symmetry

MDPI

Article

PDAM–STPNNet: A Small Target Detection Approach for Wildland Fire Smoke through Remote Sensing Images

Jialei Zhan [1,†], Yaowen Hu [1,†], Weiwei Cai [2], Guoxiong Zhou [1,*] and Liujun Li [3]

[1] College of Computer & Information Engineering, Central South University of Forestry and Technology, Changsha 410004, China; 20192778@csuft.edu.cn (J.Z.); 20192764@csuft.edu.cn (Y.H.)
[2] Graduate College, Northern Arizona University, Flagstaff, AZ 86011, USA; vivitsai@csuft.edu.cn
[3] Department of Civil, Architectural and Environmental Engineering, University of Missouri-Rolla, Rolla, MO 65401, USA; llpwc@umsystem.edu
* Correspondence: t20060599@csuft.edu.cn
† Jialei Zhan and Yaowen Hu contributed equally to this work.

Abstract: The target detection of smoke through remote sensing images obtained by means of unmanned aerial vehicles (UAVs) can be effective for monitoring early forest fires. However, smoke targets in UAV images are often small and difficult to detect accurately. In this paper, we use YOLOX-L as a baseline and propose a forest smoke detection network based on the parallel spatial domain attention mechanism and a small-scale transformer feature pyramid network (PDAM–STPNNet). First, to enhance the proportion of small forest fire smoke targets in the dataset, we use component stitching data enhancement to generate small forest fire smoke target images in a scaled collage. Then, to fully extract the texture features of smoke, we propose a parallel spatial domain attention mechanism (PDAM) to consider the local and global textures of smoke with symmetry. Finally, we propose a small-scale transformer feature pyramid network (STPN), which uses the transformer encoder to replace all CSP_2 blocks in turn on top of YOLOX-L's FPN, effectively improving the model's ability to extract small-target smoke. We validated the effectiveness of our model with recourse to a home-made dataset, the Wildfire Observers and Smoke Recognition Homepage, and the Bowfire dataset. The experiments show that our method has a better detection capability than previous methods.

Keywords: remote sensing images; forest fire; smoke detection; UAV forest fire monitoring system; component stitching data enhancement; parallel spatial domain attention mechanism; symmetry; small-scale transformer feature pyramid network

check for updates

Citation: Zhan, J.; Hu, Y.; Cai, W.; Zhou, G.; Li, L. PDAM–STPNNet: A Small Target Detection Approach for Wildland Fire Smoke through Remote Sensing Images. *Symmetry* **2021**, *13*, 2260. https://doi.org/10.3390/sym13122260

Academic Editors: Dejun Zhang, Whoi-Yul Kim and Moonsoo Ra

Received: 11 November 2021
Accepted: 24 November 2021
Published: 27 November 2021

1. Introduction

Forest fires can cause widespread forest mortality, bringing huge losses to forest ecological resources and the social economy, and serious forest fires can even lead to human casualties [1,2]. In recent years, the frequency of forest fires has increased, and the extent of the damage has been increasing year by year. Wildfires in Australia burned more than 1300 houses and approximately 6 million hectares of land in January 2020 [3]. As combustible material in forests is not usually dry, burning produces large amounts of fine solid particles that form smoke [4]. Early smoke areas are larger than flame areas, and fires can easily be covered by smoke, making monitoring smoke an effective means of conducting early forest fire monitoring [5]. If forest fires are not responded to in a timely manner, they cause greater damage and increase the cost of fire suppression [6]. If we can detect the distinctive visual feature of smoke in the early stages of a forest fire, we can control small fires that have not yet spread and reduce the damage they cause to a minimum.

To meet the demand for real-time performance and accuracy in forest fire monitoring tasks, researchers have conducted research on forest fire smoke monitoring using UAV images of smoke from burning forest fires [7]. Amiaz et al. [8] proposed a method to detect dynamic texture regions in image sequences based on the luminance produced by smoke in the images and achieved the segmentation of static and dynamic texture regions according to a level set scheme. However, texture regions are difficult to segment, and this motion estimation approach is based on an assumption of constant luminance, with which the actual scene is somewhat different, thus leading to missed and false alarms. Ugur Toreyin et al. [9] proposed a method using a spatial wavelet transform to estimate the background of the scene and monitor the reduction of the high frequency energy of the scene, based on the translucent characteristics of smoke. Furthermore, a Hidden Markov Model (HMM), which simulates the temporal behaviour of smoke, has been used to analyse the periodic behaviour in smoke boundaries [10]. The combination of the wavelet transform and the Hidden Markov Model gives good results; however, the smoke is sometimes dense and sometimes sparse and is easily confused with its background, so that it is difficult to determine accurately. Yu et al. [11] proposed the Lucas Kanade optical flow algorithm to calculate the optical flow in a candidate region and analyse the motion characteristics of the smoke based on the optical flow results, which can be used to distinguish the smoke from some other moving objects. However, it is computationally intensive and time-consuming and cannot meet the demanding real-time requirements of forest fire detection. Traditional image-based forest fire detection algorithms are often based on the RGB colour and transparency of the smoke, which is not so effective in complex forest environments.

With the evolution of the UAV industry and the rise of the field of deep learning in recent years, the combination of UAVs and deep learning has become the mainstream solution for forest fire smoke detection today [12,13]. In practical applications, fixed-wing UAVs and multi-rotor UAVs are used to work in tandem to achieve the rapid and accurate monitoring of detection areas [14]. By combining the flight characteristics of multi-rotor and fixed-wing UAVs, Kinaneva et al. used fixed-wing UAVs to cruise the Rusenski Lom National Park area at medium altitude (350–5000 m) to detect anomalies indicating suspected smoke, and fire areas were verified at low altitudes (10–350 m) using multi-rotor drones to confirm the presence of forest fires—a verification method that effectively reduced the false alarm rate [15]. Alexandrov et al. used a combination of machine learning and drones for the aerial monitoring of forest fire smoke [16]. In the aerial UAV photography of forest fire smoke, the smoke was far away from the UAV camera and the smoke target was small due to the height of the UAV. Compared to regular-sized targets, small targets often lack sufficient information about their appearance, making it difficult to distinguish them from their backgrounds or similar targets [17]. In real scenes that are intricate and complex, there are usually problems with dramatic changes in lighting, smoke obscured by trees, and dense, connected smoke patterns, and the effect of these factors on small target features, such as smoke in aerial imagery, is even more dramatic, further increasing the difficulty of detection. Small targets are characterised by a small pixel share, small coverage area, and little information, making it difficult to extract features with discriminatory power. They are highly susceptible to interference from environmental factors, which in turn makes it difficult for detection models to accurately locate and identify small targets [18].

In order to solve the above problems, some data enhancement methods have been proposed for small targets. Yu et al. proposed a scale-matching strategy for data processing, which crops according to different target sizes to reduce the gap between targets of different sizes, thus avoiding the situation that small target information is easily lost in conventional scaling operations [19]. However, this data enhancement strategy can lead to unclear smoke subjects and a loss of texture information. Kisantal et al. proposed a copy enhancement method to increase the number of training samples of small targets by copying and pasting small targets in the image several times; thus, the detection performance of small targets was improved [20]. However, the application of this method to forest fire smoke detection tasks can easily lead to unrealistic determinations of the locations at which fires start and can also result in image distortion. To date, the detection accuracy for small target objects in publicly available datasets is about half of that for regular-sized objects. There is a lack of proven solutions for the detection of small objects, including smoke.

At the same time, researchers have made many explorations into the application of deep learning models in target detection [21]. To improve the performance of deep learning models in the field of target detection, YOLO [22], proposed by Redmon et al., takes a whole image as the input of the network and regresses the position and class of BBox directly in the output layer. However, its localisation accuracy for small target objects is poor, as the effect of different target scales is not taken into account in the loss function. Subsequently, Redmon et al. proposed YOLOv2 [23], which preprocesses the size of the prior frame by the K-means algorithm and divides the prior frame into nine categories with different scales, which compensates for the shortcomings of YOLO to a certain extent. However, the model ignored the problem of image training for targets of different scales, rendering its recognition and localisation of small targets ineffective. Later, Redmon et al. proposed YOLOv3 [24], which used a feature pyramid network to improve detection performance and brought enlightenment to small target detection. Based on YOLOv3, Z Ge et al. proposed YOLOX [25], which uses a branch decoupling head and SimOTA and is anchor free to further improve the accuracy of the model. Compared with YOLOv3, YOLOX not only has a simpler structure but also exhibits good speed and accuracy in usual scenarios, which is advantageous in the context of small target detection. Nevertheless, forest fire smoke from UAV aerial photography has both complex texture features and small targets, and the direct use of YOLOX as a network for target detection addresses the complex nature of smoke in images. Therefore, in order to solve the problems of the low accuracy of small target smoke monitoring and the inadequate extraction of smoke features in actual forest fire smoke monitoring tasks, and to improve the detection effectiveness of the model for the small targets presented by forest fire smoke, this paper proposes a PDAM–STPNNet-based method for detecting forest fire smoke targets under UAV aerial photography, taking into account both speed and accuracy, and using YOLOX-L with a moderate model size as the benchmark network. Our contributions are summarised as follows.

(1) Component stitching data enhancement is used to generate images with smaller scale targets in a scaled collage. The collage generates images of the same size as the original images, ensuring that the model can effectively detect small targets of forest fire smoke without incurring additional overheads to the model.

(2) A parallel spatial domain attention mechanism is proposed, which contains a parallel local attention mechanism module and a global attention mechanism module as its sub-modules. The attention mechanism module explores the local deep texture features of smoke and the relationship between features, while the global attention mechanism module focuses on the global texture features of smoke, taking half of the number of channels of the feature map, respectively, and using concat fusion features to fully consider the smoke texture features and improve the results.

(3) The small-scale transformer feature pyramid network is proposed to capture rich global and contextual information, with the aim of improving the detection of small targets in forest fire smoke detection tasks and avoiding the misdetection of small target smoke as far as possible.

We designed PDAM–STPNNet to work on improving the model's feature extraction and feature fusion effect on smoke. The application of PDAM–STPNNet to the forest fire smoke detection task aims to improve smoke detection accuracy and reduce the error rate of detection, which is important for the timely monitoring of forest fires. The diagram of the working principle is shown in Figure 1, where the UAV forest fire monitoring system captures the input images and performs the target detection of smoke in the images. First, component stitching data enhancement is used to increase the number of small target samples. Next, basic features are extracted using traditional backbone extraction features. Then, the parallel spatial domain attention mechanism is used to combine local texture and global texture features. Finally, a small-scale transformer feature pyramid network is used to enhance the fusion effect of small target features. After these steps, the UAV forest fire monitoring system obtains the final detection results with the help of component stitching data enhancement and PDAM–STPNNet.

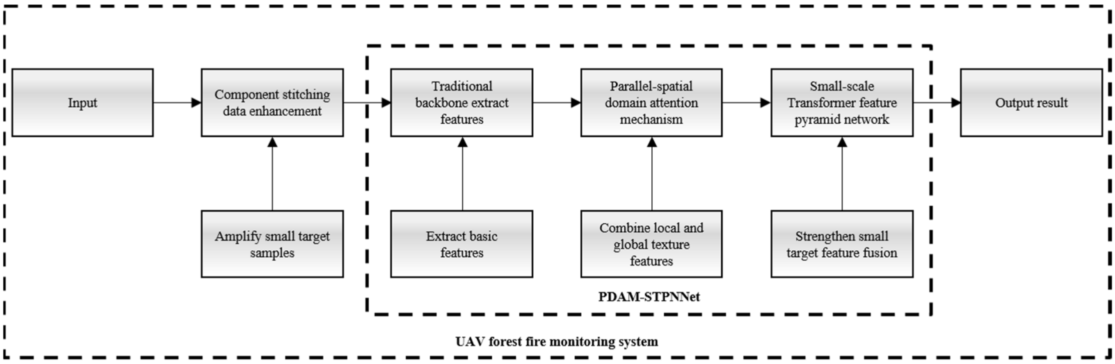

Input
Component stitching data enhancement
Traditional backbone extract features
Parallel-spatial domain attention mechanism
Small-scale Transformer feature pyramid network
Output result
Amplify small target samples
Extract basic features
Combine local and global texture features
Strengthen small target feature fusion
PDAM-STPNNet
UAV forest fire monitoring system

Figure 1. The schematic diagram of PDAM–STPNNet.

2. Materials and Methods

2.1. Study Area

We conducted our study mainly in Huangfengqiao State Forestry (E:113.34569, N:27.00023) and Qingyang Lake State Forestry (E:112.125191, N:28.149261) (see Figure 2). (1) Huangfengqiao Forestry is located in the east and west of You County, Hunan Province, and is dominated by low and medium mountainous landscapes, with a maximum elevation of 1270 m and a minimum elevation of 115 m. The forest type is mainly fir plantation. (2) Qingyang Lake State Forestry Field is located in the remnants of the Xuefeng Mountains, the main mountain system of Ningxiang, in a zone of transition from low and medium mountains to hills. The forest type is mainly subtropical deciduous broad-leaved forest. The woodland is interspersed with residential areas, with obvious traces of human activity. The different background characteristics of the two woodlands provide a good test of the effectiveness of our method in different complex environments. Moreover, the two woodlands are located in a subtropical monsoon climate with a wide variety of trees and are prone to fires in the dry summer and autumn, making them valuable for field testing.

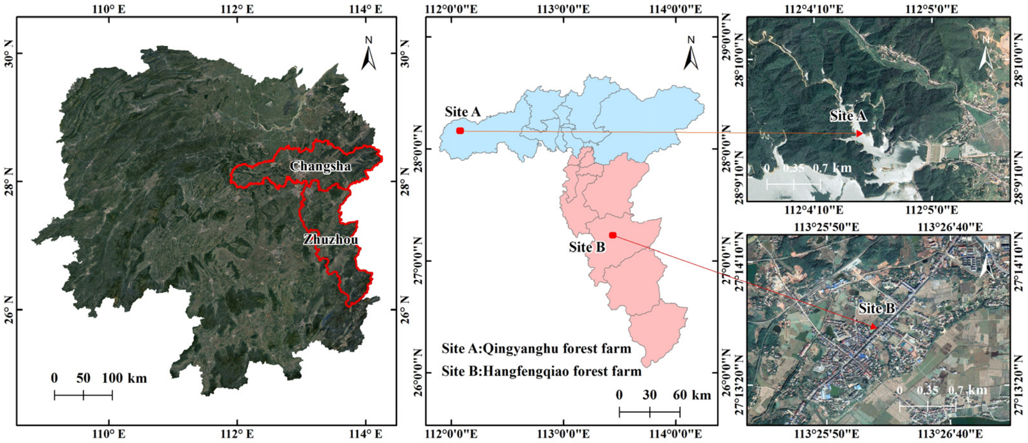


Figure 2. Study sites in Huangfengqiao Forestry, Zhuzhou, Hunan and Qingyanghu Forestry, Changsha, Hunan.

2.2. Component Stitching Data Enhancement

The definition of image enhancement is very broad. Usually, image enhancement is used to purposefully emphasise the overall or local features of an image and improve the clarity of the image. When drones monitor forest fires, they are often far away from the fire source for reasons of flight safety and the monitoring range of the drone [26]. At this time, the smoke captured by the UAV's camera usually corresponds to small targets. In order to enhance the detection of small targets of forest fire smoke, we use component stitching data enhancement to emphasise the differences between the features of different targets in the images and to balance the proportion of targets of different sizes in the dataset during model training [27].

In order to ensure that the images of objects such as smoke and trees in a forest environment are not distorted, the aspect ratio needs to be maintained when stitching the images. This is achieved by reducing and stitching together k regular images arranged in the same number of rows and columns to form a stitched image, where k is the number of squared rows/columns; for example, 1, 2^2, 3^2, etc. The spatial resolution of the original individual images is $(h{,}w)$, and the spatial resolution of each component image after the composite image is $(\frac{h}{\sqrt{k}}, \frac{w}{\sqrt{k}})$.

$$row = col \tag{1}$$

$$k = row * col \tag{2}$$

where row represents the number of rows in the collage,col represents the number of columns in the collage, and k represents the total number of components. Experiments show that a collage of two rows and two columns gives the best result—i.e., $k = 4$—as shown in Figure 3.

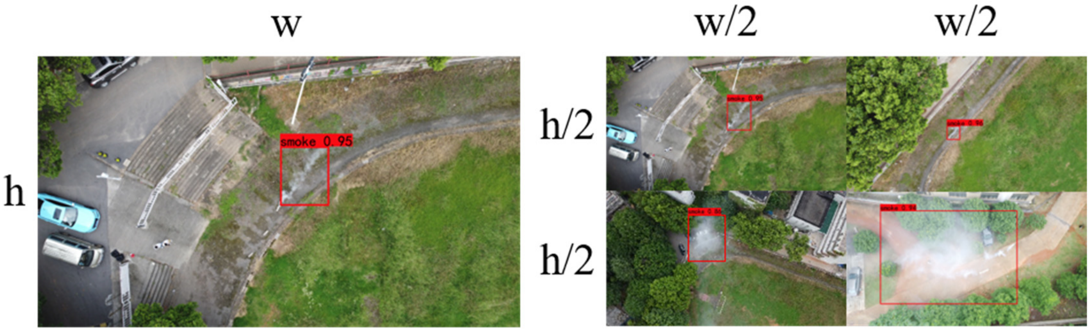


Figure 3. Image collage method.

It can be seen that the image is scaled to half of its original length and width and then collaged, with the collaged image maintaining its original size. It can be seen that the collage increases the proportion of small targets in the dataset by creating targets with smaller scales. As the composite image remains the same size as the regular image, there is no additional overhead involved in the forward propagation of the network model.

An image collage is not an infinite augmentation of the images in the dataset. To determine exactly how many images need to be collaged, a feedback paradigm is set. During the training process of PDAM–STPNNet, the proportion of loss caused by small targets can be calculated after each forward propagation. If the proportion of loss caused by small targets in the current iteration is less than a threshold, the collaged images are used in the next iteration; otherwise, no image collage is performed—i.e.,

$$I^{t+1} = \begin{cases} I^c,\ if\ r_s^t \leq \tau, \\ I,\ otherwise \end{cases} \tag{3}$$

where I^{t+1} denotes the next iteration, I^c denotes the use of a collage, I denotes the use of the original image, τ denotes the set threshold, and r_s^t denotes the percentage of loss caused by the small target in the current iteration.

2.3. PDAM–STPNNet

Given the operational characteristics of UAV forest patrols, where the UAV carries image-capture equipment to collect images of the forest in real time while flying at high altitude, the large area covered by the field of view requires models that can accurately identify forest fire features and locate them accurately. Forest fires are usually shadowy processes, with smoke being produced when a fire occurs. The timely and accurate detection of smoke is difficult due to the small size of the fire, the lightness of the smoke and the presence of trees, shifting branches, exposed grey and white rocks, and other smoke-like objects.

In order to efficiently and accurately identify smoke in different stages and states from UAV aerial images and to reduce and minimize forest fire damage, the objectives of this paper for smoke detection are to solve the difficulty of detecting the small targets presented by forest fire smoke by conventional methods and to identify the location of a fire's starting point through the positioning information obtained by UAVs. Therefore, this paper improves the model architecture based on YOLOX-L and proposes the method of PDAM–STPNNet for target detection based on UAV aerial photography of forest fire smoke. Our model structure is shown in Figure 4.

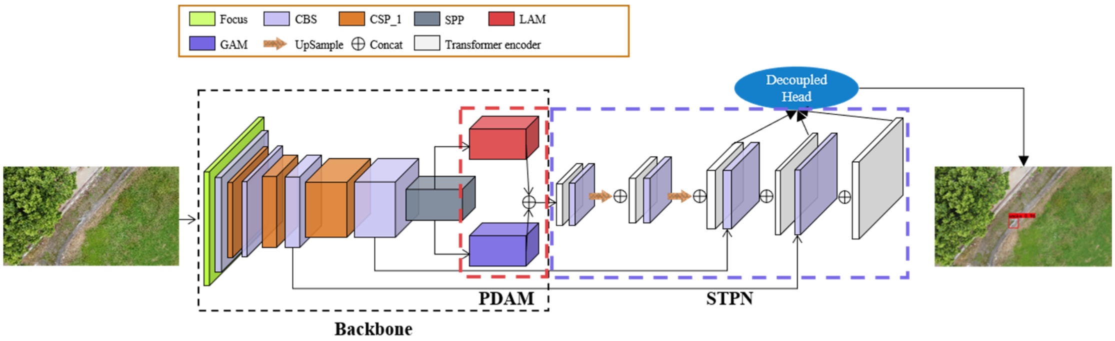


Figure 4. The schematic diagram of PDAM–STPNNet.

The convolution blocks in the dashed black box all belong to the backbone, after which comes the FPN structure. In contrast to YOLOX-L, we added the PDAM (enclosed by the red dashed box at the end of the backbone) in order to take full account of local and global texture features of smoke. The features extracted from the backbone are later fused using STPN (enclosed by the purple dashed box), which uses a transformer encoder instead of the original CSP_2 to enhance the detection of small targets of forest fire smoke. For more information, see the following three subsections.

2.3.1. Parallel Spatial Domain Attention Mechanism (PDAM)

A. Local attention mechanism module (LAM)

The local attention mechanism module is based on deep local texture features and the relationship between the features. The texture features express the spatial distribution and combination of the target, and the full extraction of the local texture features of the object improves the stability of the model and distinguishes it from nearby backgrounds that can easily be confused with smoke. The extraction effect of fully enhanced texture features is intended to be specific to the local texture features of the smoke.

As shown in Figure 5, the LAM assigns horizontal weight coefficients to each row of features through the attention mechanism and vertical weight coefficients to each column of features through the vertical attention mechanism. Each row feature obtained by dimensionality reduction is symmetrical with each column feature, so the transmission of the data stream in the LAM is symmetrical, which is more conducive to extracting complete and effective features.

$$c_i = \sum_{j=1}^{n} \frac{\exp(e_{i,j})}{\sum_{k=1}^{n} \exp(e_{ik})} h_j \tag{4}$$

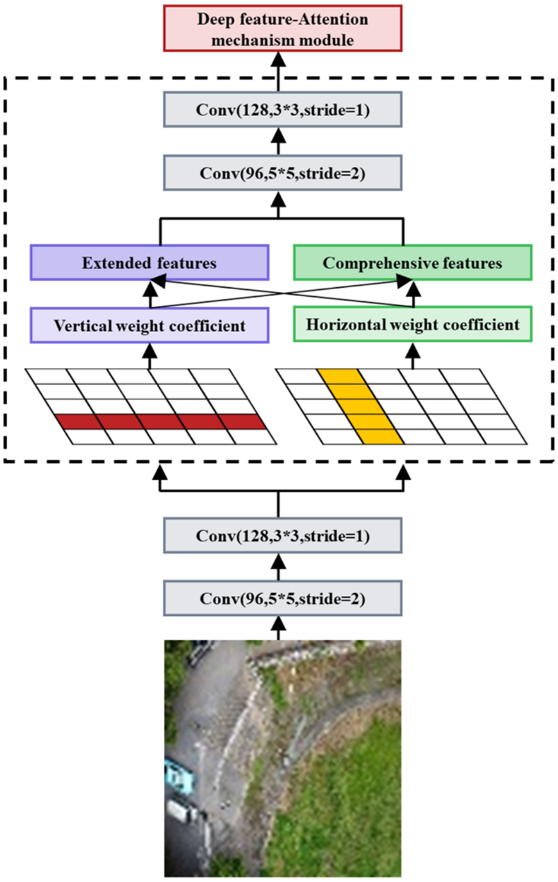


Figure 5. LAM algorithm flow chart.

We process the weighted features on rows and columns as follows, mining the deep feature information and expanding the weight coefficients by multiplication with a minimum term penalty to obtain extended features.

$$EF = c_I * c_{II} - \min(c_I, c_{II}) \tag{5}$$

Here, the local maximum feature is taken as the significant feature and summed with α multiples of the minimum value characteristic, where α is a decimal between 0 and 1. Using this method, the maximum value is used as the main factor, and another feature is taken into account to obtain comprehensive features.

$$CF = \max(c_I, c_{II}) + \alpha * \min(c_I, c_{II}) \tag{6}$$

LAM integrates the processed feature information by means of concatenation in the following steps.

$$LAM = concatenate([c_I, c_{II}, EF, CF]) \tag{7}$$

where e_{ij} is the weight coefficient of the LAM, i is the temporal feature, j is the sequence feature, h_j is the hidden layer information of the sequence feature j, $(c_I = \{c_1, c_2 \ldots c_{i-1}, c_i\})$ is the sequence of features in the column dimension, and $(c_{II} = \{c_1, c_2 \ldots c_{i-1}, c_i\})$ is the sequence of features in the row dimension. EF stands for deep feature information, max for maximum operation, min for minimum operation and CF for combined feature information. The LAM weight assignment procedure is shown in Figure 6.

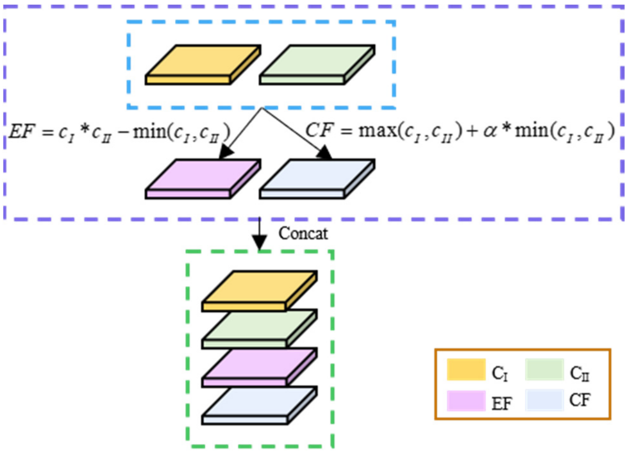

$EF = c_I * c_{II} - \min(c_I, c_{II})$
$CF = \max(c_I, c_{II}) + \alpha * \min(c_I, c_{II})$
Concat
C_I
C_{II}
EF
CF

Figure 6. LAM weighting steps.

B. Global attention mechanism module (GAM)

The global attention mechanism module focuses on spatial features from a global perspective, fully considering the contextual relationship between forest fire smoke and the forest background. This is designed to comprehensively capture information with differentiation, exclude the interference of redundant information, occlusion, and blurring in the image, improve the adaptability of the model and effectively integrate more comprehensive features of forest fire smoke and preserve similar features of different smoke targets during training. The structure of the GAM is shown in Figure 7.

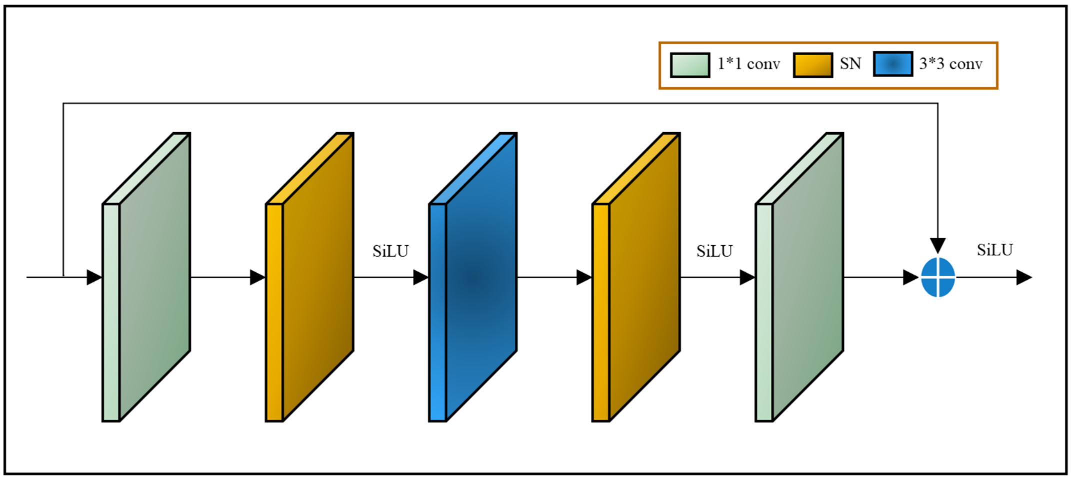


Figure 7. Global attention mechanism module structure composition.

(1) In general, larger convolution kernels are better at perceiving large target objects, while smaller-sized convolution kernels are better at extracting features from small targets. However, the diffusion direction and concentration of different kinds of smoke vary significantly; some backgrounds are complex, and targets are not easy to find. Therefore, we increase the branch of convolutional kernels of different sizes and use convolutional kernels of sizes 3 * 3, 5 * 5, and 7 * 7 to improve recognition accuracy.

(2) The GAM structure divides the feature map obtained after 1 * 1 convolution into four scales equally, where 3 * 3 convolution uses depth-separable convolution to reduce the number of parameters and computational effort.

(3) Remote sensing images of forest fire smoke have diverse scenes, and to adapt to the complex and variable forest background we use the integrated normalization method of switchable normalization (SN) instead of the traditional batch normalization (BN) layer.

The statistics of SN are used to calculate the statistics of BN, LN, and IN, and then six weighting parameters (corresponding to the mean and variance, respectively) are introduced to calculate the weighted mean and weighted variance as the mean and variance of the SN [28]. Normalisation is performed using the softmax activation function:

$$\hat{h}_{ncij} = \gamma \frac{h_{ncij} - \sum k \in \Omega w_k \mu_k}{\sqrt{\sum k \in \Omega w'_k \sigma_k^2 + \varepsilon}} + \beta \tag{8}$$

The input data of an implicit convolutional layer of PDAM–STPNNet can be represented as a feature map with four dimensions (N, C, H, W). The four dimensions represent the minibatch size, the number of channels and the height and width of the channels, respectively. h_{ncij} denotes a pixel; $\hat{h}_{ncij}$ denotes the h_{ncij} normalized result of the corresponding pixel; w_k, w'_k indicate weighting factors; μ_k is the mean; and σ_k^2 is the variance. The model learns the scaling factor γ, the offset factor β and ε.

$$w_k = \frac{e^{\lambda_k}}{\sum_{z \in \{in, ln, bn\}} e^{\lambda_z}} \tag{9}$$

where λ_k denotes the control parameters corresponding to the three-dimensional statistics. The control parameters are all initialised to 1 and optimised for learning during backpropagation. The control parameters λ_k are normalised using the softmax function and the weight coefficients w_k are calculated.

(4) The SiLU activation function is used instead of the commonly used ReLU activation function or sigmoid activation function to improve the learning convergence of the model. The SiLU activation function is calculated as follows:

$$f(x) = x * \sigma(x) \tag{10}$$

The structure allows the neural network model to focus more on the global texture features of smoke, such as granular smoke, with particles of a similar colour and size, and its spatial distribution, and also to distinguish backgrounds that are similar to smoke features, improving the accuracy of the extraction of detailed smoke features. Figure 8 shows a detailed design schematic of the GAM structure.

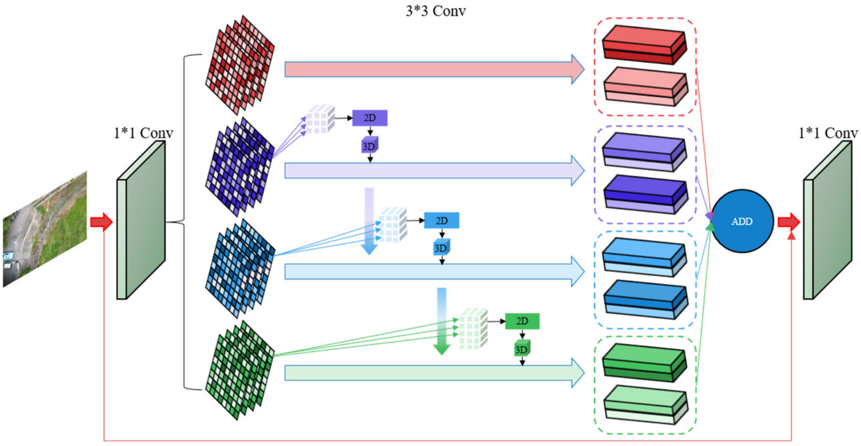


Figure 8. Detailed design structure of the global attention mechanism module.

The multi-scale convolutional structure, while expanding the field of perception, also introduces the problem of increasing the number of computational parameters. At the same time, the GAM uses a multiscale structure, as well as multiple 1 * 1 and 3 * 3 small-sized convolutional kernel structures and a deeper network model. Therefore, we use depth-wise separable convolution to build the GAM, which has the advantage of allowing complex multi-scale structures to operate efficiently. Depthwise separable convolution is based on the idea of splitting the traditional convolution operation into two steps: first, depthwise convolution—i.e., one-to-one two-dimensional convolution for each channel, where the input feature map is used to reduce the parameter computation—then, after the traditional convolution (3D convolution) operation, a 1 * 1 sized convolution kernel is used to combine the features of each channel, also known as point-wise convolution. The structure of depthwise separable convolution is shown in Figure 9.

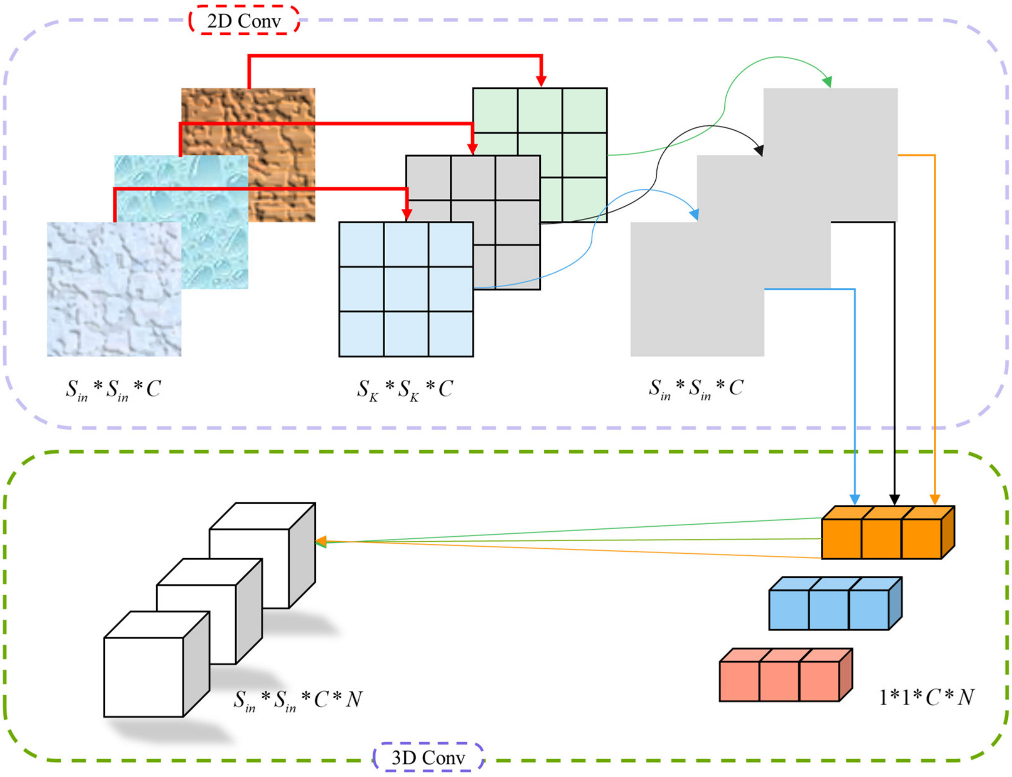


Figure 9. Structure of depthwise separable convolution.

Assuming that the size of the input feature mapping is $S_{in} * S_{in}$, the number of channels is C, the size of the convolution kernels is $S_K * S_K$, and the total number of 3D convolution kernels is N, the computational effort for conventional convolution and depth-separable convolution is as follows.

$$Traditional = S_{in} * S_{in} * C * N * S_K * S_K \tag{11}$$

$$DSC = S_{in} * S_{in} * C * S_K * S_K + S_{in} * S_{in} * C * N \tag{12}$$

Thus, the computational ratio of depth-separable convolution to conventional convolution is:

$$ratio = \frac{S_{in} * S_{in} * C * S_K * S_K + S_{in} * S_{in} * C * N}{S_{in} * S_{in} * C * N * S_K * S_K} = \frac{1}{N} + \frac{1}{S_K * S_K} \tag{13}$$

It can be seen that the reduction in computational effort for depthwise separable convolution is related to the size of the 2D convolution kernels $S_K * S_K$ and the total number N of 3D convolution kernels.

In practice, depthwise separable convolution generally uses a convolution kernel of size 3 × 3. In contrast, conventional convolution parameters are 10 times more computationally intensive than depthwise separable convolution. The PDAM is placed at the end of the backbone, and the feature channels generated by the LAM and GAM are each compressed to half of their original size, and the two are fused in a concatenation symmetrically.

2.3.2. Small-Scale Transformer Feature Pyramid Network (STPN)

In the remote sensing image dataset of forest fire smoke, many very small smoke subjects are included. In this paper, a small-scale transformer feature pyramid network, with the specific structure shown in Figure 4, is proposed to adapt to the single classification task of forest fire smoke detection and to enhance the prediction capability of small forest fire smoke targets. Different flight altitudes of drones and varying fire sizes often lead to drastic changes in the scale of smoke objects. The structure mitigates the negative effects caused by this, thus enhancing the feature fusion capability for smoke images of different scales.

Figure 10 shows the specific architecture of the transformer encoder. The transformer has achieved excellence in the fields of image recognition, target detection, and semantic segmentation [29]. We employed the transformer encoder to replace all the CSP_2 blocks in the neck section. Compared to the CSP_2 blocks in CSPDarknet53, the transformer encoder can capture global information and rich contextual information. Each transformer encoder contains two sub-layers—the multi-headed attention layer and the multilayer perceptron (MLP) layer—which are connected using residuals. The multi-headed attention layer helps the model to focus on different locations to accommodate multiple fires in the image.

$$\begin{gathered} MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \\ where\ head_i = Attention(QW_i{}^Q, KW_i^K, VW_i^V) \end{gathered} \tag{14}$$

where the projection is the parameter matrix $W_i^Q \in \mathbb{R}^{d_{model} * d_k}$, $W_i^K \in \mathbb{R}^{d_{model} * d_k}$, $W_i^V \in \mathbb{R}^{d_{model} * d_v}$, $W^Q \in \mathbb{R}^{d_{model} * hd_v}$. We set the multi-headed attention layer here to 8 layers. At the same time, we set $d_k = d_v = d_{model}/h = 64$.

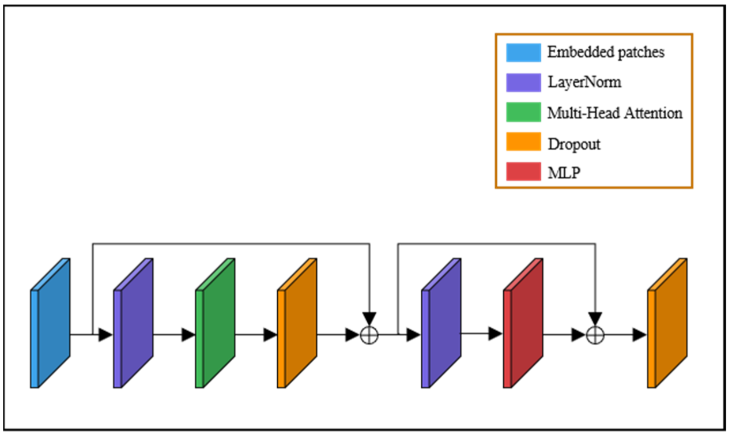


Figure 10. Schematic diagram of the transformer encoder architecture.

The transformer encoder can capture different local information and search for features through a self-attentive mechanism [30]. Therefore, the transformer encoder exhibits better performance in forest fire smoke small target detection.

Based on YOLOX-L, we use the transformer encoder to replace the original CSP_2, with the introduction of the transformer encoder and the addition of the prediction head used to form the STPN. The STPN is applied to the low-resolution feature maps to reduce computational and memory costs.

2.4. UAV Forest Fire Monitoring System

Traditional technical means of forest fire monitoring have various problems, such as blind monitoring areas, poor real-time performance, high operating costs, and high resource consumption [31]. To address these problems, this paper presents a UAV forest fire monitoring system based on PDAM–STPNNet which detects smoke from images returned by UAVs in real-time to determine fire conditions. When a fire occurs, an alarm can be issued and the UAV can be operated from a distance to photograph the situation around the fire, providing information for decision making in relation to rescue and firefighting operations.

The UAV forest fire monitoring system has three sub-systems: (1) a UAV–gimbal camera system, (2) a ground control system, and (3) a ground station terminal monitoring system. (1) The UAV–gimbal camera system uses the YK6-30 model six-rotor UAV. This model UAV is equipped with jiYi K++v2 for flight control; its maximum load is 30 kg and the effective flight time is 20 min to meet the mission requirements. The UAV will experience high-frequency vibration and angular oscillation in flight and is equipped with a gimbal camera for image stabilisation and angular compensation. (2) The ground control system is mainly used for the inspection range, inspection altitude input and inspection task start, UAV flight status view, and other functions. (3) The ground station terminal monitoring system receives high-definition images from the gimbal camera and processes the images for smoke target detection.

The UAV forest fire monitoring system designed in this paper uses a UAV equipped with a high-definition camera and a ground control system which uploads a planned route. After receiving the automatic mission command, the UAV is controlled in ortho mode (gimbal pitch axis: vertical, down) and then takes off and ascends to the route altitude, using the GPS positioning system to feed real-time position information to the ground station terminal monitoring system. While the drone is flying on the route, the pod transmits the image information collected in real time to the ground station terminal monitoring system, which uses PDAM–STPNNet to detect smoke and analyse whether a fire has occurred. When a fire is judged by the monitoring system to have occurred, the ground station sends an alert to the fire service. The workflow is shown in Figure 11.

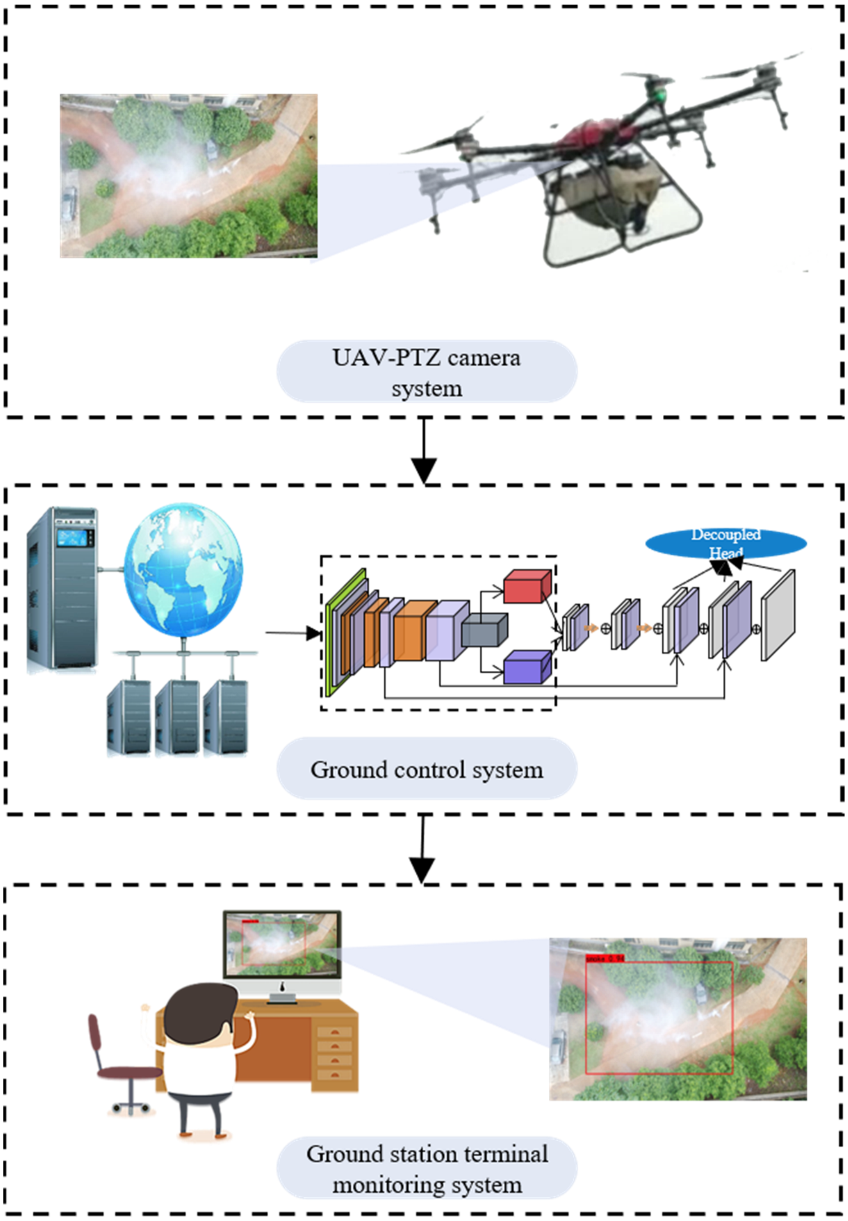


Figure 11. Schematic diagram of the UAV forest fire monitoring system based on PDAM–STPNNet.

3. Results

This section experimentally verifies the effectiveness of DRGNet for smoke detection and compares it with other related models for the same test set. This section treats of dataset acquisition, evaluation metrics, the experimental environment and setup, DRGNet performance analysis, the analysis of method effects, a comparison between different models, ablation experiments, and practical application testing.

3.1. Dataset Acquisition

In order to train the model proposed in this paper and test its effectiveness in the smoke detection task, a home-made dataset of remotely sensed images of forest fire smoke from aerial photographs taken by drones was developed. The dataset is divided into three parts.

The first part of the dataset comes from a number of publicly available video smoke datasets, such as (1) a public dataset published by the Signal Processing Group of Bilkent University, Turkey [32]; (2) a smoke dataset published by the Machine Intelligence Laboratory of the University of Salerno, Italy [33]; (3) a public dataset published by Professor Yuan Feiniu [34]; and (4) a computer vision and pattern recognition laboratory public dataset [35]. We collected 6928 smoke images from the video dataset using screenshots. Of these, we removed some images with low pixels in which it was difficult for the human eye to distinguish the smoke targets, leaving 5935 images with a total of 2638 small targets of smoke. The smoke in these images has obvious features that facilitate the extraction and learning of smoke features by the model. They cover various forms of smoke on the ground and can better reflect the high transparency and lack of obvious edges of the smoke itself, but the ground dataset of the close observation scene has differences in texture, colour, and background from the images taken by the UAV, and it is difficult to adapt to the overhead characteristics of aerial photography using ground data alone.

The second part of the dataset comes from the FLAME dataset [36]—a UAV aerial smoke video dataset which was produced as a selected image dataset with video frame draws (5814 images in total). Of these, we removed some images with blurred cameras and low-quality frame draws, leaving 4652 images and a total of 2394 images of small target smoke. These images were captured with a UAV camera in the air, and the captured images have the characteristics of remote sensing images, which can better reflect the characteristics of small smoke targets and long-distance overhead views under UAV remote sensing conditions and are more suitable for the target detection task of UAV aerial photography scenes. However, there is only one data scene, from a single location in an Arizonan pine forest.

The third part of the dataset is derived from images obtained from aerial photography of simulated forest fire smoke scenarios using drones. In order to improve the robustness of the model, we lit smoke cakes made of fresh branches, dried grass, flour, rosin, and ammonium chloride in various scenarios ranging from school woods to the tops of buildings to open fields in the countryside. We used a Phantom 4 Pro, manufactured by DJI. We flew the drone between 10 m and 90 m in the air and took a total of 1399 images of the smoke. Of these, we removed some poorly angled, poorly lit and blurred images, leaving 1093 images and a total of 539 images of small target smoke. These images simulate the production of smoke when a forest fire occurs, with a background similar to the forest. The purpose of capturing these images was to enrich the forest fire scenario under UAV remote sensing and to enhance the generalisation of the model. The different flight altitudes of the UAV and the multiple smoke situations can better simulate a realistic UAV inspection scenario. The captured scenes are shown in Figure 12.

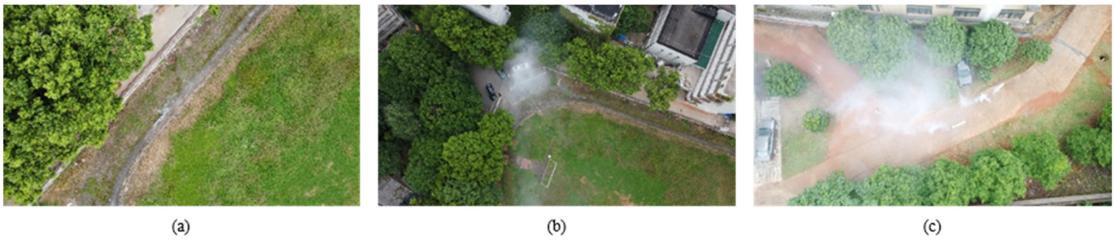


Figure 12. Forest fire smoke simulation scene using a drone. (**a**) Smoke where the target is small and similar to the background. (**b**) Smoke that is easily confused with surrounding trees. (**c**) Smoke that is diffuse.

As can be seen from the figure, the smoke target in Figure 12a is too small and too similar to the background, making it difficult for YOLOX-L to extract features and detect the presence of smoke. The smoke in Figure 12b is not well defined against the surrounding trees, making it difficult for YOLOX-L to extract global features of the image. The smoke in Figure 12c is diffuse and its local texture features are more complex.

The analysis shows that the YOLOX-L detection method has difficulty in accurately detecting smoke similar to that in Figure 12a–c. Therefore, it is necessary to design a forest fire smoke detection algorithm to extract global and local textures from the images and to enhance the small target smoke handling capability. In Section 3.6, we show a comparison of the visualisation results for the model we designed.

3.2. Experimental Preparation

3.2.1. Assessment Indicators

When the IoU between the detection box and the labelled box is greater than the threshold, we consider it to be detected as a positive sample by the model. Otherwise, it is considered to be detected as a negative sample by the model. Based on the above settings, we can classify the sample results of the target detection model as $true\ positive(TP)$, $false\ positive(FP)$, $true\ negative(TN)$, and $false\ negative(FN)$.

In this paper, the performance of the model is evaluated by precision (P), recall (R), mAP, AR, FPS, parameter size, and FLOPs. To compare the performance of PDAM–STPNNet with other models, we use the commonly used evaluation metrics of: precision (P), recall (R), F_1-score (F_1), parameter size, and FLOPs.

$$P = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{17}$$

where precision indicates the percentage of true positives among the successfully detected images in the test set, and recall indicates the percentage of images in the test set that were correctly judged as positive samples. The F_1-score is also used to evaluate the performance of the model. The F_1-score represents the average precision (AP) as the area under the P–R curve—i.e., the mean of the precision values of the P–R curve. The number of parameters is the sum of the parameters of each layer of the architecture of the neural network model and reflects the size of the model. FLOPs refer to the number of floating-point operations, which reflects the amount of computation in the model and can be used to measure the complexity of the model.

Frames per second (FPS) is an important measure of the speed of detection. In this paper, it represents the average number of images detected per second. The formula for calculating FPS is as follows:

$$FPS = \frac{1}{t} \tag{18}$$

where t is the average time taken to process each image.

mAP is averaged for multiple categories. In contrast, the target detection method studied in this paper focuses on the detection of a single category of smoke targets, on the basis of which mAP and AP have the same meaning. For convenience, in this paper, we only refer to mAP. mAP is the single most important measure of performance in target detection and is calculated as:

$$mAP = \int_0^1 p(r)dr \tag{19}$$

The average recall rate (AR) is mainly used to measure the degree of model detection failure. AR is calculated using the following formula:

$$AR = \frac{Recall}{n} \tag{20}$$

3.2.2. Experimental Environment

In order to verify the performance of the proposed PDAM–STPNNet, all experiments in this paper were conducted in the same hardware environment and software environment and with the same imaging equipment, with the specific environmental parameters shown in Table 1.

Table 1. Hardware and software environment.

Hardware environment	CPU	AMD Ryzen 7 5800H with Radeon Graphics
	RAM Video memory GPU	64 GB 16 GB NVIDIA GeForce RTX 3080Ti
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	

Imaging device	Angle jitter: ±0.02° Effective pixel: 36 million Phase detection focus: 567 Contrast detection focus: 425 Viewfinder coverage: 74%	

3.2.3. Experimental Setup

To verify the accuracy and effectiveness of the PDAM–STPNNet model, the model was trained using the Pytorch framework, and the trained model was used to predict aerial smoke images. The hardware environment for this experiment was an NVIDIA GeForce GTX 3080Ti GPU and the software environment was Windows 10.

Prior to the start of this experiment, we produced the available aerial forest fire smoke dataset. To ensure that the model could fully extract the features of smoke, we divided the home-made dataset into 70% for the training set, 15% for the validation set, and 15% for the test set to train the model. During training, each layer of the model was initialised by a Gaussian distribution. Considering the GPU memory size and time cost, we set the batch_size to 16, the momentum to 0.9, and the initial learning rate to 0.005, and we adjusted the learning rate according to the ADAM optimizer. We also set the decay to 0.002 and the number of iterative rounds to 150 epochs (see Table 2).

Table 2. Experimental setup.

Size of Input Images	Batch_Size	Momentum	Initial Learning Rate	Decay	Iterations
608 * 608	16	0.9	0.005	0.002	150 epochs

3.3. Comparison with YOLOX-L

We conducted a series of performance evaluation experiments on a home-made dataset to verify the advantages of PDAM–STPNNet over YOLOX-L for aerial forest fire smoke detection tasks. To prevent the interference of different IoU thresholds in the experiments, an IoU threshold of 0.5 was set in this paper. The results of the comparison experiments between PDAM–STPNNet and YOLOX-L on the home-made dataset are shown in Table 3.

Table 3. Performance comparison of PDAM–STPNNet and YOLOX-L.

Method	YOLOX-L	PDAM–STPNNet
mAP (%)	67.34	77.86
mAP^{50}(%)	81.15	88.01
mAP^{75}(%)	71.87	80.54
AR	41.19	49.23
FPS	77.1	58.6
Parameter	54.18 M	57.19 M
GFLOPs	155.61	164.31
Infertime	14.5 ms	16.1 ms

This paper compares the accuracy of PDAM–STPNNet and YOLOX-L at two IoU thresholds (see Figure 13).

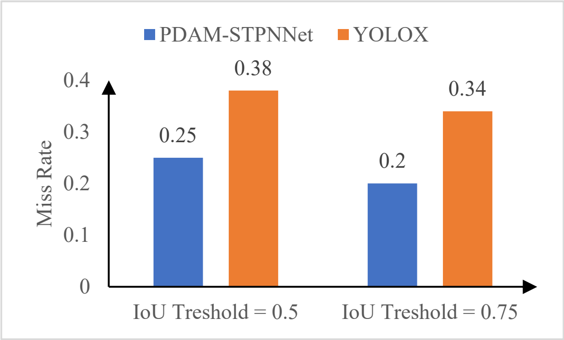


Figure 13. Comparison of the miss rates under different IoUs.

The accuracy of PDAM–STPNNet is significantly improved compared to YOLOX-L for both threshold values of 0.5 and 0.75. The reason for this is that PDAM–STPNNet has a stronger feature extraction capability and feature fusion capability.

3.4. Comparison between Different Models

To validate the detection performance of the PDAM–STPNNet model for the aerial photography of forest fire smoke scenes, this paper compares some target detection models that have performed well in recent years with the PDAM–STPNNet proposed in this paper on the same test environment and home-made dataset. The mAP, mAP^{50}, mAP^{75}, AR, and FPS of different models for the test set are shown in Table 4.

Table 4. Test results of PDAM–STPNNet and other single-stage detectors and two-stage detectors for home-made datasets.

Method	Backbone	mAP	mAP^{50}	mAP^{75}	AR	FPS
Two-stage detectors						
Fast R-CNN	ResNet-101	61.18	70.24	65.05	36.31	-
Faster R-CNN	VGG16	63.11	72.51	66.42	37.02	-
R-FCN	ResNet-101	64.91	74.52	66.98	38.74	-
D-FCN	Aligned-Inception-Resnet	69.01	79.26	71.57	42.99	-
CoupleNet	ResNet-101	63.05	72.47	66.49	36.88	-
FPN	ResNet-101	68.15	80.83	69.94	40.43	-
Mask R-CNN	ResNet-101	72.11	81.96	75.00	45.19	-
Regionlets	ResNet-101	70.97	81.68	74.26	44.82	-
Libra R-CNN	RseNext-101	70.23	80.86	73.40	44.19	-
SINPER	ResNet-101	71.89	81.46	74.62	44.86	-
Cascade Mask R-CNN	ResNet-152	75.04	85.65	78.79	46.83	-
D-RFCN + SNIP	DPN-98	76.73	86.88	80.19	47.29	-
One-stage detectors						
SSD512	VGG16	59.41	70.30	61.89	37.31	81.2
DSSD513	ResNet-101	61.62	72.79	64.53	37.96	68.1
FSAF	ResNext-101	76.12	85.86	76.73	46.94	22.8
NAS-FPN	AmoebaNet	79.86	90.03	81.02	49.76	19.6
YOLOv3 + ASFF	Darknet53	65.81	77.10	69.52	41.64	31.4
YOLOv4-L	CSP-Darknet53	66.21	79.26	69.52	41.73	34.8
YOLOv5-L	CSP-Darknet53	65.38	78.93	69.73	41.40	82.9
YOLOX-S	CSP-Darknet53	62.14	75.98	66.24	39.35	96.8
YOLOX-M	CSP-Darknet53	63.58	77.12	67.38	39.83	87.6
YOLOX-L	CSP-Darknet53	67.34	81.15	71.87	41.19	77.1
YOLOX-L	AmoebaNet	69.12	82.75	73.54	42.51	24.7
YOLOX-X	CSP-Darknet53	68.58	82.37	73.19	41.97	62.9
Ours						
PDAM–STPNNet	AmoebaNet	76.87	87.21	79.68	48.75	17.7
PDAM–STPNNet	CSP-Darknet53	77.86	88.01	80.54	49.23	58.6

In an emergency situation during a forest fire, the two-stage detector has significant shortcomings compared to the single-stage detector due to its poor real-time performance.

Compared to single-stage detectors, dual-stage detectors show good accuracy but have difficulty in meeting the rapid response requirements of forest fire detection tasks. For single-stage detectors, SSD512, DSSD513, and YOLOv5-L excel in speed but lack in accuracy. FSAF and NAS-FPN have high detection accuracy but have poor FPS index performance and are not suitable given the urgency of actual forest fire monitoring scenarios. YOLOX-L outperformed YOLOv5-L for mAP by 1.96%, for mAP^{50} by 2.22%, and for mAP^{75} by 2.14%. Compared to the other YOLOX models, YOLOX-L achieves a trade-off in terms of speed and accuracy. We improve on YOLOX-L and show that PDAM–STPNNet is second only to NAS-FPN in terms of accuracy and significantly better than the other models, while meeting the criteria for real-time detection in terms of speed. We found that the accuracy of NAS-FPN is higher than that of PDAM–STPNNet. To further explore the optimal model, we conducted experiments after replacing the backbone of YOLOX-L and PDAM–STPNNet using AmoebaNet (the backbone of NAS-FPN) to see whether the model can be further optimised. The experimental results show that replacing the YOLOX-L's backbone with AmoebaNet can slightly improve its performance by 1.78% on mAP, 1.60% on mAP50, and 1.67% on mAP75 compared to CSP-Darknet53, but at a significantly lower speed. After replacing the backbone of PDAM–STPNNet with AmoebaNet, the model's accuracy was not as good as before the replacement and the speed was significantly reduced. It was thought that the reason for the reduced model accuracy might be that AmoebaNet searched the network architecture with the evolutionary strategy of aging evolution, which caused the disorder of PDAM's weight assignment. In summary, PDAM–STPNNet is the most suitable model for aerial forest fire smoke detection. We analyse the reasons why our proposed model outperforms other models: (1) PDAM–STPNNet is improved based on YOLOX-L, integrating many speed-optimised solutions, and its model architecture is simple and outstanding in terms of real-time performance. (2) The four improvement strategies proposed in this paper are all designed according to the smaller target characteristics of the aerial forest fire smoke task, which are highly targeted and have significant improvement effects. (3) The home-made dataset in this paper eliminates some blurred and low-quality images, and its images are clear and conducive to the training of the model.

3.5. Exploration of Methodological Effects and Ablation Experiments

3.5.1. Exploring the Effects of a Single Approach

In this section, the experimental results of component stitching data enhancement, LAM, GAM, and STPN in this paper are presented in detail to demonstrate the process of investigating the effects of each method and the way in which they were combined in our experiments. The comparison of their model performance before and after the addition of YOLOX-L is also analysed. The results of the comparative experiments are presented in the following four subsections.

(a) Component stitching data enhancement

YOLOX-L uses Mosaic, and MixUp achieves better data enhancement results. However, the forest fire smoke studied in this paper was taken by an unmanned aircraft, and the targets were often far away from the camera. In this paper, component stitching data enhancement is added to the original data enhancement method to improve the image enhancement method. To demonstrate the effectiveness of this method and to investigate the optimal parameters, we added component stitching data enhancement to the YOLOX-L model and set different k values for comparison experiments to investigate the effect of component stitching data enhancement in the image enhancement session. The results are shown in Table 5.

Table 5. Comparison results before and after adding component stitching data enhancement.

Method	k = 1^2	k = 2^2	k = 3^2
mAP (%)	67.34	69.21	69.07
mAP^{50}(%)	81.15	83.06	82.95
mAP^{75} (%)	71.87	73.74	73.60
AR	41.19	42.07	41.91
FPS	77.1	76.2	75.9
Parameter	54.18 M	54.18 M	54.18 M
GFLOPs	155.61	157.39	157.17

The experimental results show that collage augmentation is beneficial in balancing the proportion of small targets and improving the training effect. It can be seen that when $k = 2^2$, the model has optimal accuracy and the speed is not significantly reduced compared to $k = 1^2$, indicating that component stitching data enhancement can alleviate the imbalance of the dataset due to the under-representation of small targets. Therefore, PDAM–STPNNet uses the component stitching data enhancement approach set to $k = 2^2$.

(b) Local attention mechanism module (LAM)

In Section 2.3.1, the process of obtaining comprehensive features is conducted through Equation (6), where the value of α represents the extent to which the local feature information is taken into account and has an important impact on the effectiveness of LAM, which involves the allocation of the weights. In order to investigate the appropriate value of α, we added YOLOX-L to LAM and set different values of α for testing. The experimental results are shown in Figure 14.

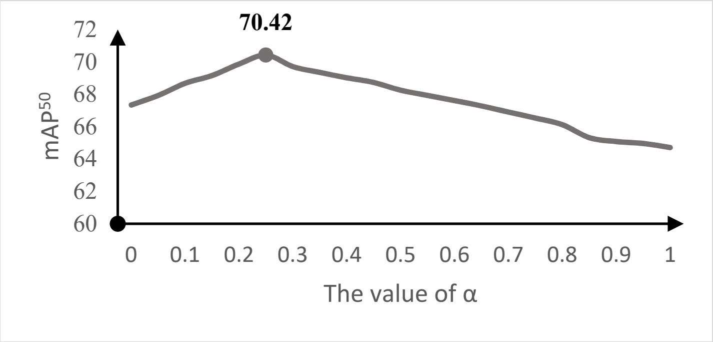


Figure 14. Correspondence between mAP^{50} and the α values taken.

Experiments have verified that LAM can enhance the local texture features of smoke. It can be seen that the optimal value is $\alpha = 0.25$. The reason for this is that when the value of α is too large, the smaller value feature vector is over-considered, and attention is not fully focused on the important information. α is neglected and the attention is over-focused on the global information, which affects the effectiveness of LAM feature extraction.

(c) Global attention mechanism module (GAM)

GAM employs the SN and SiLU activation functions to achieve optimal accuracy enhancement. To verify the feasibility and effectiveness of this choice of GAM, this paper adds GAM to YOLOX-L and conducts ablation experiments on GAM to investigate the effect of GAM performance under different batch processing methods and activation functions, the results of which are shown in Table 6.

Table 6. Experimental investigation of the ablation of GAM.

Method	mAP	$\mathbf{mAP^{50}}$	$\mathbf{mAP^{75}}$	AR
BN + ReLU	69.64	83.74	74.22	42.47
BN + Sigmoid	69.58	83.57	74.17	42.39
BN + SiLU	69.96	83.91	74.53	42.68
SN + ReLU	70.10	84.16	74.81	42.94
SN + Sigmoid	70.06	84.11	74.78	42.92
SN + SiLU	70.42	84.31	75.03	43.16

The experiments verify that GAM can enhance the global texture features of smoke. From the experimental results, it can be seen that the GAM selects the appropriate batching method and activation function to ensure that more effective feature maps are retained, which facilitates the improvement of accuracy. The optimal SN + SiLU combination improves mAP by 0.84%, mAP^{50} by 0.74%, and mAP^{75} by 0.86% compared to the BN + Sigmoid combination.

(d) Small-scale transformer feature pyramid network (STPN)

In this paper, a detection head was added all the way through, and a transformer encoder was used to replace all the CSP_2 blocks in the neck section of the YOLOX-L. To investigate its effectiveness, three improvement experiments were carried out on YOLOX-L, and the test results are shown in Table 7.

Table 7. Experimental investigation of ablation of STPN.

Method	mAP	$\mathbf{mAP^{50}}$	$\mathbf{mAP^{75}}$	AR	Param
No improvement	67.34	81.15	71.87	41.19	54.18 M
Add SE Attention after CSP_2	67.46	81.29	72.15	41.37	55.95 M
Add CBAM Attention after CSP_2	67.59	81.34	72.28	41.54	56.17 M
Replace CSP_2 with transformer encoder	69.26	82.97	74.38	42.56	53.56 M

When the SE attention and CBAM blocks are added after the CSP_2 block, it is found that detection accuracy is slightly improved by the addition of the attention mechanism module, but the improvement is not significant, and the introduction of the attention mechanism module inevitably leads to an increase in the number of parameters. From the results, it is clear that the replacement of the CSP_2 block with the transformer encoder is an effective and feasible way to improve the accuracy of aerial forest fire smoke detection while reducing the number of parameters in the network.

3.5.2. Ablation Experiments with PDAM–STPNNet

In order to verify the overall effectiveness of the method proposed in this paper with optimal parameters, we designed ablation experiments for PDAM–STPNNet, based on YOLOX-L, and we used the control variables method to add component stitching data enhancement, LAM, GAM, and STPN for the combination of these four improvement points for 16 sets of experiments. The results of the experiments are shown in Table 8.

Table 8. Ablation experiment results (A, B, C, and D, respectively, represent the addition of different modules to YOLOX-L. A: Component stitching data enhancement, B: Local attention mechanism module, C: Global attention mechanism module, D: Small-scale transformer feature pyramid network).

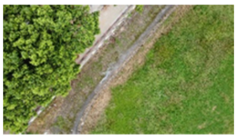
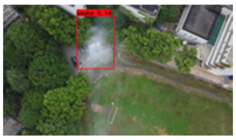
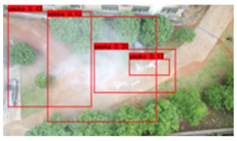
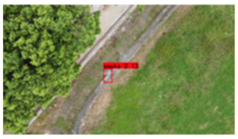
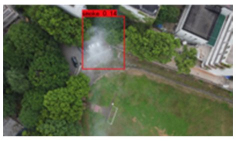
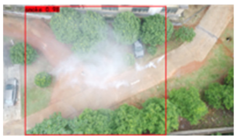
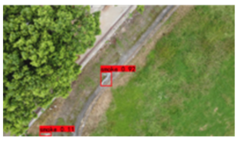
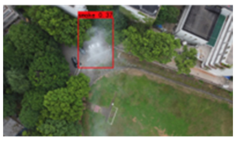
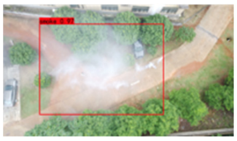
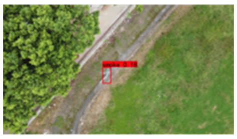
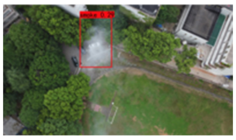
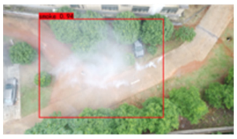
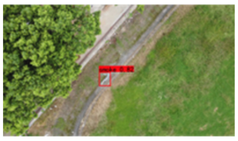
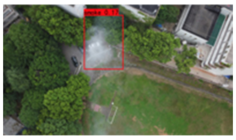
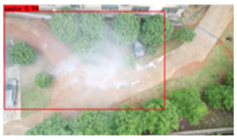
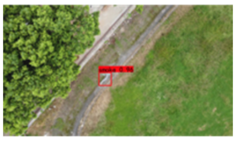
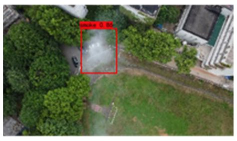
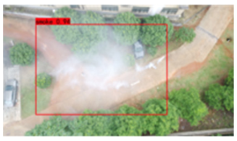
Method	mAP	mAP^{50}	mAP^{75}	AR	FPS	Param	GFLOPs
YOLOX-L	67.34	81.15	71.87	41.19	77.1	54.18 M	155.61
A	69.21	83.06	73.74	42.07	76.2	54.18 M	157.39
B	70.56	84.23	75.11	43.20	66.8	56.94 M	159.13
C	70.42	84.31	75.03	43.16	66.3	56.02 M	158.76
D	69.26	82.97	74.38	42.56	78.4	53.56 M	154.27
A + B	72.64	86.15	76.99	44.08	65.9	56.94 M	161.09
A + C	72.23	85.98	77.04	43.95	65.4	56.02 M	160.87
A + D	71.39	84.68	76.56	43.42	77.3	53.56 M	156.10
B + C	73.60	87.11	78.25	45.36	57.9	58.78 M	163.83
B + D	72.95	86.10	77.94	45.03	68.1	56.32 M	158.19
C + D	72.48	85.92	77.63	44.89	67.6	55.98 M	157.72
A + B + C	75.83	89.26	80.43	46.84	56.7	58.78 M	166.68
A + B + D	74.96	88.35	79.56	46.30	66.8	56.32 M	159.94
A + C + D	74.81	88.24	79.58	46.22	66.1	56.13 M	159.27
B + C + D	75.91	86.12	78.49	48.11	59.4	57.19 M	162.50
A + B + C + D (PDAM–STPNNet)	77.86	88.01	80.54	49.23	58.6	57.19 M	164.31

PDAM can improve the feature discrimination ability of the feature extraction network and effectively prevent the smoke and its background being confused with each other. Among them, LAM mainly extracts local texture features, and adding LAM to YOLOX-L can improve mAP by 3.22%, mAP^{50} by 3.08%, and mAP^{75} by 3.24%. GAM mainly extracts global texture features and adding GAM to YOLOX-L can improve mAP by 3.08%, mAP^{50} by 3.16%m and mAP^{75} by 3.16%. Comparing the two, LAM can improve detection accuracy more, while GAM can help the model to be more accurate in localisation. STPN focuses on the detection of small targets designed for forest fire smoke and replaces the original YOLOX-L block to reduce the number of parameters in the model, resulting in improvements in speed, the number of parameters and accuracy. In summary, PDAM–STPNNet improved by 10.52% for mAP, 6.86% for mAP^{50}, and 8.67% for mAP^{75} compared to YOLOX-L. There was a small reduction in speed, but it still met the criteria for real-time detection. The results of the 16 sets of experiments demonstrate the role of component stitching data enhancement, LAM, GAM, and STPN. Based on the above experiments, we conclude that PDAM–STPNNet is the most suitable target detection model for forest fire smoke detection tasks using UAV aerial photography.

3.6. Comparison of Visualisation Results

For a more visual analysis of PDAM–STPNNet, we visualise the detection results for YOLOX-L, PDAM–STPNNet, and YOLOX-L, with component stitching data enhancement, LAM, GAM, and STPN added, respectively. The detection frames, categories, and confidence levels are shown in the detection results graph in Table 9.

Table 9. Visual comparison of test results.

Experimental Method	Detection Result		
YOLOX-L			
YOLOX-L with component stitching data enhancement			
YOLOX-L with LAM			
YOLOX-L with GAM			
YOLOX-L with STPN			
PDAM–STPNNet			
	a	**b**	**c**

From this, we can see that for Figure a in Table 9, the smoke is far away from the UAV camera and has a high similarity to the surrounding background. YOLOX-L does not recognise the target but identifies it as background, and thus a misdetection occurs. After component stitching data enhancement, the smoke is detected successfully, but with low confidence. When LAM was used to enhance the local texture extraction, the target is detected, but a false detection occurs because there is a spot where the background is similar to the smoke. The problem was solved when GAM was used for global texture extraction, or when STPN was used to enhance the detection of small targets in forest fire smoke. The PDAM–STPNNet integrating these four improved strategies demonstrates good results in detection and yields accurate localisations and high confidence levels. For Figure b in Table 9, the smoke density is low and the local features are not obvious. YOLOX-L is inaccurately localised, and a low confidence level is obtained. After component stitching data enhancement, the results were not significantly improved. The use of LAM, GAM, and STPN all improved the detection of smoke at this low density, and PDAM–STPNNet gave accurate localisation and high confidence. For Figure c in Table 9, the smoke spreads over a large area and is strongly illuminated. YOLOX-L has difficulty in recognising the exact localisation of the smoke and produces more detection frames. After component stitching data enhancement, the results improve, but the localisation is still inaccurate; LAM, GAM, and STPN all improved the results, but the localisation was not accurate enough, while PDAM–STPNNet accurately localised it and obtained a high confidence level.

3.7. Model Performance Comparison on Public Datasets

The metrics obtained from the evaluation of home-made datasets alone may not be able to fully objectively evaluate the performance of PDAM–STPNNet. In this section, in order to verify the superiority of PDAM–STPNNet over the commonly used target detection models in recent years, we used 85% of home-made datasets as the training set, 15% of home-made datasets as the validation set to train the model, and the Wildfire Observers and Smoke Recognition Homepage [37] and Bowfire datasets [38], two public datasets, were used as test sets for a comprehensive evaluation experiment.

3.7.1. Wildfire Observers and Smoke Recognition Homepage

The Wildfire Observers and Smoke Recognition Homepage dataset was created and is maintained by the Wildfire Research Centre, which is part of a university separate from the School of Electrical Engineering, Mechanical Engineering and Shipbuilding. While early fire detection was traditionally based on human wildfire monitoring, this dataset uses modern information and communication technologies (ICT) that can replace wildfire observers to perform human wildfire observations. The site's dataset is divided into two categories: (1) a wildfire smoke image database and (2) a wildfire smoke video database. The image data in the wildfire smoke video database are frame-by-frame screenshots of the video data, so we dropped the wildfire video database and selected one of the wildfire smoke images from the database for every two images as test data. In total, 3482 data points were selected, and six metrics—mAP, mAP^{50}, mAP^{75}, AR, FPS, and GFLOPs—were measured; the results are shown in Table 10.

Table 10. Wildfire Observers and Smoke Recognition Homepage assessment results.

Method	YOLOX-L	YOLOv4	YOLOv5	YOLOR	PDAM–STPNNet
mAP(%)	67.13	66.79	65.77	66.26	73.81
mAP^{50}(%)	79.15	78.61	77.78	78.94	86.21
mAP^{75}(%)	67.77	66.75	65.89	66.94	76.54
AR	41.39	41.02	40.49	40.69	45.23
FPS	59.78	35.75	112.4	50.78	51.6
GFLOPs	183.15	202.55	101.87	167.15	186.31

The analysis of the above model evaluation parameters shows that YOLOX-L still excels in terms of speed, but its accuracy is no longer sufficient to meet the requirements of forest fire smoke detection and it is prone to missed and false detections. YOLOv4, YOLOv5 and YOLOR are lacking too much in terms of accuracy to meet the accuracy requirements of forest fire smoke detection. PDAM–STPNNet achieved values of 73.86% for mAP, 86.21% for mAP^{50}, 76.54% for mAP^{75}, and 51.6 FPS when tested on the Wildfire Observers and Smoke Recognition Homepage. PDAM–STPNNet improved upon YOLOX-L by 6.68% for mAP, 7.06% for mAP^{50}, and 8.77% for mAP^{75}. Compared to the other models in the table, PDAM–STPNNet has the best overall performance on the Wildfire Observers and Smoke Recognition Homepage set. The experimental results for the Wildfire Observers and Smoke Recognition Homepage show that PDAM–STPNNet has a strong feature extraction capability and a strong generalization capability, even in scenes with varying smoke features.

3.7.2. Bowfire Dataset

The Bowfire dataset is a classical dataset of only 227 images and contains many negative samples that are easily confused with smoke. We used the Bowfire dataset directly as a test set to examine the resistance of PDAM–STPNNet to interference. Due to the small size of this dataset, we trained some commonly used target detection models of recent years on a home-made dataset and tested them on Bowfire, measuring six metrics: mAP, mAP^{50}, mAP^{75}, AR, FPS, and GFLOPs. The results are shown in Table 11.

Table 11. Bowfire dataset assessment results.

Method	YOLOX-L	YOLOv4	YOLOv5	YOLOR	PDAM–STPNNet
mAP(%)	55.86	56.39	50.81	55.24	60.52
mAP^{50}(%)	65.18	66.57	66.84	67.58	74.22
mAP^{75}(%)	53.47	54.93	53.78	54.23	63.12
AR	35.14	35.41	33.44	34.96	38.03
FPS	56.37	34.67	100.49	41.14	43.62
GFLOPs	196.97	221.19	111.38	179.74	200.04

The Bowfire dataset contains many objects that can be easily confused with smoke, making detection difficult and prone to misses and false detections. Nevertheless, our PDAM–STPNNet has significantly improved its mAP, mAP^{50}, and mAP^{75} scores compared to YOLOX-L, YOLOv4, YOLOv5, and YOLOR and has obtained good detection results. The experimental results on the Bowfire dataset show that PDAM–STPNNet also has a strong detection capability in difficult detection scenarios, which is important for forest fire smoke detection in some complex forest terrains. PDAM–STPNNet achieved values of 60.52% for mAP, 74.22% for mAP^{50}, and 63.12% for mAP^{75}, and 43.62 for FPS when tested on the Bowfire dataset.

3.8. Practical Application Tests

To test the generalisation ability and practicality of the model, we conducted field tests in Huangfengqiao State Forestry Farm, and the results were good. Huangfengqiao State Forestry is located in a hilly area in the east and west of You County, Hunan Province, China, as shown in Figure 15. The area is dominated by fir plantations with relatively low water content, which are densely distributed and prone to large fires, and where foliage and smoke obscure each other, resulting in poor detection. There are also bare rocks in the hilly areas that are similar in colour to the smoke and can be easily misidentified. The complex conditions for the occurrence of actual forest fires are met.

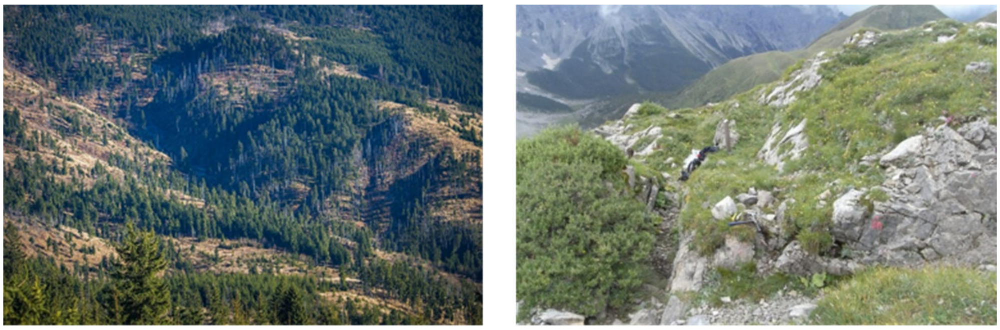


Figure 15. Dense cedar forest and easily misidentified rocks.

A 40-day simulation was carried out at Huangfengqiao State Forest, creating smoke by lighting a smoke cake with flour, rosin, and ammonium chloride as the main ingredients. STPNNet was used to detect the captured images.

A comparison of the detection results of YOLOX-L and PDAM–STPNNet for the three categories of smoke is shown in Figure 16. For each category, 50 images of real scenes were selected for testing, and the detection results were considered to be accurate if the IoU with the actual location of the smoke was greater than 0.7. As can be seen from the figure, the recognition accuracies of YOLOX-L and PDAM–STPNNet were 86% and 98% in category A, 58% and 84% in category B, and 16% and 74% in category C, respectively.

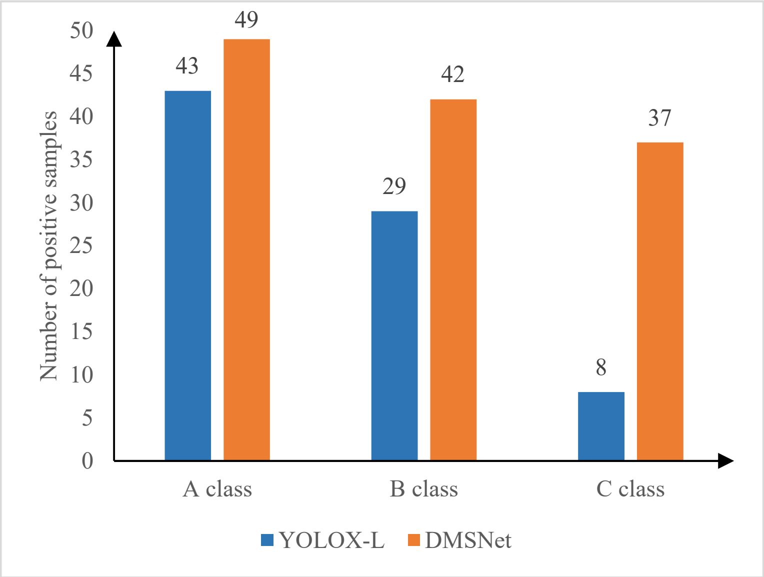


Figure 16. Comparison of identification results for PDAM–STPNNet and YOLOX-L in Huangfengqiao state-owned forestry site: Class A refers to cases with medium smoke concentration and size, while Class B refers to cases with low smoke concentration and less distinctive features; Class C refers to cases where smoke is distant and in the early stages of burning.

4. Discussion

4.1. Training and Test Datasets

In order to investigate the effect of PDAM–STPNNet on forest fire smoke detection in practical application scenarios, this paper developed a home-made dataset of remote sensing images of forest fires taken by UAVs and applied the dataset to the training and testing of PDAM–STPNNet. In addition, to further illustrate the detection effectiveness of PDAM–STPNNet, we tested PDAM–STPNNet on the home-made datasets, Wildfire Observers and Smoke Recognition Homepage and Bowfire, respectively, and achieved good evaluation results, proving PDAM–STPNNet's model performance and generalisation capability. Due to the nature of the smoke cake material used in the production of the simulated forest fire smoke, the smoke images in this dataset are usually light-coloured and transparent white smoke. When actual forest fires occur, this may lead to a lack of oxygen in the burned area, and the smoke contains large-scale carbon particles and takes on a dark black colour character. In the future, the darker black smoke should be included in the dataset to enhance the model's ability to detect when there is insufficient oxygen in the burning area.

4.2. Application and Future Work Directions

In this paper, PDAM–STPNNet is deployed in the UAV forest fire monitoring system, and three sub-systems—the UAV–gimbal camera system, the ground control system, and the ground station terminal monitoring system—are built and put into practical application for forest fire smoke monitoring. The UAV forest fire monitoring system built in this paper uses GPS positioning, combined with image target detection, to roughly locate the UAV and the fire location. The operational airspace for UAVs in forest fire monitoring is mainly concentrated in mountainous woodlands. Due to signal problems, how to establish long-lasting, reliable, and timely communication with UAVs and to determine the precise location of UAVs has always troubled UAV-related practitioners. In the future, the signal in mountainous woodlands should be further strengthened to precisely locate the UAV and fire locations in order to enhance the practicality and feasibility of the UAV forest fire monitoring system in practical applications.

4.3. Advantages of the Method in This Paper

In the training session of the model, a portion of the images used had a large smoke coverage area and occupied a large number of pixels in the image. We use component stitching data enhancement to pre-process the images and input k images into the neural network model for training, which reduces the scale of the smoke target while ensuring that the length and width of the collaged image is the same as the original image, allowing the model to learn the smoke features of multiple fire sources, which is conducive to improving the performance of the model in practical applications. PDAM is proposed, which mines deep features through both local and global textures, which is enabled by the original backbone feature extraction capability and is of great significance for practical applications. STPN is proposed, which uses the transformer encoder to replace CSP_2 on FPN, aiming to make the model perform well in the detection of small targets of forest fire smoke. The methods proposed in this paper are all designed according to the needs of the UAV aerial photography forest fire smoke detection task, taking into account factors such as the long distance of the target during aerial photography, the unclear colour characteristics of the smoke, and the complexity of the forest environment, and focus on improving the detection capability for remote sensing images of forest fire smoke. In this paper, we propose a PDAM–STPNNet for aerial photography of forest fire smoke detection by UAV, and the feasibility and practicality of the PDAM–STPNNet is demonstrated in experiments. However, in practice, forest fires do not only occur during the daytime; smoke is a prominent visual feature during the day and is difficult to detect at night. In the future, in order to improve the forest fire detection task and to ensure that forest fires can be detected in a timely manner at all times of day and night, a deep learning model should be trained and deployed on the UAV forest fire monitoring system to complete the task of detecting flames at night.

4.4. Analysis and Outlook

Analyzing the falsely detected data is very important to improve the performance of our network. Therefore, we analysed a sample of smoke images that were misdetected by PDAM–STPNNet. Figure 17 shows a blurred white transparent object formed by the reflection of car glass, which has an irregular shape and white transparency similar to that of smoke. It can be seen that PDAM–STPNNet is not good at distinguishing small targets that have a similar shape, colour and transparency to smoke. To solve this problem, our future work will consider studying the properties of object reflections to further address the problem of false detection of small targets. Although the PDAM–STPNNet proposed in this paper has achieved good results in smoke detection from UAV aerial photography, in reality forest fires do not only occur during daytime; smoke is a prominent visual feature during daytime, but difficult to detect at night. In the future, in order to improve forest fire detection and ensure that forest fires can be detected in time in all weathers, flames, which are prominent features at night, should be adopted as the research object and figure in deep learning models for deployment in a UAV forest fire monitoring system combined with sensors, such as thermal cameras, to complete the forest fire detection task at night.

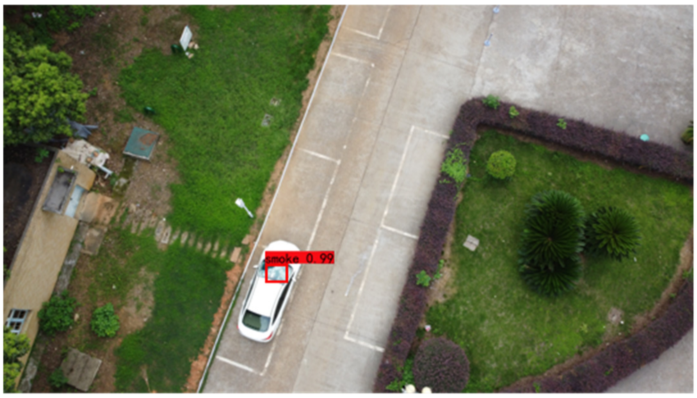


Figure 17. White transparent object created by reflection from car glass.

5. Conclusions

In order to solve the problem that small target smoke in UAV remote sensing images is under-represented in datasets and easily confused with its background, this paper proposes a forest fire smoke detection model—PDAM–STPNNet—for a UAV forest fire monitoring system. At the same time, we constructed a forest fire smoke target detection dataset based on UAV images, containing a total of 11,680 forest fire smoke images, of which there are 5571 small target smoke images. PDAM–STPNNet has been improved based on YOLOX-L. Component stitching data enhancement can balance the proportion of small targets in the dataset, increasing mAP by 1.87%, mAP^{50} by 1.91%, and mAP^{75} by 1.87% without changing the original network structure. PDAM can improve the feature discrimination ability of the feature extraction network and effectively prevent the smoke and background from being confused with each other. Among them, adding LAM to YOLOX-L can improve mAP by 3.22%, mAP^{50} by 3.08%, and mAP^{75} by 3.24%. Adding GAM to YOLOX-L can improve mAP by 3.08%, mAP^{50} by 3.16%, and mAP^{75} by 3.16%. STPN can mitigate the negative effects caused by drastic scale changes and has stronger feature fusion capability. Adding STPN to YOLOX-L was able to improve mAP by 1.92%, mAP^{50} by 1.82%, and mAP^{75} by 2.51%. We chose two public datasets to validate the effectiveness of PDAM–STPNNet, and it is clear from the experimental results that PDAM–STPNNet has high detection accuracy and speed.

In forest fire control, PDAM–STPNNet can be mounted on a UAV forest fire monitoring system to detect smoke based on remote sensing images captured by UAVs to locate forest fire smoke, which is of great importance for the protection of forest ecology.

Author Contributions: J.Z.: methodology, writing—original draft preparation, conceptualisation. Y.H.: software, data acquisition, formal analysis. W.C.: model guidance, resources. G.Z.: validation, project administration, funding acquisition, supervision. L.L.: visualisation, writing—review and editing. We are grateful to all members of the Forestry Information Research Centre for their advice and assistance in the course of this research. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Changsha Municipal Natural Science Foundation (Grant No. kq2014160), in part by the National Natural Science Foundation in China (Grant No. 61703441), in part by the key projects of the Department of Education Hunan Province (Grant No. 19A511) and in part by Hunan Key Laboratory of Intelligent Logistics Technology (2019TP1015).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to partial author disagreement.

Acknowledgments: We are grateful to all members of the Forestry Information Research Centre for their advice and assistance in the course of this research.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

The abbreviations in this paper are as follows:

YOLO	You Only Look Once
IoU	Intersection over Union
mIoU	Mean Intersection over Union
CNN	Convolutional Neural Network
R-CNN	Region-based Convolutional Neural Network
SSD	Single Shot Multibox Detector
FPN	Feature Pyramid Network
FCN	Full Convolutional Network
ResNet	Residual Network
UAV	Unmanned Aerial Vehicle
SPP	Spatial Pyramid Pooling
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
AP	Average Precision
mAP	Mean Average Precision
AR	Average Recall
FPS	Frames Per Second
PDAM	Parallel spatial domain attention mechanism
STPN	Small-scale transformer feature pyramid network

References

1. Agus, C.; Azmi, F.F.; Ilfana, Z.R.; Wulandari, D.; Rachmanadi, D.; Harun, M.K.; Yuwati, T.W. The impact of Forest fire on the biodiversity and the soil characteristics of tropical Peatland. In *Handbook of Climate Change and Biodiversity*; Springer: Cham, Switzerland, 2019; pp. 287–303.
2. Fachrie, M. Indonesia's forest fire and haze pollution: An analysis of human security. *Malays. J. Int. Relat.* **2020**, *8*, 104–117. [CrossRef]
3. Gramling, C. Here's How Climate Change May Make Australia's Wildfires More Common. 2020. Available online: https://www.sciencenews.org/article/how-climate-change-may-make-australia-wildfires-more-common (accessed on 4 June 2021).
4. Cascio, W.E. Wildland fire smoke and human health. *Sci. Total Environ.* **2018**, *624*, 586–595. [CrossRef]
5. Chowdary, V.; Gupta, M.K. Automatic forest fire detection and monitoring techniques: A survey. In *Intelligent Communication, Control and Devices*; Springer: Singapore, 2018; pp. 1111–1117.
6. Jiang, W.; Wang, F.; Fang, L.; Zheng, X.; Qiao, X.; Li, Z.; Meng, Q. Modelling of wildland-urban interface fire spread with the heterogeneous cellular automata model. *Environ. Model. Softw.* **2021**, *135*, 104895. [CrossRef]
7. Shah, R.; Satam, P.; Sayyed, M.A.; Salvi, P. Wireless Smoke Detector and Fire Alarm System. *Int. Res. J. Eng. Technol.* **2019**, *6*, 1407–1412.
8. Amiaz, T.; Fazekas, S.; Chetverikov, D.; Kiryati, N. Detecting regions of dynamic texture. In *International Conference on Scale Space and Variational Methods in Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 848–859.
9. Toreyin, B.U.; Dedeoglu, Y.; Cetin, A.E. Contour based smoke detection in video using wavelets. In Proceedings of the 2006 14th European Signal Processing Conference, Florence, Italy, 4–8 September 2006; pp. 1–5.
10. Ghassempour, S.; Girosi, F.; Maeder, A. Clustering multivariate time series using hidden Markov models. *Int. J. Environ. Res. Public Health* **2014**, *11*, 2741–2763. [CrossRef] [PubMed]
11. Chunyu, Y.; Jun, F.; Jinjun, W.; Yongming, Z. Video fire smoke detection using motion and color features. *Fire Technol.* **2010**, *46*, 651–663. [CrossRef]
12. Jiao, Z.; Zhang, Y.; Xin, J.; Mu, L.; Yi, Y.; Liu, H.; Liu, D. A deep learning based forest fire detection approach using UAV and YOLOv3. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 23–27 July 2019; pp. 1–5.
13. Kim, H.W. A Study on Application Methods of Drone Technology. *J. Korea Inst. Inf. Electron. Commun. Technol.* **2017**, *10*, 601–608.
14. Roldán-Gómez, J.J.; González-Gironda, E.; Barrientos, A. A Survey on Robotic Technologies for Forest Firefighting: Applying Drone Swarms to Improve Firefighters' Efficiency and Safety. *Sciences* **2021**, *11*, 363. [CrossRef]
15. Kinaneva, D.; Hristov, G.; Raychev, J.; Zahariev, P. Early forest fire detection using drones and artificial intelligence. In Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 20–24 May 2019; pp. 1060–1065.

16. Alexandrov, D.; Pertseva, E.; Berman, I.; Pantiukhin, I.; Kapitonov, A. Analysis of machine learning methods for wildfire security monitoring with an unmanned aerial vehicle. In Proceedings of the 2019 24th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 8–12 April 2019; pp. 3–9.
17. Tian, G.; Liu, J.; Zhao, H.; Yang, W. Small object detection via dual inspection mechanism for UAV visual images. *Appl. Intell.* **2021**, 1–14. [CrossRef]
18. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
19. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale match for tiny person detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 1257–1265.
20. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
21. Sitaula, C.; Xiang, Y.; Aryal, S.; Lu, X. Scene image representation by foreground, background and hybrid features. *Expert Syst. Appl.* **2021**, *182*, 115285. [CrossRef]
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 21–23 June 1994; pp. 779–788.
23. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
24. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
26. Chen, Y.; Zhang, M.; Yang, X.; Xu, Y. The research of forest fire monitoring application. In Proceedings of the 2010 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010; pp. 1–5.
27. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Xiang, S.; Sun, J.; Jia, J. Stitcher: Feedback-driven data provider for object detection. *arXiv* **2020**, arXiv:2004.12432.
28. Luo, P.; Ren, J.; Peng, Z.; Zhang, R.; Li, J. Differentiable learning-to-normalize via switchable normalization. *arXiv* **2018**, arXiv:1806.10779.
29. Zhang, J.; Li, C.; Grzegorzek, M. Applications of Artificial Neural Networks in Microorganism Image Analysis: A Comprehensive Review from Conventional Multilayer Perceptron to Popular Convolutional Neural Network and Potential Visual Transformer. *arXiv* **2021**, arXiv:2108.00358.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Hook, NY, USA, 2017; pp. 5998–6008.
31. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A Forest Fire Detection System Based on Ensemble Learning. *Forests* **2021**, *12*, 217. [CrossRef]
32. Cetin, E. Computer Vision Based Fire Detection Dataset. 2015. Available online: http://signal.ee.bilkent.edu.tr/VisiFire/Demo/SmokeClips/ (accessed on 20 December 2015).
33. University of Salerno. Smoke Detection Dataset. 2015. Available online: http://mivia.unisa.it/ (accessed on 20 December 2015).
34. University of Science and Technology of China, State Key Lab of Fire Science. Available online: http://staff.ustc.edu.cn/,yfn/vsd.html (accessed on 20 December 2015).
35. Keimyung University. Wildfire Smoke Video Database (CVPR Lab, Keimyung University). 2012. Available online: https://cvpr.kmu.ac.kr/ (accessed on 20 December 2015).
36. Shamsoshoara, A.; Afghah, F.; Razi, A.; Zheng, L.; Fulé, P.Z.; Blasch, E. Aerial Imagery Pile burn detection using Deep Learning: The FLAME dataset. *Comput. Netw.* **2021**, *193*, 108001. [CrossRef]
37. Wildfire Observers and Smoke Recognition Homepage. Available online: http://wildfire.fesb.hr/index.php?option=com_content&view=article&id=59&Itemid=55 (accessed on 20 December 2015).
38. Bowfire Dataset. Available online: https://bitbucket.org/gbdi/bowfire-dataset/downloads/ (accessed on 20 December 2015).