frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 03 February 2023
DOI 10.3389/fphy.2023.1136021

OPEN ACCESS

EDITED BY
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China

REVIEWED BY
Puhong Duan,
Hunan University, China
Guanqiu Qi,
Buffalo State College, United States

*CORRESPONDENCE
Kaizheng Wang,
kz.wang@foxmail.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 02 January 2023
ACCEPTED 23 January 2023
PUBLISHED 03 February 2023

CITATION
Zhou F, Wen G, Ma Y, Wang Y, Ma Y, Wang G, Pan H and Wang K (2023), Multilevel feature cooperative alignment and fusion for unsupervised domain adaptation smoke detection.
Front. Phys. 11:1136021.
doi: 10.3389/fphy.2023.1136021

Multilevel feature cooperative alignment and fusion for unsupervised domain adaptation smoke detection

Fangrong Zhou[1], Gang Wen[1], Yi Ma[1], Yifan Wang[1], Yutang Ma[1], Guofang Wang[1], Hao Pan[1] and Kaizheng Wang[2]*

[1]Joint Laboratory of Power Remote Sensing Technology, Electric Power Research Institute, Yunnan Power Grid Company Ltd., China Southern Power Grid, Kunming, China, [2]Faculty of Electrical Engineering, Kunming University of Science and Technology, Kunming, China

Early smoke detection using Digital Image Processing technology is an important research field, which has great applications in reducing fire hazards and protecting the ecological environment. Due to the complex changes of color, shape and size of smoke with time, it is challenging to accurately recognize smoke from a given image. In addition, limited by domain shift, the trained detector is difficult to adapt to the smoke in real scenes, resulting in a sharp drop in detection performance. In order to solve this problem, an unsupervised domain adaptive smoke detection algorithm rely on Multilevel feature Cooperative Alignment and Fusion (MCAF) was proposed in this paper. Firstly, the cooperative domain alignment is performed on the features of different scales obtained by the feature extraction network to reduce the domain difference and enhance the generalization ability of the model. Secondly, multilevel feature fusion modules were embedded at different depths of the network to enhance the representation ability of small targets. The proposed method is evaluated on multiple datasets, and the results show the effectiveness of the method.

KEYWORDS

smoke detection, unsupervised domain adaptive object detection, domain alignment, small object detection, feature fusion

1 Introduction

Natural disasters have always been the main cause of power grid failures. Among them, forest fires are easy to cause serious failures of multiple transmission lines due to coupling, causing irreparable losses to power equipment, posing a great threat to the safe operation of the power system, and even affecting people's normal life. The early occurrence of wildfire is often accompanied by the rise of smoke. Therefore, smoke detection is an important method to effectively avoid fire hazards.

Thanks to the rapid development of deep learning [1–6] and the wide applications in other computer vision tasks such as image fusion [7], image dehazing [8] and semantic segmentation [9], the performances of smoke detection have been remarkably improved in recent year, there are still many difficulties to detect smoke in real time. Usually, the training of deep learning models requires a large amount of data, it is extremely difficult to collect thousands of smoke images and manually label in actual scenes. Some researchers have proposed synthetic smoke datasets [10] to make up for this defect. However, due to the domain gap between the synthetic smoke and the real scene smoke, the performance of the detection model is limited. Figure 1

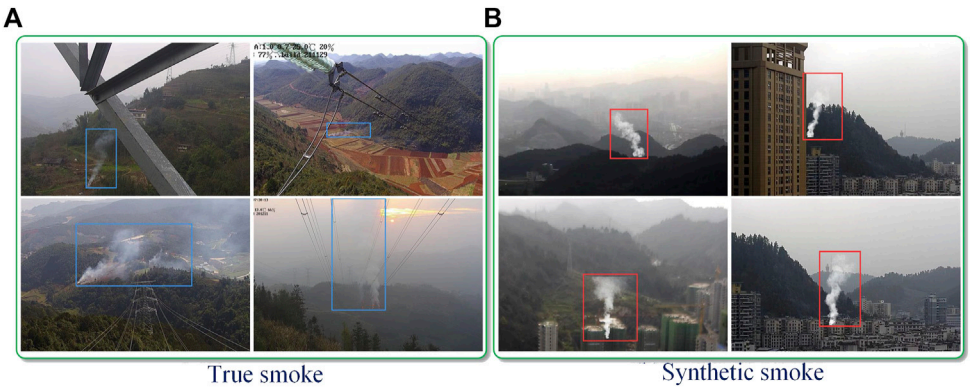

FIGURE 1
(A) Real smoke and **(B)** synthetic smoke. Real smoke has variable color and shape, while synthetic smoke [10] has relatively fixed shape and color.

shows the synthetic smoke and real smoke, there is a large difference between them. In the real scene, due to the unknown weather conditions, it is easy to cause the color, shape and transparency of smoke to change, so that the smoke obtained by the acquisition equipment has differences in resolution, view and brightness, which further increases the difficulty of real-time detection of smoke and fire. It is necessary to study an algorithm that can transfer the knowledge learned from a labeled dataset (source domain) to another unlabeled dataset (target domain).

One approach to solve this problem is Unsupervised Domain Adaptive Object Detection (UDAOD) [11], which aims to adapt the detector using labeled source data and unlabeled target data to alleviate the performance degradation by learning a feature representation that is not affected by domain gap. Existing UDAOD methods can be classified into: style transfer based methods [12–14], self-training based methods [15,16], and domain alignment based methods [17–19].

The method based on style transfer usually uses GAN [20] to transfer the style of the target domain image to the source domain image, and then uses the transformed image to supervise training the detection network, to reduce the domain shift caused by the style difference. However, the smoke image obtained in the real scene has complex background, the image generated by style transfer is different from the real image to some extent, and GAN increases the calculation amount of the model, the final detection performance is highly dependent on the quality of the generated image. Therefore, such methods cannot be well applied to the smoke detection task in the actual scene.

The method based on self-training generally trains the detection model with the source data, then inputs the target data to predict pseudo-label, and finally fine-tunes the model with the pseudo-labels. However, the shape, color and background of smoke in the real scene are not fixed, which is easy to make the predicted pseudo-labels have noise. Fine-tuning the model with noisy pseudo-labels will reduce the detection performance of the model.

Domain alignment based methods achieve feature alignment by adversarial learning. Although such methods have achieved considerable improvement, they align the boundary distribution of the two domains without considering the category information, which may lead to incorrect alignment of samples from different categories of the source domain and the target domain, thus failing to train the best model. The detection category of this work is only smoke, so the above problem does not exist. Existing methods consider the alignment of global features, while this paper aligns features of different scales.

The existing smoke detection work [21,22] pays little attention to cross-domain detection. In order to solve the problems faced by smoke detection in real scenes, this paper proposes a domain adaptive smoke detection algorithm based on multi-level feature fusion and alignment. Specifically, considering the problem of small and fuzzy smoke caused by long shooting distance, the algorithm proposes a multi-level and multi-scale feature fusion strategy to enhance the feature representation ability of the model for small targets. In addition, in order to reduce the domain difference, the algorithm proposes a multi-level feature alignment strategy to reduce the distribution difference between the source domain and the target domain on different levels of features. Compared with the two-stage object detection method, the proposed method is based on YOLOv5 [23], which does not require candidate box prediction and screening, and improves the detection speed.

The main contributions of this paper contain:

- A Multilevel Feature Cooperative domain Alignment method is proposed to reduce the data distribution difference between the source domain and the target domain at the multi-scale feature level.
- A Multilevel Feature Fusion method is proposed to enhance the feature representation ability of small target smoke by fusing features of different scales at different levels of the network.
- The proposed method can perform end-to-end training and detection without additional candidate box calculation and screening, which ensures the training and detection efficiency of the model.

2 Related work

2.1 Object detection

Object detection is a task to classify and locate objects for a given image, which is one of the important research contents in Computer Vision. Recently, deep learning based methods can be divided into two-stage object detection and single-stage object detection.

The two-stage object detection algorithm include two steps: the first step generates the candidate regions, and the second step classifies the candidate regions and regress their positions. The basic idea is to generate regions with high recall such that all objects on the image belong to at least one candidate region. In the second step, the candidate regions generated in the first step are classified by a deep model. Typical two-stage object detection algorithms include R-CNN [24], SPP-net [25], Fast R-CNN [26], Faster R-CNN [27], etc. Due to the large number of candidate regions generated by these algorithms, there are more repeated information and more invalid regions, which leads to large amount of calculation and slow detection speed.

Because the two-stage method needs to process a large number of candidate regions in turn, the detection speed is generally slow. To solve this problem, the one-stage object detection algorithm came into being. Compared with the two-stage object detection algorithm, the one-stage object detection algorithm does not need to generate candidate regions, and directly returns the object category and location on the input image, so the detection speed is faster, but the accuracy is slightly worse. Typical one-stage object detection algorithms include YOLO [28], SSD [29], etc.

2.2 Domain adaptation for object detection

Domain adaptation has been widely studied in Computer Vision. The object detection method based on deep learning is affected by the domain shift, and the network trained on one dataset often performs poorly on other datasets, which is often encountered in real scenarios. Unsupervised Domain Adaptive Object Detection (UDAOD) [30] aims to reduce the domain gap between training data and test data and improve detection performance. Existing UDAOD methods can be divided into: style transfer based methods, self-training based methods, and domain alignment based methods.

Object detection based on style transfer is a popular method in the past few years, and the representative literatures of this kind of method are [12–14]. Hsu et al. [12] proposed a progressive domain adaptation method, which decomposed the problem into two subtasks. Firstly, based on Cycle-GAN [31], synthesized an intermediate domain located in the distribution of the source domain and the target domain, and then adopted a progressive adaptation strategy to gradually narrow the domain gap through the intermediate domain. Inoue et al. [13] believe that the

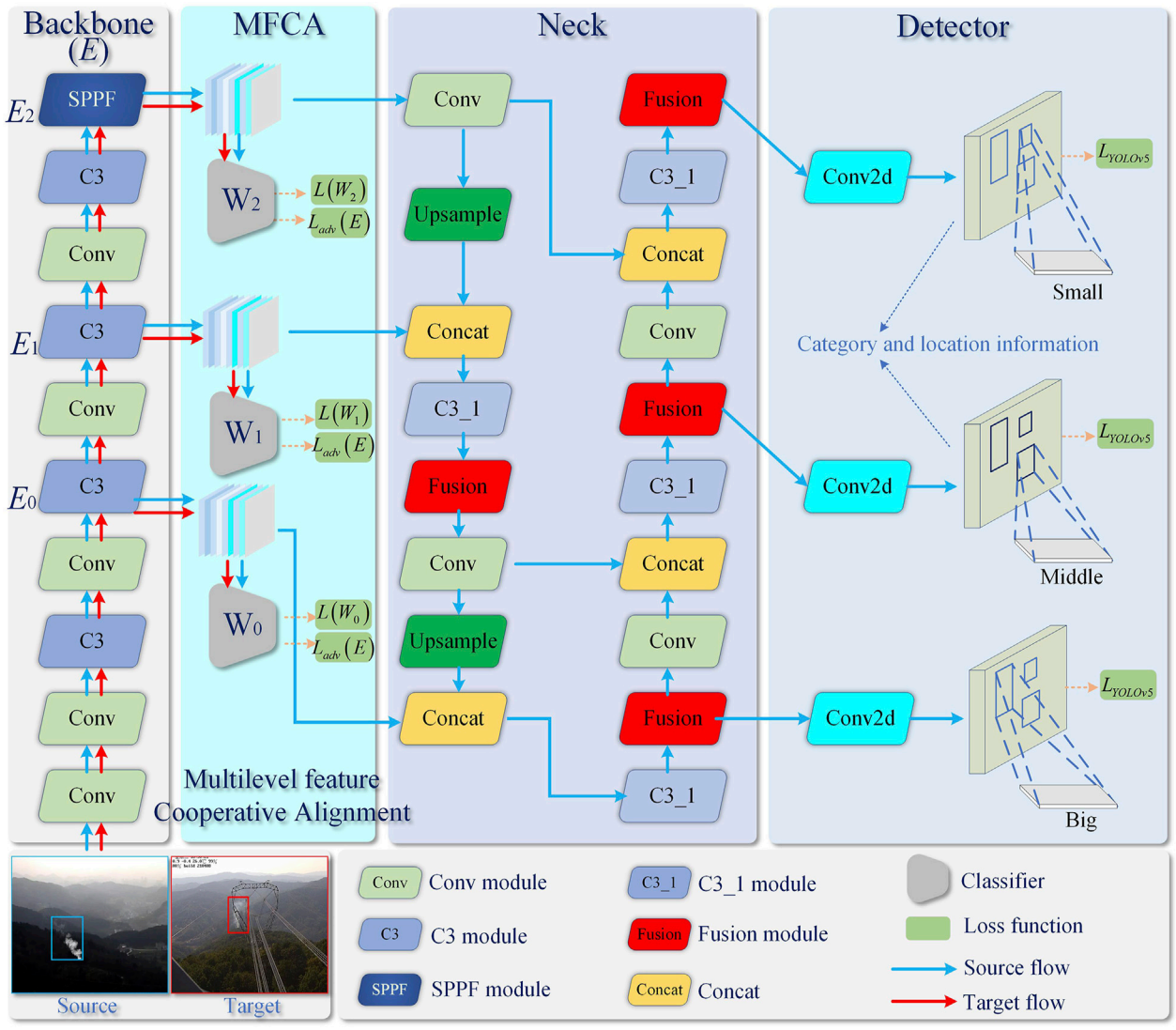

FIGURE 2
The pipeline of our proposed method. See the experiment section for more network details.

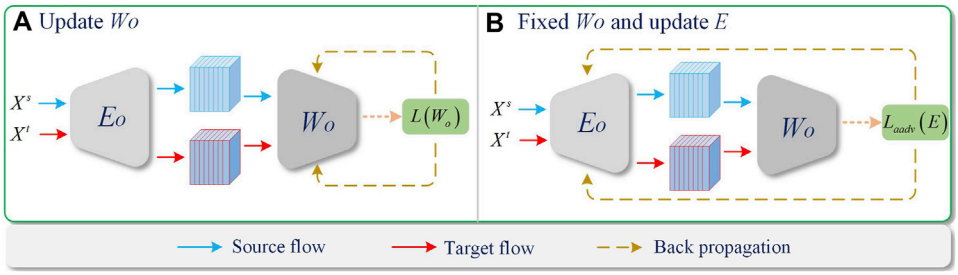

FIGURE 3
Training process. **(A)** Update W_o, **(B)** Fixed W_o and update E.

differences between the source and target domains mainly lie in their underlying features, such as color and texture. By generating similar images with the target domain images based on Cycle-GAN to capture these differences, and then fine-tuning the fully supervised trained detector use the generated images to make the detector robustly to these differences. Kim et al. [14] proposed a two-stage method of Domain Diversification and Multi-domain Invariant Representation Learning to alleviate pixel-level and feature-level domain differences at the same time. In the Domain Diversification stage, samples with different domain differences are generated from labeled source domain data to improve the adaptability of the model. Such methods alleviate the impact of domain differences to a certain extent, but the introduced GAN network increases the amount of computation, and the accuracy of the detector highly depends on the quality of the generated image, which is not suitable for real scenes.

Self-training based object detection [15,16] generally predicts pseudo-labels in the predicted target domain, and then uses the predicted pseudo-labels to fine-tune the model. However, the noise contained in the pseudo-labels will have a negative impact on the performance of the model. Kim et al. [15] proposed a weak self-training method to reduce the adverse effects of inaccurate pseudo-labels and stabilize the learning process. RoyChowdhury et al. [16] proposed an improved knowledge distillation loss by using existing high-confidence detectors to directly obtain the pseudo-labels of the target domain, and studied several methods to assign soft labels to the training samples of the target domain.

Object detection based on domain alignment [17–19] is one of the more commonly used methods. Wu et al. [17] proposed a disentangled representation method based on vector decomposition, attempting to disentangle the representation of domain-invariant features and domain-specific features, to realize domain alignment. Saito et al. [18] proposed a weakly alignment model, which uses adversarial learning to focus the adversarial alignment loss on globally similar images and pays less attention to globally dissimilar images. Zhu et al. [19] believe that the traditional domain adaptive method to align the whole image, while the object detection essentially focuses on the region of interest (local region), and propose to only focus on the relevant region and perform domain alignment.

2.3 Smoke detection

In recent years, researchers have proposed many smoke detection algorithms based on deep learning. Yin et al. [32] proposed a convolutional neural network with depth normalization to automatically extract smoke features and classify them, which reduced the influence of smoke shape and color to a certain extent. In order to further solve the problem of smoke shape and color changes, Gu et al. [33] proposed a dual-channel neural network by successively connecting multiple convolutional layers and Max pooling layers. A batch normalization layer is then selectively attached to each convolutional layer to prevent overfitting and speed up training. zhao et al. [34] proposed that the depth-wise separable method with fixed convolution kernel instead of training iteration was used for smoke detection, which could improve the speed of the algorithm and meet the requirements of real-time fire propagation for speed detection [21]. proposed a Convolutional Neural Network (CNN) -based smoke detection and segmentation framework for clear and hazy environments, employing EfficientNet for better smoke detection [35]. proposed a smoke detection method in normal and foggy weather that combines attention mechanism with feature-level and decision-level fusion modules. An attention mechanism module combining spatial attention and channel attention was proposed to solve the problem of small smoke detection. Secondly, a lightweight feature-level and decision-level fusion module is proposed, which can not only improve the recognition ability of similar objects such as smoke and fog, but also ensure the real-time performance of the model. Zhan et al. [36] proposed a recursive pyramid network with deconvolution and dilated convolution to solve the problem of low detection accuracy caused by high smoke transparency and unclear edges.

Most of the existing smoke detection methods are trained and tested on the same dataset, combined with practical application, this paper focuses on smoke detection in real scenes, and proposes a domain adaptive smoke detection algorithm based on Multilevel feature Cooperative Alignment and Fusion.

3 Proposed method

The structure of Multilevel feature Cooperative Alignment and Fusion (MCAF) algorithm is shown in Figure 2. The algorithm takes YOLOv5 [23] object detection network as the baseline, and is composed of Multilevel Feature Cooperative Alignment module (MFCA) and Multilevel Feature Fusion module (MFF). Specifically, for the given smoke image in the source domain and the target domain, Backbone (denoted as E) is used to extract smoke-related features, and then domain alignment is achieved through the cooperation between multi-scale classifiers W_0, W_1, W_2 in MFCA

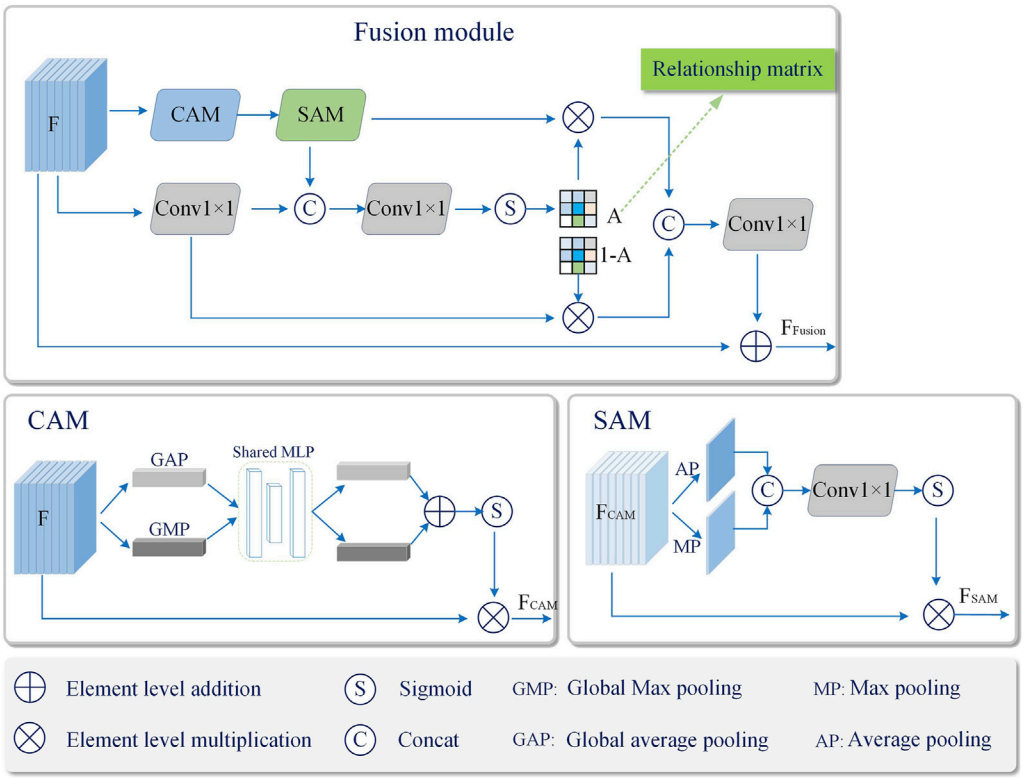

FIGURE 4
The structure of Fusion model.

to reduce the domain differences between the source and target domain features, where, W_0, W_1, W_2 has the same structure, which are consists of a global average pooling, fully connected layer. The difference is that their input feature sizes are not the same. Finally, in the Neck of the detector network (The object detectors developed in recent years often insert some layers between the backbone and the detector, people usually call this part the Neck of the detector) embed feature fusion module at different positions to make the obtained features better adapt to targets of different sizes. Through the end-to-end training of MCAF algorithm, a better detection effect is obtained.

3.1 Model pretrain

In order to make the smoke detection network adapt to real scenarios, pre-training is carried out first. The smoke data in the source and target domains are defined as $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{N^s}$ and $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{N^t}$, where $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ denote the ith sample of the training set in the source and target domains, respectively, N^s and N^t denote the number of training samples in the source and target domains, respectively. During training, the object detection network is optimized by minimizing the following loss function,

$$L_{YOLOv5} = L_{class} + L_{obj} + L_{loc} \tag{1}$$

where L_{YOLOv5} denote the total loss function of YOLOv5 [23]. L_{class}, L_{obj}, L_{loc} denote the classification loss, confidence loss and localization loss, respectively.

In addition, in order to make W_0, W_1, W_2 have the ability to distinguish features from the source domain or the target domain, the following loss function is minimized to optimize,

$$L_{id}(W_o) = \sum_{o=0}^{2} CE\left(W_o\left(E_o\left(\mathbf{x}_i^s\right)\right), y_0\right) + CE\left(W_o\left(E_o\left(\mathbf{x}_i^t\right)\right), y_1\right) \tag{2}$$

where $CE(\cdot)$ denotes the cross-entropy loss, E_o ($o = 0, 1, 2$) denote the features output by different network depths of E (the details can be seen in Figure 2), $y_0 = 0$ and $y_1 = 1$ denote the domain labels of the source and target domains, respectively.

Through model pre-training, the object detection network has a basic detection ability and W_0, W_1, W_2 can distinguish the source domain and target domain samples. However, when testing on unseen smoke datasets, the detection performance will drop sharply due to the domain shift between different datasets. In order to reduce the data distribution difference between the source domain and the target domain, this paper proposes Multilevel Feature Cooperative Alignment module.

3.2 Multilevel feature cooperative alignment (MFCA)

In order to mitigate the impact of inter-domain differences, this paper proposes a Multilevel Feature Aooperative domain Alignment module. The gap between the source and target domains is narrowed by adversarial learning between E and W_0, W_1, W_2. In theory, if the extracted features from source and target domain do not have

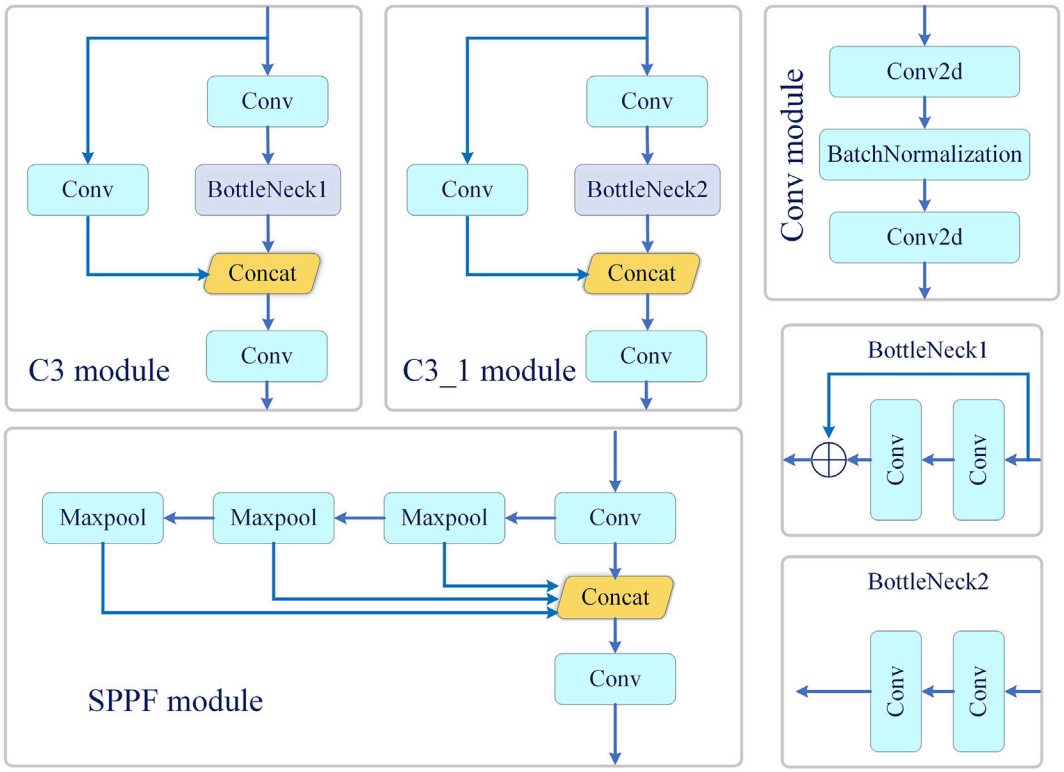

FIGURE 5
The details of the C3 module, SPPF module and other module in Figure 2.

differences, W_0, W_1, W_2 should not be able to distinguish the source domain and the target domain. By minimizing the following loss function to realize,

$$L_{adv}(E) = \sum_{o=0}^{2} CE\left(W_o\left(E_o\left(\mathbf{x}_i^s\right)\right), y_1\right) + CE\left(W_o\left(E_o\left(\mathbf{x}_i^t\right)\right), y_0\right) \quad (3)$$

At this time, the parameters of E (where E include E_0, E_1, E_2) are updated by fixing the parameters of W_0, W_1, W_2, and the source domain and target domain features are cross-constrained by domain labels. In this way, the extracted features are trained to adapt to the source domain and target domain, and the effect of domain alignment is achieved. The training process is shown in Figure 3. It is worth noting that, this paper not only uses the above way to mitigate the domain differences in the final features of Backbone, but also constrains the intermediate features of Backbone at the same time, and finally alleviates the impact of domain differences through the cooperative alignment of multi-level features.

3.3 Multilevel feature fusion (MFF)

In real scenes, smoke changes with time in different colors and shapes, which increases the difficulty of feature extraction. In order to improve the robustness of smoke features, this paper proposes to embed a Multilevel Feature Fusion module (MFF) in the Neck part of the detection network, the design of this module is shown in Figure 4.

Specifically, for the feature map $\boldsymbol{F}$ output by C3_1 module in Neck, it is divided into two branches. Follow CBAM [37], the first branch is enhanced by Channel Attention Module (CAM) and Spatial attention module (SAM), and the second branch is enhanced by 1×1 convolutional layer (Conv1×1) to adjust the feature dimensions and increase the non-linear mapping ability of the network. The deep features are further extracted, and then the correlation matrix $\boldsymbol{A}$ is calculated for the features of the two branches,

$$\boldsymbol{A} = Sigmoid\left(Conv\left[SAM\left(CAM\left(\boldsymbol{F}\right)\right); Conv\left(\boldsymbol{F}\right)\right]\right) \quad (4)$$

where, $Sigmoid\ (\cdot)$ represents the sigmoid activation function, $\boldsymbol{F}$ represents the output of C3_1 module in Neck, $[a; b]$ denotes concatenation of a and b, Conv denotes 1×1 convolution, SAM and CAM denote spatial attention module and channel attention module, respectively. The relation matrix $\boldsymbol{A}$ reflects the relationship between the corresponding positions of the feature maps obtained by the two branches, and the larger the value is, the more important it is. Finally, the feature maps of the two branches are fused by the following operations, and the fused features F_{Fusion} are used as the input of the later network layer.

$$F_{Fusion} = Conv\left(\left[\boldsymbol{A} \otimes F_{SAM}; (1 - \boldsymbol{A}) Conv\left(\boldsymbol{F}\right)\right]\right) \oplus \boldsymbol{F} \quad (5)$$

where, $F_{SAM} = SAM(CAM(\boldsymbol{F}))$ denotes the output of the spatial attention module, $\otimes$ denotes element-wise multiplication, and $\oplus$ denotes element-wise addition. It can be seen that the proposed fusion module adaptively adjusts the contribution of the two branches at the corresponding positions of $\boldsymbol{A}$ and 1-$\boldsymbol{A}$, so as to

TABLE 1 Comparison with other methods on RF → TS, RF → SF, SF → TS, and SF → RF. P and R denote Precision and Recall (%),respectively.

Methods	RF → TS			RF → SF			SF → TS			SF → RF		
	P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP
YOLOv3	1.79	17.4	12.7	3.69	71.3	78.2	1.33	15.9	8.72	2.85	85.7	59.0
YOLOv5s	21.4	18.6	14.3	84.4	74.6	80.3	16.9	9.77	9.20	64.5	64.8	58.4
Faster-RCNN	—	—	8.72	—	—	65.95	—	—	6.23	—	—	44.54
MCAF	33.6	23.9	21.6	86.9	82.8	88.5	30.6	17.6	14.3	67.4	66.7	66.0

TABLE 2 The running time of Faster-RCNN and MCAF in several setting.

Methods	RF → TS		SF → TS	
	Training time (/h)	Testing time (/s)	Training time (/h)	Testing time (/s)
Faster-RCNN	≈ 14.3	102.6	≈ 13.5	103.4
MCAF	≈ 1.5	7.2	≈ 1.3	7.1

achieve a better fusion effect, and finally improve the robustness of smoke features and enhance the representation ability of small target smoke.

3.4 Optimization

By considering all the loss functions jointly, the objective function in this paper is as follows,

$$L_{total}(E, W_o) = L_{YOLOv5} + L_{id}(W_o) + L_{adv}(E) \tag{6}$$

Firstly, the detection network and classifier are trained to have the basic smoke detection ability and the ability to distinguish the source domain and the target domain by L_{YOLOv5} and $L_{id}(W_o)$, respectively. Then, the adversarial learning strategy is used to alleviate the differences between the source domain and the target domain through $L_{adv}(E)$.

4 Experiments

In order to prove the effectiveness of the proposed method (MCAF), this chapter carries out a large number of experiments. Firstly, the data set used in the experiment is introduced, and then the performance of the proposed method is compared with that of classical object detection algorithms. Finally, the effectiveness and superiority of the proposed method are demonstrated by ablation experiments.

4.1 Datasets and evaluation protocol

The real scene dataset *True_smoke* (*TS*) used in this experiment contains a total of 4,128 images. Among them, 1,275 images were taken from real transmission lines, 2,853 images were taken from google search engine and State Key Laboratory of Fire Science of University of Science and Technology of China. The training set and test set were divided according to a ratio of 7:3, and LabelImg was used for annotation, and the annotation format was the same as that of the popular dataset PAS-CAL VOC [38], the annotation information was stored in the. *xml* file. In addition, the synthetic datasets *RFdataset* (*RF*) [10] and *SFdataset* (*SF*) [10] are also used for experiments. *RFdataset* (*RF*) is synthesized from real smoke and forest background and contains 12,620 images, where, 3,155 images are used for training and 6,310 images are used for testing. The *SFdataset* (*SF*) is synthesized from simulated smoke and forest background and contains 12,620 images, where, 3,155 images are used for training and 6,310 images are used for testing.

In this paper, Precision, Recall and mean average precision (mAP) are used as performance evaluation indicators. It is calculated as follows,

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

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

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

where, *TP* is the number of samples correctly predicted as smoke, *FP* is the number of samples correctly predicted as smoke, *FN* is the number of samples correctly predicted as smoke, *N* is the total number of classes, and AP_i is the Average Precision of class *i*.

4.2 Implementation details

Considering that the actual application needs to deploy the algorithm to mobile devices, the YOLOv5s object detection network is used as the basic framework, and the detailed structure of each module is shown in Figure 5. In the training phase, the

10.3389/fphy.2023.1136021

TABLE 3 Ablation study. P and R denote Precision and Recall (%),respectively. MCAF represents the proposed method.

Methods	RF → TS			SF → TS		
	P	R	mAP	P	R	mAP
Baseline	21.4	18.6	14.3	16.9	9.77	9.2
Baseline+MFCA	33.1	20.2	17.2	18.1	16.8	11.2
Baseline+MFF	33.8	20.4	18.6	26.6	14.0	12.3
Baseline+MFCA+MFF (MCAF)	34.6	23.9	21.6	30.6	17.6	14.3

FIGURE 6
Detection result display. The left is the result of the proposed method, and the right is the result of baseline.

maximum of epochs is set to 100, in which the first 20 epochs are fore pretrain. Being similarly to YOLOv5, Mosaic, random cropping, horizontal flipping, etc., are used for data augmentation. the size of the input image is uniformly resized to 640 × 640×3 (for length, width and channel), and the feature maps output by the layer concatenated with Neck in backbone are used as the input of MFCA module (specifically, the outputs of the 4th, 6th and 9th layers of backbone are used as the input of MFCA module). The feature map sizes of the input three domain classifiers are 80 × 80×128, 40 × 40×256, and 20 × 20×512, respectively. The three classification networks have the same structure, consisting of global average pooling and fully connected layers. In addition, the MFF module is embedded behind the C3_1 module in Neck. For the training of Backbone, the SGD optimizer was used with the learning rate set to 0.01 and momentum to 0.3, and for the training of the three classification networks, the SGD optimizer was used with the learning rate set to 0.1 and momentum to 0.9. The batchsize for training and testing both set to 16. This experiment was performed on pytorch 1.13 [39], and all experiments were done on a Linux server for NVIDIA GEForce RTX3090Ti.

4.3 Comparison to other methods

At present, there is no public dataset for cross-domain smoke detection. In addition, existing works are trained under supervised conditions and cannot be directly compared. The comparison method in this paper uses more mature object detection methods. These methods include YOLOv3 [40], YOLOv5s [23], Faster-RCNN [27]. The comparative experimental Settings are RF → TS, SF → TS, RF → SF, SF → RF, (a → b represents a as the source domain and b as the

target domain cross-domain task), and the target domain category and location labels is unknown during training.

The experimental results are shown in Table 1. It can be seen that the mAP of the proposed method is much higher than that of classical object detection methods such as YOLOv3 and faster-RCNN. Such methods do not consider the inter-domain differences and thus perform poorly. Compared with YOLOv5s in the four experimental Settings, the mAP of the method in this paper is increased by 7.3%, 8.2%, 5.1%, 7.6% respectively, indicating that the method in this paper indeed enhances the ability of the model to extract smoke robust features.

Table 2 shows the comparison of training and testing time between the proposed method and Faster-RCNN. It can be seen that comparing with Faster-RCNN, the proposed method is more suitable for real-time smoke detection and more efficient.

4.4 Ablation study

This section discusses the ablation study. Firstly, the L_{YOLOv5}-guided optimized network is considered as Baseline. On the basis of Baseline, the Multilevel Feature Cooperative Alignment module and the Multilevel Feature Fusion module are gradually added, which proves that they are helpful to improve the performance, the results can be seen in Table 3. Ablation experiments were performed at RF → TS and SF → TS.

4.4.1 The effectiveness of multilevel feature cooperative alignment module (MFCA)

In order to alleviate the domain gap existing in the source domain and the target domain, a multilevel feature cooperative alignment module is proposed. By following the adversarial strategy of E and the domain classifier, removing the gap between the source and target domain. As shown in Table 3, Baseline on RF → TS, SF → TS Precision/Recall/mAP respectively is 21.4%/18.6%/14.3% and 16.9%/9.77%/9.2, When the Multilevel Feature Cooperative Alignment module is added, the performance is significantly improved, which proves the effectiveness of this module.

4.4.2 The effectiveness of multilevel feature fusion module (MFF)

In order to improve the feature representation ability of small target smoke, a multilevel feature fusion module is proposed. The feature fusion module is embedded in different depths of Neck in the detection network to enhance the features of different scales. As can be seen from Table 3, after adding the Multilevel Feature Fusion module on the basis of baseline, the Precision/Recall/mAP on RF → TS and SF → TS Raised to 33.8%/20.4%/18.6% and 26.6% 14.0%/12.3%. The validity of the module is proved.

As shown in Table 3, when the Multilevel Feature Cooperative Alignment module and the Multilevel Feature Fusion module are added to baseline at the same time, the overall performance is improved, indicating that the proposed method is effective.

In addition, Figure 6 shows the visualization of the detection results. Clearly, with the embedding of the proposed technique, the models become more powerful in terms of detection, confirming their significance.

5 Conclusion

Fire prevention is of great significance to the protection of human property safety, natural environment and industrial equipment. Smoke detection is helpful in the early warning of fire, and many researchers continue to improve the detection algorithm to meet the needs of this field. In order to adapt to smoke detection in real scenes, this paper proposes an unsupervised domain adaptive smoke detection algorithm based on multi-level feature fusion and cooperative alignment. On the one hand, the difference between the source domain and the target domain data is reduced by the cooperative alignment of features at different scales. On the other hand, by embedding fusion modules at different depths of Neck, the representation ability of features is enhanced. In this paper, the module structure, training method, loss function and network parameter setting of the proposed method are introduced in detail. The effectiveness of each module is proved by ablation experiments.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://smoke.ustc.edu.cn/datasets.htm.

Author contributions

FZ responsible for paper scheme design, experiment and paper writing. GW and YiM collectiong data. YW and YuM annotating data. GW and HP guide to do experiments and write papers. KW guide the paper scheme design and revision.

Funding

This work was supported by the Major scientific and technological projects of Yunnan Province Research on Key Technologies of ecological environment monitoring and intelligent management of natural resources in Yunnan (202202AD080010); Fundamental Research Fund of Science and Technology Department of Yunnan Province (202201AU070172).

Conflict of interest

Authors FZ, GW, YiM, YW, YuM, GW, and HP were employed by the Company Yunnan Power Grid Company Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Li H, Yan S, Yu Z, Tao D. Attribute-identity embedding and self-supervised learning for scalable person re-identification. *IEEE Trans Circuits Syst Video Technol* (2019) 30: 3472–85. doi:10.1109/tcsvt.2019.2952550

2. Li H, Chen Y, Tao D, Yu Z, Qi G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Trans Inf Forensics Security* (2020) 16:1480–94. doi:10.1109/tifs.2020.3036800

3. Li H, Dong N, Yu Z, Tao D, Qi G. Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. *IEEE Trans Circuits Syst Video Technol* (2021) 32:2814–30. doi:10.1109/tcsvt.2021.3099943

4. Li H, Xu K, Li J, Yu Z. Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification. *Knowledge-Based Syst* (2022) 251:109315. doi:10.1016/j.knosys.2022.109315

5. Lin X, Li J, Ma Z, Li H, Li S, Xu K, et al. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18-24 June 2022; New Orleans, LA, USA (2022). p. 20973–20982.

6. Wang Y, Qi G, Li S, Chai Y, Li H. Body part-level domain alignment for domain-adaptive person re-identification with transformer framework. *IEEE Trans Inf Forensics Security* (2022) 17:3321–34. doi:10.1109/tifs.2022.3207893

7. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

8. Berman D, Avidan S, Treibitz T. Non-local image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 27-30 June 2016; Las Vegas, NV, USA (2016). p. 1674–1682.

9. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. *Inf Fusion* (2023) 91:376–87. doi:10.1016/j.inffus.2022.10.022

10. Zhang Q, Lin G, Zhang Y, Xu G, Wang J. Wildland forest fire smoke detection based on faster r-cnn using synthetic smoke images. *Proced Eng* (2018) 211:441–6. doi:10.1016/j.proeng.2017.12.034

11. Li Y-J, Dai X, Ma C-Y, Liu Y-C, Chen K, Wu B, et al. Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18-24 June 2022; New Orleans, LA, USA (2022). p. 7581–7590.

12. Hsu H-K, Yao C-H, Tsai Y-H, Hung W-C, Tseng H-Y, Singh M, et al. Progressive domain adaptation for object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision; March 1-5, 2020; Snowmass Village, CO (2020). p. 749–757.

13. Inoue N, Furuta R, Yamasaki T, Aizawa K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 18 2018 to June 23 2018; Salt Lake City, UT, USA (2018). p. 5001–5009.

14. Kim T, Jeong M, Kim S, Choi S, Kim C. Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 12456–12465.

15. Kim S, Choi J, Kim T, Kim C. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 27 October 2019 - 02 November 2019; Seoul, Korea (South) (2019). p. 6092–6101.

16. RoyChowdhury A, Chakrabarty P, Singh A, Jin S, Jiang H, Cao L, et al. Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 780–790.

17. Wu A, Liu R, Han Y, Zhu L, Yang Y. Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 10-17 October 2021; Montreal, QC, Canada (2021). p. 9342–9351.

18. Saito K, Ushiku Y, Harada T, Saenko K. Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 6956–6965.

19. Zhu X, Pang J, Yang C, Shi J, Lin D. Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15-20 June 2019; Long Beach, CA, USA (2019). p. 687–696.

20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* (2020) 63:139–44. doi:10.1145/3422622

21. Khan S, Muhammad K, Hussain T, Del Ser J, Cuzzolin F, Bhattacharyya S, et al. Deepsmoke: Deep learning model for smoke detection and segmentation in outdoor environments. *Expert Syst Appl* (2021) 182:115125. doi:10.1016/j.eswa.2021.115125

22. Liu H, Lei F, Tong C, Cui C, Wu L. Visual smoke detection based on ensemble deep cnns. *Displays* (2021) 69:102020. doi:10.1016/j.displa.2021.102020

23. ultralytics. yolov5 (2020). Available at: https://github.com/ultralytics/yolov5 (accessed on may 18, 2020).

24. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 23-28 June 2014; Columbus, OH, USA (2014). p. 580–587.

25. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans pattern Anal machine intelligence* (2015) 37:1904–16. doi:10.1109/tpami.2015.2389824

26. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision; December 7 - 13, 2015; Santiago, Chile (2015). p. 1440–1448.

27. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: C Cortes N Lawrence, editors. *Advances in neural information processing systems*. New York: Curran Associates, Inc (2015).

28. 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; 27-30 June 2016; Las Vegas, NV, USA (2016). p. 779–788.

29. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: Single shot multibox detector. In: European conference on computer vision; 8-16 October; Amsterdam, The Netherlands. Springer (2016). p. 21–37.

30. Chen Y, Li W, Sakaridis C, Dai D, Van Gool L. Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 18-23 June 2018; Salt Lake City, UT, USA (2018). p. 3339–3348.

31. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 22-29 October 2017; Venice, Italy (2017). p. 2223–2232.

32. Yin Z, Wan B, Yuan F, Xia X, Shi J. A deep normalization and convolutional neural network for image smoke detection. *Ieee Access* (2017) 5:18429–38. doi:10.1109/access.2017.2747399

33. Gu K, Xia Z, Qiao J, Lin W. Deep dual-channel neural network for image-based smoke detection. *IEEE Trans Multimedia* (2019) 22:311–23. doi:10.1109/tmm.2019.2929009

34. Zhao Y, Zhang H, Zhang X, Chen X. Fire smoke detection based on target-awareness and depthwise convolutions. *Multimedia Tools Appl* (2021) 80:27407–21. doi:10.1007/s11042-021-11037-1

35. He L, Gong X, Zhang S, Wang L, Li F. Efficient attention based deep fusion cnn for smoke detection in fog environment. *Neurocomputing* (2021) 434:224–38. doi:10.1016/j.neucom.2021.01.024

36. Zhan J, Hu Y, Zhou G, Wang Y, Cai W, Li L. A high-precision forest fire smoke detection approach based on argnet. *Comput Electron Agric* (2022) 196:106874. doi:10.1016/j.compag.2022.106874

37. Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); September 8-14, 2018; Munich, Germany (2018). p. 3–19.

38. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. *Int J Comput Vis* (2010) 88:303–38. doi:10.1007/s11263-009-0275-4

39. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in neural information processing systems*. New York: Curran Associates, Inc (2019).

40. Redmon J, Farhadi A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).