Knowledge-Based Systems 241 (2022) 108219

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Fast forest fire smoke detection using MVMNet

Yaowen Hu [a,1], Jialei Zhan [a,1], Guoxiong Zhou [a,*], Aibin Chen [a], Weiwei Cai [a], Kun Guo [a], Yahui Hu [b], Liujun Li [c]

[a] College of Computer & Information Engineering, Central South University of Forestry and Technology, Changsha 410004, China
[b] Plant Protection Research Institute, Henan Academy of Agricultural Sciences, Changsha 410125, China
[c] Department of Civil, Architectural and Environmental Engineering, University of Missouri-Rolla, Rolla, MO 65401, USA

ARTICLE INFO

Article history:
Received 14 September 2021
Received in revised form 11 January 2022
Accepted 11 January 2022
Available online 21 January 2022

Keywords:
Forest fire smoke detection
Multioriented
Value conversion-attention mechanism module
Softpool-spatial pyramid pooling
Mixed-NMS

ABSTRACT

Forest fires are a huge ecological hazard, and smoke is an early characteristic of forest fires. Smoke is present only in a tiny region in images that are captured in the early stages of smoke occurrence or when the smoke is far from the camera. Furthermore, smoke dispersal is uneven, and the background environment is complicated and changing, thereby leading to inconspicuous pixel-based features that complicate smoke detection. In this paper, we propose a detection method called multioriented detection based on a value conversion-attention mechanism module and Mixed-NMS (MVMNet). First, a multioriented detection method is proposed. In contrast to traditional detection techniques, this method includes an angle parameter in the data loading process and calculates the target's rotation angle using the classification prediction method, which has reference significance for determining the direction of the fire source. Then, to address the issue of inconsistent image input size while preserving more feature information, Softpool-spatial pyramid pooling (Soft-SPP) is proposed. Next, we construct a value conversion-attention mechanism module (VAM) based on the joint weighting strategy in the horizontal and vertical directions, which can specifically extract the colour and texture of the smoke. Ultimately, the DIoU-NMS and Skew-NMS hybrid nonmaximum suppression methods are employed to address the issues of smoke false detection and missed detection. Experiments are conducted using the homemade forest fire multioriented detection dataset, and the results demonstrate that compared to the traditional detection method, our model's mAP reaches 78.92%, mAP^{50} reaches 88.05%, and FPS reaches 122.

1. Introduction

Forests are an important component of the Earth's ecosystem, and their protection is everyone's responsibility [1]. A forest fire is an uncontrolled and unrestricted natural disaster that threatens forested land [2]. In recent years, forest fires have become more frequent, and their scope of damage has increased every year [3]. Forest fires destroy millions of hectares of forest every year and cause a series of environmental disasters, which mainly include global warming, and governments have spent tens of billions of dollars fighting these fires [4]. Forest fires not only damage the ecosystem but also pose a potential danger to human survival and development [5]. Forest fires are unpredictable and usually occur in isolated forest areas [6]. If a forest fire is not detected in time, it can easily spread, thereby causing greater damage to the ecosystem and posing a greater challenge to firefighting efforts [7]. Smouldering is the most common cause of forest fires. Smoke is the first sign of a forest fire. It is critical to shorten the response time of forest firefighting and the severity of forest fire damage by effectively detecting smoke in time and pinpointing the region where smoke occurs.

Several issues need to be addressed immediately in the current forest fire smoke detection scene (as shown in Fig. 1): (1) Inclined smoke. When a forest fire occurs, the smoke tends to tilt due to the impacts of the surrounding environment and weather, which reveals useful information such as the wind direction and fire source location. Previously, the horizontal detection box that was employed for target detection overlooked useful information and lacked consistency. (2) Misdetection of objects that are similar to smoke. In the complicated and changeable forest environment, there are many smoke-like phenomena, such as moving clouds in the sky and fog in the atmosphere. These phenomena have similar features to smoke, and traditional feature extraction networks struggle to detect the differences. (3) The smoke is too far away. When the burning point is far from the camera, the detection

* Corresponding author.
E-mail addresses: huyaowen_12345@163.com (Y. Hu), 522987225@qq.com (J. Zhan), zhougx01@163.com (G. Zhou), 5708111@qq.com (A. Chen), vivitsai@csuft.edu.cn (W. Cai), 317557750@qq.com (K. Guo), llphuyahui@hnppi.com (Y. Hu), llpwc@umsystem.edu (L. Li).

[1] Yaowen Hu and Jialei Zhan contribute equally to this work.

https://doi.org/10.1016/j.knosys.2022.108219

(a) Inclined smoke (b) Misdetected Objects similar to smoke (c) The smoke is too far

Fig. 1. Smoke detection interference factors.

box's confidence is low, and it is filtered out, thereby increasing the difficulty of smoke detection.

To handle the problem of smoke tilt, L Tian et al. proposed a detection framework that consists of an image enhancement module and a dense feature reuse module for addressing the densely arranged features of objects in remote sensing scenarios [8]. In the remote sensing setting, W Huang et al. suggested a cross-scale feature fusion pyramid network and employed a multioriented detection box to adapt to targets such as slanted ships [9]. The remote sensing scene's multioriented detection method has inspired us, and the multioriented detection box is well fitted to the peculiarities of smoke flying in the wind. As a result, we propose a multioriented detection method in which the target box can adaptively describe the direction of the smoke and has reference relevance for determining the direction of the fire source. We use the PolyIOU method to calculate the degree of overlap between anchor boxes to adapt to multioriented detection and create the forest fire multioriented detection dataset.

L He et al. proposed an attention mechanism module that integrates spatial attention and channel attention to address the problem of false detection of smoke-like items in foggy weather [10]. D Sheng et al. proposed pixel oversegmentation to assist the convolutional neural network in extracting smoke features and lower the likelihood of false detection [11]. Although these methods are simple and effective, they have difficulty overcoming the problem of information loss during the feature extraction process, and they still result in false detections against complicated backgrounds, such as clouds in the sky being misidentified as smoke. As a consequence, we propose a Soft-SPP, which can alleviate the problem of feature information loss that is caused by the inconsistent input sizes of smoke images while preserving more feature information by substituting MaxPool with SoftPool operations. Following that, a value conversion-attention mechanism (VAM) is proposed, which employs the colour and texture feature information of the smoke image while enhancing the weight distribution of texture information based on the joint weight assignment strategy in the horizontal and vertical directions.

E Zhao et al. introduced an adaptive detection box method for limiting the occurrence of missed detection in forest fire smoke detection tasks, which substantially increased the detection speed of Faster-RCNN and reduced the occurrence of missed detection [12]. This method, however, has limitations and is only appropriate for two-stage target detection models. Based on Skew-NMS and DIoU-NMS, we redesign the Mixed-NMS method in this paper to overcome the problem of missed detection that is caused by similar smoke at long distances. Skew-NMS considers the angle and threshold information in smoke target multioriented detection to determine the anchor box information, whereas DIoU-NMS considers the overlap area and the centre point distance. When the burning point is far from the camera or multiple clouds of smoke are close to each other, the combination of these two NMS methods can better fit the actual form of the smoke and solve the problem of missed detection.

The contributions of this paper are summarized as follows:

(1) A multioriented detection method is proposed for describing the directional state of smoke. In addition, to accommodate multioriented detection, we create the forest fire multioriented detection dataset.

(2) A Soft-SPP is proposed. This module substitutes the SPP's three parallel MaxPool operations with SoftPool operations to preserve more distinctive information.

(3) A value conversion-attention mechanism is proposed. To efficiently extract the smoke texture features, this attention mechanism uses the feature information of the colour and texture of the smoke image while enhancing the weight distribution of the texture information.

(4) Mixed-NMS is proposed while taking angle information and centre point distance into account in the multioriented detection of smoke targets.

2. Related work

Because combustibles in the forest are rarely dry, a substantial volume of smoke is frequently produced during the start of a forest fire. The ability to detect smoke effectively is critical for the prevention and management of forest fires. Methods for smoke detection can be classified as (1) traditional methods, which include manual and sensor detection methods, and (2) image-based methods, which include traditional machine learning and deep learning. Traditional methods have drawbacks such as high cost, sluggish response time, and a narrow application range. Traditional machine learning methods have the problems of insufficient feature extraction, a high missed-positive rate, and a high false-positive rate among image-based algorithms. Deep learning methods not only address the flaws of traditional methods and machine learning methods but also have the advantages of high precision and precise identification in complicated environments. As a corollary, using deep learning methods to extract deep-level smoke features and effectively distinguish smoke from complicated backgrounds is an important goal of forest fire smoke monitoring tasks, which is also the main objective of this paper.

The currently popular methods rely primarily on manual detection or sensor detection. The manual detection method entails forest guards conducting ground patrols in the forest and reporting to the appropriate department as soon as a fire arises to carry out firefighting and extinguishing measures. The forest fire detection method has flaws such as excessive error, a small patrol area, a high cost burden, and an insufficient coverage area, and it cannot meet the speed or accuracy criteria of forest fire detection. Thermal imaging technologies, such as infrared sensors, multispectral sensors, and hyperspectral sensors, are examples of sensor methods [13]. Because of the sensor's angle, distance, and occlusion, sampling that is based on smoke particles or humidity

is easily affected by irrelevant objects in the surrounding region, thereby resulting in disadvantages such as extremely long time delays, high cost, limited scope, and difficulty of operation [14]. Traditional smoke detection systems are ineffective in the early warning of forest fires for the reasons that are stated above.

To detect forest fires early, researchers have conducted extensive research on smoke recognition using images of smoke and extracting the colour and textural features of the smoke [15]. Initially, typical digital image processing algorithms and pattern recognition methods [16] were used to extract and evaluate smoke properties to identify smoke. D Krstini et al. investigated numerous colour space transformations by measuring the separability of smoke and nonsmoke pixels to find a suitable combination of colour space and pixel-level smoke segmentation methods for smoke pixel-level segmentation [17]. Gubbi, J. et al. proposed a method for smoke characterization based on wavelets and support vector machine that effectively minimizes the smoke false-alarm rate [18]. Surapong et al. devised a forest fire smoke detection method that processes digital images based on an examination of static and dynamic smoke features [19]. This method locates the region of interest using the Unicom component algorithm, uses the convex hull algorithm to calculate the area of the region and divide the changed area, analyses and calculates the static and dynamic characteristics, and determines whether the changed object in the image is smoke. Chen et al. analysed early smoke video footage, estimated smoke colour distribution rules, and built a decision tree for recognizing smoke based on smoke colour variations and dynamic diffusion [20]. Yu et al. processed smoke images in blocks, evaluated and computed the changes in the texture of smoke videos, categorized them, and verified that the texture features of smoke varied significantly from those of nonsmog areas [21]. Yuan et al. proposed local and global textures for image textural features to improve smoke detection, but the local and global boundaries were difficult to measure. Setting a uniform threshold based on the pixel level in this research method is difficult because it is challenging to reconcile near smoke with distant smoke and missed fires or false alarms are likely [22]. Using the principle of fractal coding, Fujiwara and Terada devised a method for extracting smoke regions from images. This method does not extract shape information or the shape itself but rather uses the smoke shape's self-similarity and the pixel position connections to extract the smoke area from a single bright image [23]. Although this method reduces the reliance on the dataset, it takes a long time to analyse the features of smoke and is ineffective for detecting smoke that is far away from the lens. Gubbi et al. proposed a smoke detection method that is based on wavelets and support vector machines and uses a three-level wavelet decomposition for different levels of features. Sixty features, such as entropy, kurtosis and skewness, were calculated as inputs to the classifier model based on coefficients [24]. Finally, a support vector machine is used for classification. This method is based on detecting high-quality smoke images that are captured by a lower fixed camera and does not take into account the impact of weather on image quality. Verstockt et al. proposed a multimodal flame and smoke detector for large open spaces that obtains visual and magnitude images based on the time of flight [25]. If an object has a high probability of being a flame feature, it is labelled as a candidate flame region. Amplitude disorder was also investigated, and regions with high cumulative amplitude differences and high values in all detailed images of the discrete wavelet transform of the amplitude image are also labelled candidate flame regions. A fire alarm goes off when at least one of the visual and one of the amplitude candidate flame regions overlaps. This solution is immature due to its low resolution and the high power consumption of the technique, which requires the receiver and transmitter to be close to the fire source. This is difficult to realize in vast forests.

In recent years, the development of deep learning techniques has enabled us to detect forest fires based on images, which has brought new ideas to forest fire early warning [26]. Deep learning methods can extract features from many labelled images and identify changing smoke images [27]. Object detection is generally divided into two-stage detection and one-stage detection. Two-stage detection is represented by RCNN [28], which has the advantage of being highly accurate; however, two-stage detection has high hardware requirements, is not easy to widely deploy, and has a low detection rate. One-stage detection is represented by YOLO [29], which has an efficient detection rate and is suitable for fast fire detection. YOLO, which was proposed by Redmon et al. is an end-to-end real-time object detection algorithm that uses convolutional neural networks to perform feature extraction and classify objects and localized images, which has the advantage of high efficiency but lacks accuracy. YOLOv2 [30], which was proposed by Redmon and Frahadi, uses K-means clustering to preprocess the size of the prior box and clusters the prior box into 9 classes with different lengths and widths to detect targets with different size scales, thereby compensating for the shortcomings of YOLO in terms of accuracy. Nevertheless, YOLOv2 does not focus on the training problem of different image sizes and does not perform well in detecting small targets. In a later study, Redmon and Frahadi proposed YOLOv3 [31], which builds on YOLOv2 and features a pyramid network [32] to improve detection performance, especially in small object detection. However, the performance of YOLOv3 on objects with complex features still needs to be improved, and YOLOv3 is still far from being applicable for practical forest fire detection. YOLOv4 [33], which was proposed by Bochkovskiy et al. optimizes various aspects of YOLOv3 in terms of data processing, the backbone network, network training, the activation function, and the loss function, among other factors, and realizes different degrees of optimization in practical applications to improve model performance and provide guidelines for practical applications to forest fire detection. YOLOv5, which was proposed by Glenn Jocher, uses mosaic data augmentation on the input side; CSPDarknet53, the Mish activation function and Dropblock on the backbone; and the SPP structure in the neck to improve upon YOLO. Compared with YOLOv4, YOLOv5 is a lighter model that realizes a higher speed and still ensures accuracy. The proposed YOLOv5 model is of light weight and high accuracy in practical applications compared to other YOLO models [34].

Deep learning methods outperform traditional methods in detecting forest fire smoke and are frequently based on simple futures such as the colour and contour of the smoke. More research in the field of forest fire smoke detection is required to investigate diverse smoke features and improve the detection performance in complicated forest environments [35]. In contrast to prior work, our study is devoted to collecting deeper smoke features for differentiating smoke from complicated backdrops (such as clouds and fog). Simultaneously, a multidirectional sensing box is employed to identify the smoke's direction.

Consequently, the method in this paper employs the standard backbone of YOLOv5 to extract fundamental features, Softpool-spatial pyramid pooling to retain more feature information, and the value conversion attention mechanism module to characterize smoke's spatial texture information. The original YOLOv5 feature pyramid is utilized for feature fusion at the feature fusion step. Mixed-NMS is proposed to replace the weighted NMS algorithm of YOLOv5 in the image postprocessing step to reduce missing and false detections of smoke. Fig. 2 illustrates the MVMNet methodology that is presented in this paper.

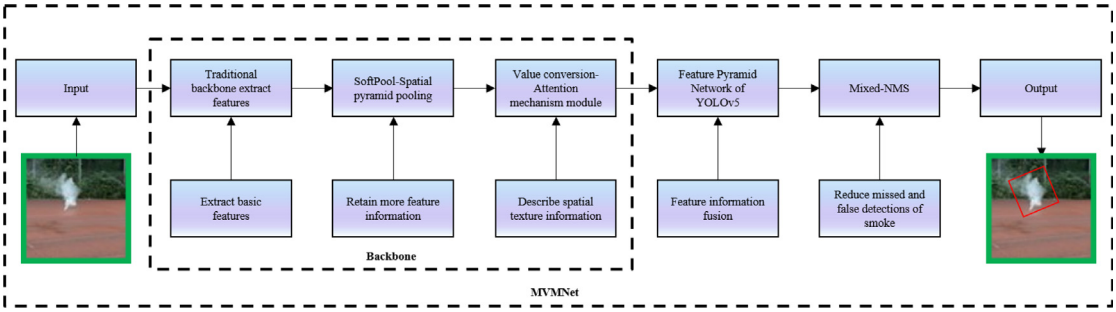


Fig. 2. Workflow of MVMNet.

3. Method

3.1. Data acquisition

No smoke detection data are currently publicly available since few smoke images have been captured, as smoke images are difficult to obtain, and the number of researchers in the field of forest fire smoke target detection is limited. As a result, prior to conducting the research that is outlined in this paper, smoke data must be collected. First, we employ Hikvision cameras to photograph 10,532 images of smoke in various burning states and climate settings in forests and farmland. The direction and shape of the smoke vary depending on the environment in which the fire is located and the magnitude of the wind. Next, to supplement the dataset, this paper replicates the forest habitat and generates 5377 images. We eventually obtain the forest fire multioriented detection dataset, which includes 15,909 images (see Fig. 3 for example images). The collections of the above real and simulated scenes without smoke, which is closer to the real situation, because the forest is not constantly burning.

The three images in Fig. 4 are all simulated data that were obtained in the field in a clearing on the outskirts of a city. As shown in the images, the clearing contained plants and trees, and the surroundings were relatively open to simulate a forest background. Smoke was produced by burning dry branches and leaves. Fig. 4 **(a)** was captured in the morning, Fig. 4 **(b)** in the middle of the day and Fig. 4 **(c)** in the evening to simulate smoke that is produced by a forest fire under different weather conditions and at different times of day.

The images that were gathered on the spot are inconsistent with the initial images that were collected in a simulated environment, and their quality is not consistent. To facilitate research, high-definition image standards are unified through screening, cropping, and standardization to depict the true forest environment. To label the multioriented data, we utilize LabelImg2 software, and the labels include position coordinates, length, height, and angle information. LabelImg2 labels the dataset only in VOC format. After the labelling is completed, the dataset in VOC format is obtained; therefore, the dataset must also be converted to the YOLO format. Fig. 5 depicts the steps for converting the dataset.

3.2. YOLOv5 backbone

YOLOv3 employs Darknet53 as its backbone and a feature pyramid network to fuse image features; hence, it is a classic target detection network. YOLOv5 adheres to the core structure of YOLOv3 and divides the complete network structure into four parts: input, backbone, neck, and prediction. In contrast to YOLOv3, YOLOv5 employs CSPDarknet53 as the backbone, along with the Focus structure and two kinds of constructed CSP structures. Prior to passing the image to the backbone, the Focus structure executes a slicing operation on it and obtains a value for every other pixel in the image, similar to adjacent downsampling, to produce four images that have not lost any information at this time. As a result, the original image's W and H information is concentrated in the channel space, and the input channel is expanded by four times, thereby resulting in a spliced image with 12 channels as opposed to the original RGB three-channel mode. Finally, the image that is obtained by the focus slice is subjected to the convolution operation, which yields a double downsampled feature map with complete information. The backbone network employs the CSP1_X structure, whereas the neck network employs the CSP2_X structure. The distinction between the two is that CSP1_X employs a residual component, and the residual connection of the residual component facilitates the reuse of extracted feature information, which renders it suitable for the backbone. CSP2_X makes use of the CBL module, which is made up of three parts: a convolutional layer, a batch processing layer, and a LeakyReLU activation function that facilitates feature map fusion. Smoke has irregular diffusion features, and its pixel-based features are not visible. It is easy to lose critical information when employing a single-channel neural network to extract features, which decreases the detection effect. YOLOv5's CSPDarknet53 solves the problem of severe gradient degradation via residual connection and has produced good results in the field of common target detection.

Actual forest fire monitoring tasks require high-performance real-time monitoring methods. Furthermore, smoke has strong transparency properties, which distinguishes it from ordinary objects, and most neural networks do not take transparency into account. Through the slicing operation, the Focus module of YOLOv5 converts the RGB three-channel mode to the twelve-channel mode, which not only boosts the speed but also helps extract the high-transparency features of the smoke. To take into account the real-time and high-transparency properties of smoke, this paper uses YOLOv5 as the basic network. YOLOv5 offers advantages in forest fire smoke detection tasks; however, the properties of smoke and typical objects are significantly different. By designing according to the features of smoke, it is possible to considerably enhance the model's detection rate. As a result, we propose the MVMNet target detection method, which improves upon YOLOv5.

3.3. Multioriented detection based on the value conversion-attention mechanism module and mixed-NMS (MVMNet)

MVMNet is upgraded on the basis of YOLOv5, and the network structure changes are illustrated in Fig. 6.

YOLOv5's feature extraction network is composed of the Focus, CBL, SPP, and CSP1_X modules. This paper builds on YOLOv5 by replacing the SPP module with the Soft-SPP module and adding the VAM module. This paper improves the IoU calculation method, the loss function calculation method, and the NMS strategy, in addition to the network structure. Additional details are provided in the following chapters.

Fig. 3. Examples of the experimental data where the images show (a) long-range smoke, (b) medium-range smoke, (c) close-range smoke and (d) no smoke.

Fig. 4. Simulated forest fire scenes that were captured in the field.

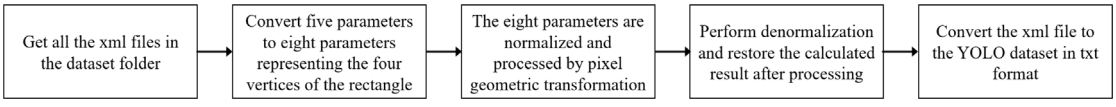


Fig. 5. Dataset format conversion steps.

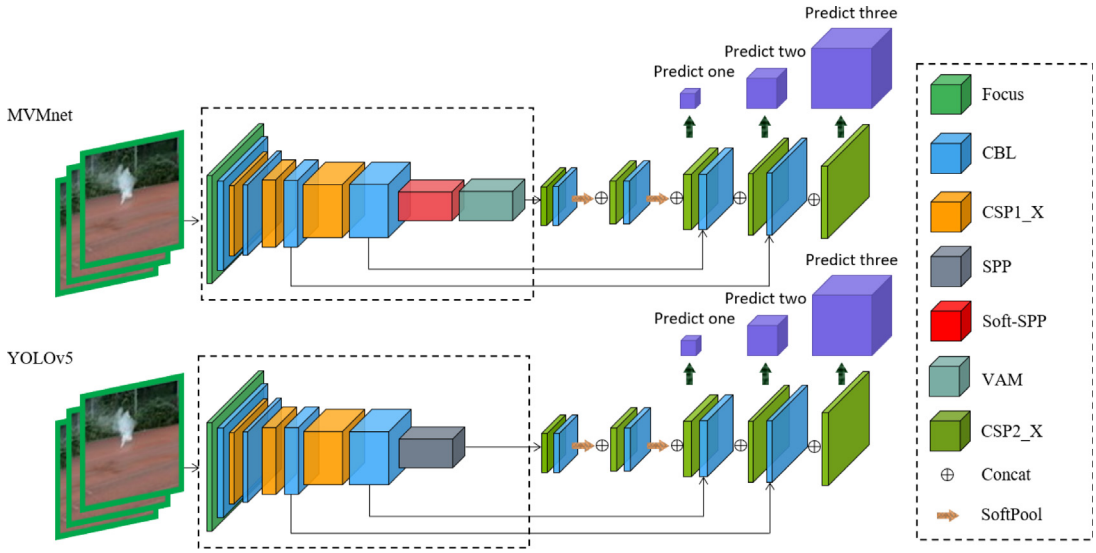


Fig. 6. Structures of MVMNet and YOLOv5.

3.3.1. Multioriented detection

a) Anchor box regression and the loss function

During the training process, each anchor box is described by a five-parameter representation. To accommodate the properties of smoke in multiorientation detection, a regression approach is used to determine the location of each multiorientation anchor box. The regression is calculated as follows [36]:

$$t_x = (x - x_a)/\varpi_a,\ t_y = (y - y_a)/h_a \tag{1}$$

$$t_w = \log(\varpi/\varpi_a),\ t_h = \log(h/h_a),\ t_\theta = \theta - \theta_a \tag{2}$$

$$t'_x = (x' - x_a)/\varpi_a,\ t'_y = (y' - y_a)/h_a \tag{3}$$

$$t'_w = \log(\varpi'/\varpi_a),\ t'_h = \log(h'/h_a),\ t'_\theta = \theta' - \theta_a \tag{4}$$

The variables x, y, w, h, θ represent the horizontal position, vertical position, width, height and angle of the detection box, respectively, respectively, and the variables x, x_a, x' represent the ground-truth box, anchor box and predicted box, respectively (likewise for y, w, h, θ).

A loss function is employed during the model training phase. For backpropagation, we combine the regression loss function, the VAM loss function, and the angle loss function. The total loss function that we establish is presented as formula (5):

$$L = \frac{\lambda_1}{N}\sum_{n=1}^{N} t'_n \sum_{i\in\{x,y,w,h,\theta\}} \frac{|-\lg(IoU)|L_{reg}(v'_{ni}, v_{ni})}{|L_{reg}(v'_{ni}, v_{ni})|} + \frac{\lambda_2}{h*w}\sum_{i}^{h}\sum_{j}^{w} L_{VAM}(u'_{ij}, u_{ij}) \tag{5}$$

Here, λ_1, λ_2, and λ_3 are the hyperparameters; N is the number of candidate frames; and t'_n is a binary value that indicates whether a pixel point is covered by a candidate frame (coverage is indicated by 1 and noncoverage by 0). $\frac{L_{reg}(v'_{ni}, v_{ni})}{|L_{reg}(v'_{ni}, v_{ni})|}$ determines the direction of the gradient propagation, and IoU is the intersection ratio of the prediction box and the ground truth. h is the height of the candidate frame, w is the width of the candidate frame, and $L_{VAM}(u'_{ij}, u_{ij})$ is the loss function of the VAM.

(a) PolyIoU for a multioriented anchor box

The direction of smoke fluttering in its natural condition is related to air thermal convection, which has the characteristics of upward fluttering and irregular shifting. Fig. 7 depicts smoke with multiple orientations.

We aim to determine the direction of the smoke as accurately as possible to facilitate the determination of the fire location, which will help prevent as much damage as possible from the forest fire. In addition, we choose a suitable anchor box representation to reduce the amount of redundant information that is provided to the network during training, thereby reducing the number of learning options that are available to the network, facilitating the constraint of the network's training direction, and reducing the convergence time of the network. We use a five-parameter long-edge representation (x, y, w, h, θ) to represent the multioriented anchor box θ, where x denotes the angle through which the longest edge is encountered under clockwise rotation in the axial direction and $\theta \in [0, \pi)$. The traditional IoU calculation method for multioriented anchor boxes uses axis-aligned bounding boxes to calculate the IoU; however, the IoU calculation on axis-aligned bounding boxes may lead to an inaccurate IoU for multioriented detection and further corrupt the final prediction.

To accurately calculate the intersection ratio of a multioriented anchor box, we use the PolyIoU calculation, in which the angles are considered. First, the intersection points that are formed by the edges of two rectangular boxes are identified with the vertices of a rectangular box that is contained within another rectangular box. From this, the problem is transformed into a problem of finding the area of the polygon that these vertices surround. The polygon area problem can be solved using the fork product between two of the vertices. The right-hand rule is used to determine the positive and negative aspects of the area and to determine whether the calculated area is positive or negative. The calculation of PolyIoU for the three crossover methods is illustrated in Fig. 8.

The intersecting area of the rectangular boxes is calculated using formula (6).

$$S_{Inter\,sec\,tion} = \sum_{i=0}^{n-1}(x[i]*y[(i+1)\%n] - y[i]*x[(i+1)\%n]) \tag{6}$$

Next, we determine the union area employing the tolerance repulsion theorem, which yields Formula (7).

$$S_{Union} = S_1 + S_2 - S_{Inter\,sec\,tion} \tag{7}$$

Finally, the PolyIoU calculation method directly calculates the intersection ratio between two anchor boxes, where the intersection area of the two inclined boxes is the numerator and the

Fig. 7. Smoke in three directions.

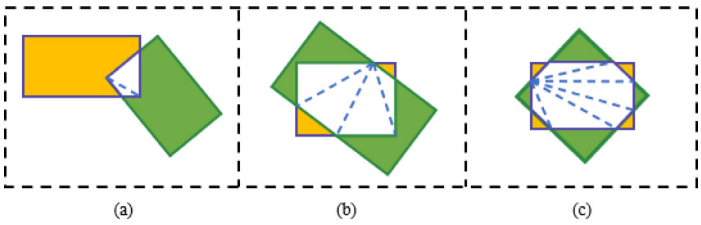


Fig. 8. PolyIoU calculation: The white part is the intersection of two rectangular boxes, which is divided into several triangles with a blue dotted line. Then, the areas of the triangles are calculated separately and added together to obtain the total area of the intersection.

merged area is the denominator, as expressed in Formula (8).

$$IoU = \frac{S_{Inter\ sec\ tion}}{S_{Union}} \tag{8}$$

When the ground truth is horizontal, an inaccurate prediction is obtained. For this reason, we move each point forwards or backwards by one bit, calculate the losses of the three methods when moving forwards and moving backwards, and select the smallest of these values as the result. This calculation, which is expressed by the position loss function, is presented in Formula (9).

$$l_{mr}^{8p} = \min \begin{cases} \sum_{i=0}^{3} \left(\left| x_{(i+3)\%4} - x_i^* \right| + \left| y_{(i+3)\%4} - y_i^* \right| \right) \\ \sum_{i=0}^{3} \left(\left| x_i - x_i^* \right| + \left| y_i - y_i^* \right| \right) \\ \sum_{i=0}^{3} \left(\left| x_{(i+1)\%4} - x_i^* \right| + \left| y_{(i+1)\%4} - y_i^* \right| \right) \end{cases} \tag{9}$$

3.3.2. Softpool-spatial pyramid pooling (Soft-SPP)

The images in the smoke dataset are of different sizes, but the input images need to be of the same fixed size for the network training. In previous object detection methods, random cropping or warping was often used. Random cropping solves the problem of varying input image sizes, but when the cropped region is repeated, the weight of the repeated region is inadvertently increased. Object distortion and image distortion can occur. To address these issues, YOLOv5 employs spatial pyramid pooling (SPP) [37] at the end of the feature extraction network. Three parallel MaxPool operations are used in this structure. However, the introduction of these MaxPool operations results in a significant loss of feature map information. It is critical to preserve as much feature map information as feasible in a single-target detection problem, such as forest fire smoke detection. Therefore, this paper designs Softpool-spatial pyramid pooling (Soft-SPP), which is a module that replaces all MaxPool operations in SPP with SoftPool operations [38]. This has the advantage of retaining as much feature map information as possible and performs better in single-category object detection. SoftPool is a fast and efficient exponential weighting method, and compared with other pooling methods, SoftPool retains more information in downsampled activation mapping and operates at a finer scale. This facilitates the solution of single-category object detection problems such as smoke detection. The SoftPool steps are illustrated in Fig. 9.

P_1, P_2, P_3, P_4 form an area of size 2*2 in the original figure. First, we use Formula (10) to convert P_1, P_2, P_3, and P_4 in Area1 into four areas in Area2. Then, the four areas in Area1 and Area2 are multiplied separately and added to the results that are obtained by multiplying them together to obtain the final result.

$$\frac{e^{P_i}}{\sum_{j=1}^{4} e^{P_j}} (i = 1, 2, 3, 4) \tag{10}$$

On ImageNet, for a range of popular CNN architectures, replacing the original pooling operation with SoftPool resulted in 1%–2% improvements in consistency and accuracy, thereby demonstrating that significant results can be obtained by replacing other pooling operations with SoftPool [37]. In addition, SoftPool has fewer parameters and realizes better convergence in the first few training rounds than downsampling using convolution. The SoftPool operation is expressed as follows:

$$a = \sum_{i \in R} \frac{e^{a_i} * a_i}{\sum_{j \in R} e^{a_j}} \tag{11}$$

In each grid of the feature map, we respond to each convolution using SoftPool. The number of grids is set to M, the number of filters for the previous convolution layer is k, and the output of the spatial pyramid pooling is a k*M-dimensional vector. The fixed-dimensional vector is the input to the fully connected layer. The Soft-SPP module is illustrated in Fig. 10.

3.3.3. Value conversion-attention mechanism module (VAM)

Smoke diffusion is irregular, and the forest environment is complicated and changing, which results in inconspicuous pixel-based features. YOLOv5 uses only CSP-Darknet as the feature extraction network for the backbone network, which makes it difficult to adequately take into account the textural features of smoke. Textural features are used to describe the spatial distribution in combination with the object colours and have advantages in terms of stability. To enhance the extraction performance of the texture feature extraction network, we propose an attention mechanism value conversion-attention mechanism module (VAM) method that is based on a joint weight allocation strategy

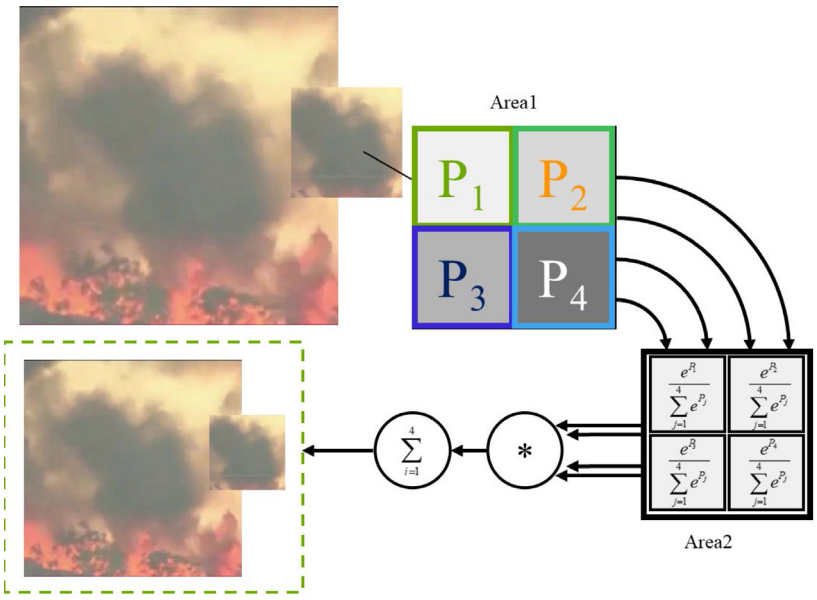


Fig. 9. SoftPool exponential weighting process.

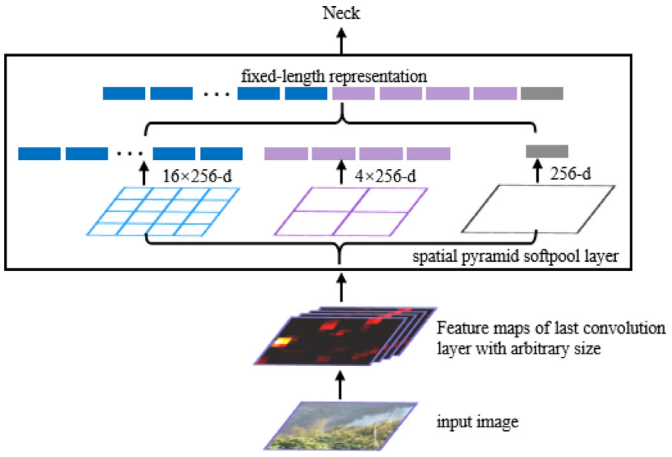


Fig. 10. Soft-SPP module.

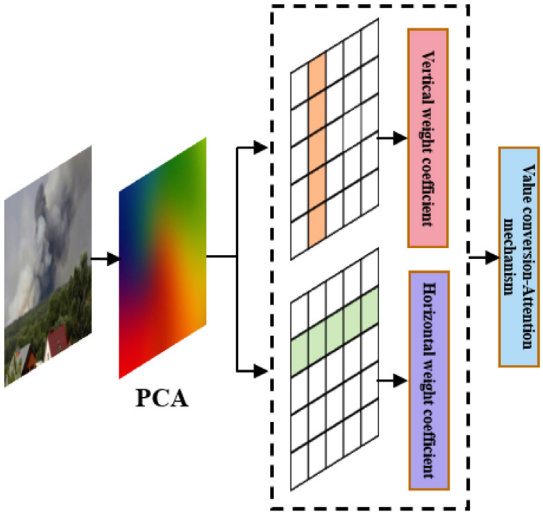


Fig. 11. VAM algorithm flow.

for the horizontal and vertical directions. It targets the colour, texture and specificity of the smoke for feature extraction. The first part is an attention mechanism that generates weighted features in the horizontal and vertical directions. The second part sums the two types of weights and further expands the weight coefficients. The third part matches the weighting features, considers the two types of weighting features, and selects the larger weighting features to determine the larger weighting coefficients. Then, these are used to complement the weighting coefficient results of the second part and to direct the focus onto objects in the image with colours and textures that are similar to those of smoke.

In smoke feature processing using VAM, first, we use principal component analysis (PCA) for dimensionality reduction to obtain low-dimensional image features. For each image, we obtain image features for each row of pixels and image features for each column of pixels, thereby reducing the computational cost. Then, the features that are obtained by dimensionality reduction are fed into the VAM to obtain more representative deeper features and relationships between features. The VAM algorithm flow is illustrated in Fig. 11.

Three strategies are used in the VAM to amplify the differences in the weight coefficients of the features. They are described below.

(a) Horizontal and vertical distribution strategy

We assign horizontal weight coefficients to each row of features by using the horizontal attention mechanism and assign vertical weight coefficients to each column of features by using the vertical attention mechanism.

$$c_i = \sum_{j=1}^{n} \frac{\exp(e_{i,j})}{\sum_{k=1}^{n} \exp(e_{ik})} h_j \tag{12}$$

(b) Weighted coefficient additive strategy

We sum the two types of weighting features to further extend the weighting factors.

$$add = (c_I + c_{II}) \tag{13}$$

Fig. 12. Concatenation and merging process.

(c) Weight allocation strategy

To use the maximum value as the main factor and to take into account the other features, the strategy is matched with two types of weighting features, which are used to complement the results of the weighting addition strategy in the second step.

$$distribution = \alpha^* \max(c_I, c_{II}) + \beta^* \min(c_I, c_{II}) \tag{14}$$

In the experiments in Section 4.2.2, we confirm that the optimal values of the weight assignment parameters are $\alpha = 0.8$ and $\beta = 0.2$.

The three strategies in this method are combined in the following formula, and the concatenation procedure is illustrated in Fig. 12.

$$VAM = concatenate([c_I, c_{II}, add, distribution]) \tag{15}$$

In the above formula, e_{ij} represents the weight coefficients of the attention mechanism, i represents the temporal features, j represents the sequence features, h_j represents the hidden-layer information of the feature sequence, c_I represents the vertical attention mechanism feature sequence ($c_I = \{c_1, c_2 \ldots c_{i-1}, c_i\}$), and ($c_{II} = \{c_1, c_2 \ldots c_{i-1}, c_i\}$) represents the horizontal attention mechanism feature sequence. *add* represents the additive weighting factor. max represents the maximum value operation. min represents the minimum value operation. *distribution* represents the weight assignment strategy. The VAM weight assignment procedure is illustrated in Fig. 13.

3.3.4. Mixed-NMS algorithm for processing candidate frames

Nonmaximum suppression (NMS) is extensively employed in the prediction stage of target detection based on deep learning models to overcome the problem of multiple repeating detection boxes around the object. The forest environment is complex, with small changes in smoke pixel values early in smoke generation and when the smoke is far away. Using a single threshold for smoke detection can easily lead to an increase in missed reports, thereby making it difficult to overcome the impact of complex background environments on smoke detection. The variable forest environment also makes it difficult to predict the shape and size of the smoke that is produced, along with the distribution and density of the smoke. As the main strategy, YOLOv5 uses IoU values in NMS to select prediction boxes and uses a weighted NMS approach. However, for smoke, it is difficult to choose a suitable threshold because real smoke targets vary in shape, thereby resulting in a range of confidence levels for detection. The use of a low NMS threshold may lead to overfiltering of candidate frames, thereby resulting in a decrease in mAP. The reason for this is that there may be a detection box b_i that is very close to the real target but has a confidence level that is slightly below M; this closer target box would be incorrectly filtered out. Additionally, the use of a high NMS threshold may also introduce false detections. In this case, repeated false smoke detections would occur in large numbers. This is because high thresholds enable different features of the same cloud of smoke to be detected repeatedly as multiple objects.

In addition, the method for object detection of smoke using a multioriented anchor box approach that is described above has an added angle parameter, but this angle needs to be considered in the NMS phase. In multioriented object detection, the IoU is substantially affected when the angle of the two objects changes. Weighted NMS of YOLOv5 is not suitable for multiorientation detection in forest fire smoke scenarios. As illustrated in Fig. 14, the IoU for two preselected boxes is significantly smaller when the boxes are oriented at larger angles than at smaller angles.

Inaccurate calculation of the IoU can lead to missed or false smoke detections. When missed detections occur, people are not informed of fires in time, thereby leading to uncontrollable fires. When false detections occur, a heavy toll is placed on the fire-fighting system. To obtain excellent forest fire detection results and reduce the occurrence of missed and false detections, this paper proposes Mixed-NMS by combining the Skew-NMS [36] and DIoU-NMS [39] methods.

Skew-NMS considers angle information based on NMS. If there are candidate frames with IoUs of greater than 0.7, the largest candidate frame is retained; if all candidate frames are located in [0.3, 0.7], the smallest candidate frame with an angle of less than 30° is retained.

The DIoU is calculated based on the traditional IoU such that if the centroid of an adjacent box is closer to the centroid of the current maximum scoring Box M, it is more likely to be a redundant box. The calculation formula is as follows:

$$DIoU = IoU - \frac{d^2}{c^2} \tag{16}$$

where d is the distance between the centres of the two a priori boxes and c is the distance between the furthest vertices of the two a priori boxes, as illustrated in Fig. 15.

Mixed-NMS combines the DIoU with the Skew-NMS method and uses two different thresholds for the interval division, with the angle as the parameter that should be considered. Mixed-NMS sorts the preselected boxes in descending order according to the confidence level, sets two thresholds, and uses different penalty operations for different DIoU intervals. If the DIoU is less than 0.3, no penalty operation is applied; if the DIoU is in [0.3, 0.7], then the preselected box with the highest confidence level is retained if the angle is less than $\frac{\pi}{6}$; and if the DIoU is greater than 0.7, then the preselected box with the lowest confidence level is filtered out directly.

4. Results and analysis

This section is separated into subsections to (1) describe the experimental environment and settings, including the hardware and software environment, and hyperparameter settings; (2) evaluate the effectiveness and identify the critical parameters of each MVMNet module; (3) demonstrate the efficacy of the new method proposed in this paper via ablation experiments on MVMNet; (4) compare MVMNet to other methods and demonstrate that MVMNet outperforms other methods in forest fire smoke monitoring tasks; and (5) describe tests in real-world application scenarios. According to the test results, MVMnet overcomes the challenges of inclined smoke, misdetection of objects that are similar to smoke, and smoke that is too far away from the camera.

4.1. Experimental environment and settings

All of the tests in this work are carried out on the same hardware and software platform. The environmental parameters are listed in Table 1.

We create and publish a useable forest fire smoke dataset. To avoid aspect ratio mismatch and ensure a satisfactory learning

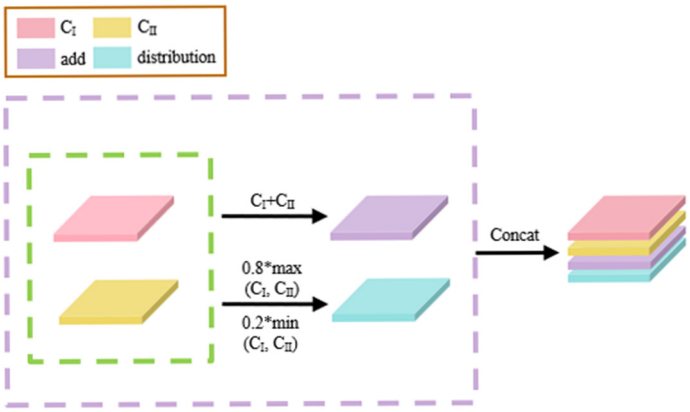


Fig. 13. Weight allocation strategy.

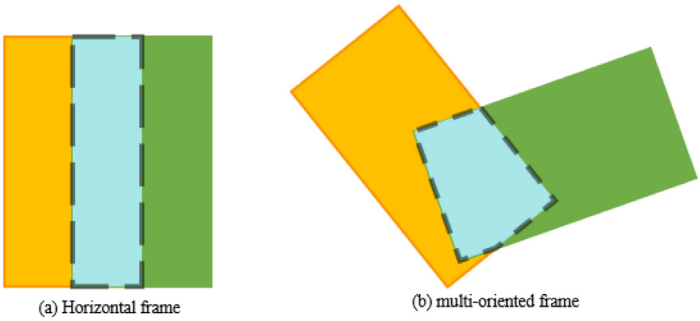


Fig. 14. Effect of angle on the IoU.

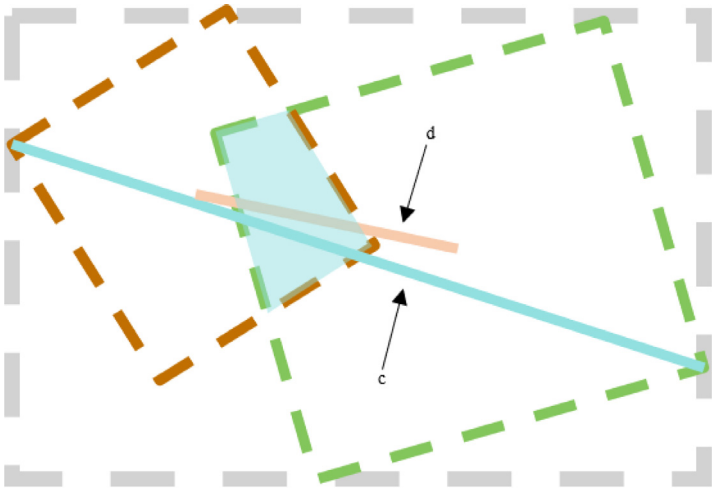


Fig. 15. Meanings of d and c in the DIoU.

effect, we chop all the images in the dataset into images with height equal to width. During training, the model resizes the input images to 512 * 512. The Gaussian distribution is used to initialize each layer of the model. Next, we set the batch size to 1, momentum to 0.8, initial learning rate to 0.005, decay to 0.0005, and length of each training to 150 epochs, taking into consideration the GPU memory size and the time consumption of the experiment. Table 2 contains the settings and hyperparameters. In the experiment, we use a 7:2:1 ratio to divide the 15909 images from the forest fire multioriented detection dataset into a training set, a validation set, and a test set.

4.2. Module effectiveness analysis

The parameters and functions of each module that we built are examined in detail in this section, and the optimal pooling method in the SPP module is detailed in Section 4.2.1. The weight distribution coefficient of VAM is determined in 4.2.2. Experiments in 4.2.3 demonstrate that PolyIOU is better suited for multioriented detection than axis-alignment IOU. In 4.2.4, the performance of YOLOv5 is evaluated when greedy-NMS, DIoU-NMS, Skew-NMS, and Mixed-NMS are employed, and it is concluded that Mixed-NMS is the best NMS method for multioriented detection.

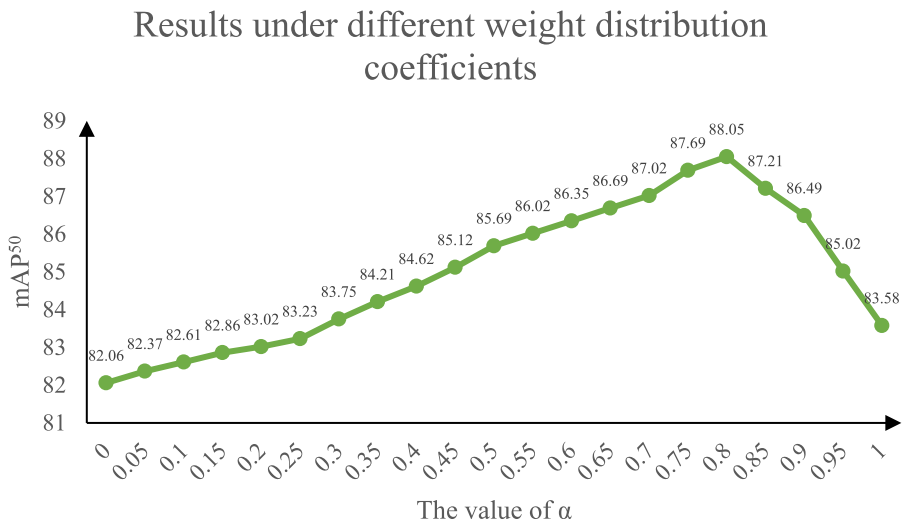


Fig. 16. Correspondence between mAP^{50} and the weight assignment factor.

Table 1
Hardware and software parameters.

Hardware environment	CPU	AMD Ryzen 7 5800H with Radeon Graphics
	RAM	16 GB
	Video memory	16 GB
	GPU	NVIDIA GeForce RTX 3060 Laptop GPU
Software environment	OS	Windows 10
	CUDA Toolkit V11.1; CUDNN V8.0.4; Python 3.8.8; torch 1.8.1; torchvision 0.9.1	

4.2.1. Effectiveness of Soft-SPP

To assess the efficacy of SoftPool for Soft-SPP and determine the appropriate pooling method for retaining smoke features, we employ average pooling, median pooling, average + maximum pooling, and SoftPool to boost the SPP in YOLOv5. Table 3 displays the experimental results:

The experimental results demonstrate that compared to the original SPP of YOLOv5, employing average and median pooling rather than maximum pooling results in a minor loss in accuracy. It can be demonstrated that simply replacing the maximum pooling layer does not enhance the accuracy. In contrast, the average + maximum pooling combination method can boost the mAP^{50} by 0.25%, which does not considerably improve the smoke detection accuracy. Using SoftPool instead of MaxPool's Soft-SPP can considerably increase the model's detection accuracy, which improves to the accuracy of forest fire monitoring.

4.2.2. Effectiveness of VAM

In the weight assignment strategy of Section 3.3.3, the values of α and β affect the performance of VAM in relation to the weight assignment of c_1 and c_2. For convenience, we specify that

$$\alpha + \beta = 1. \tag{17}$$

A comparison experiment is carried out. The interval between values is set to $\alpha = 0.05$, and the values of β are always related to the values of α.

To identify the optimal values of α and β, we modify the weight assignment coefficients of VAM in MVMNet for testing. The experimental results are shown in Fig. 16.

The optimal values are $\alpha = 0.8$ and $\beta = 0.2$. When the value of α is too large, the weight of the minimum value is neglected, thereby resulting in the image features at locations with smaller values being ignored and the extraction of image features losing its global nature. When the value of α is too small, excessive consideration of the global features makes the attention mechanism unable to focus sufficiently on important information, thereby affecting the performance of VAM.

We introduce the SE Attention and CBAM Attention modules to the backbone end of YOLOv5 for comparative tests to investigate the efficiency of VAM. Table 4 displays the test results.

The attention mechanism module improves the detection accuracy after the addition of SE Attention and CBAM Attention after the CSP_2 block at the conclusion of the backbone. However, the improvement effect is subtle, and the addition of the attention mechanism module increase the number of parameters. When VAM is applied, the accuracy improves dramatically, and it results in approximately the same increase in the number of parameters as SE Attention and CBAM Attention. As a result, we select VAM to enhance MVMNet's feature extraction capability.

4.2.3. Effectiveness of PolyIoU

We add a multioriented anchor box to the traditional YOLOv5 and utilize the two distinct IoU calculation methods to demonstrate the advantages of the PolyIoU calculation method over the traditional axis-aligned IoU calculation method for calculating the IoU. Fig. 17 presents the experimental results.

The axis-aligned IoU calculation method has evident flaws, and the target detection of the multioriented anchor box differs from that of the horizontal box. If the original method is still used when utilizing a multioriented anchor box, the mAP will decrease substantially. As a result, the original calculation approach should be abandoned in favour of the PolyIoU method, which increases the accuracy by 1.23% while employing the multioriented anchor box.

4.2.4. Effectiveness of Mixed-NMS

Typically, conventional YOLOv5 methods use greedy NMS as a nonextreme value suppression method. In this paper, we propose Mixed-NMS, which mixes DIoU-NMS and Skew-NMS. We perform comparative experiments on models that incorporate only multioriented anchor box and PolyIoU methods for the traditional YOLOv5, and the experimental results are shown in Fig. 18.

The experimental results show that Skew-NMS improves the mAP^{50} by 1.63% and DIoU-NMS improves the mAP^{50} by 1.21%

Table 2
Experimental settings.

Size of input images	Batch_size	Momentum	Initial learning rate	Decay	Iterations
512 × 512	1	0.8	0.005	0.0005	150 epochs

Table 3
Performance of SoftPool for Soft-SPP.

Method	mAP^{50} (%)	mAP^{75} (%)	AR	FPS	Params	FLOPs
SPP	77.12	69.25	41.52	156	22.0 M	51B
Avg-SPP	76.98	69.11	41.34	150	22.1 M	51B
Med-SPP	76.81	68.97	41.21	155	22.0 M	51B
Avg+Max-SPP	77.37	69.51	41.79	148	22.3 M	52B
Soft-SPP	79.98	70.93	43.86	146	22.4 M	53B

Table 4
Performance of VAM.

Method	mAP^{50} (%)	mAP^{75} (%)	AR	FPS	Params	FLOPs
No improvement	77.12	69.25	41.52	156	22.0 M	51B
SE Attention	79.28	71.79	43.16	139	22.5 M	55B
CBAM Attention	79.69	72.14	43.88	132	22.7 M	57B
VAM	84.39	76.12	47.63	134	22.7 M	56B

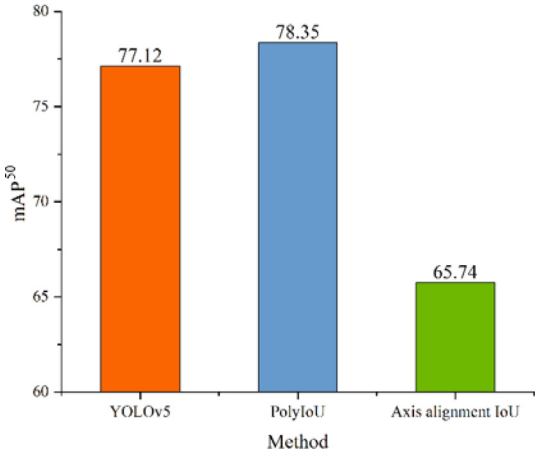


Fig. 17. mAP^{50} results of three methods.

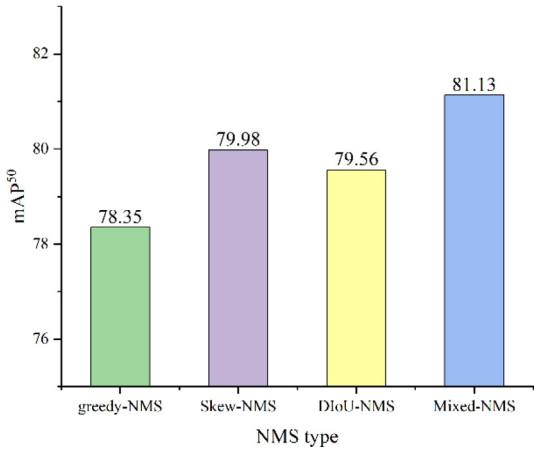


Fig. 18. mAP^{50} results of four NMS methods.

compared to the greedy-NMS method. Mixing the two nonmaximum suppression methods and setting a reasonable threshold constitute the most beneficial approach for maintaining model accuracy, which improves the mAP^{50} by 2.77% compared to the greedy NMS. Although the addition of these two NMS methods would theoretically reduce the prediction speed, in practice, the prediction phase only accounts for a small fraction of the time consumption and has little impact on the prediction speed of the model in actual deployment. Therefore, this paper uses a combination of these two nonmaximum suppression methods for smoke detection.

4.3. Ablation experiments

To evaluate the performance of the method that is proposed in this paper, we perform MVMNet ablation experiments and deploy various MVMNet variants. We remove either Soft-SPP, VAM, or Mix-NMS from each implementation. To facilitate comparison, we add angle parameters to each set of tests and use the PolyIoU calculation method. Table 5 presents the experimental results. The following observations are made:

Based on MVMNet, we apply the control variable approach to erase VAM, Soft-SPP, and Mixed-NMS one at a time and replace them with SPP and NMS (the two modules are included in the YOLOv5 method). Finally, YOLOv5 is compared to multioriented detection-YOLOv5.

- Comparative experiments between the eighth and ninth groups reveal that using an angled detection box can modestly improve mAP, which is more noticeable on the easy-to-detect smoke dataset than on the difficult-to-detect smoke dataset. Simultaneously, the FPS is reduced by 2 due to the insertion of the angle parameter.
- Comparison of the seventh and eighth groups reveals that the Mixed-NMS method can raise the smoke detection index mAP while decreasing the FPS by 5.
- Comparison between the sixth and eighth groups demonstrates that Soft-SPP can successfully improve mAP while reducing the speed and number of parameters.
- Comparison between the fifth and eighth groups demonstrates that VAM can significantly boost mAP but is also the primary cause of an increased number of MVMNet parameters and decreased FPS. The VAM module may ensure that all smoke feature information is extracted completely.

The results of nine groups of experiments thoroughly demonstrate the contributions of VAM, Soft-SPP, and Mixed-NMS to mAP improvement. MVMNet is more suitable for forest fire smoke target detection than YOLOv5.

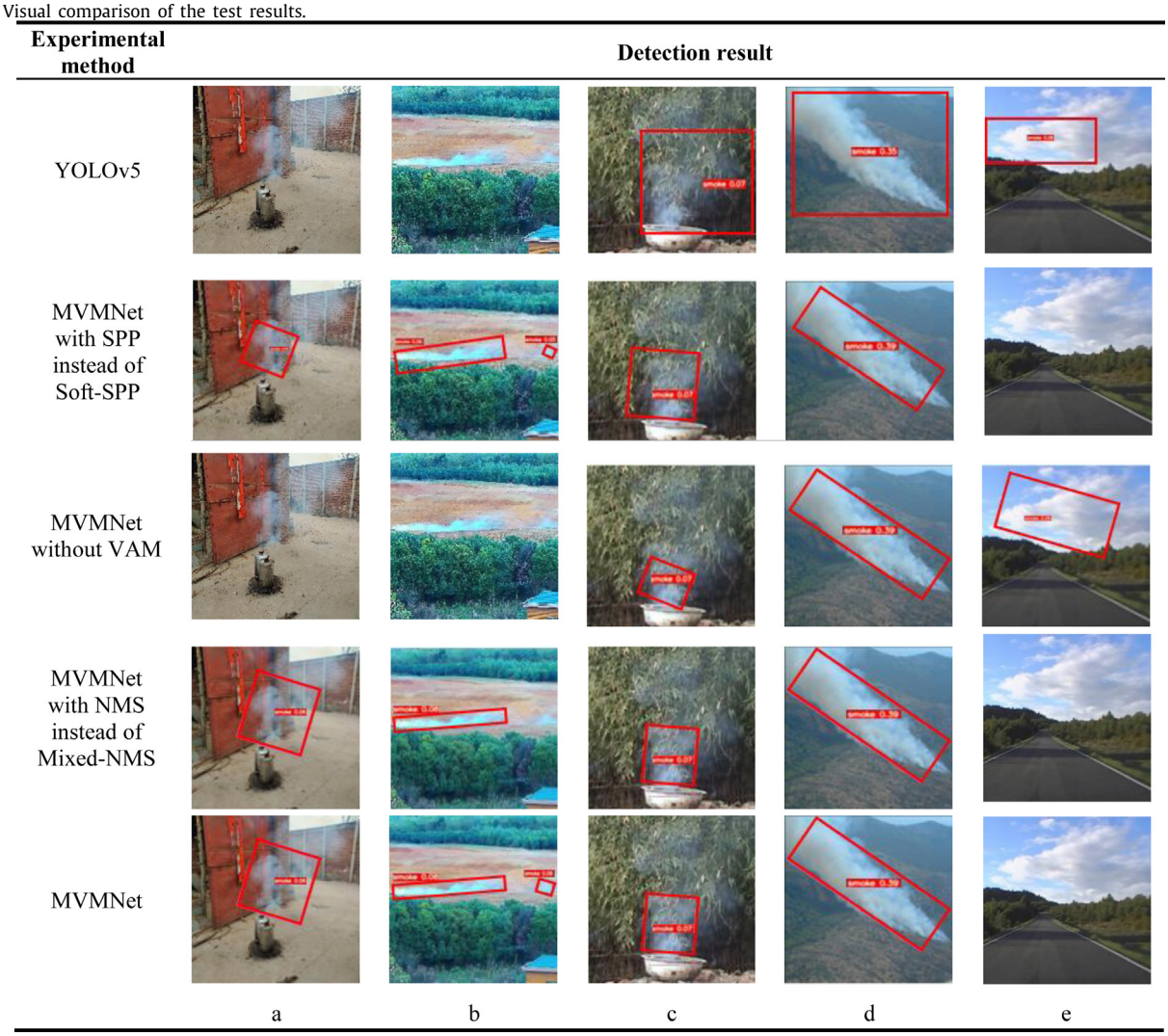
We examine the detection results of MVMNet with VAM removed, SPP replacing Soft-SPP, and NMS replacing Mixed-NMS to analyse the performance of the method that is proposed in this paper. The detection box, category, and confidence are all displayed on the detection result graph, as illustrated in Table 6.

As shown in Table 6, the smoke concentration in Fig. a and Fig. b is small, and YOLOv5 is unable to detect this smoke, while MVMNet is able to detect it successfully. If SPP is used to replace Soft-SPP in MVMNet, the accuracy of the results is reduced. If VAM is removed, the smoke features are difficult to extract. If conventional NMS is used instead of Mixed-NMS, the smaller cloud of smoke in Fig. b is missed, thereby demonstrating that Mixed-NMS can reduce missed detections. In Fig. d, the smoke is tilted, the YOLOv5 detection using the horizontal box does not match as well as that of MVMNet, and in both cases, the detection loses its match when Soft-SPP in MVMNet is replaced by SPP and VAM is removed. The absence of Mixed-NMS has no effect. The colour and morphology of the clouds in the sky in Fig. e are

Table 5
Ablation experiment results.

Number	Method	mAP^{50} (%)	mAP^{75} (%)	AR (%)	FPS	Param	FLOPs
1	VAM+Soft-SPP+Mixed-NMS(MVMNet)	88.05	80.75	48.91	122	23.1M	59B
2	VAM+Soft-SPP+NMS	87.01	80.17	48.32	126	23.1M	58B
3	VAM+SPP+Mixed-NMS	86.38	79.91	48.01	132	22.7M	57B
4	Soft-SPP+Mixed-NMS	82.94	74.63	48.85	143	22.4M	54B
5	VAM+SPP+NMS	84.39	76.12	47.63	134	22.7M	56B
6	Soft-SPP+NMS	79.98	70.93	43.86	146	22.4M	53B
7	SPP+Mixed-NMS	81.13	71.05	43.27	149	22.1M	52B
8	SPP+NMS(multioriented detection-YOLOv5)	78.35	69.37	41.76	154	22.1M	51B
9	YOLOv5	77.12	69.25	41.52	156	22.0M	51B

Table 6
Visual comparison of the test results.



very similar to those of smoke, and the YOLOv5 model misdetects these clouds as smoke, whereas MVMNet does not cause misdetection. This misdetection also occurs when VAM is removed. Thus, the important role of VAM in the feature extraction session is demonstrated.

4.4. Comparison of MVMNet with other methods

We compare the performance of MVMNet with those of other target detection networks in Section 4.4.1 and demonstrate that it outperforms other models in forest fire smoke monitoring tasks. Then, in Section 4.4.2, we apply MVMNet in a complex context and demonstrate that MVMNet can fulfil difficult forest fire smoke monitoring tasks.

4.4.1. Comparison of the MVMNet method with other target detection networks

To further analyse the performance of MVMNet, we compare it with RCNN, SPPNet, Fast RCNN, Faster RCNN, OverFeat, YOLO, YOLOv2, YOLOv3, YOLOv4, YOLOv5, and SSD in smoke object detection. RCNN extracts category-independent values from the input smoke image candidate regions, and for each region, a fixed-length feature vector is extracted using CNN, and the extraction is uniformly transformed to a fixed size of 227*227, regardless of the size and shape of the candidate regions. To remove the constraint of fixed image size, SPPNet [65] introduces a spatial pyramid pooling layer. After the last convolutional layer, the SPP layer is inserted. To avoid distortion of the smoke image

Table 7
MVMNet versus other models on the forest fire multioriented detection dataset test-dev. MVMNet outperforms all one-stage detectors and achieves results that are competitive with those of two-stage detectors.

Method	Backbone	AP	AP^{50}	AP^{75}	AR	FPS
Two-stage detectors						
DeNet	ResNet-101	61.52	70.38	65.12	36.49	–
CoupleNet	ResNet-101	63.27	72.63	66.59	37.11	–
Fast RCNN	ResNet-101	65.05	74.69	67.10	38.89	–
Faster R-CNN by G-RMI	Inception-ResNet-v2	66.52	76.26	68.99	39.76	–
Faster R-CNN+++	ResNet-101	67.68	77.49	69.23	40.19	–
Faster R-CNN w/FPN	ResNet-101	68.24	81.13	70.75	40.73	–
Faster R-CNN w/TDM	Inception-ResNet-v2	68.80	78.87	71.06	42.27	–
D-FCN	Aligned-Inception-Resnet	69.54	80.06	72.33	43.43	–
Regionlets	ResNet-101	70.87	81.49	74.05	44.64	–
Mask R-CNN	ResNet-101	72.31	82.14	75.21	45.29	–
LH R-CNN	ResNet-101	74.62	84.12	77.13	46.29	–
Cascade R-CNN	ResNet-101	75.54	85.96	78.86	47.16	–
D-RFCN + SNIP	DPN-98	77.22	87.31	80.82	48.01	–
One-stage detectors						
YOLO	GoogLeNet	46.11	54.10	48.63	33.02	44
YOLOv2	DarkNet-19	59.01	69.95	61.79	37.41	67
YOLOv3	DarkNet-53	65.97	77.25	69.63	41.73	32
YOLOv4	CSP-Darknet53	77.36	86.62	79.02	47.94	35
YOLOv5	CSP-Darknet53	66.02	77.12	69.25	41.52	**156**
DSOD300	DS/64-192-48–1	62.95	74.20	65.14	38.85	26
GRP-DSOD320	DS/64-192-48–1	63.72	75.03	68.86	39.38	18
SSD513	ResNet-101	59.63	70.57	62.14	37.58	89
DSSD513	ResNet-101	61.89	73.01	64.76	38.16	52
RefineDet512(single scale)	ResNet-101	64.52	77.81	67.31	39.86	81
RetinalNet800	ResNet-101	68.72	78.35	70.19	42.08	68
RefineDet512(multi scale)	ResNet-101	70.25	80.96	73.29	43.97	67
YOLOX [40]	CSP-Darknet53	69.17	82.71	73.46	42.68	25
FireNet [41]	–	77.15	87.23	80.34	48.48	23
DCNN [42]	–	77.64	87.64	80.40	48.60	30
DeepSmoke [43]	EfficientNet	78.16	87.31	80.80	48.93	33
Two-stage detectors (multioriented detection method)						
SCRDet [44]	SF-Net+MDA-Net	70.19	80.36	73.02	43.10	–
R^2CNN [45]	ResNet-101	58.17	–	–	–	–
GLSNet [46]	ResNet-101	70.81	–	–	–	–
FADet [47]	ResNet-152	71.31	–	–	–	–
SARD [48]	ResNet-152	71.64	–	–	–	–
FFA [49]	ResNet-101	73.79	–	–	–	–
APE [50]	ResNext-101	73.81	–	–	–	–
F^3Net [51]	ResNet-152	74.26	–	–	–	–
CSL [52]	ResNet-152	74.73	–	–	–	–
MRDet [53]	ResNet-101	74.99	–	–	–	–
SCRDet++ [54]	ResNet-101	75.12	–	–	–	–
FR-EST [55]	ResNet-101-DCN	77.15	–	–	–	–
One-stage detectors (multioriented detection method)						
TOSO [56]	ResNet-101	57.72	67.48	60.94	37.13	53
A^2SDet [57]	ResNet-101	67.85	77.16	71.01	44.18	49
DRN [58]	H-104	70.99	80.08	74.32	46.02	61
R^4Det [59]	ResNet-152	74.48	83.59	78.26	47.39	88
R^3Det [60]	ResNet-152	72.14	81.19	75.78	46.86	106
PolarDet [61]	ResNet-101	74.26	83.95	77.15	47.26	68
RDD [62]	ResNet-101	76.19	86.02	79.39	48.11	79
S^2A-Net [63]	ResNet-101	77.69	87.08	79.92	48.26	46
GWD [64]	ResNet-152	78.15	87.69	80.46	48.82	29
MVMNet	CSP-Darknet-53	**78.92**	**88.05**	**81.15**	**49.08**	122

that is caused by “flatness" or “hypertrophy” that is induced by cropping or warping at the beginning, the SPP layer pools the feature map and generates a fixed-length output. Fast RCNN [66] convolves the entire smoke image directly, thereby substantially reducing the number of repetitive calculations. Faster RCNN [67] combines multiple steps of feature extraction, proposal extraction, and bounding box regression into one network, thereby improving the performance of smoke detection. OverFeat [68] uses multiscale training and moves away from traditional non-maximal suppression to a cumulative prediction approach. YOLO uses the whole map as input to the network and regresses the location and class of the detection box in the output layer, with varying degrees of improvement in each of its five versions. SSD [69] uses one CNN network for detection as YOLO does but with two multiscale feature maps, which enables more granular smoke features to be extracted. The above methods all use a deep convolutional neural network model, namely, MVMNet, and their comparative experimental results on the forest fire multioriented detection dataset are presented in Table 7.

The results show that the one-stage detectors are faster than the two-stage target detectors. Compared to the traditional horizontal box single-stage target detector, YOLOv2 and SSD are substantially faster, but their accuracies are insufficient. The subsequent single-stage target detection model shows improvements

Table 8
MVMNet and the current popular methods ranked in descending order of AP, AR and FPS. The performance difference of popular methods relative to MVMnet are specified in parentheses.

Rank	AP	AR	FPS
1	**MVMNet (78.92)**	**MVMNet (49.08)**	YOLv5 156 (+34)
2	DeepSmoke (78.16(−0.66))	DeepSmoke (48.93(−0.15))	**MVMNet 122**
3	GWD (78.15(−0.67))	GWD (48.82(−0.26))	R^3Det (106(−16))
4	S^2A-Net (77.69(−1.23))	DCNN (48.6(−0.48))	SSD513 (89(−33))
5	DCNN (77.64(−1.28))	FireNet (48.48(−0.6))	R^4Det (88(−34))
6	YOLOv4 (77.36(−1.56))	S^2A-Net (48.26(−0.82))	RefineDet512(single scale) (81(−41))
7	D-RFCN + SNIP (77.22(−1.70))	RDD (48.11(−0.97))	RDD (79(−43))
8	FireNet (77.15(−1.77))	D-RFCN + SNIP (48.01(−1.07))	RetinalNet800 (68(−54))
9	FR-EST (77.15(−1.77))	YOLOv4 (47.94(−1.14))	PolarDet (68(−54))
10	RDD (76.19(−2.73))	R^4Det (47.39(−1.69))	YOLOv2 (67(−55))

Table 9
Comparison of the degree of variation in classification ability across models.

Rank	Method	Q statistics	Rank	Method	Q statistics
1	**MVMNet**	1	6	YOLOv4	0.89177
2	DeepSmoke	0.92762	7	D-RFCN + SNIP	0.88922
3	GWD	0.92237	8	FireNet	0.88764
4	S^2A-Net	0.89812	9	FR-EST	0.88764
5	DCNN	0.89513	10	RDD	0.85671

in terms of both accuracy and speed. YOLOv4 has a mAP^{50} of 86.62% and YOLOv5 has an FPS of 156. In addition, three up-to-date fire smoke detection methods are compared in the table. Since some of the more effective smoke monitoring algorithms employ image classification, we compare target detection using MVMNet's head rather than the fully connected layer in the original study. Our model improves the AP^{50} by 10.93% compared to YOLOv5 with little speed reduction and by 1.53% compared to YOLOv4 with an improvement in the detection speed by more than three times. MVMNet outperforms most detectors in terms of accuracy compared to the standard horizontal box two-stage target detector. MVMNet performs somewhat better overall than D-RFCN+SNIP, although it performs significantly worse in small target detection. However, in terms of speed, the two-stage detector falls short. Even in smoke detection for small targets, it is not as effective as MVMNet. In terms of accuracy, MVMNet outperforms all of the two-stage detectors in the table compared to newer multioriented detection algorithms that were developed in recent years. The single-stage detector MVMNet has a clear speed benefit. MVMNet outperforms novel single-stage detection algorithms such as RDD and GWD in terms of speed and accuracy. Specifically, the proposed MVMNet relies on the VAM and Soft-SPP modules to better extract complex and variable smoke features and fits the smoke morphology with a multioriented detection box. In summary, an accuracy of 88.05% is achieved in the experiments while maintaining an FPS of 122. Although slightly inferior in terms of speed compared to YOLOv5, it still meets the criteria for real-time forest fire smoke detection. Compared to the other models in the table, MVMNet is the most suitable model for forest fire smoke detection. We identify the following possible reasons why our proposed MVMNet model outperforms the other deep neural network models:

(1) Because our forest fire smoke dataset contains only 15,909 images, which is modest in comparison to large-scale open public datasets such as COCO and VOC, it puts the model's feature extraction capabilities to the test. In addition, large-scale open public datasets are multiclass and dominated by objects that are usually common in everyday life, such as people and cars, while smoke is an object that is usually rare and difficult to detect. While these traditional deep neural network models are designed for large-scale common public datasets, MVMNet is a well-designed structure for forest fire smoke detection research.

(2) VAM focuses on extracting features from smoke texture. The effect of feature extraction is improved by taking into account the horizontal and vertical dimensions, and it makes a significant contribution to the enhancement of mAP.

(3) MVMNet employs a multioriented detection and PolyIoU calculation method that is more in line with the features of smoke's indeterminate direction, which can improve the IoU during detection and, as a result, the mAP.

(4) A hybrid nonmaximum suppression method is used to effectively reduce false detections and missed detections that are caused by inappropriate threshold settings of traditional nonmaximum suppression methods, and the combination of two nonmaximum suppression methods further improves the positioning accuracy of smoke detection.

To more concisely reflect the advantages of MVMNet over popular solutions, we rank the top 10 classes of methods in descending order of AP, AR, and FPS.

According to Table 8, the proposed MVMNet is able to maintain TOP-1 AP and AR in the forest fire multioriented detection dataset. At the same time, it achieves 122 in the metric of FPS, second only to YOLOv5. Currently, popular high-precision models usually have extremely high numbers of layers and parameters; thus, the perform poorly in terms of FPS. Compared to popular methods, MVMNet shows a speed-accuracy trade-off.

In Table 9, we calculated Q statistics to indicate the difference between the top 10 AP methods and the MVMNet classification abilities. In general, the smaller the degree of difference, the closer the Q statistics is to 1 [70].

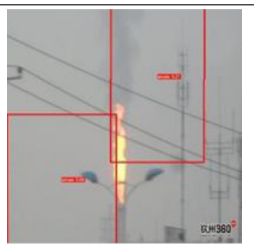
Table 9 demonstrates that the Q statistics of DeepSmoke and GWD are relatively close to one, indicating that they differ from MVMNet to a smaller amount. However, these two methods have just 33 FPS and 29 FPS, respectively, whereas the remainder of the methods differs significantly from MVMNet. Popular high precision models typically have extremely high layer and parametric counts. As a result, it performs poorly in terms of FPS. MVMNet has a speed-accuracy trade-off when compared to popular methods.

4.4.2. Visualization experiments of MVMNet under complex backgrounds

Three representative images from the forest fire multioriented detection dataset were chosen to exhibit the visualization results to demonstrate that MVMNet is up to the challenging task of forest fire smoke detection. The smoke in image a is next to a cloud, which is also a smoke-like object; the smoke in image b is very thin and translucent; and the smoke in image c is photographed against a hazy background. The smoke and the background are almost the same. In addition, for comparison, FireNet, DCNN, and DeepSmoke are trained individually. Table 10 displays the results of the comparative experiment.

The table shows that the clouds near the smoke in Figure a are misidentified as smoke by the FireNet, DCNN, and DeepSmoke

Table 10
Visual comparison with other fire smoke monitoring methods under complex backgrounds.

Experimental method	Detection result		
FireNet [41]			
DCNN [42]			
DeepSmoke [43]			
MVMNet			
	a	b	c

models but MVMNet identifies the smoke more accurately and does not misidentify the cloud. FireNet and DCNN miss the thin and transparent smoke in Figure b. DeepSmoke detects fuzzy grass as smoke, which could be due to a lack of extraction of smoke texture features. MVMNet performs fantastically. Because the background colour is highly similar to the smoke in Figure c, FireNet, DCNN, and DeepSmoke fail, but MVMNet detects the smoke appropriately. VAM is extremely likely to have played a role in the extraction of smoke features.

4.5. Testing of real applications

We replicate a forest fire combustion experiment in Zhuzhou Forest Farm and test it for 4 months using MVMNet's forest fire smoke detection method. We create a wildfire smoke IoT detection system based on MVMNet by combining MVMNet and hardware equipment. A smoke data gathering module, a host computer, and a smoke image processing module are the key components of the system. The smoke image is captured with a Hikvision DS-2DYH277IDU high-definition panhead camera, and one frame of the image is captured as test data every 4 s. Then, the gathered image data are sent over the network to the server for image processing and target detection. On the host computer, the test results can be displayed in real time. Fig. 19 presents a schematic diagram of the system.

When smoke is detected, our cameras can capture it and sound an alarm. Fig. 20 shows a comparison of YOLOv5 and MVMNet on the three recognized smoke categories. For testing on each category, we select 20 real-life scenes. As shown in the figure, the recognition accuracies of YOLOv5 and MVMNet are 100% in category A, 85% and 65% in category B, and 60% and 20% in category C, respectively.

MVMnet performs admirably in the actual application test and successfully handles the three scenarios that are depicted in Fig. 1. The smoke target in Fig. 21**.a** is modest and at a 45-degree angle to the ground. Fig. 21**.b** depicts a situation in which the smoke area is large but the smoke concentration is low and the colour is light. Fig. 21**.c** depicts a case in which the smoke target is tiny. MVMNet is capable of detecting the smoke in all the above scenarios. The application results in these real-world scenarios demonstrate MVMNet's superior performance in the task of smoke detection.

5. Discussion

A forest fire multioriented detection dataset for multioriented detection of fire smoke is produced, which covers a wide range of smoke features that occur during a fire. The effectiveness of the proposed MVMNet model for forest fire smoke detection is verified through the comparison and analysis of several sets of experiments. In particular, three main problems are well solved

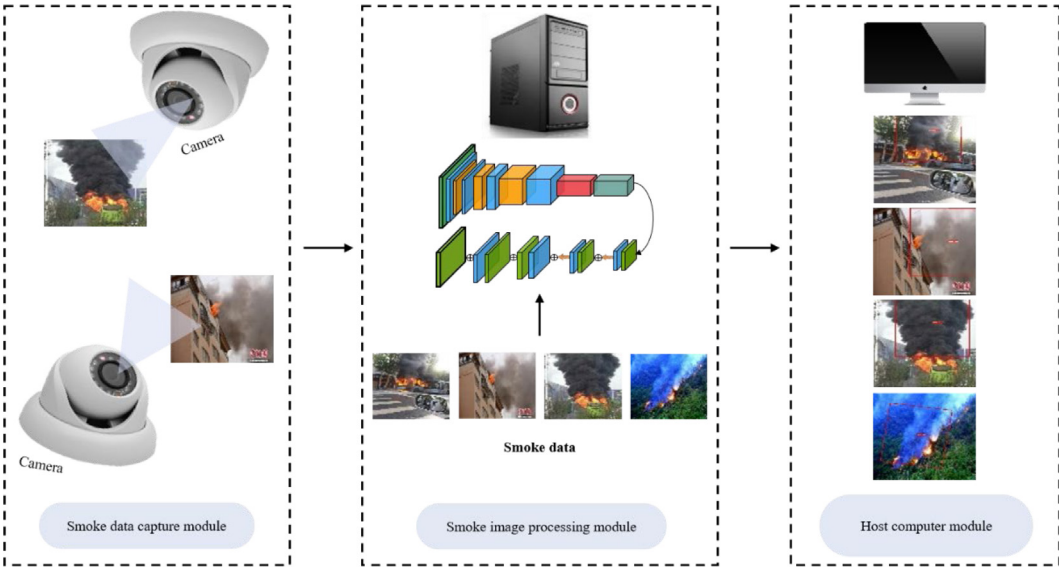


Fig. 19. Schematic diagram of the IoT wildfire smoke detection system that is based on MVMNet.

Fig. 20. Recognition results of MVMNet and YOLOv5 for three types of smoke. Class A refers to cases with a medium smoke density and size, Class B refers to cases with a low smoke density and poor features, and Class C refers to cases in which the smoke target is small and difficult to find.

by the model, namely, confusion of smoke with clouds and other smoke-like objects, tilting of smoke, and long distance between the burning point and the camera.

Many smoke video datasets are already in public use. However, previous studies in the field of forest fire smoke detection have not yet made image datasets publicly available. As a result, the forest fire multioriented detection dataset that we offer can help support the use of image-based deep learning methods in ecological environment preservation.

Analysing the incorrectly detected samples provides useful insights for improving the performance of our proposed network. The MVMNet model is designed as a perspective vector where θ is divided into 180 classes, namely, $\theta \in [0°, 179°]$. Since the model is designed as a classification task, we can avoid the boundary problem that is posed in the regression task. However, in our dataset, the number of angles that correspond to distinct classes varies widely, thereby giving rise to an interclass imbalance problem. As shown in Fig. 22, the interclass imbalance problem can lead to the dominance of the more numerous categories in the backpropagation, with the detector tending to predict the more numerous categories. As a result, the predicted angles will differ from the true angles.

To address such difficulties, we will consider mitigating this problem in future work by determining angular information using a circular smooth label. At the same time, the number of images in the dataset should be increased as much as possible to ensure that it is further improved upon and more balanced.

6. Conclusions and outlook

Deep learning-based smoke detection algorithms have become increasingly popular in the detection of forest fires in recent years. We proposed an MVMNet for forest fire smoke detection to improve the accuracy and effectiveness of forest fire smoke target detection. First, we proposed multioriented detection to fit the direction of the smoke, followed by VAM to improve the capacity to extract smoke features. To preserve more information, we replaced MaxPool in SPP with SoftPool. Finally, in the postprocessing stage of YOLOv5 we replaced the original NMS with Mixed-NMS, thereby solving the problems of misdetection and missed detection that commonly occur in forest fire smoke detection and improving the target detection accuracy. More importantly, MVMNet improves the accuracy and effectiveness of the YOLO model in the detection of forest fire smoke, and it is simple to train and utilize. This opens up new avenues for using deep learning in forest fire prevention and management. In our experiment, we used a 7:2:1 ratio to divide the 15909 image data from the forest fire multioriented detection dataset into a training set, a validation set, and a test set. MVMNet attained a mAP^{50} of 88.05% and an FPS of 122, thereby outperforming the compared methods and demonstrating the effectiveness of the model that we developed.

The experimental results showed that the proposed MVMNet object detection method could effectively identify smoke in the smoke dataset, but in practice, there are still gaps in its application to forest fire prevention and control.

In the future, as many images in complex environments as possible should be collected to discover more correlations between the morphological and detailed features of smoke and its environment and to combine semantic information in the environment with smoke feature information to improve the accuracy of forest fire smoke recognition. Furthermore, it would be useful to deploy drones to obtain forest images and create drone models. The methods that were proposed in this paper can be applied to forest fire prevention and control to promote rapid and effective forest fire detection and an organized response to fires.

a. Low concentration smoke b. Remote smoke c. Small target smoke

Fig. 21. Real applications.

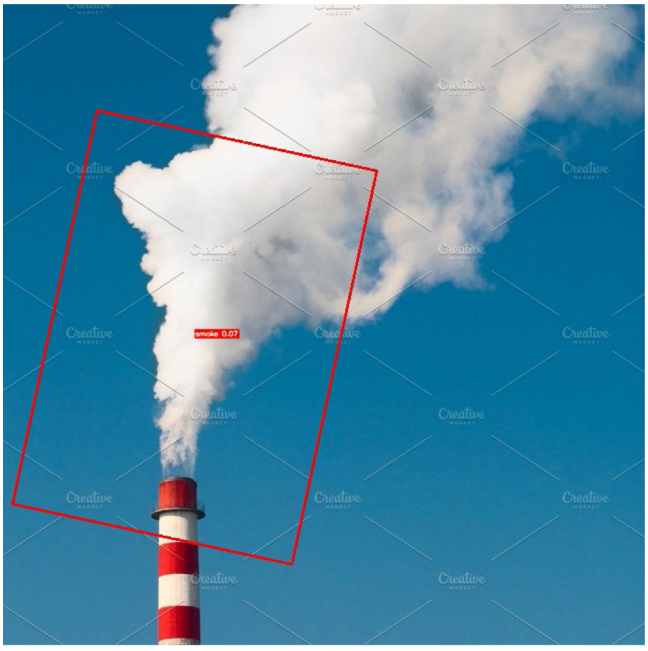


Fig. 22. Angular deviation due to interclass imbalance.

Funding

This work was supported by National Natural Science Foundation in China (Grant No. 61703441); in part by the Key Project of Education Department of Hunan Province (Grant No. 21A0179); in part by the Changsha Municipal Natural Science Foundation (Grant No. kq2014160); in part by Hunan Key Laboratory of Intelligent Logistics Technology (2019TP1015).

CRediT authorship contribution statement

Yaowen Hu: Methodology, Writing – original draft, Conceptualization. **Jialei Zhan:** Software, Data acquisition, Investigation. **Guoxiong Zhou:** Validation, Project administration. **Aibin Chen:** Supervision, Funding acquisition. **Weiwei Cai:** Model guidance. **Kun Guo:** Data curation. **Yahui Hu:** Formal analysis, Resources. **Liujun Li:** Visualization, Writing – review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of supporting data

Some of the datasets that were used and/or analysed in this study have been uploaded to the website https://github.com/guokun666/Forest_Fire_Smoke_DATA.git. In addition, we have attached the format conversion code for xml-YOLO. All the homemade datasets in this study (15909 sheets in total) can be obtained by contacting the corresponding author.

Acknowledgements

We are grateful to all members of the Forestry Information Research Centre for their advice and assistance in the course of this research. The language of our manuscript have been refined and polished by Elsevier Language Editing Services (Serial number: **LE-229528-CD224E4C9C79**)

References

[1] M.E. Lucas-Borja, J. González-Romero, P.A. Plaza-Álvarez, J. Sagra, M.E. Gómez, D. Moya, J. Heras, et al., The impact of straw mulching and salvage logging on post-fire runoff and soil erosion generation under mediterranean climate conditions, Sci. Total Environ. 654 (2019) 441–451, http://dx.doi.org/10.1016/j.scitotenv.2018.11.161.

[2] M.E. Lucas-Borja, J. Hedo, A. Cerdá, D. Candel-Pérez, B. Viñegla, Unravelling the importance of forest age stand and forest structure driving microbiological soil properties, enzymatic activities and soil nutrients content in mediterranean spanish black pine (pinus nigra ar. ssp. salzmannii) forest, Sci. Total Environ. 562 (2016) 145–154, http://dx.doi.org/10.1016/j.scitotenv.2016.03.160.

[3] Y. Pan, R.A. Birdsey, O.L. Phillips, R.B. Jackson, The structure, distribution, and biomass of the world's forests, Annu. Rev. Ecol. Evol. Syst. 44 (2013) 593–622, http://dx.doi.org/10.1146/annurev-ecolsys-110512-135914.

[4] J.R. Martinez-de Dios, B.C. Arrue, A. Ollero, L. Merino, F. Gómez-Rodríguez, Computer vision techniques for forest fire perception, Image Vis. Comput. 26 (4) (2008) 550–562, http://dx.doi.org/10.1016/j.imavis.2007.07.002.

[5] M.E. Harrison, S.E. Page, S.H. Limin, The global impact of Indonesian forest fires, Biologist 56 (3) (2009) 156–163, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.709.6831&rep=rep1&type=pdf.

[6] M. Ertugrul, T. Varol, H.B. Ozel, M. Cetin, H. Sevik, Influence of climatic factor of changes in forest fire danger and fire season length in Turkey, Environ. Monit. Assess. 193 (1) (2021) 1–17, http://dx.doi.org/10.1007/s10661-020-08800-6.

[7] M.E. Lucas-Borja, D.A. Zema, .B.G. Carrà, A. Cerdà, P.A. Plaza-Alvarez, J.S. Cózar, J. Heras, et al., Short-term changes in infiltration between straw mulched and non-mulched soils after wildfire in mediterranean forest ecosystems, Ecol. Eng. 122 (2018) 27–31, http://dx.doi.org/10.1016/j.ecoleng.2018.07.018.

[8] L. Tian, Y. Cao, B. He, Y. Zhang, C. He, D. Li, Image enhancement driven by object characteristics and dense feature reuse network for ship target detection in remote sensing imagery, Remote Sens. 13 (7) (2021) 1327, http://dx.doi.org/10.3390/rs13071327.

[9] W. Huang, G. Li, Q. Chen, M. Ju, J. Qu, CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection, Remote Sens. 13 (5) (2021) 847, http://dx.doi.org/10.3390/rs13050847.

[10] L. He, X. Gong, S. Zhang, L. Wang, F. Li, Efficient attention based deep fusion CNN for smoke detection in fog environment, Neurocomputing 434 (2021) 224–238, http://dx.doi.org/10.1016/j.neucom.2021.01.024.

[11] D. Sheng, J. Deng, J. Xiang, Automatic smoke detection based on SLIC-DBSCAN enhanced convolutional neural network, IEEE Access 9 (2021) 63933–63942, http://dx.doi.org/10.1109/ACCESS.2021.3075731.

[12] E. Zhao, Y. Liu, J. Zhang, Y. Tian, Forest fire smoke recognition based on anchor box adaptive generation method, Electronics 10 (5) (2021) 566, http://dx.doi.org/10.3390/electronics10050566.

[13] H. Kaufmann, K. Segl, S. Chabrillat, S. Hofer, T. Stuffler, A. Mueller, H. Bach, et al., EnMAP a hyperspectral sensor for environmental mapping and analysis, in: In 2006 IEEE International Symposium on Geoscience and Remote Sensing, IEEE, 2006, pp. 1617–1619, http://dx.doi.org/10.1109/IGARSS.2006.417.

[14] J. Cheon, J. Lee, I. Lee, Y. Chae, Y. Yoo, G. Han, A single-chip CMOS smoke and temperature sensor for an intelligent fire detector, IEEE Sens. J. 9 (8) (2009) 914–921, http://dx.doi.org/10.1109/JSEN.2009.2024703.

[15] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: A survey, Knowl.-Based Syst. 80 (2015) 14–23, http://dx.doi.org/10.1016/j.knosys.2015.01.010.

[16] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under concept drift: A review, IEEE Trans. Knowl. Data Eng. 31 (12) (2018) 2346–2363, http://dx.doi.org/10.1109/TKDE.2018.2876857.

[17] D. Krstinić, D. Stipaničev, T. Jakovčević, Histogram-based smoke segmentation in forest fire detection system, Inf. Technol. Control 38 (3) (2009) https://www.itc.ktu.lt/index.php/ITC/article/view/12105.

[18] J. Gubbi, S. Marusic, M. Palaniswami, Smoke detection in video using wavelets and support vector machines, Fire Saf. J. 44 (8) (2009) 1110–1115, http://dx.doi.org/10.1016/j.firesaf.2009.08.003.

[19] S. Surit, W. Chatwiriya, Forest fire smoke detection in video based on digital image processing approach with static and dynamic characteristic analysis, in: 2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering, 2011, pp. 35–39, http://dx.doi.org/10.1109/CNSI.2011.47.

[20] T. Chen, Y. Yin, S. Huang, et al., The smoke detection for early fire-alarming system base on video processing[C].information hiding, 2006, http://dx.doi.org/10.1109/IIH-MSP.2006.265033.

[21] C.Y. Yu, Z. Yongming, F. Jun, et al., Texture analysis of smoke for real-time fire detecti on[C], in: International Workshop on Computer Science & Engineering, IEEE, 2010, http://dx.doi.org/10.1109/WCSE.2009.864.

[22] F. Yuan, X. Xia, J. Shi, Holistic learning-based high-order feature descriptor for smoke recognition, Int. J. Wavelets Multiresolut. Inf. Process. 17 (02) (2019) 1940005, http://dx.doi.org/10.1142/S0219691319400058.

[23] N. Fujiwara, K. Terada, Extraction of a smoke region using fractal coding, in: IEEE International Symposium on Communications and Information Technology, Vol. 2, ISCIT 2004, IEEE, 2004, pp. 659–662, http://dx.doi.org/10.1109/ISCIT.2004.1413797.

[24] J. Gubbi, S. Marusic, M. Palaniswami, Smoke detection in video using wavelets and support vector machines, Fire Saf. J. 44 (8) (2009) 1110–1115.

[25] S. Verstockt, S.Van. Hoecke, T. Beji, B. Merci, B. Gouverneur, A.E. Cetin, R. Walle, et al., A multi-modal video analysis approach for car park fire detection, Fire Saf. J. 57 (2013) 44–57, http://dx.doi.org/10.1016/j.firesaf.2012.07.005.

[26] L. Tian, J. Wang, H. Zhou, J. Wang, Automatic detection of forest fire disturbance based on dynamic modelling from MODIS time-series observations, Int. J. Remote Sens. 39 (12) (2018) 3801–3815, http://dx.doi.org/10.1080/01431161.2018.1437294.

[27] A. Gaur, A. Singh, A. Kumar, A. Kumar, K. Kapoor, Video flame and smoke based fire detection algorithms: A literature review, Fire technology 56 (5) (2020) 1943–1980.

[28] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 58, 2014, pp. 0–587, http://dx.doi.org/10.1109/CVPR.2014.81.

[29] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 77, 2016, pp. 9–788.

[30] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 726, 2017, pp. 3–7271.

[31] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[32] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[33] A. Bochkovskiy, C.Y. Wang, H.Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection, 2020, arXiv preprint arXiv:2004.10934.

[34] Ultralytics, Available at https://github.com/ultralytics/yolov5.

[35] X. Qiang, G. Zhou, A. Chen, X. Zhang, W. Zhang, Forest fire smoke detection under complex backgrounds using trpca and tsvb, Int. J. Wildland Fire http://dx.doi.org/10.1071/WF20086.

[36] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, K. Fu, et al., Scrdet: Towards more robust detection for small, cluttered and rotated objects, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8232–8241, http://dx.doi.org/10.1109/ICCV.2019.00832.

[37] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.

[38] A. Stergiou, R. Poppe, G. Kalliatakis, Refining activation downsampling with softpool, 2021, arXiv preprint arXiv:2101.00440.

[39] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Trans. Multimed. 20 (11) (2018) 3111–3122, http://dx.doi.org/10.1109/TMM.2018.2818020.

[40] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, 2021, arXiv preprint arXiv:2107.08430, https://arxiv.org/abs/2107.08430.

[41] A. Jadon, M. Omama, A. Varshney, M.S. Ansari, R. Sharma, FireNet: a specialized lightweight fire & smoke detection model for real-time IoT applications, 2019, arXiv preprint arXiv:1905.11922, https://arxiv.org/pdf/1905.11922.pdf.

[42] K. Gu, Z. Xia, J. Qiao, W. Lin, Deep dual-channel neural network for image-based smoke detection, IEEE Trans. Multimed. 22 (2) (2019) 311–323, http://dx.doi.org/10.1109/TMM.2019.2929009.

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

[44] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, K. Fu, et al., Scrdet: Towards more robust detection for small, cluttered and rotated objects, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8232–8241, https://openaccess.thecvf.com/content_ICCV_2019/papers/Yang_SCRDet_Towards_More_Robust_Detection_for_Small_Cluttered_and_Rotated_ICCV_2019_paper.pdf.

[45] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, Z. Luo, et al., R2cnn: rotational region cnn for orientation robust scene text detection, 2017, arXiv preprint arXiv:1706.09579, https://arxiv.org/ftp/arxiv/papers/1706/1706.09579.pdf.

[46] R. Bao, K. Palaniappan, Y. Zhao, G. Seetharaman, W. Zeng, GLSNet: Global and local streams network for 3D point cloud classification, in: 2019 IEEE Applied Imagery Pattern Recognition Workshop, AIPR, IEEE, 2019, pp. 1–9, http://dx.doi.org/10.1109/AIPR47015.2019.9174587.

[47] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, J. Yang, Feature attentioned object detection in remote sensing imagery, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 3886–3890, http://dx.doi.org/10.1109/ICIP.2019.8803521.

[48] Y. Wang, Y. Zhang, Y. Zhang, L. Zhao, X. Sun, Z. Guo, Sard:Towards scale-aware rotated object detection in aerial imagery, IEEE Access 7 (2019) 173855–173865, http://dx.doi.org/10.1109/ACCESS.2019.2956569.

[49] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, X. Sun, Rotation aware and multi-scale convolutional neural network for object detection in remote sensing images, ISPRS J. Photogramm. Remote Sens. 161 (2020) 294–308, http://dx.doi.org/10.1016/j.isprsjprs.2020.01.025.

[50] Y. Zhu, X. Wu, J. Du, Adaptive period embedding for representing oriented objects in aerial images, 2019, arXiv preprint arXiv:1906.09447, http://dx.doi.org/10.1109/TGRS.2020.2981203.

[51] J. Wei, S. Wang, Q. Huang, F^3net: Fusion, feedback and focus for salient object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, 2020, pp. 12321–12328, http://dx.doi.org/10.1609/aaai.v34i07.6916.

[52] X. Yang, J. Yan, Arbitrary-oriented object detection with circular smooth label, 2020, arXiv preprint arXiv:2003.05597, https://arxiv.org/pdf/2003.05597.pdf.

[53] R. Qin, Q. Liu, G. Gao, D. Huang, Y. Wang, MRDet: A multi-head network for accurate oriented object detection in aerial images, 2020, arXiv preprint arXiv:2012.13135, https://arxiv.org/pdf/2012.13135.pdf.

[54] X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, T. He, Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing, 2020, arXiv preprint arXiv:2004.13316, https://arxiv.org/pdf/2004.13316.pdf.

[55] Kun Fu, Zhonghan Chang, Yue Zhang, Xian Sun, Pointbased estimator for arbitrary-oriented object detection in aerial images, IEEE Trans. Geosci. Remote Sens. (2020) http://dx.doi.org/10.1109/TGRS.2020.3020165.

[56] Pengming Feng, Youtian Lin, Jian Guan, Guangjun He, Huifeng Shi, Jonathon Chambers, Toso: Student'st distribution aided one-stage orientation target detection in remote sensing images, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 4057–4061, https://arxiv.org/pdf/2101.11952.pdf.

[57] Zhifeng Xiao, Kai Wang, Qiao Wan, Xiaowei Tan, Chuan Xu, Fanfan Xia, A2s-det: Efficiency anchor matching in aerial image oriented object detection, Remote Sens. 13 (1) (2021) 73, http://dx.doi.org/10.3390/rs13010073.

[58] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, Changsheng Xu, Dynamic refinement network for oriented and densely packed object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11207–11216, https://arxiv.org/pdf/2101.11952.pdf.

[59] Peng Sun, Yongbin Zheng, Zongtan Zhou, Wanying Xu, Qiang Ren, R4det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images, Image Vis. Comput. 103 (2020) 104036, http://dx.doi.org/10.1016/j.imavis.2020.104036.

[60] Xue Yang, Junchi Yan, Ziming Feng, Tao He, R3det: Refined single-stage detector with feature refinement for rotating object, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, https://www.aaai.org/AAAI21Papers/AAAI-364.YangX.pdf.

[61] Pengbo Zhao, Zhenshen Qu, Yingjia Bu, Wenming Tan, Ye Ren, Shiliang Pu, Polardet: A fast, more precise detector for rotated target in aerial images, 2020, arXiv preprint arXiv:2010.08720, http://dx.doi.org/10.1080/01431161.2021.1931535.

[62] Bo Zhong, Kai Ao, Single-stage rotation-decoupled detector for oriented object, Remote Sens. 12 (19) (2020) 3262, http://dx.doi.org/10.3390/rs12193262.

[63] Jiaming Han, Jian Ding, Jie Li, Gui-Song Xia, Align deep features for oriented object detection, 2020, arXiv preprint arXiv:2008.09397, http://dx.doi.org/10.1109/TGRS.2021.3062048.

[64] Djalil Chafaï, Wasserstein distance between two gaussians. Website, 2010, https://djalil.chafai.net/blog/2010/04/30/wassersteindistance-between-two-gaussians/.

[65] P. Purkait, C. Zhao, C. Zach, SPP-Net: Deep absolute pose regression with synthetic views, 2017, arXiv preprint arXiv:1712.03452.

[66] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448, http://dx.doi.org/10.1109/ICCV.2015.169.

[67] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[68] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013, arXiv preprint arXiv:1312.6229.

[69] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, Cham, 2016, pp. 21–37.

[70] Ning Zhang, Qin Chen, Ensemble learning training method based on AUC and Q statistics, J. Comput. Appl. 39 (04) (2019) 935–939, http://dx.doi.org/10.11772/j.issn.1001-9081.2018102162.