MMFNet: Forest Fire Smoke Detection Using Multiscale Convergence Coordinated Pyramid Network With Mixed Attention and Fast-Robust NMS

Liangji Zhang, Chao Lu, Haiwen Xu, Aibin Chen, Liujun Li, *Member, IEEE*, and Guoxiong Zhou

***Abstract*—There is a problem in the field of early automatic detection of forest fire smoke that due to low concentration or tiny size, some smoke is difficult to capture. This article proposes a multiscale convergence coordinated pyramid network (MCCPN) with mixed attention and Fast-robust NMS (MMFNet) for the fast detection of forest fire smoke. First, an MCCPN is designed, which combines a dual-attention feature pyramid network and a coordinated convergence module. It improves the detection rate of targets of different sizes. Second, a mixed attention module is designed to focus more on the smoke in the image and enhance the extraction of horizontal and vertical features of smoke. Then, a Fast-robust nonmaximum suppression is proposed to accelerate the convergence of bounding boxes and increase the accuracy of the prediction box. Finally, a forest fire detection system of the Internet of Things using MMFNet is built. The experimental results show that our method achieves 80.72% AP, 88.52% AP^{50}, 83.45% AP^{75}, 46.88% AR, and 154 FPS, which is superior to the state-of-art forest fire smoke detection methods.**

***Index Terms*—Dual-attention feature pyramid network (DAFPN), forest fire smoke detection, Internet of Things (IoT), mixed attention module, multiscale convergence coordinated pyramid network (MCCPN).**

I. Introduction

FORESTS cover one third of the Earth's land area. It performs an important function across the globe. Forests are not only valuable resources necessary for human survival and development but also important safeguards for the Earth's ecological environment and biodiversity. Forest fire is a serious natural disaster that occurs more frequently around the world in recent years, seriously endangering the forest environment ecosystem and posing a great threat to the safety of human life and property.

Therefore, the monitoring and rapid positioning of forest fires are of great significance. In the early stages of forest fires, flames are not easy to detect because they often obscured by tall and dense trees. Comparatively, due to the slow ascendance of temperature and evaporation of large amount of water accompanied with producing of plenty of smoke in the initial stage, the rising smoke is easier to be captured with highly deployed cameras. Therefore, the detection of forest fire smoke can give out early warning information and helps locate the accident as soon as possible, which is useful when controlling the fire from becoming large-scale forest fires.

Early fire identification was realized by sensors based on temperature, smoke particles, and other physical data sampled [1]. It performs well in the fixed environment, but it will be influenced by outdoor interferences and the complexity of the environment, resulting in a significant decrease in the detection accuracy. Meanwhile, this method cannot obtain real-time on-site images. The traditional visual sensor fire detection method uses the artificial design of flame detection features, such as color, texture, dynamic, and so on. Among them, color is the most important feature of fire detection, and there are effective color models, such as RGB, HIS, YCbCr, etc. Zaidi et al. [2] proposed a fire detection method based on RGB and YCbCr features. Li et al. [3] proposed a video self-image based on the Gaussian color mixing model of the Dirichlet process main flame detection model. Wu et al. [4] carried out feature extraction on forest fires and established a fire identification model through comprehensive extraction of color, texture, shape, and other features.

Compared to traditional manual extraction of features in computer vision detection, deep learning methods can extract more abstract and deeper features. Frizzi et al. [5] were the first to use CNNs to identify flames and smoke in videos and achieved results comparable to traditional methods, proving that deep learning methods are very promising in the field of flame and smoke detection. Wu and Zhang [6] performed flame detection with the mainstream target detection frameworks Faster R-CNN [7], YOLOv1 [8], and SSD [9] and found that SSD had the highest accuracy and the best detection

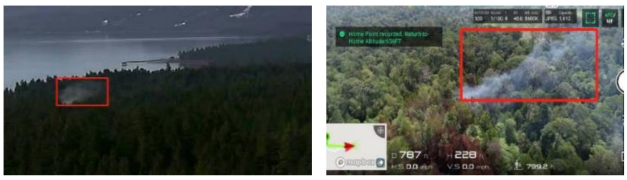
Manuscript received 28 October 2022; revised 2 February 2023 and 31 March 2023; accepted 15 May 2023. Date of publication 18 May 2023; date of current version 9 October 2023. This work was supported by the Natural Science Foundation of China under Grant 61902436. *(Chao Lu is co-first author.) (Corresponding author: Guoxiong Zhou.)*

Liangji Zhang, Chao Lu, Aibin Chen, and Guoxiong Zhou are with the School of Computer Information and Engineering, Central South Forestry University of Technology, Changsha 410018, China (e-mail: ioo0614001x@163.com; 1244414754@qq.com; hotaibin@163.com; t20060599@csuft.edu.cn).

Haiwen Xu is with the Department of Forest and Grassland Fire Prevention Monitoring and Dispatching Evaluation Center of Hunan Province, Changsha 410004, China (e-mail: hivenit@163.com).

Liujun Li is with the Precision Agriculture and Intelligent Robotics, Department of Soil and Water Systems, University of Idaho, Moscow, ID 83844 USA (e-mail: liujunl@uidaho.edu).

Digital Object Identifier 10.1109/JIOT.2023.3277511



(a) (b)

Fig. 1. Smoke detection interference factors. (a) Smoke of tiny size. (b) Smoke in low concentration.

efficiency and proposed the lightweight network tiny-YOLO for mobile devices. Li and Zhao [10] performed flame detection with Faster R-CNN, R-FCN [11], SSD, YOLOv3 [12], and other target detection algorithms. The performance of target detection algorithms and computer vision methods on flame detection indicates that the accuracy rate of the target detection algorithm is higher than that of the manual feature extraction algorithm, while the YOLOv3 algorithm has the highest detection accuracy and speed.

Two issues are to be addressed in the current forest fire smoke detection scene (as shown in Fig. 1). (a) *Smoke of Small Size:* Small target detection is always a difficult task. In terms of the forest fire smoke, the captured target is too tiny to recognize owing to the long distance between smoke and camera or the small size of the smoke itself. (b) *Smoke in Low Concentration:* When forest fire occurs, the smoke rises quickly. But in the very early stage, due to the high density and the height of vegetation, the smoke rises and appears in low concentration. As a result, the textural and contour features of the smoke are less pronounced.

The limitation of image resolution and contextual makes small target detection a challenging task. Although some previous work [13], [14], [15], [16], [17] has achieved good results in forest fire smoke detection, they are inability to optimize the detection of small targets. The methods in [18], [19], and [20] also failed in yielding high accuracy in detecting lighter concentrations of smoke. Existing deep-learning-based methods for optimal small target detection can be divided into five directions: 1) data augmentation; 2) training strategies; 3) contextual background information-based; 4) multiscale learning networks; and 5) generative adversarial networks. Data augmentation is a common method used to improve model performance in computer vision. The mainly used methods for data augmentation dealing small targets are replication augmentation [21], adaptive sampling [22], scale matching [23], scaling and stitching [24], and self-learning data augmentation [25]. In terms of training strategies, Singh and Davis [26] proposed scale normalization of image pyramids to the problem of recognition and detection due to extreme scale variations. In terms of contextual background information, Bell et al. [27] proposed an Inside–outside Network (IoN) to fuse the internal and external information of the region of interest, IoN, however, is complex and requires careful initialization for stable training. In terms of multiscale learning feature learning, Lin et al. [28] used it to construct high-level semantic feature maps at multiple scales by using feature pyramids network (FPN), and FPN fuses two adjacent feature maps by a top-down structure, but this method suffers from inadequate feature fusion. To make full use of both deep and shallow features, PANet [29] adds an additional bottom-up structure to FPN. ASFF [30] adjusts the contribution of feature maps at different scales by learning the weight parameters. NASFPN [31] uses neural architecture search to automatically search an FPN for fusing multiscale features while increasing much computational effort. In generative adversarial network-based detection, Rabbi et al. [32] proposed a new edge-enhanced super-resolution generative adversarial network for the small target detection problem.

At the early stage of the fire, the smoke is quite low in concentration, while in some complex context of the forests, there could be many smoke-like objects, such as white clouds floating in the sky, exposed gray and white rocks, which have similar semantic information with the smoke in the image. Traditional networks have difficulty in making full use of context information and often unable to distinguish feature differences, which result in low detection accuracy. To solve this problem, Ghiasi et al. [31] proposed an attention-enhanced bidirectional long and short-term memory network to distinguish the difference between smoke and smoke-like objects by spatial and temporal feature analysis of image sequences. Hu et al. [18] proposed a new value conversion-attention module to extract deep features of smoke by a joint weighting strategy for horizontal and vertical directions to distinguish smoke from smoke-like objects. We propose a new mixed attention to solve the problem that the network has difficulty in distinguishing smoke and nonsmoke due to low smoke concentration.

The main goal of our work is to achieve a fast and accurate forest fire smoke detection system that can detect low-concentration and small-size smoke. We hope that the designed object can be easily deployed, easily operated, and can be adapted to various forest environments. Our contributions are summarized as follows.

1) We take CSPDarkNet as the backbone and propose a multiscale convergence coordinated pyramid network (MCCPN) to make full use of semantic information of high-level features and fine-grained features of low-level features of smoke, thus realize effective extraction of features at different scales, improving the detection rate of small smoke targets. a) Dual-attention feature pyramid network (DAFPN) was proposed to enhance the networks' perception ability of large- and small-size targets and b) coordinated convergence module (CCM) was proposed to enable four scales of feature maps with different information integrate into three predictions.
2) A mixed attention which combined feature-highlighting module and cross attention is proposed in this article. It improved the feature expression of smoke target in regions where the contour features are not obvious caused by low concentration. The feature-highlighting module enables the network to be more focused on the smoke in the image, and the cross-attention module coordinates to enhance the extraction of horizontal and vertical features of smoke.

3) We proposed to use Fast-robust nonmaximum suppression (NMS) to make the target box regression more accurate and stable and accelerate the iterative regression.
4) A forest fire detection system of the Internet of Things (IoT) MMFNet is built. The experimental results indicate that our proposed method achieves 75.72% AP, 83.52% AP^{50}, 77.45% AP^{75}, 46.88% AR, and 154 FPS, which is competent for fast detection of forest fire smoke.

II. Related Work

Object detection can be divided into three steps: 1) target area selection; 2) feature extraction; and 3) classification and positioning. In object detection and other computer vision tasks, attention mechanisms and Fast NMS has been proved to be effective and widely used in many key steps. In this section, we will briefly introduce attention mechanisms in feature extraction, feature fusion, Intersection over Union (IoU), and NMS.

A. Feature Extraction

Generally, the data set of object detection task contains target objects of different sizes and different shapes. From the view of size, large objects can be easily detected while small ones are not. As to the view of shape, it is up to the object itself. Like our object smoke, it does not even have a fixed shape. Limited by kernel size, traditional convolution networks can only capture relationships in local area rather than considering more details or overall tradeoffs. Thus, it is difficult for computers to effectively learn features. There are many means for enhancing certain attributes such as introducing an attention mechanism to supplement convolution so that to improve the feature extraction ability. The attention modules commonly used in target detection are mainly divided into channel-wise attention and point-wise attention. The representatives of these two attention models are squeeze-and-excitation (SE) [34] and spatial attention module (SAM) [35]. However, there are problems companied with the models such as more time cost or slight improvement.

B. Feature Fusion

During the process of feature extraction, information sharing of different scales is significant to small target object detection. Feature fusion on different image scales is an important method to improve the performance of image recognition network. Low-level features have higher resolution and contain more location and detail information, but due to less convolution, they have lower semantics and more noise. High-level features have stronger semantic information, but low resolution and poor perception of detail. The key to improve the segmentation model is how to combine them efficiently, take advantage of their advantages and discard their disadvantages. FPN is an effective operation to integrate feature information of different scales. One way is to construct a pyramid structure by means of multiple branches of dilated convolution with different dilation rates, like ASPP [36], RFB [37], etc. The other is to generate layers with different resolutions through multiple downsampling to form a pyramid structure. For instance, FPN first added top-down pathway in the network, which assigned the low-level feature map better semantic information. Besides, FPN aims at dividing objects into three scales (big, medium, and small) by size and perform detection, respectively. It solves the problem of low accuracy in small cases due to the insufficiency of semantic information in shallow-layer feature maps to a certain extent. Based on that point, PANet adds an additional bottom-up pathway and proposes adaptive feature pooling, different from FPN, it uses information from candidate boxes generated in all levels to participate in the prediction. Moreover, ASFF fused the feature information of each layer adaptively by weighted fusion and it was also divided into three scales for detection. It was applied into the YOLOv3 backbone and there was 5–10 mAP improvement on the COCO data set. Feature fusion of different scales, like ASFF, is exactly an effective way to improve the performance, but there is still room for improvement in the expression of fine-grained features.

C. Bounding Box Regression

In the prediction phase of target detection, many alternative anchor boxes are output, among which there are many apparently overlapping boxes around the same target, and the job of NMS is to remove the redundant boxes. Once there are other boxes in the list with its IoU greater than the threshold value, these boxes will be deleted, and the first-sorted box will be taken out from the sorted list as the list of candidate boxes. This process is repeated until it is empty. In the filtering process of boxes by NMS, IoU is used as an indicator of box overlap, and IoU cannot contribute to the filtering of the box when two boxes do not intersect or when they overlap.

Some enhancements on IoU can effectively improve the detection of NMS. For example, GIoU [38], IoU is not sensitive to the size of the object and cannot optimize the nonoverlapping part. GIoU includes the shape and orientation of the object and focuses not only on the overlapping region but also on the nonoverlapping region. However, one of the existing problems of GIoU is that when two boxes intersect, GIoU degenerates into IoU loss leading to difficulties in predicting BBox and GT BBox. DIoU [39] is more in line with the regression mechanism of the target box than GIoU, considering the distance between target and anchor, overlapping rate, and size. In addition to the improvement of IoU, there are also some optimizations of NMS, such as Soft-NMS setting attenuation function on adjacent detection boxes based on the size of the overlap, instead of discarding them completely. The Fast NMS [40] is proposed in the instance segmentation model YOLCAT, using GPU to parallel computing IoU to reduce time cost. Like Fast NMS, Matrix NMS [41] realizes parallel computation of Mask IoU, which is valuable for box-free intensive prediction instance segmentation model. Two methods require only one iteration. Cluster NMS [42] performs best in speed due to the more basic coding while it is less flexible to be applied with other methods. There are other ways to improve the efficiency of an NMS, such as dynamically regulating thresholds. The combination of IoU

TABLE I
WILDFIRE SMOKE DATA SET OF THIS ARTICLE

Type	a	b	c	d	e
Image					
Quantity	430	157	428	332	153

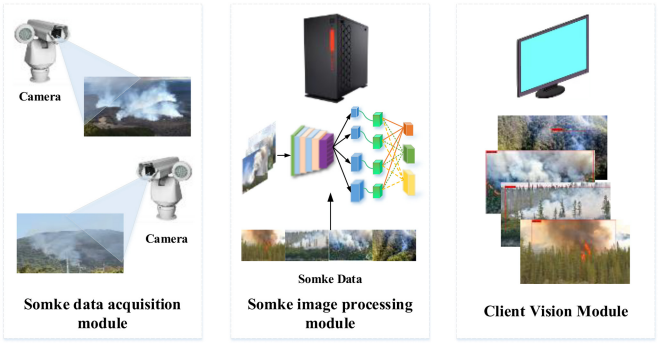


Fig. 2. Schematic of MMFNet-based IoT system for wildfire smoke detection.

and NMS accelerates the iterative regression and makes the generated BBox position more accurate. So, we combined to use CIoU [38] and Fast NMS and proposed Fast-robust NMS.

III. MATERIALS AND METHODS

A. MMFNet-Based IoT System for Wildfire Smoke Detection

This article combines MMFNet and hardware equipment to build an MMFNet-based IoT system to detect wildfire smoke, as shown in Fig. 2. The system's main components are the smoke data gathering, image processing, and computer vision parts. A Hikvision DS-2DYH277I-DU high-definition pan-head camera with 2 million pixels and a 1920 × 1080 resolution was used to capture smoke photographs. It can shoot in both horizontal and vertical directions. The shooting parameters of the camera are configured, and images are continuously collected in the surveillance area. The image data is subsequently sent across the network to a server for image processing and object detection.

B. Data Acquisition

The original data set of this article is composed of three parts: 1) obtained from the public data set [43]; 2) *Web Crawler:* we utilize the crawler program to automatically download the images in the search results from Bing Pictures, Google Pictures, Baidu Pictures, and so on according to keywords (smoke, forest fire, and wildfire); and 3) *Forest Farm Shooting:* we artificially ignite wildfire burning in Forest Farm of You Country, Zhuzhou and use HikVision DS-2DYH277I-DU to shoot smoke images. The collected images were then manually filtered as irrelevant images, duplicate images, too blurry images, and low-resolution images. Finally, the wildfire smoke data set of this article with a total of 1500 smoke pictures is obtained. The data set was made public (*https://pan.baidu.com/s/1UrL6hdq5mmklFzXExVgbDQ?pwd=igwa*). In this article, we have two categories of target detection: 1) smoke and 2) nonsmoke. As shown in Table I, we classify the types of smoke into five categories according to the characteristics of the captured images: 1) smoke in common color and of moderate size in the image; 2) smoke in a relatively fuzzy image; 3) smoke of low concentration and small size in the image; 4) smoke of high concentration and even in darker color; and 5) smoke in a relatively darker environment and of small size. It is difficult for common target detection methods to obtain a good detection effect when the shooting distance is long, the smoke is small, and low concentration. MMFNet can effectively detect the smoke of small targets at a long shooting distance and obtain a good effect when the smoke is light.

In the experiment, 1445 available images were selected from 1500 images, some blurred or repeated images were screened out, and the annotation tool Labeling was used to annotate the images. The K-fold cross-validation method was used to evaluate the model, in which $(k-1)$ folds were used for training and the rest for verification. The whole process was repeated for k times, and finally, the average value of k models was taken as the evaluation result. In this manuscript, the value of k is 5.

C. MMFNet for Fast Detection of Wildfire Smoke

To improve the detection speed and solve the difficulties caused by smoke characteristics, we propose MMFNet for fast detection of wildfire smoke, as shown in Fig. 3. We choose CSPDarkNet as the backbone network to extract features, we fuse the high-level features after backbone with the low-level features, and then use mixed attention to assign weights to the obtained feature maps so that the network can fully utilize the semantic features, after that we use four convolutions of different sizes to extract features at different scales for the feature maps. The multiscale convolution can effectively perform feature extraction on different sizes of smoke. Finally, we use DAFPN to fully utilize the semantic information of high-level features and the fine-grained features of low-level features to achieve effective feature extraction at different scales. DAFPN has 3

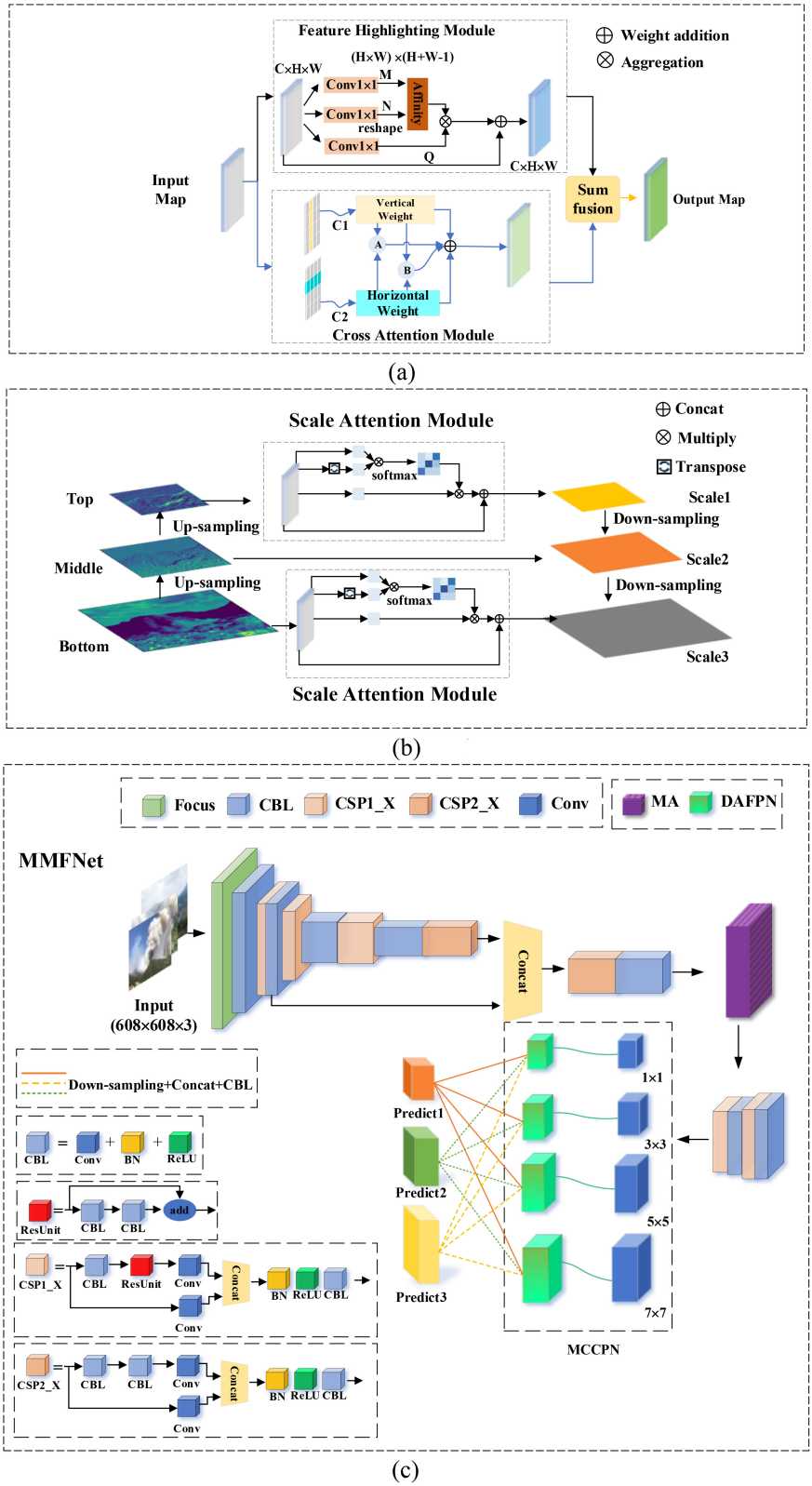


Fig. 3. Structure of MMFNet. (a) Mixed attention. (b) DAFPN. (c) MMFNet.

outputs of different sizes, and we up-sample the feature maps of the same size and then splice and fuse them to finally obtain the prediction results. Along the process, Fast-robust NMS was used to improve the precision and speed of BBox generation.

1) Feature Extraction With Mixed Attention: We proposed a mixed attention mechanism composed of the features-highlighting module and the cross-attention module to reduce the computational resource consumption of the model, obtain

intensive context information, and coordinate the overall feature extraction of forest fire smoke image. The features-highlighting module focuses on the main characteristics of the forest fire smoke image, while cross attention focuses on coordinating the extraction of local and overall features.

a) Feature highlighting module: Generally, the methods based on self-attention need to calculate the correlation degree between pixels, which is accompanied by the time complexity and space complexity as high as $O((H \times W) \times (H \times W))$. In that, $H \times W$ is the spatial dimension of the input feature map, H represents the height of the feature map, and W is for the weight. They are computationally intensive and occupy a large amount of GPU memory of deep learning server. The overall context information of all pixels can be effectively gathered by the feature highlighting module, obviously reducing the value of time and space complexity to $O((H \times W) \times (H + W - 1))$.

In this feature highlighting module, I represents the input local feature and $I \in R^{C \times W \times H}$. Two convolutions of kernel size 1×1 are used for I to generate two feature maps M and N. $\{M, N\} \in R^{C' \times W \times H}$, where C' is the channel number of the feature map and $C' < C$. In the space dimension of feature map M, a vector $M_p \in R^{C'}$ can be obtained for each position p, and M_p is the vector composed of positions p on each channel. Extract the feature vector from the feature map N to obtain the set Ω_p, and p in the set Ω_p corresponds to the position p in the same row and column as M_p. After that, attention map $A = R^{(H+W-1) \times (H \times W)}$ was further generated through correlation operation. The association operation is defined as follows:

$$d_{i,p} = M_p \Omega_{i,p}^T \tag{1}$$

where $d_{i,p} \in R^{(H+W-1) \times (H \times W)}$ represents the relevance of M_p and the various characteristics of its cross paths. Then, another convolution of kernel size 1×1 is used for I to generate a feature adaption map $Q = R^{C \times W \times H}$. We can get a cross-path vector at each position p of feature map Q's spatial dimension. To get the set $\phi_p = R^{(H+W-1) \times C}$. Remote context information is obtained through aggregation operations

$$O_p = \sum_{i \in |\phi_p|} A_{i,p} \phi_{i,p} + I_p. \tag{2}$$

b) Cross attention: Traditional attention mechanism algorithms tend to assign weights in one direction, only distinguishing differences between feature sequences from left to right, ignoring the upper and lower feature sequences, which may lead to the loss of feature information. As shown in the figure, vertical weight coefficient and horizontal weight coefficient were obtained through the input. Let $c1$ represent the value of vertical weight and $c2$ represent the value of horizontal weight. The two types of weights are multiplied to get A. Then, the weight coefficients are further expanded. The two weight features were matched to obtain the maximum value B. The four weights $c1$, $c2$, A, and B were finally concatenated. The whole process can be explained as the following formulas:

$$c_i = \sum_{j=1}^{n} \frac{\exp(e_{i,j})}{\sum_{k=1}^{n} \exp(e_{ik})} \tag{3}$$

$$A = \text{mul}(c1, c2) \tag{4}$$

$$B = \max(c1, c2) \tag{5}$$

$$CA = \text{concatenate}([c1, c2, A, B]). \tag{6}$$

The effectiveness of the mixed attention module is verified in Section IV.

2) Feature Fusion Based on Multiscale Convergence Coordinated Pyramid Network: To obtain dense context information of forest fire smoke image and improve the detection rate of different size targets, we proposed MCCPN composed of DAFPN and CCM. The DAFPN is used to coordinate pay more attention on small and big targets. CCM is used for fusing the enhanced features to increase the feature information that each scale object contains.

a) DAFPN: In applications, forest fire smoke is not stable due to the heat and weather condition. Sometimes, when it is far or the smoke is just coming into being, it will be hard to detect for the target is small. There are other common cases that the smoke target was detected of a huge size, under which circumstance the precision of the prediction box will be low. Smoke is a fuzzy target, and its shape and size may change in the image. In YOLO, small targets may be partitioned into a smaller grid, while large targets may be partitioned into a larger grid. Since the number of bounding boxes predicted for each grid is limited, small targets may be incorrectly combined into one bounding box, while large targets may be divided into multiple bounding boxes, which may result in poor detection accuracy. To solve these problems, we take the idea of FPN to generate prediction heads of three scales for object detection. FPN is a one-way fusion method that can only integrate smoke features from top to bottom and does not fully utilize them. Thus, we proposed DAFPN with a scale attention module to emphasize the feature expression, especially to the top and bottom features.

The channel map of each high-level feature can be considered as a specific class of responses with different semantic responses interrelated. By exploiting the interdependencies between the channel maps, we can emphasize the feature maps of different scales that are interdependent on the channels and improve the feature representation of specific semantics, thus improving the feature representation across scales. Therefore, we construct a scale attention module to explicitly model the interdependencies between channels at different scales.

As shown in the figure, we first reshape the input map I to $R^{C \times K}$. Then, a matrix multiplication is performed to I and I^T. Finally, we use a softmax layer to obtain the attention map $M \in R^{C \times C}$

$$m_{ji} = \frac{\exp(I_i \cdot I_j)}{\sum_{i=1}^{C} \exp(I_i \cdot I_j)} \tag{7}$$

where m_{ji} measures the influence of the ith channel on the jth channel. In addition, we reshape the result of the multiplication between X^T and I to $R^{C \times H \times W}$. Then, we concat the result with I to obtain the final output $O \in R^{C \times H \times W}$

$$O_j = \sum_{i=1}^{C} (m_{ji} A_i) + A_j. \tag{8}$$

b) MCCPN: The weakness of FPNs is that the three scales prediction heads share no information among each other. They are not capable of dealing with the feature inconsistency of different scales. This makes the feature information not well shared between feature maps at different levels, which affects the prediction accuracy of targets of different sizes. Inspired by Inception [44] and ASFF, we use four convolutions of different sizes to separately perceive features at different scale levels. Different from ASFF's direct weighted fusion method of feature maps at different scales, we first divide the feature maps into four scales by using four convolutions with different kernel sizes. The 1×1 convolution is used to reduce the number of channels and aggregate information, along with the 3×3 convolution, which focus on the features of tiny and normal targets. Considering that some smoke features will exceed the receptive field of 3×3 convolution, we use 5×5 and 7×7 convolution to extract features of larger targets to strengthen context connection. In this process, we control the step size to make the output feature map consistent in size. After that, the four feature maps are feed into DAFPN to extract information at different levels. And then we concatenate the feature maps at corresponding scales to finally obtain the three scales YOLO heads. The effectiveness of the MCCPN module is verified in Section IV.

3) Bounding Box Regression With Fast-Robust NMS: In the prediction phase of target detection, many candidate anchor boxes are output, many of which are obviously overlapping prediction bounding boxes all around the same target, and it is time to use NMS to merge similar bounding boxes of the same target, or to keep the best one of these bounding boxes. The main problems about NMS are as follows.

1) The threshold value needs to be set manually. The setting of the threshold value will directly affect the detection of overlapping targets. Sometimes it is too large and leads to misdetection, some are too small to achieve an ideal situation. Liu et al. [45] aimed at the special application scenario of pedestrian detection in crowds, an adaptive NMS is proposed, which enlarges the NMS threshold in dense crowds and smaller in sparse crowds.
2) It can only run on CPU, whose performance becomes an important factor affecting the speed.
3) It is evaluated by IoU. The IoU approach has different effects on the scale and distance of the target frame. In the classic NMS algorithm, IoU is the only factor considered, but when two target frames are overlapped or not, IoU cannot contribute to the screening of target frames. Zheng et al. [39] applied DIoU to NMS and is more effective in suppressing redundant target frames.
4) It is too straightforward to delete the target frames those who are larger than the threshold value directly. Bodla et al. [46] proposed Soft-NMS, which sets an attenuation function for adjacent detection frames based on the size of the overlapping part instead of completely setting their confidence to zero.

We combined to use CIoU and Fast NMS to form a Fast-robust NMS which is of high speed and stability. To improve the prediction precision of forest-fire smoke, CIoU was used to consider more influence factors to make the target box regression more accurate and stable. And to accelerate the iterative regression, Fast NMS was used to suppress redundant BBoxes.

a) CIoU: As we know, IoU is the most common use index in object detection. It can intuitively tell whether the sample is positive or negative one and evaluate the distance between the predicted BBox and the ground truth

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

where A represents the area of predicted BBox, and B represents the area of ground-truth BBox. Besides the intuitiveness, IoU is scale invariant, which means that the performance of IoU is not associated with the object size. However, when the two boxes do not intersect, the value of IoU turns out to be 0. It means that IoU cannot reflect the distance between them anymore and IoU regression loss will not converge. Also, different ways of overlapping may have the same value of IoU, but the regression effects may vary greatly. DIoU includes the distance between the target and the anchor, the overlap rate, and the scale, which makes the BBox regression more stable and does not have problems like scattering during training as IoU and GIoU

$$\mathrm{DIoU} = \mathrm{IoU} - \frac{\rho^2\left(b, b^{gt}\right)}{c^2} \tag{10}$$

where b and b^{gt}, respectively, represent the center point of predicted BBox and ground-truth BBox. (b, b^{gt}) represents the calculation of the Euclidean distance between the two center points. c represents the diagonal distance of the minimum closure interval that can contain both two boxes. The formula of DIoU Loss is as follows:

$$L_{\mathrm{DIoU}} = 1 - \mathrm{DIoU}. \tag{11}$$

Compared to IoU Loss, DIoU Loss can still provide the direction of movement for the bounding box and speed up the convergence of the network. But it depends on the weight ρ. When ρ is 0, DIoU degenerates to IoU. Considering that the aspect ratio has not been considered in the calculation, CIoU added the penalty term based on DIoU

$$R_{\mathrm{CIoU}} = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \partial\upsilon \tag{12}$$

where ∂ is the weight parameter, and υ is used to measure the similarity of aspect ratio

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{13}$$

where h and w are the height and weight of BBox. The formula of CIoU loss is defined as

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + R_{\mathrm{CIoU}}. \tag{14}$$

With complete factors considered in, CIoU better leads the regression of the BBox to more accurate stage.

TABLE II
HARDWARE AND SOFTWARE ENVIRONMENT

Hardware Environment	CPU	Intel Core I9-10980XE
	RAM	64G
	GPU	NVIDIA GeForce RTX 2080 Ti
	Video memory	16G
	OS	Windows 10
Software Environment	CUDA Toolkit 10.0; CUDNN V7.5.0; Python 3.6; Pytorch-GPU 1.6	

b) Fast NMS: Given that n BBoxes were obtained in one smoke image, due to sequential processing, a certain box computes CIoU with other boxes for at least once and at most $n-1$ times. In addition to sequential iterative suppression, the NMS algorithm needs to calculate CIoU at least n–1 times, and the maximum number of times is

$$\text{times} = \frac{1}{2}n^2 - \frac{n}{2}. \tag{15}$$

The time complexity of the above calculation is $O(n^2)$. In order to accelerate NMS, parallel computation is performed on CIoU. Arrange BBox set $B = \{B_i\}_{i=1,2,\ldots,n}$ in descending order according to score, B_1 is the highest and B_n is the lowest score, the CIoU matrix is as follows:

$$X = \text{CIoU}(B, B) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \tag{16}$$

$$x_{ij} = \text{CIoU}\big(B_i, B_j\big). \tag{17}$$

Due to the symmetry of the CIoU matrix, $\text{CIoU}(B_i, B_j) = \text{CIoU}(B_j, B_i)$, it can be seen that X is a symmetric matrix. Besides, it is meaningless for a BBox to calculate CIoU with itself. So, Fast NMS first performs the upper triangulation to X and obtains a CIoU matrix X with both diagonal and lower triangulation elements being 0

$$\begin{pmatrix} 0 & x_{12} & x_{13} & \cdots & x_{1n} \\ 0 & 0 & x_{23} & \cdots & x_{2n} \\ 0 & 0 & 0 & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{18}$$

Then, maximization by column is performed to X, and 1-D tensor $b = [b_1, b_2, \ldots, b_n]$ is obtained, where b_i represents the maximum value of the ith column. Finally, the tensor b is binarized using the threshold of NMS.

The effectiveness of Fast-robust NMS is verified in Section IV.

IV. EXPERIMENTAL ANALYSES

A. Experimental Environment

To examine and verify the performance of MMFNet this article proposed, necessary experiments were carried out. The hardware and software environment are all the same and specific environmental information is shown in Table II.

The data set was first shuffled, and then we used tenfold cross-validation to train the model. Besides, we resized all the input images to 608 × 608. To accelerate the convergence speed and improve the stability of the training process, an Adam optimizer was used and a cosine annealing algorithm was adapted to adjust the learning rate. For the first 300 of 1000 epochs, the learning rate was set [1e-6,1e-3]. Meanwhile, Mosaic data augmentation was used to expand background complexity and increase the number of small targets.

B. Evaluation Index

In this article, the performance of the proposed model is evaluated by metrics AP, average recall (AR), frames per second (FPS), parameter size, and GFLOPs.

There are two commonly used indexes precision (P) and recall (R). P denotes the proportion of correct classification, and R denotes the proportion of relevant information detected to the total. They can be defined as

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{20}$$

We use average accuracy (AP) rather than mean average accuracy (mAP) because there is only one target object, smoke. The metric AP can be calculated as

$$\text{AP} = \sum_{i=1}^{n} P(i)\Delta r(i) \tag{21}$$

where $P(i)$ represents the precision at the threshold i, and $\Delta r(i)$ represents the change of recall between i and $i+1$.

AR is an index that indicates the missed detection of the detector. It can be calculated as

$$\text{AR} = \frac{\text{Recall}}{n} \tag{22}$$

where n is the number of detected object frames.

FPS is a crucial indicator to measure the detection speed. FPS can be calculated as

$$\text{FPS} = \frac{1}{t} \tag{23}$$

where t is the time cost required to process each picture frame.

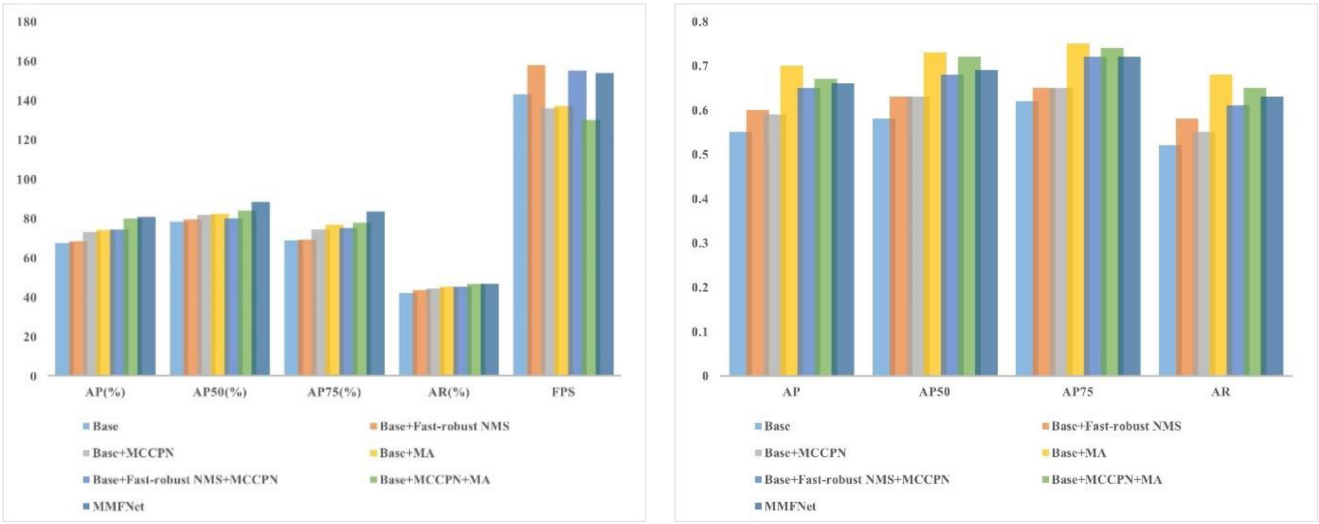
Parameter size and GFLOPs are indicators for measuring the complexity of the network model.

C. Module Effectiveness Experiments

To verify the effectiveness of our proposed approaches, a series of experiments were carried out.

First, we compared the parameter size and model complexity of our proposed model with some other YOLO series models. The result is shown in Table III.

From Table III, we can obviously see that, compared to the YOLO series, the parameter size of our model is bigger than any others, and the value of GFLOPs is even smaller. This indicates that although the addition of Mixed Attention and MCCPN increased the complexity of the network, Robust NMS makes the model significantly faster when selecting BBox.



(a) (b)

Fig. 4. (a) Results of module effectiveness ablation experiments. (b) Standard deviation of each index.

TABLE III
PARAMETER SIZE

Methods	Backbone	Resolution of Input	GFLOPs	Params
YOLO V3	DarkNet-19	416×416	60	25.6M
YOLO V4	CSPDarkNet53	608×608	52	28.4M
YOLO5s	CSPDarkNet53	608×608	49	21.4M
YOLOX	CSPDarkNet53	608×608	70.8	25.3M
MMFNet	CSPDarkNet53	608×608	40	29.3M

Then, we verified the performance of our proposed method by ablation experiments and tests. The results are shown in Fig. 4.

As a result, MMFNet reaches a high index of AP (80.72%), AP^{50} (88.52%), AP^{75} (83.45%), AR (46.88%), and FPS (154). From a separate view, with only the addition of the robust NMS module, there is a tremendous improvement in FPS. This is because, compared with the traditional IOU and DIOU, CIOU can play an effect on the convergence of the model even when the B-boxes do not intersect or coincide. At the same time, compared with the traditional NMS, Fast-NMS speeds up the selective process of multiple B-boxes in the model. When the MCCPN module is added only, the feature extraction ability of the network is effectively enhanced, and the detection ability of small targets is strengthened, and the indicators increased significantly. However, the addition of the feature extraction module also increased the complexity of the network more and the FPS is reduced. When the MCCPN and mixed attention module were added, the FPS is lower than that of the base, it indicates that the complexity of the model is increased, but other indexes were higher which means that these two modules indeed improved the feature expression ability of the network. Finally, the FPS of MMFNet is lower than that of only joining Fast-robust NMS, other indexes are significantly higher.

D. Performance Test

We tested the detection performance of MMFNet and compared it with that of YOLO5s. The results are shown in Fig. 5.

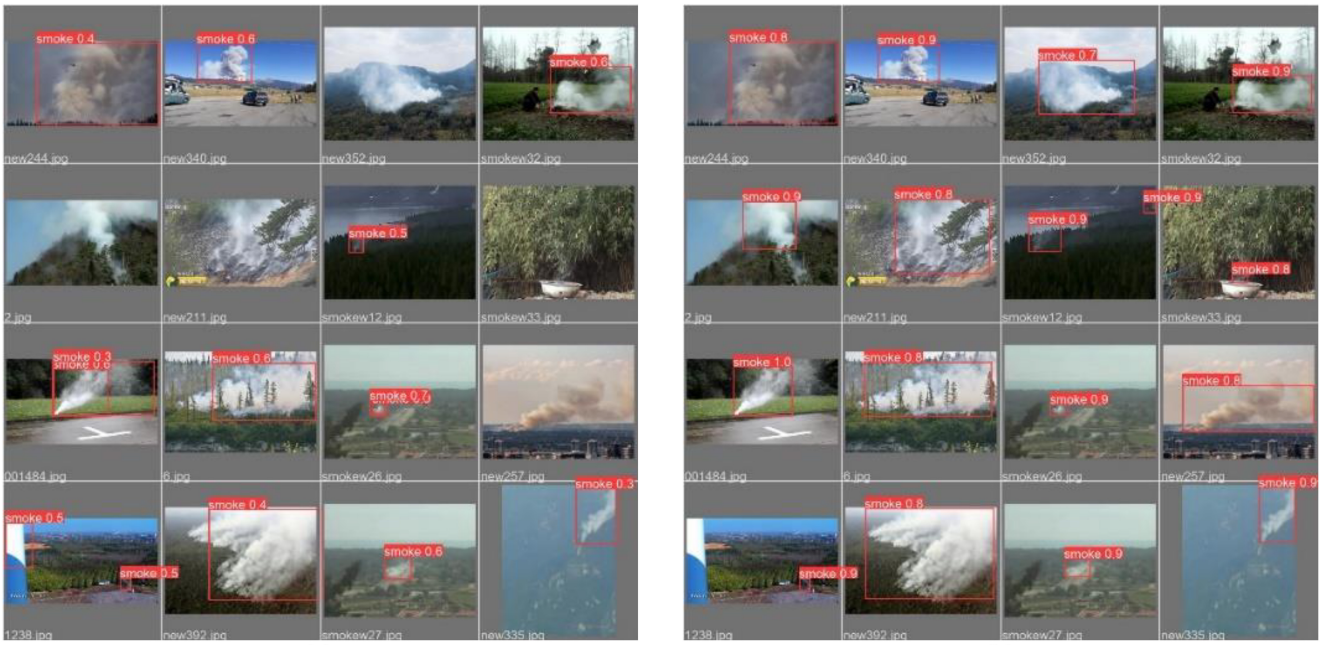
As the training process we did, we reshaped the test images into 608 × 608 and take every 16 images as one batch. Fig. 5(a) is the result of YOLO5s, and Fig. 5(b) is the result of MMFNet. It can be clearly seen from Fig. 5(a) that the network has poor feature extraction ability. For the smoke in low concentration like Somkew33.jpg, it is difficult for the network to identify it and the target cannot be labeled in the image. In the bounding box selection of the network, we can see that the prediction boxes of 001484.jpg are duplicated because the traditional IOU cannot promote the convergence of the network when the B-boxes are overlapped or noninter-sected. Take 392.jpg as a reference, we can clearly see that our model is also more accurate in predicting the position of the box, which can almost cover the whole smoke target. In general, MMFNet has increased significantly in the feature extraction ability and detection rate of remote small targets. We also recorded the standard deviation of each index. As can be seen from Fig. 5(b), the standard deviation is mostly within the range of 0.5–0.75, and the experimental results after the fivefold cross-validation show no obvious fluctuation of each index.

We also compared some curves that related to the stability of the training process, the results are shown in Figs. 6 and 7.

The selected data were the indexes of the first 100 epochs. And it is obvious that the curves of our method are relatively smooth, which indicates that the training process of MMFNet is more stable.

E. Comparison Experiments of Different Detection Models

To further validate the overall performance of MMFNet, we compared it with some classical and advanced object detection methods. We conducted experiments on a homemade data set



(a) (b)

Fig. 5. Test results comparison. (a) Result of YOLO5s. (b) Result of MMFNet.

TABLE IV
DETECTION PERFORMANCE COMPARISON OF MMFNET AND STATE-OF-THE-ART NETWORKS

Method	Backbone	AP (%)	AP^{50} (%)	AP^{75} (%)	AR (%)	FPS
YOLOv3	DarkNet19	64.41(0.65)	75.33(0.69)	67.36(0.71)	40.25(0.63)	29
YOLOv4	CSPDarknet53	73.36(0.59)	82.26(0.62)	76.02(0.64)	45.49(0.58)	34
YOLOv5	CSPDarknet53	77.56(0.66)	83.01(0.69)	80.88(0.70)	42.22(0.64)	68
YOLOX	CSPDarknet53	69.32(0.55)	82.71(0.58)	73.36(0.62)	42.38(0.54)	30
YOLOv7[47]	CSPDarknet53	80.12(0.70)	83.30(0.72)	79.36(0.73)	45.83(0.68)	149
RetinaNet[48]	ResNet-101	67.02(0.66)	78.31(0.73)	69.23(0.72)	41.25(0.65)	68
Fast R-CNN++	ResNet-101	66.86(0.62)	78.05(0.65)	69.05(0.68)	40.19(0.60)	/
MVMNet[18]	CSPDarknet53	77.68(0.57)	83.13(0.59)	77.35(0.62)	45.54(0.56)	130
ALFRNet[17]	DarkNet19	71.96(0.59)	80.33(0.62)	74.31(0.63)	43.27(0.57)	56
ARGNet[19]	ResNet50	**81.19**(0.56)	86.12(0.58)	81.15(0.62)	46.62(0.53)	123
BCMNet[20]	/	80.21(0.58)	88.42(0.62)	83.21(0.67)	46.16(0.55)	39
MMFNet	CSPDarknet53	80.72(0.66)	**88.52**(0.69)	**83.45**(0.72)	**46.88**(0.63)	**154**

with optimal parameter tuning for each method, with a training epoch of 1000, and a five-fold cross-validation. The code implementations of the methods were obtained from open-source websites. The results in Table IV are the average of 5 times results.

The data in parentheses are standard deviations to verify the generalization ability of the model. It can be seen from Table IV that MMFNet has a high accuracy and speed, which is better than other algorithms in the YOLO series and it outperforms traditional object detection algorithms (Fast R-CNN++ and RetinaNet) in terms of detection speed and accuracy. Though the index AP of ARGNet is higher than our method, but other indexes are lower. As a result, it can be concluded that MMFNet has a high accuracy and speed for real-time testing and can meet the requirement of wildfire smoke detection.

F. Testing of Real Application

At Zhuzhou Forest Farm, we ran a wildfire combustion simulation experiment. We utilized the Hikvision DS-2DYH277I-DU camera's original resolution (1920 × 1080) as the input size to guarantee the integrity of the image data. MMFNet's detection speed can reach 103 FPS which satisfies the demands of real-time detection in the case of high resolution. Additionally, we replicated smoke 50 times in three different situations and examined how many times YOLO5s

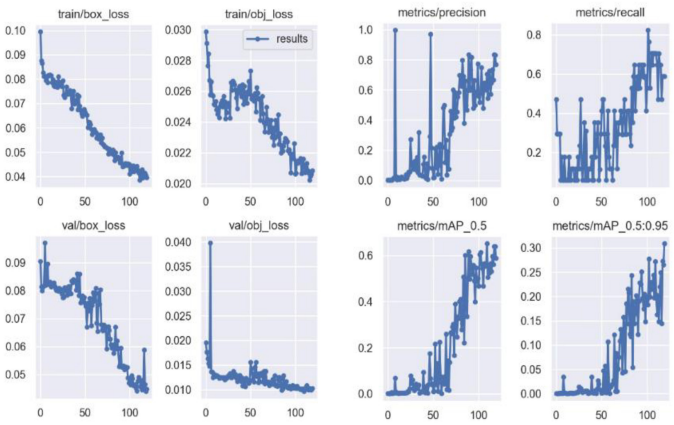


Fig. 6. Indexes curves in the training process of YOLO5s.

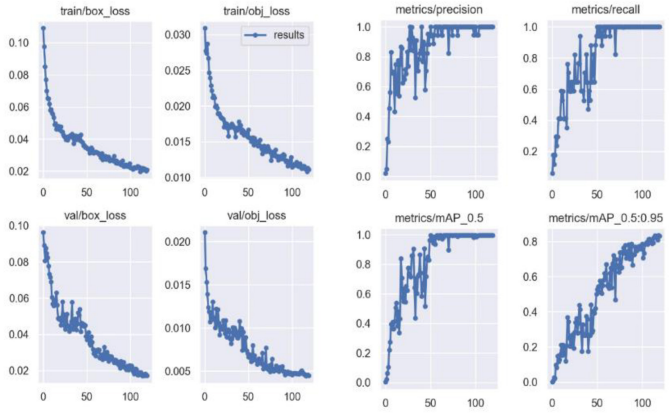


Fig. 7. Indexes curves in the training process of MMFNet.

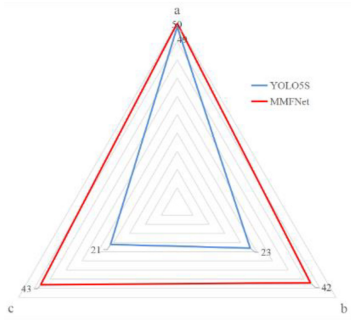


Fig. 8. Experimental comparison. In the figure, a refers to the situation that the smoke is of moderate size and concentration, while b is the case that the concentration of smoke is low, and c refers to the small size smoke detection.

and MMFNet were able to recognize smoke. As shown in Fig. 8, MMFNet has a greater identification rate for wildfire smoke and offers a more practical solution to preventing wildfires.

We present some challenging-to-identify smoke images in Fig. 9. The case depicted in Fig. 9(a) illustrates how the method suggested in this article can efficiently detect it when the smoke area is large, but the concentration is low. The case where the smoke object is small is depicted in Fig. 9(b), demonstrating how well our approach can find tiny objects.

(a)

(b)

Fig. 9. Real application. (a) Smoke in low concentration. (b) Smoke of small size.

They effectively demonstrate the superiority of MMFNet for smoke detection task.

V. Conclusion

To improve the effectiveness and speed of forest fire smoke detection, an MMFNet for fast and accurate detection of wildfire smoke is proposed in this article. Compared with the model YOLO5s, the improvements of the proposed method are as follows.

1) On the self-built smoke data set, MMFNet achieves 75.72% AP, 83.52% AP50, 77.45% AP75, 46.88% AR, and 154 FPS in accuracy. Visualization of test results and comparison of experiments indicate that the detection accuracy of forest fire smoke is significantly increased by using MMFNet. This is because Mixed Attention enhanced the feature extraction of both local and overall features. Meanwhile, MCCPN can make full use of semantic information of high-level features and fine-grained features of low-level features of smoke, thus realize effective extraction of features at different scales, improving the detection rate of small targets. The

improvement in FPS is because the use of Robust NMS greatly increased the convergence of BBoxes.

2) An MMFNet-based forest fire smoke detection system is designed and implemented in practice. It can detect smoke in real time and accurately predict the occurrence of a fire, which is of great significance to protect ecological resources and reducing losses.

Compared with BCMNet, our model does not have an obvious improvement in accuracy, but it is superior to BCMNet in running speed. To enhance the network extraction of smoke features, BCMNet proposes a multiscale downsampling module to address the loss of semantic information in traditional downsampling, and a bidirectional transposition FPN to address the underutilization of shallow semantic information. To reduce the dependence on computational resources, a cross-layer extraction module is proposed to reduce network training. However, BCMNet is ineffective in detecting smoke at lower concentrations, while our proposed mixed attention can aggregate long-distance contextual information, realizes weight assignment for different channels, and allocates more weights on lighter parts of smoke in the image to achieve effective detection of low-concentration smoke. As can be seen in Table IV, compared to YOLOv3, YOLOv4, and YOLOv5, the speed boost of BCMNet operation is smaller, while our utilization of Fast-robust NMS makes the network run faster and more applicable to actual hardware devices and complex natural environments.

Although MMFNet performs well in smoke detection, the interference factors of hardware equipment in different natural environments, such as sunlight exposure at a specific Angle and water mist on the lens, will cause misjudgment. In future work, we will deeply study the characteristic differences between smoke and these interfering objects. Also, we will avoid the formation of interference factors from the hardware perspective to further improve the accuracy while ensuring faster detection speed. At the same time, we will expand the sources of smoke images, and further study and explore the forest fire smoke detection using infrared images and hyperspectral images.

REFERENCES

[1] S.-J. Chen, D. C. Hovde, K. A. Peterson, and A. W. Marshall, "Fire detection using smoke and gas sensors," *Fire Safety J.*, vol. 42, no. 8, pp. 507–515, 2007.

[2] N. I. B. Zaidi, N. A. A. B. Lokman, M. R. B. Daud, H. Achmad, and K. A. Chia, "Fire recognition using RGB and YCbCr color space," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 21, pp. 9786–9790, 2015.

[3] Z. Li, L. S. Mihaylova, O. Isupova, and L. Rossi, "Autonomous flame detection in videos with a Dirichlet process Gaussian mixture color model," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1146–1154, Mar. 2018.

[4] D. Wu, C. Zhang, L. Ji, R. Ran, H. Wu, and Y. Xu, "Forest fire recognition based on feature extraction from multi-view images," *Traitement du Signal*, vol. 38, no. 3, pp. 775–783, 2021.

[5] S. Frizzi, R. Kaabi, M. Bouchouicha, J.-M. Ginoux, E. Moreau, and F. Fnaiech, "Convolutional neural network for video fire and smoke detection," in *Proc. IECON 42nd Annu. Conf. IEEE Ind. Electron. Soc.*, 2016, pp. 877–882.

[6] S. Wu and L. Zhang, "Using popular object detection methods for real time forest fire detection," in *Proc. 11th Int. Symp. Comput. Intell. Des. (ISCID)*, vol. 1, 2018, pp. 280–284.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[9] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Sci.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[10] P. Li and W. Zhao, "Image fire detection algorithms based on convolutional neural networks," *Case Stud. Thermal Eng.*, vol. 19, Jun. 2020, Art. no. 100625.

[11] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[13] D. P. Lestari, R. Kosasih, T. Handhika, Murni, I. Sari, and A. Fahrurozi, "Fire hotspots detection system on CCTV videos using you only look once (YOLO) method and tiny YOLO model for high buildings evacuation," in *Proc. 2nd Int. Conf. Comput. Inform. Eng. (IC2IE)*, 2019, pp. 87–92.

[14] S. Saponara, A. Elhanashi, and A. Gagliardi, "Real-time video fire/smoke detection based on CNN in antifire surveillance systems," *J. Real-Time Image Process.*, vol. 18, pp. 889–900, Jun. 2021.

[15] D. Shen, X. Chen, M. Nguyen, and W. Q. Yan, "Flame detection using deep learning," in *Proc. 4th Int. Conf. Control Autom. Robot. (ICCAR)*, 2018, pp. 416–420.

[16] F. Shi, H. Qian, W. Chen, M. Huang, and Z. Wan, "A fire monitoring and alarm system based on YOLOv3 with OHEM," in *Proc. 39th Chin. Control Conf. (CCC)*, 2020, pp. 7322–7327.

[17] J. Li et al., "Adaptive linear feature-reuse network for rapid forest fire smoke detection model," *Ecol. Inform.*, vol. 68, May 2022, Art. no. 101584.

[18] Y. Hu et al., "Fast forest fire smoke detection using MVMNet," *Knowl. Based Syst.*, vol. 241, Apr. 2022, Art. no. 108219.

[19] J. Zhan, Y. Hu, G. Zhou, Y. Wang, W. Cai, and L. Li, "A high-precision forest fire smoke detection approach based on ARGNet," *Comput. Electron. Agr.*, vol. 196, May 2022, Art. no. 106874.

[20] J. Li, G. Zhou, A. Chen, C. Lu, and L. Li, "BCMNet: Cross-layer extraction structure and multiscale downsampling network with bidirectional transpose FPN for fast detection of wildfire smoke," *IEEE Syst. J.*, vol. 17, no. 1, pp. 1235–1246, Mar. 2023.

[21] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, *arXiv:1902.07296*.

[22] C. Chen et al., "RRNet: A hybrid detector for object detection in drone-captured images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 100–108.

[23] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for Tiny person detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1246–1254.

[24] Y. Chen et al., "Stitcher: Feedback-driven data provider for object detection," 2020, *arXiv:2004.12432*.

[25] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 566–583.

[26] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection—SNIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3578–3587.

[27] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2874–2883.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

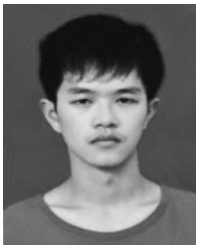
[29] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[30] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[31] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.

[32] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, p. 1432, 2020.

[33] Y. Cao, F. Yang, Q. Tang, and X. Lu, "An attention enhanced bidirectional LSTM for early forest fire smoke recognition," *IEEE Access*, vol. 7, pp. 154732–154742, 2019.
[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[37] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 404–419.
[38] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[39] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, pp. 12993–13000, 2020.
[40] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9156–9165.
[41] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17721–17732.
[42] Z. Zheng et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022.
[43] Q.-X. Zhang, G.-H. Lin, Y.-M. Zhang, G. Xu, and J.-J. Wang, "Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images," *Procedia Eng.*, vol. 211, no. 3, pp. 441–446, 2018.
[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4278–4284.
[45] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6452–6461.
[46] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5562–5570.
[47] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.



Liangji Zhang received the B.Sc. degree in computer science and technology from Hunan Institute of Engineering, Xiangtan, China, in 2016. He is currently pursuing the M.Sc. degree in electronic information engineering with the Central South University of Forestry and Technology, Changsha, China.

His main research interests include deep learning and graphics and image processing.

Chao Lu received the B.Sc. degree in software engineering from Anhui Normal University, Wuhu, China, in 2019. He is currently pursuing the M.Sc. degree in software engineering with the Central South University of Forestry and Technology, Changsha, China.

His main research interests include deep learning and graphics and image processing.

Haiwen Xu received the M.Sc. degree in computer science and technology from Hunan University, Changsha, China, in 2016.

He is currently working with the Department of Forest and Grassland Fire Prevention Monitoring and Dispatching Evaluation Center of Hunan Province, Changsha. He is specialized in information management and research of government departments and is proficient in intelligent forest fire prevention theory and technology.

Aibin Chen received the Ph.D. degree in computer application technology from Central South University, Changsha, China, in 2010.

He is currently a Professor with the Central South University of Forestry and Technology, Changsha. His main research interests include artificial intelligence and forest information engineering.

Liujun Li (Member, IEEE) received the B.Sc. degree in automation from Hunan Agricultural University, Changsha, China, in 2002, and the M.Sc. degree in materials engineering and the Ph.D. degree in mechanical engineering from Central South University, Changsha, in 2005 and 2012, respectively.

Guoxiong Zhou received the Ph.D. degree in control science and engineering from Central South University, Changsha, China, in 2010.

He is currently an Associate Professor with the Central South University of Forestry and Technology, Changsha. His research interests include forest fire prevention and robotics.