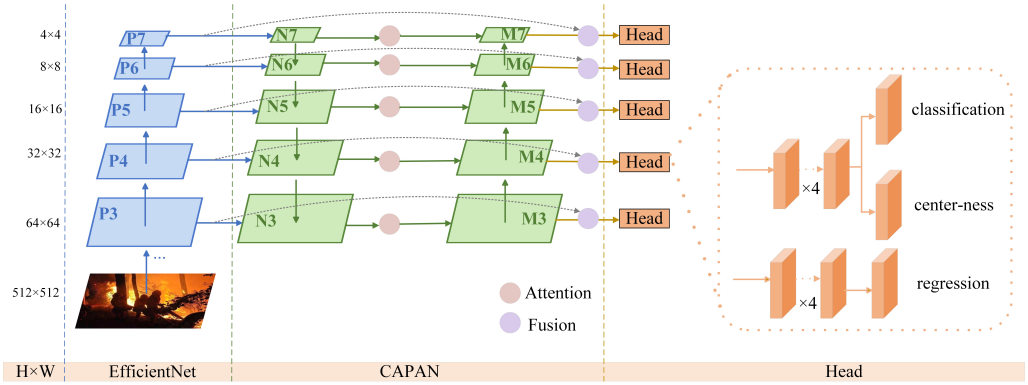


Figure 1. The network architecture of the anchor-free smoke and flame recognition algorithm proposed in this paper. The backbone network is replaced with EfficientNet to extract features from the input images.

The training samples are classified as positive and negative samples. The anchor-free smoke and flame recognition network considers the location (x, y) as a positive sample if it falls into any ground-truth box, and the class label of the location is the class label of the ground-truth box. Otherwise, it is a negative sample. Since the region inside the ground-truth box near the edge is often still the background rather than the foreground object that is to be predicted, these positive samples near the edges tend to produce low-quality predicted bounding boxes. To suppress the low-quality predicted bounding boxes, the anchor-free smoke and flame recognition network proposed in this paper uses the centerness branch in FCOS; the centerness branch is parallel to the classification branch, and binary cross-entropy loss is used during training and added to the total loss function. During prediction, the predicted centerness is multiplied by the classification score; with high probability, these low-quality bounding boxes might be filtered out by the final NMS process.

The outputs of the classification branch represent the foreground and background scores, and the 4D vector $T = (l, t, r, b)$ of the regression branch outputs depicts the distances from the locations to the four sides of the ground-truth bounding box, as shown in Figure 2. Let (x_0, y_0) and (x_1, y_1) be the coordinates of the top-left and bottom-right corners of the ground-truth bounding box, respectively, and if location (x, y) is associated with the ground-truth bounding box, the four regression targets for the locations can be formulated as:

$$T^* = \left(\begin{array}{c} l^* = x - {x_0}^{(i)}, t^* = y - {y_0}^{(i)} \\ r^* = x - {x_1}^{(i)}, b^* = y - {y_1}^{(i)} \end{array} \right). \quad (2)$$

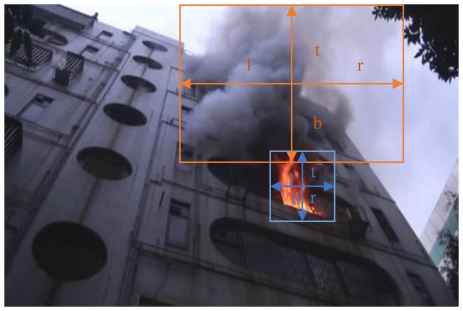


Figure 2. The anchor-free smoke and flame recognition algorithm works by predicting a 4D vector (l, t, r, b) depicting the relative offsets from the four sides of a bounding box to the location.

3.2. CAPAN

In the smoke and flame recognition scenario, smoke and flame objects in fire images have different scales. PANet combines multiscale contextual information by applying up-and-down sampling and multiscale feature fusion through top-down and bottom-up paths. However, the smoke and flame feature maps of different channels tend to focus on

different scene categories, as shown in Figure 3. To enhance the foreground features and suppress the background features, we propose CAPAN for more efficient feature fusion. PANet and CAPAN are shown in Figure 4. CAPAN constructs a multiscale feature fusion network based on the multiscale features P3 to P7 extracted by the backbone. Different from PANet, CAPAN adds Efficient Channel Attention (ECA) [32] after the first top-down feature fusion from P3 to P7, where the attention weights of each level are independent, and the attention modules of different levels are responsible for the extraction of the channel relationship in their respective levels. The multiscale features obtained by the attention module are then subjected to the second bottom-up feature fusion. In addition, the new connections are added to the second feature fusion to avoid the information loss caused by channel reduction.

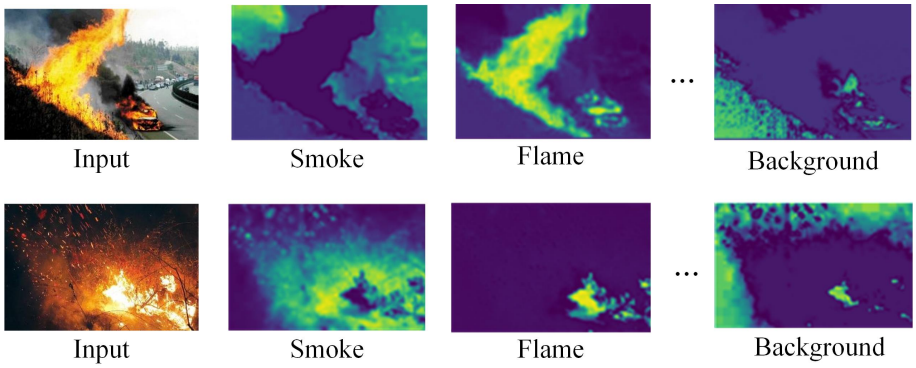


Figure 3. The smoke and flame feature maps of different channels tend to focus on different scene categories.

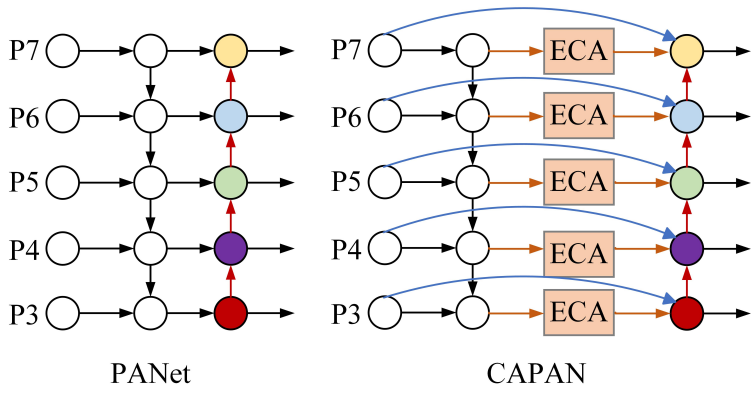


Figure 4. The architecture of PANet and CAPAN.

The ECA is improved from the Squeeze-and-Excitation Networks (SENet) [33]. SENet first proposed a channel attention learning method. SENet maps channel features to a low-dimensional space and then maps them back, making the channel relationship and its weights indirect. Unlike SENet, ECA enhances the cross-channel information exchange, while keeping the channel dimension constant. The ECA architecture is shown in Figure 5. The ECA learns more effective channel attention while reducing the model complexity. The channel weights in the ECA are calculated as follows:

$$\omega_i = \sigma(\sum_{j=1}^{k} \alpha_i^j y_i^j), y_i^j \in \Omega_i^k, \tag{3}$$

where Ω_i^k is the k-domain channel of y_i, y_i is the feature representation of channel i after global averaging pooling, α_i^j is the shared parameter, σ is the activation function, and ω_i represents the weights of channel i. The value of k as a key parameter can be adjusted in

size to determine the range of interactions between channels, and the range of interactions increases with the increasing channel dimension.

CAPAN improves the fusion performance of low-level and high-level features by adding channel attention and residual connections to the feature fusion network, forcing the network to focus on the channel features with foreground information and weakening the background features.

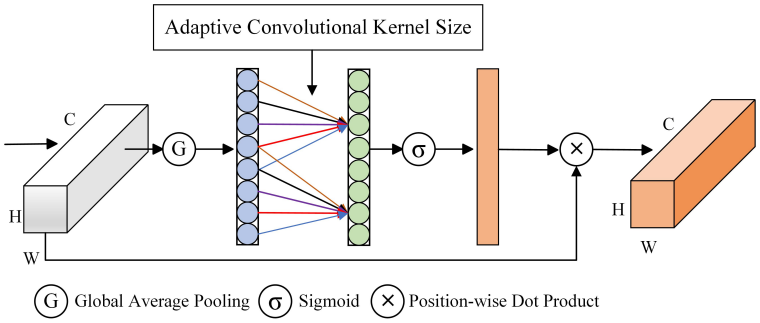


Figure 5. The architecture of the ECA module, which enhances the cross-channel information exchange while keeping the channel dimension constant.

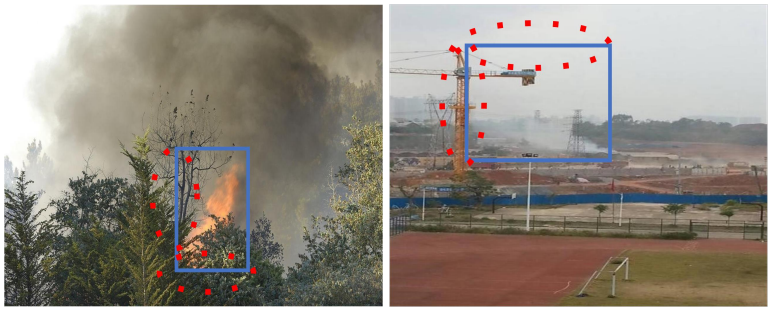
3.3. Distribution Focal Loss

The issue often encountered with flame and smoke objects is the lack of distinctly recognized boundaries. The ground-truth labels are sometimes not credible, resulting in low-quality predictions of the bounding boxes. In previous research work, most models focused on model optimization, such as better detection accuracy, lighter weight models, and faster detection speed [34–36]. However, the smoke and flame boundary situation that exists in reality was not considered, and the algorithm proposed in this paper takes into account the smoke and flame boundary situation. The algorithm proposed in this paper takes the distances from the locations to the four sides of the bounding box as the regression targets. The representation of bounding boxes can be viewed as Dirac delta distribution without considering the ambiguity and uncertainty of the smoke and flame boundaries in the fire images, as shown in Figure 6. To solve this problem, on the basis of the original loss function of the object detection algorithm, the distribution focal loss [37] is introduced to supervise the smoke and flame recognition network model during the training, which enables the model to learn a more accurate distribution of the locations of bounding boxes, while reducing the gap between the predicted bounding boxes and the ground-truth bounding boxes.

The locations of the bounding boxes are no longer modeled as Dirac delta distributions but arbitrary distributions $P(x)$ due to the use of the DFL. Given a label range $y(y_0 \leq y \leq y_n, n \in N^+)$, the values of $[y_0, y_n]$ are discretized into a set $[y_0, y_1, y_2, \ldots, y_{n-1}, y_n]$. According to the discrete distribution property $\sum_{i=0}^{n} P(y_i) = 1$, the estimated regression value can be calculated as $\hat{y} = \sum_{i=0}^{n} P(y_i) y_i$. The regression branch has $n+1$ predictions for each distance from the location to the four sides of the bounding box, and $P(y_i)$ can be implemented through a softmax classifier. In addition, the predicted locations will not be far away from the labels. To make the model focus on y_i and y_{i+1} near label y, enlarging the probabilities of nearest two values to label y in the form of a cross-entropy loss function. Denoting $P(y_i)$ as S_i, the DFL is defined as follows:

$$L_{DFL} = -((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})), \tag{4}$$

where the aim of L_{DFL} is to enlarge the probabilities of the values near the label y. When $S_i = \frac{y_{i+1}-y}{y_{i+1}-y_i}$ and $S_{i+1} = \frac{y-y_i}{y_{i+1}-y_i}$, L_{DFL} obtains the global minimum solution; in this case, the predicted value $\hat{y}$ is infinitely close to the label y.



(**a**) (**b**)

Figure 6. (**a**) Uncertainty of flame boundaries. (**b**) Ambiguity of smoke boundaries. Due to the ambiguity and uncertainty of the smoke and flame boundaries in the fire images, the ground-truth bounding boxes (blue boxes) are sometimes not credible (see red circles).

The DFL models bounding box locations as arbitrary distributions, thereby offering a more precise representation of the boundary associated with the smoke and flame perimeters. This leads to more precise bounding box predictions for anchor-free networks and better localization accuracy of smoke and flame objects.

3.4. Multi-Loss Fusion

3.4.1. Classification Loss

In the training process of the classification network, the location is a positive sample if it falls into any ground-truth bounding box, and the class label of the location is the class label of the ground-truth box. Otherwise, it is a negative sample. In most fire images, the area of the effective region only accounts for a small part of the feature map. To solve this problem, the classification loss function in this paper adopts the focal loss [38], and the focal loss is defined as follows:

$$L_{FL} = \begin{cases} -(1-p)^{\gamma}\log(p), \text{when } y = 1 \\ -p^{\gamma}\log(1-p), \text{when } y = 0 \end{cases} \tag{5}$$

$$p_t = \begin{cases} p, \text{when } y = 1 \\ 1-p, \text{when } y = 0 \end{cases} \tag{6}$$

$$L_{FL} = -(1-p_t)^{\gamma}\log(p_t), \tag{7}$$

where p_t denotes the category probability, and $p \in [0,1]$ denotes the estimated probability when $y = 1$. When $y = 1$, $p_t = p$, and when $y = 0$, $p_t = 1 - p$. γ is an adjustable parameter. The focal loss consists of a standard cross-entropy component $-\log(p_t)$ and a dynamic scaling factor $(1-p_t)^{\gamma}$. The scaling factor $(1-p_t)^{\gamma}$ automatically reduces the contribution of the easy-to-classify samples during the training and quickly focuses the model on the difficult samples.

The focal loss was proposed to address the extreme imbalance between the foreground and background classes during the training of single-stage object detection networks. In this paper, the focal loss solves the problem of the unbalanced positive and negative samples during the training of the smoke and flame recognition model, allowing the model to learn to focus on the difficult-to-classify samples and achieve a better smoke and flame recognition performance.

3.4.2. Regression Loss

In this paper, the regression loss function of the anchor-free network adopts the more comprehensive CIoU loss function [39]. In the field of object detection, the IoU is commonly used to calculate the difference between the predicted bounding boxes and the ground-truth bounding boxes; the IoU is defined as follows:

$$IoU = \frac{A \cap B}{A \cup B}, \tag{8}$$

where A and B are the predicted box and the target box, respectively. CIoU is a more comprehensive optimization of IoU, taking into account the distance, scale, and aspect ratio between the predicted box and the target box. The CIoU loss function is defined as follows:

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{9}$$

$$CIoU = IoU - \frac{\rho^2(b, b^{gt})}{c^2} - av \tag{10}$$

$$L_{CIoU} = 1 - CIoU, \tag{11}$$

where a is the weight value, v is used to measure the similarity of the aspect ratio between the predicted box and the target box, b and b^{gt} are the predicted box and the target box, respectively, w and h are the width and height of the predicted box, respectively, w^{gt} and h^{gt} are the width and height of the target box, respectively, c is the diagonal of the smallest outer rectangle of the two rectangular boxes, ρ is the Euclidean distance between the center points of the two rectangular boxes, and L_{CIoU} is the CIoU target loss.

The CIoU loss function directly optimizes the distance between the two rectangular boxes for faster convergence. In addition, it considers the aspect ratio to quickly reduce the difference between the prediction boxes and the target boxes.

3.4.3. Loss for Centerness Branch

The centerness branch is used to predict the normalized distance from the location to the center of the bounding box for which the location is responsible. Low-quality predicted bounding boxes would be suppressed by the predicted centerness, thus improving the recognition accuracy. The centerness is defined as follows:

$$centerness = \sqrt{\frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)}} \tag{12}$$

where l,r,t, and b are the distances from the location to the four sides of the bounding box, respectively. The closer the location is to the center of the bounding box, the higher the centerness score and vice versa. The centerness branch uses a binary cross entropy (BCE) loss function, and the loss function for the centerness branch can be presented as:

$$L_{center} = BCE(centerness, \hat{centerness}) \tag{13}$$

where $\hat{centerness}$ is the predicted value and $centerness$ is the target value.

3.4.4. Multi-Loss Fusion

The total loss consists of the focal loss, CIoU loss, DFL, and the loss for the centerness branch, and the weighted fusion of the above four components of the loss is used to optimize the parameters of the smoke and flame recognition model; the total loss is defined as follows:

$$Loss = L_{FL} + \lambda_1 L_{CIoU} + \lambda_2 L_{center} + \lambda_3 L_{DFL}, \tag{14}$$

where L_{FL} is the focal loss, L_{CIoU} is the CIoU loss, L_{center} is the loss for the centerness branch, and L_{DFL} is the DFL. λ_1, λ_2, and λ_3 are scaling factors to balance the four parts of the loss, which can theoretically be chosen as a whole real number greater than 0. Therefore, it is impossible to enumerate all cases. In this paper, the ratio of λ_1, λ_2, and λ_3 is set to 2:1:1, respectively, based on experience and practical debugging.

4. Experiments

This section introduces the experimental environment, dataset, evaluation indicators of the model effect, and analysis of the experimental results of the training network.

4.1. Smoke and Flame Dataset

The dataset used in the experiment was constructed by collecting smoke and flame pictures from public websites and videos shot in the field. In total, 8540 high-quality images were obtained, classified into flame and smoke. The images came from different scenes, such as forests, factories, cities, warehouses, houses, and electrical and construction sites. The image dataset consisted of fire incidents occurring during day and night, spanning various stages of fire development, from ignition to extinction. The completed dataset is shown in Table 2.

Table 2. The number of images and lables in the three categories: only smoke, only flame, smoke and flame.

Dataset	Only Smoke	Only Flame	Smoke and Flame	Total
Number of images	1924	2693	3923	8540
Number of annotated samples	6389	10,702	-	17,091

After numbering the above dataset images, we used the LabelImg tool to manually label the images, including drawing the bounding boxes and classifying categories; the dataset was annotated in PASCAL VOC dataset format. The dataset was randomly divided into a training set, validation set, and testing set. The number of images in the training and validation sets was 7686, and the number in the test set was 854. The ratio of the training set to the validation set was 9:1.

4.2. Experimental Environment

The experiments were conducted on computing platforms comprising NVIDIA RTX2080 SUPER and Intel i7-9750H, with a memory capacity of 32 G. Pytorch was the development framework, the batch size was set to 16, the learning rate was set to 0.01, the learning rate decay adopted the linear decay mode, and each model trained 100 epochs. Moreover, the input image size was set at a resolution of 512 × 512.

4.3. Evaluation Indicators

In order to directly compare these models, the average detection accuracy mAP, mAP@0.5, flame AP50, and smoke AP50 were selected in this paper as the evaluation metrics for the model. For the binary classification problem, according to the combination of the true category and the predicted category, samples can be divided into four types: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The confusion matrix of the classification results is shown in Table 3.

Table 3. The confusion matrix of the real and predicted categories for dichotomous problems.

Labeled Name	Predicted	Confusion Matrix
Positive	Positive	TP
Positive	Negative	FN
Negative	Positive	FP
Negative	Negative	TN

Precision and recall are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

AP and mAP are defined as follows:

$$\mathrm{AP} = \int_0^1 \mathrm{Precision}(\mathrm{Recall}) d(\mathrm{Recall}) \tag{17}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i} \tag{18}$$

4.4. Algorithm Comparison Analysis

To verify the effectiveness of the algorithm proposed in this paper, our method was compared with seven representative methods, including Faster-RCNN, multiple YOLO series detection algorithms, and three other single-stage object detection algorithms, where the FCOS was used as the baseline. The comparison results are listed in Table 4.

Table 4. The comparison between the model proposed in this paper and the existing six object detection network model on the testing set in terms of the mAP, mAP@0.5, AP50 on flames, AP50 on smoke, and FPS.

Method	mAP	mAP@0.5	Flame AP50	Smoke AP50	FPS
Faster-RCNN	26.8	62.5	58.1	67.7	7
SSD	28.1	63.2	62.5	63.5	23
YOLOv3	28.9	63.5	63.4	68.1	30
YOLOv4	33.4	72.1	68.2	76.9	28
YOLOv5m	46.7	76.9	70.1	83.7	55
EfficientDet	36.9	70.8	61.7	80.5	26
FCOS	47.5	78.2	69.2	87.2	36
Our method	52.5	83.4	77.5	89.3	33

From the comparison with the six object detection algorithms and our method, it can be seen that the algorithm proposed in this paper had the highest average detection accuracy, which was improved by 5% for the mAP compared with the baseline, and showed higher recognition accuracy in terms of both flame and smoke, which improved by 8.3% and 2.1% for the AP50, respectively, with a best comprehensive performance. The algorithm proposed in this paper ran at a speed of 33 frames per second, and can more easily be applied to smoke and flame recognition in real scenes.

To verify the performance of our algorithm in the scenes with irregular shapes and unclear contours of smoke and flames, as shown in Figure 7, some representative images were selected to visualize and compare the recognition effect of our algorithm with the baseline. The recognition results showed that the proposed algorithm was significantly better than the baseline model regarding both localization and classification accuracy. Compared with Figure 7a, all objects were detected, as shown in Figure 7e. Compared with Figure 7b,c, Figure 7f,g showed more accurate predicted bounding boxes. Compared with Figure 7d, the obscured flame boundary was better identified in Figure 7h. The recognition results show that the CAPAN improved the fusion efficiency of low-level features and high-level features, the network paid more attention to foreground object information, and the multi-loss fusion approach enabled the network to learn a target box location distribution that is closer to the real distribution.

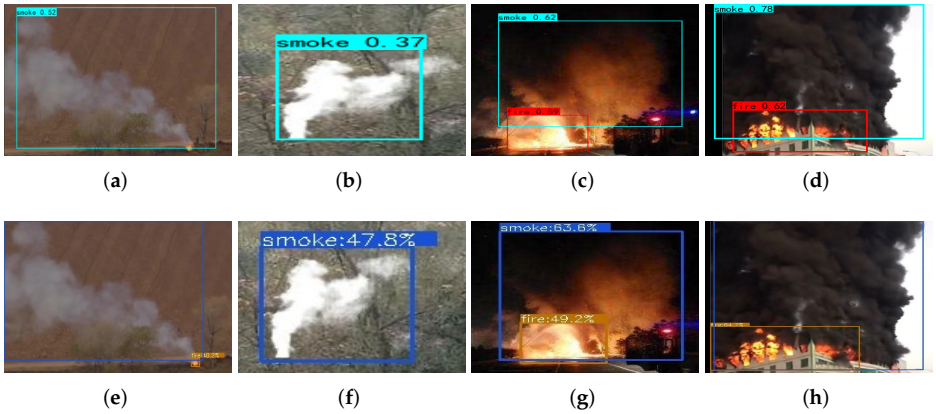


Figure 7. (**a**–**d**) Baseline. (**e**–**h**) Our method. The recognition effect of the baseline and our method in terms of the smoke and flame boundaries and the small target smoke and flames.

4.5. Effect of Adding the ECA at Different Positions

The algorithm proposed in this paper improved the model performance by CAPAN. To verify the effect of embedding the ECA module at different positions on the model, the FCOS incorporated with PANet and DFL was used as the baseline. Three experimental setups were established for comparative analysis. As shown in Figure 8, prior to the feature fusion layers N3 to N7 was the first addition of a position, between the feature fusion layers N3 to N7 and the feature fusion layers M3 to M7 was the second addition of a position, and the third additional position came after the feature fusion layers M3 to M7. As listed in Table 5, I, II, and III denote the first, second, and third positions, respectively. The experimental results show that adding the ECA module at the second position had the best recognition effect.

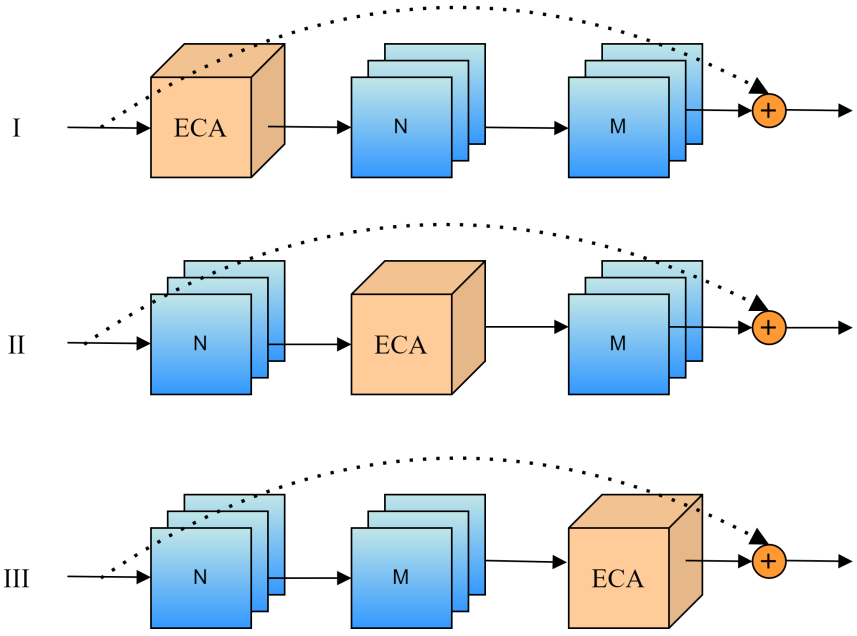


Figure 8. The different additional positions for the ECA module.

Table 5. The recognition effect of embedding an ECA module at different positions in the model in terms of the mAP and mAP@0.5.

Method	Adding Location	mAP	mAP@0.5
Our method	I	51.3	82.7
	II	52.5	83.4
	III	52.2	82.1

4.6. Algorithm Ablation Study

To verify the improvement in the ECA channel attention module and the DFL regarding the performance of the anchor-free smoke and flame recognition algorithm, this work validated the advantages of the proposed anchor-free smoke and flame recognition algorithm by conducting ablation experiments on the different improvements; the FCOS was used as the baseline. As listed in Table 6, after combining the ECA channel attention module with the multiscale feature fusion network, the network paid more attention to the important foreground information of the fire images, and the mAP@0.5 was increased from 78.2% to 80.3%. A multi-loss fusion method was used to learn the arbitrary distribution of the target box locations so that the distribution of predicted box locations was closer to the real distribution, and the mAP@0.5 increased from 78.2% to 81.7%. The results of the ablation experiment showed that both the ECA channel attention module and the DFL improved the recognition accuracy of the smoke and flame recognition algorithm based on the anchor-free network architecture. When the two methods were combined, the recognition accuracy was significantly improved. The recognition proficiency of the model satisfies the requirements of smoke and flame recognition tasks in video surveillance settings.

Table 6. The ablation experimental results of the different improvements proposed in this paper in terms of the mAP and mAP@0.5.

Method	ECA	DFL	mAP	mAP@0.5
FCOS			47.5	78.2
Our method	✓		51.8	80.3
		✓	52.1	81.7
	✓	✓	52.5	83.4

5. Discussion

In Sections 4.4–4.6, we designed different experiments to verify the effectiveness of the method proposed in this paper.

The smoke and flame recognition algorithm proposed in this paper achieved satisfactory results on the recognition of smoke and flame targets with ambiguous and uncertain boundaries and can provide real-time smoke and flame recognition. Nevertheless, our experiments indicate that the proposed algorithm is susceptible to misclassifying fire-like targets as fire targets. This occurrence can be attributed to the resemblance between fire-like targets and fire targets with regard to shape and color. Thus, this poses a significant challenge to the accurate recognition of fires, as shown in Figure 9. Encouragingly, these difficulties are not insurmountable. During the training of the object detection model, the ability of the model to recognize fire-like objects can be improved by adding fire-like images to the training samples, which showed good results. In forthcoming research, our proposed model will be further optimized with the objective of constructing a dataset for fire-like images, thereby enhancing the smoke and flame recognition ability of the model.

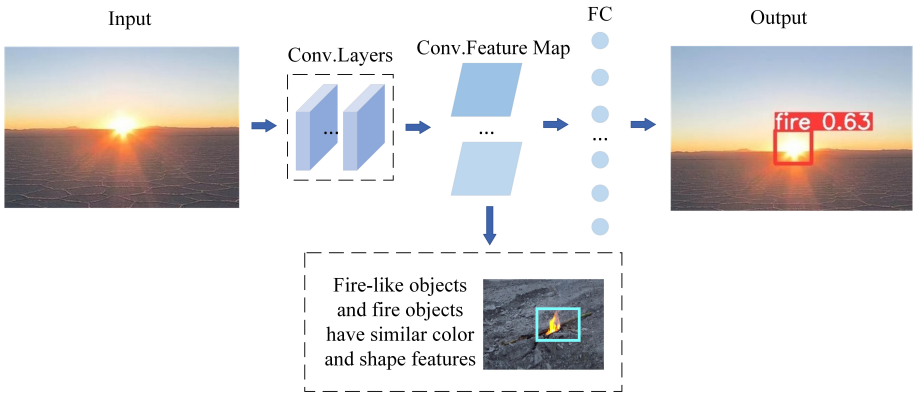


Figure 9. When the model proposed in this paper detects fire-like images, it is easy to produce inaccurate detection results.

6. Conclusions

This paper proposes an effective and reliable method for recognizing smoke and flames with no fixed shape in the early stages of fire. The procedure for developing the anchor-free smoke and flame recognition model has been clearly described, as well as the all improvements that promote the algorithm's ability to recognize smoke and flames for the model. The algorithm was improved by 5% for the mAP compared with the baseline and showed a higher recognition accuracy for both flame and smoke, with improvements of 8.3% and 2.1% for the AP50, respectively. By incorporating the ECA channel attention module and residual connections into the multiscale feature fusion network, the model can concentrate on foreground object information, which improves the recognition of smoke and flame objects. Moreover, the algorithm utilizes a multi-loss fusion method to address the issue of ambiguous and uncertain smoke and flame boundaries, leading to more accurate regression branch output. Our experimental results demonstrate that the proposed algorithm outperforms other existing methods with a higher object recognition performance, and the detection speed satisfies the requirements of real-time detection. Thus, it has practical applications in high-fire-risk scenes, such as forests and chemical plants. The deployment of this algorithm in the industry has the potential to significantly enhance fire safety and emergency management.

Future work will focus on applying existing models to detect smoke and flames in videos. In addition, the detection of fire-like targets will be optimized to further improve the accuracy of the detection model.

Author Contributions: Conceptualization, G.L. and P.C.; methodology, P.C.; software, G.L. and P.C.; validation, G.L., P.C. and C.X.; formal analysis, P.C.; investigation, P.C.; resources, G.L. and P.C.; data curation, P.C.; writing—original draft preparation, P.C.; writing—review and editing, G.L., P.C., C.X., C.S. and Y.M.; visualization, P.C., C.S. and Y.M.; supervision, G.L. and C.X.; project administration, G.L. and P.C.; funding acquisition, G.L. and C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the China Chongqing Science and Technology Commission under Grants cstc2020jscx-msxmX0086, cstc2019jscx-zdztzx0043, cstc2019jcyj-msxmX0442, and stc2021jcyj-msxmX0605; the China Chongqing Banan District Science and Technology Commission project under Grant 2020QC413; and the China Chongqing Municipal Education Commission under Grant KJQN202001137.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in the experiment was constructed by collecting smoke and flame pictures from public websites and videos shot in the field. The original 8540 images include images of flames and smoke in different weather and light lines.

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

References

1. Muhammad, K.; Ahmad, J.; Baik, S.W. Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing* **2018**, *288*, 30–42. [CrossRef]
2. binti Zaidi, N.I.; binti Lokman, N.A.A.; bin Daud, M.R.; Achmad, H.; Chia, K.A. Fire recognition using RGB and YCbCr color space. *ARPN J. Eng. Appl. Sci.* **2015**, *10*, 9786–9790.
3. Li, Z.; Mihaylova, L.S.; Isupova, O.; Rossi, L. Autonomous flame detection in videos with a Dirichlet process Gaussian mixture color model. *IEEE Trans. Ind. Inform.* **2017**, *14*, 1146–1154. [CrossRef]
4. Wang, Y.; Dang, L.; Ren, J. Forest fire image recognition based on convolutional neural network. *J. Algorithms Comput. Technol.* **2019**, *13*, 1748302619887689. [CrossRef]
5. Muhammad, K.; Khan, S.; Elhoseny, M.; Ahmed, S.H.; Baik, S.W. Efficient fire detection for uncertain surveillance environment. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3113–3122. [CrossRef]
6. Sharma, J.; Granmo, O.C.; Goodwin, M.; Fidje, J.T. Deep convolutional neural networks for fire detection in images. In Proceedings of the Engineering Applications of Neural Networks: 18th International Conference, EANN 2017, Athens, Greece, 25–27 August 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 183–193.
7. Wu, S.; Zhang, L. Using popular object detection methods for real time forest fire detection. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; Volume 1, pp. 280–284.
8. Xie, Y.; Zhu, J.; Cao, Y.; Zhang, Y.; Feng, D.; Zhang, Y.; Chen, M. Efficient video fire detection exploiting motion-flicker-based dynamic features and deep static features. *IEEE Access* **2020**, *8*, 81904–81917. [CrossRef]
9. Yang, Y.; Pan, M.; Li, P.; Wang, X.; Tsai, Y.T. Development and optimization of image fire detection on deep learning algorithms. *J. Therm. Anal. Calorim.* **2022**, *148*, 5089–5095 . [CrossRef]
10. Zhang, Z.; Zhao, J.; Zhang, D.; Qu, C.; Ke, Y.; Cai, B. Contour based forest fire detection using FFT and wavelet. In Proceedings of the 2008 International Conference on Computer Science and Software Engineering, Beijing, China, 12–14 December 2008; Volume 1, pp. 760–763.
11. Jiang, Q.; Wang, Q. Large space fire image processing of improving canny edge detector based on adaptive smoothing. In Proceedings of the 2010 International Conference on Innovative Computing and Communication and 2010 Asia-Pacific Conference on Information Technology and Ocean Engineering, Macao, China, 30–31 January 2010; pp. 264–267.
12. Dimitropoulos, K.; Barmpoutis, P.; Grammalidis, N. Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 339–351. [CrossRef]
13. Toulouse, T.; Rossi, L.; Celik, T.; Akhloufi, M. Automatic fire pixel detection using image processing: A comparative analysis of rule-based and machine learning-based methods. *Signal Image Video Process.* **2016**, *10*, 647–654. [CrossRef]
14. Pincott, J.; Tien, P.W.; Wei, S.; Calautit, J.K. Indoor fire detection utilizing computer vision-based strategies. *J. Build. Eng.* **2022**, *61*, 105154. [CrossRef]
15. Ahn, Y.; Choi, H.; Kim, B.S. Development of early fire detection model for buildings using computer vision-based CCTV. *J. Build. Eng.* **2023**, *65*, 105647. [CrossRef]
16. Baduge, S.K.; Thilakarathna, S.; Perera, J.S.; Arashpour, M.; Sharafi, P.; Teodosio, B.; Shringi, A.; Mendis, P. Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. *Autom. Constr.* **2022**, *141*, 104440. [CrossRef]
17. Ba, R.; Chen, C.; Yuan, J.; Song, W.; Lo, S. SmokeNet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sens.* **2019**, *11*, 1702. [CrossRef]
18. Wu, H.; Wu, D.; Zhao, J. An intelligent fire detection approach through cameras based on computer vision methods. *Process Saf. Environ. Prot.* **2019**, *127*, 245–256. [CrossRef]
19. Park, M.; Ko, B.C. Two-step real-time night-time fire detection in an urban environment using Static ELASTIC-YOLOv3 and Temporal Fire-Tube. *Sensors* **2020**, *20*, 2202. [CrossRef] [PubMed]
20. Sharma, J.; Granmo, O.C.; Goodwin, M. Emergency analysis: Multitask learning with deep convolutional neural networks for fire emergency scene parsing. In Proceedings of the Advances and Trends in Artificial Intelligence, Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, 26–29 July 2021; Proceedings, Part I 34; Springer: Berlin/Heidelberg, Germany, 2021; pp. 101–112.
21. Masoom, S.M.; Zhang, Q.; Dai, P.; Jia, Y.; Zhang, Y.; Zhu, J.; Wang, J. Early Smoke Detection Based on Improved YOLO-PCA Network. *Fire* **2022**, *5*, 40. [CrossRef]
22. Majid, S.; Alenezi, F.; Masood, S.; Ahmad, M.; Gündüz, E.S.; Polat, K. Attention based CNN model for fire detection and localization in real-world images. *Expert Syst. Appl.* **2022**, *189*, 116114. [CrossRef]
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. Densebox: Unifying landmark localization with end to end object detection. *arXiv* **2015**, arXiv:1509.04874.
25. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.

26. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
27. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
28. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
29. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
30. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
31. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
34. Zhang, X.; Qian, K.; Jing, K.; Yang, J.; Yu, H. Fire detection based on convolutional neural networks with channel attention. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 3080–3085.
35. Saponara, S.; Elhanashi, A.; Gagliardi, A. Real-time video fire/smoke detection based on CNN in antifire surveillance systems. *J. Real-Time Image Process.* **2021**, *18*, 889–900. [CrossRef]
36. Li, W.; Yu, Z. A lightweight convolutional neural network flame detection algorithm. In Proceedings of the 2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC) , Beijing, China, 18–20 June 2021; pp. 83–86.
37. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
38. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
39. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.