forests MDPI

Article

A Semi-Supervised Method for Real-Time Forest Fire Detection Algorithm Based on Adaptively Spatial Feature Fusion

Ji Lin [1], Haifeng Lin [1,*] and Fang Wang [2,*]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[2] College of Electronic Engineering, Nanjing Xiaozhuang University, Nanjing 211171, China
[*] Correspondence: haifeng.lin@njfu.edu.cn (H.L.); wangfang0182217@njxzc.edu.cn (F.W.);
Tel.: +86-25-8542-7827 (H.L.); +86-25-8617-5539 (F.W.)

Abstract: Forest fires occur frequently around the world, causing serious economic losses and human casualties. Deep learning techniques based on convolutional neural networks (CNN) are widely used in the intelligent detection of forest fires. However, CNN-based forest fire target detection models lack global modeling capabilities and cannot fully extract global and contextual information about forest fire targets. CNNs also pay insufficient attention to forest fires and are vulnerable to the interference of invalid features similar to forest fires, resulting in low accuracy of fire detection. In addition, CNN-based forest fire target detection models require a large number of labeled datasets. Manual annotation is often used to annotate the huge amount of forest fire datasets; however, this takes a lot of time. To address these problems, this paper proposes a forest fire detection model, TCA-YOLO, with YOLOv5 as the basic framework. Firstly, we combine the Transformer encoder with its powerful global modeling capability and self-attention mechanism with CNN as a feature extraction network to enhance the extraction of global information on forest fire targets. Secondly, in order to enhance the model's focus on forest fire targets, we integrate the Coordinate Attention (CA) mechanism. CA not only acquires inter-channel information but also considers direction-related location information, which helps the model to better locate and identify forest fire targets. Integrated adaptively spatial feature fusion (ASFF) technology allows the model to automatically filter out useless information from other layers and efficiently fuse features to suppress the interference of complex backgrounds in the forest area for detection. Finally, semi-supervised learning is used to save a large amount of manual labeling effort. The experimental results show that the average accuracy of TCA-YOLO improves by 5.3 compared with the unimproved YOLOv5. TCA-YOLO also outperformed in detecting forest fire targets in different scenarios. The ability of TCA-YOLO to extract global information on forest fire targets was much improved. Additionally, it could locate forest fire targets more accurately. TCA-YOLO misses fewer forest fire targets and is less likely to be interfered with by forest fire-like targets. TCA-YOLO is also more focused on forest fire targets and better at small-target forest fire detection. FPS reaches 53.7, which means that the detection speed meets the requirements of real-time forest fire detection.

Keywords: forest fire detection; deep learning; adaptively spatial feature fusion; attention mechanism; semi-supervised learning

check for updates

Citation: Lin, J.; Lin, H.; Wang, F. A Semi-Supervised Method for Real-Time Forest Fire Detection Algorithm Based on Adaptively Spatial Feature Fusion. *Forests* **2023**, *14*, 361. https://doi.org/10.3390/f14020361

Academic Editor: Viacheslav I. Kharuk

Received: 12 January 2023
Revised: 8 February 2023
Accepted: 9 February 2023
Published: 11 February 2023

1. Introduction

Forest fires occur frequently around the world these years, resulting in serious economic losses and human casualties. Forest fires spread quickly and are difficult to fight. Therefore, forest fire detection is especially important. Traditional forest fire detection methods mainly include manual inspection, sensor technology [1–3], infrared technology [4], and remote sensing satellite [5] images for fire monitoring. Forest fire monitoring through manual inspection requires huge human and material resources and is inefficient. Traditional smoke and temperature sensors have a limited detection range and are difficult to deploy in large-scale forest areas due to problems with power, communication, and networking. The traditional infrared monitoring technology is easily affected by the environment and requires high monitoring distance, which is prone to omission and misdetection. Although the monitoring of forest fires through remote sensing satellite images has a wide range, the infrared band and visible light used by satellites are easily disturbed by clouds and fog as well as weather conditions, making it difficult to detect forest areas around the clock without any dead angle.

With the continuous development of image processing technology, the use of images for forest fire detection has become a mainstream trend in forest fire monitoring. Traditional image processing methods are mainly used for forest fire detection by extracting flame color features, edge features, geometric features, etc. For example, Celik et al. [6] constructed a flame color classification model based on YCbCr, separated brightness from chromaticity, obtained flame color motion pixels using an adaptive background subtraction algorithm as well as an RGB color model and used a statistical model to achieve flame color classification. Habiboglu et al. [7] used color and spatial information and, using a covariance matrix approach, achieved the detection of forest fire flames. Jin et al. [8] achieved fast detection of forest fires using features such as the size and color of flames based on a logistic model and time domain smoothing. Dimitropoulos et al. [9] used the linear dynamic texture method for the calculation of flame regions of interest. In summary, the traditional image processing methods mainly achieve forest fire detection by manually extracting forest fire features, and the feature extraction directly determines the result of forest fire detection. Traditional image processing methods are not only affected by human subjective factors, but also by environmental factors, such as different lighting and weather conditions, which can affect the comprehensive extraction of forest fire features.

In recent years, deep learning techniques based on convolutional neural networks (CNNs) have developed rapidly, providing new ideas and methods for forest fire detection. CNNs have a powerful feature extraction capability to obtain deeper semantic information about images and have an end-to-end model training process, which effectively avoids the complexity and limitations of manual feature selection. For example, Yin et al. [10] improved the convolutional layer of traditional CNNs by batch regularization, which solved the overfitting problem that traditional CNNs are prone to and improved the detection accuracy. Zhang et al. [11] detected synthetic forest fire images with a Faster R-CNN model, and they combined real fire images as well as simulated fire images with a forest background. Avula et al. [12] used CNN for forest fire detection and improved the accuracy of forest fire detection by introducing a spatial transformer network and entropy function thresholding. Wang et al. [13] took SqueezeNet as the backbone feature extraction network to segment forest fires, fused multi-scale context information to improve accuracy, and solved the problem of the difficult segmentation of small targets in early forest fires. Jiao et al. [14] detected forest fires based on the YOLOv3 [15] target detection network and labeled the specific location of the fire area as well as the confidence probability of forest fires, improving the accuracy and efficiency of UAV forest fire detection. Shamsoshoara et al. [16] performed pixel-level segmentation based on the semantic segmentation network U-Net [17] for forest fire images; however, semantic segmentation requires high accuracy for dataset annotation.

Although CNNs have solved the problem of intelligent recognition of forest fires to some extent, due to the limitations of convolutional operations, its extracted forest fire image features are only limited to local regions, which lack long-range dependencies. Long-range dependence is important for the network to focus on the forest fire target region and ignore the noise in the whole feature map [18]. In addition, convolutional neural network target detection models also have a small field of perception and lack sufficient global and contextual information. While Transformer has recently made a big splash in the image field [19–22], the key factor of its success is that Transformer's self-attention mechanism can capture long-range dependencies in images, has a powerful global modeling capability, and can expand the receptive field to obtain more contextual information. Therefore, this paper proposes a forest fire target detection model, TCA-YOLO (T, C, and A are the acronyms of the respective modules used for improvement), with the convolutional neural network-based target detector YOLOv5 [23] as the basic framework and a series of improvements. Firstly, in order to make up for the deficiency of CNNs in the global feature extraction of forest fires, the Transformer encoder [24] is used in combination with a CNN as a feature extraction network, which fully combines the advantages of CNN and Transformer's self-attention mechanism to enhance the extraction of global information of forest fire targets and has a more powerful receptive field to obtain more contextual information. Secondly, in order to enhance the model's focus on forest fire targets, we integrate the Coordinate Attention (CA) [25] mechanism in the neck part of YOLOv5. CA not only obtains inter-channel information but also considers the information on the direction-related location, which helps TCA-YOLO to better identify and locate forest fire targets. In addition, we integrate the adaptively spatial feature fusion (ASFF) technique [26]. ASFF improves the multi-scale fusion of forest fire features by adaptive methods to adjust the fusion ratio between different feature layers. It can effectively suppress the interference of invalid features in the complex background of the forest area on forest fire detection, which further improves the accuracy of TCA-YOLO. Since the model performance of forest fire detection tasks using deep learning techniques depends on the number of training forest fire image samples, and the samples need to be annotated [27], forest fire images are obtained from a wide range of sources with a huge amount of data. Usually, the annotation of forest fire images is mostly carried out manually [28–30]. Faced with a massive forest fire image dataset, manual annotation takes a lot of time and requires specialized personnel to perform the annotation. To solve the drawbacks of manual labeling, this paper adopts semi-supervised learning to train the proposed forest fire target detection model TCA-YOLO. Only a small number of manually labeled forest fire datasets are used for training. The remaining unlabeled dataset is annotated by the proposed automatic annotation method and filtered based on the confidence level as a pseudo-label [31] to the training set to retrain the model. Continuous iteration improves accuracy and saves a lot of manual annotation work. If new unlabeled forest fire datasets become available in the future, this strategy can also be used to automatically label the unlabeled datasets and add them to the training set to retrain the model so as to continuously improve TCA-YOLO's detection accuracy.

2. Materials and Methods

2.1. Forest Fire Data Set

As shown in Figure 1, the dataset used in this paper includes 3000 forest fire images of different scenes from forest fire images captured by video surveillance devices and drones in forest areas, publicly available forest fire datasets, and forest fire datasets crawled from the Internet using crawlers [32]. Of these, 1000 were manually labeled and converted to YOLO dataset format. The 1000 labeled datasets were randomly divided into 700 for training an initial forest fire target detection model. The remaining 300 were a test set to verify the accuracy of the model. The unlabeled 2000 forest fire images were added to the training by the semi-supervised learning method proposed in Section 2.4. This method avoids a lot of manual annotation work.

Figure 1. Schematic diagram of forest fire data set.

2.2. Data Enhancement

In this paper, we used the mosaic online data enhancement method in the training process. The data samples were processed before each epoch training; multiple forest fire images were randomly cropped, scaled, and rotated, and other operations were stitched into one image as training data, which enriches the background of the forest fire dataset. The mosaic online data enhancement method also increases the number of small target samples by randomly reducing the large target samples to small target samples. To a certain extent, this can improve the convergence speed of the model as well as the detection accuracy. The online data enhancement effect is shown in Figure 2.

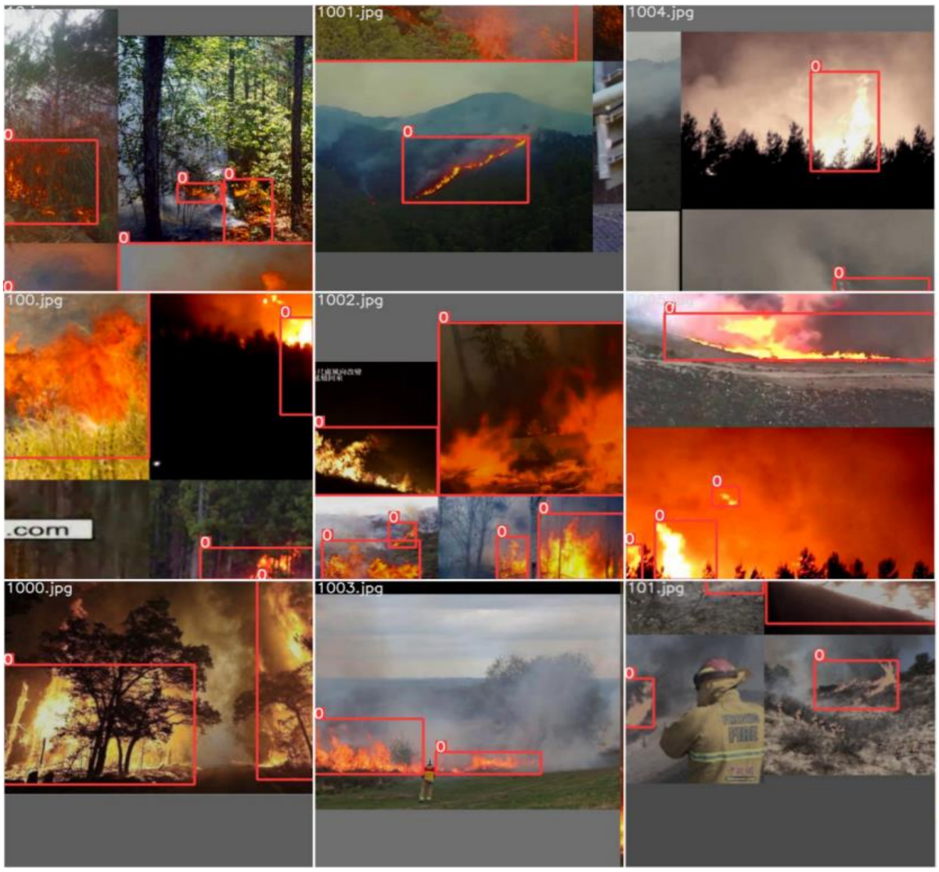


Figure 2. Schematic Diagram of Mosaic Online Enhancement Effect.

2.3. The Proposed Forest Fire Target Detection Model, TCA-YOLO

2.3.1. Basic Frame, YOLOv5

YOLOv5 is an excellent target detection model with high precision and fast detection speed. In this paper, YOLOv5 was selected as the basic framework of the forest fire target detection model. A series of improvements were made to propose an improved forest fire target detection model, TCA-YOLO. The network structure of YOLOv5 is shown in Figure 3. The whole network structure consists of the input, backbone, neck, head (prediction part), etc. The backbone part mainly consists of basic network modules such as CBS, CSP and SPPF, whose main function is to extract image feature information. The CSP module uses a residual network structure to learn more feature information. The SPPF is a spatial pyramid pooling module, which is also the output of the backbone network, and its main function is to convert the extracted feature information of arbitrary size into a fixed-size feature vector. The neck network mainly adopts a feature pyramid structure network based on PAFPN [33], which can transfer the feature information of targets of different sizes. The head part uses three feature layers to predict targets of different scales.

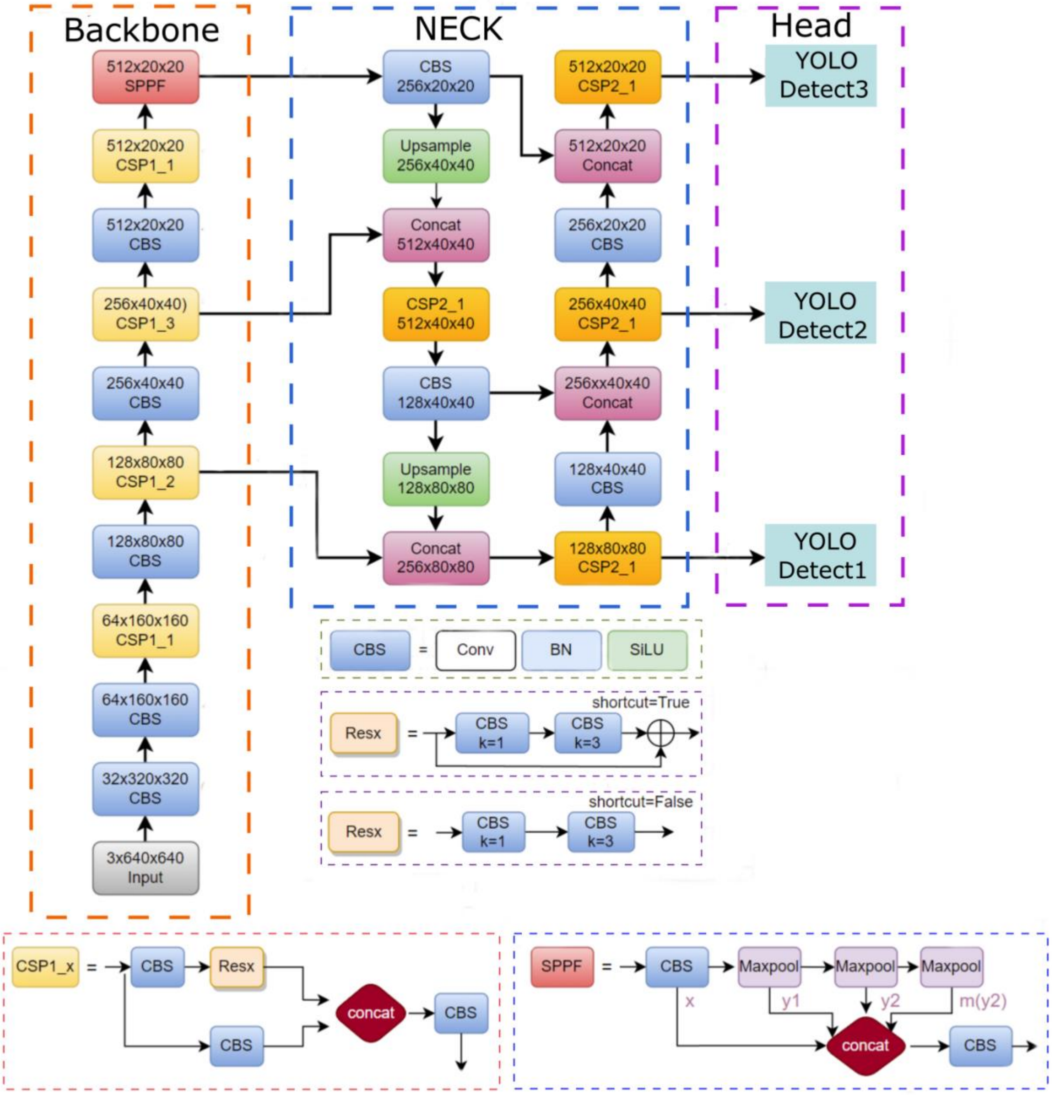


Figure 3. YOLOv5 network architecture diagram.

2.3.2. Using Self-Attention Mechanism to Enhance the Extraction of Global Information of Forest Fires

The main feature extraction network of YOLOv5 is a convolutional neural network (CNN), which can extract the features of forest fires effectively to a certain extent. However, due to the limitations of convolutional operations, the convolutional layer mainly focuses on local information by establishing relationships between neighboring pixels; its perceptual field size is limited, and it has limitations in capturing remote interaction information. Therefore, the pure CNN architecture is not sufficient for the global feature extraction of forest fires. In recent years, self-attention mechanisms have started to be introduced into the field of computer vision in order to overcome the limitations of the inherently local nature of convolutional operations. One of the best performers is Transformer, and Transformer was first applied in natural language processing [34,35]. ViT first applied the Transformer encoder to computer vision, and excellent results were achieved in various target detection and segmentation tasks. The Transformer encoder can extract global image information and rich contextual information. The Transformer encoder structure is shown in Figure 4. Firstly, the image is sliced into patches of a given size and combined with position encoding to obtain a one-dimensional vector, which forms the input to the encoder module. The encoder consists of two layers, a multi-headed attention sub-layer consisting of multiple self-attention mechanisms and an MLP fully connected sub-layer, each using residual connections, and a norm layer before and after the two sub-layers to prevent the overfitting of the network.

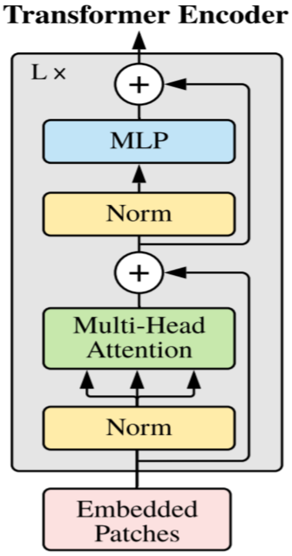


Figure 4. Transformer encoder structure.

The multi-headed attention mechanism is an important part of the encoder module, which can compute multiple sets of data from the input in parallel. The self-attention feature output is calculated by dot product attention, as shown in Equation (1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

V, Q, K are the input features, representing the value vector, query vector, and key vector, respectively, and d_k is the dimension of the input features. The correlation matrix of the vectors is obtained by multiplying with the transposition Q and K. To avoid gradient disappearance caused by the activation function of $softmax$, normalization is used, i.e., dividing by $\sqrt{d_k}$. This is then multiplied with the matrix V to obtain the weighted output. Thus, the multi-headed attention mechanism module not only focuses on the current pixel but also fuses the features of other pixels in the context.

In this paper, the Transformer encoder is embedded in the CSP module of the original feature extraction network of YOLOv5 to form the Transformer module, which forms a CNN+Transformer architecture in the original feature extraction network, which can make up for the fact that the original CNN architecture cannot fully extract the global features of forest fires. The Transformer module has a self-attention mechanism, which can solve the problem of long-distance dependence and obtain global information and contextual information on forest fire targets. Thus, it can enhance the extraction of global features of forest fire targets and achieve better detection results for forest fire targets.

2.3.3. Using Coordinate Attention Mechanism to Focus on Forest Fire Targets

We integrated the Coordinate Attention (CA) mechanism into the forest fire target detection model to further improve the attention on forest fire targets. CA is a lightweight attention mechanism that considers the channel dimension and the spatial dimension in parallel. The CA attention mechanism solves two problems: first, the SE [36] attention mechanism, although excellent, only focuses on the information of channel dimension and does not consider the spatial location information; second, the CBAM [37] attention mechanism focuses on both the channel dimension and spatial dimension, but its spatial note dimension branch attention does not address the long-distance dependence problem. CA not only acquires inter-channel information but also considers the information of direction-dependent location, which can help the model locate and identify forest fire targets accurately. The specific flow of CA is shown in Figure 5.

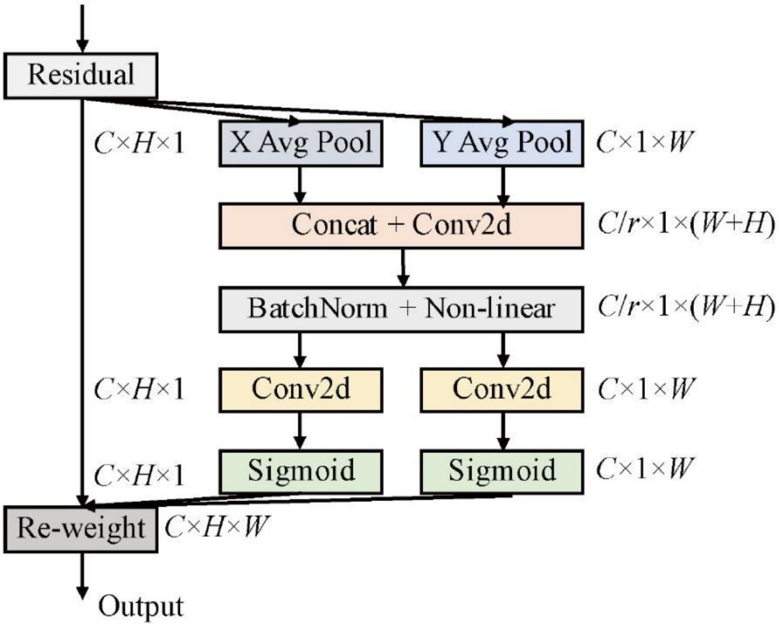


Figure 5. CA Mechanism Schematic.

As shown in the figure above, the CA attention mechanism includes two steps: coordinate information embedding and coordinate attention generating. The output of the previous layer of convolution is used as the input feature map X of the CA attention module for the information embedding operation. Using an average pooling operation of pooling kernel size (H, 1) or (1, W) on the level, each channel in the vertical coordinate direction is encoded to obtain the output characteristic diagram with channel C, height H, and channel C width W. These two transformations lead to a feature map perceived for both spatial orientations. This is very different from the SE attention mechanism, which produces a single-channel attention feature map, as shown in Equations (2) and (3).

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h,i) \tag{2}$$

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{3}$$

The above transformations can well obtain the global perceptual field and encode the location information, then perform the coordinate attention generation operation. The two spatially oriented perceptual feature maps transformed by the above two equations are subjected to the concat join operation, and then the fused feature map of spatial information in high and wide dimensions f is generated by a 1 × 1 convolution F_1, whose feature map size is C/r × 1 × (H + W), as shown in Equation (4), where δ is the nonlinear activation function.

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{4}$$

f^h and f^w are two independent feature maps divided by f along the two spatial dimensions of height and width, and then f^h and f^w are transformed using two convolution kernels of size 1 × 1, F_h and F_w, to obtain feature maps of different spatial dimensions with the same number of channels as the original input feature maps, as shown in Equations (5) and (6), where σ is the Sigmoid activation function.

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{5}$$

$$g^w = \sigma(F_w(f^w)) \tag{6}$$

According to the above calculation, the attention weight in the height direction g^h and the attention weight in the width direction g^w are obtained, and finally, the input feature map X is calculated by multiplicative weighting to obtain a feature map where the attention weights have been rescaled in the height and width dimensions, as shown in Equation (7).

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

We insert the CA attention mechanism module after each CSP module in front of YOLOv5's head. The attention mechanism is computed for each CSP module before the head part of YOLOv5 to adjust the weight of the target location information in the feature map to enhance the extraction of the main location features of the forest fire target. The position information bias is reduced, and the model's attention to the forest fire target is improved.

2.3.4. Multi-Scale Feature Fusion with Adaptively Spatial Feature Fusion

The forest fire targets in the images are often obscured, resulting in some features being missed. The scale of forest fire targets also varies. In addition, the forest fire detection process is also easily interfered with by forest fire-like targets. In order to improve the feature extraction ability of multi-scale forest fire targets and effectively suppress the interference of invalid features in the complex background of forest areas, it is necessary to enhance the fusion of multi-scale features. The unimproved YOLOv5 mainly uses PAFPN. PAFPN cannot fully utilize the features of forest fire targets at different scales because PAFPN simply transforms the feature maps to the same size and then sums them up. Therefore, in order to perform feature fusion more rationally, we integrate adaptively spatial feature fusion (ASFF) technology, which can enhance the fusion of multi-scale features, enhance the expression of relevant features and reduce the interference of irrelevant features. It enables the model to learn how to retain only useful information in space and filter out useless information from other layers. ASFF adaptively learns the corresponding weights for each layer′s feature map, multiplies the feature map with the obtained weights and then fuses them. The conflicting information can be spatially filtered to suppress the inconsistency in gradient back-propagation, which can solve the problem of conflicting image spatial information in the process of the multi-scale feature fusion of forest fire targets.

Figure 6 shows how to perform feature fusion. Taking ASFF-3 as an example, Level 1, Level 2, and Level 3 are the three feature layers output from the neck part of the YOLO network. x^1, x^2, and x^3 are the features of Level 1, Level 2, and Level 3, respectively. We then multiply the parameters α^3, β^3, and r^3 and sum them up to obtain the feature ASFF-3 after feature fusion. This process is shown in Equation (8).

$$y_{ij}^{l} = \alpha_{ij}^{l} * x_{ij}^{1\to l} + \beta_{ij}^{l} * x_{ij}^{2\to l} + r_{ij}^{l} * x_{ij}^{3\to l} \tag{8}$$

where y_{ij}^{l} denotes the new feature map obtained by ASFF. a_{ij}^{l}, β_{ij}^{l}, r_{ij}^{l} are the weight parameters of the three feature layers, which are made to satisfy Equation (9) by the Softmax function:

$$a_{ij}^{1} + \beta_{ij}^{1} + r_{ij}^{1} = 1 \tag{9}$$

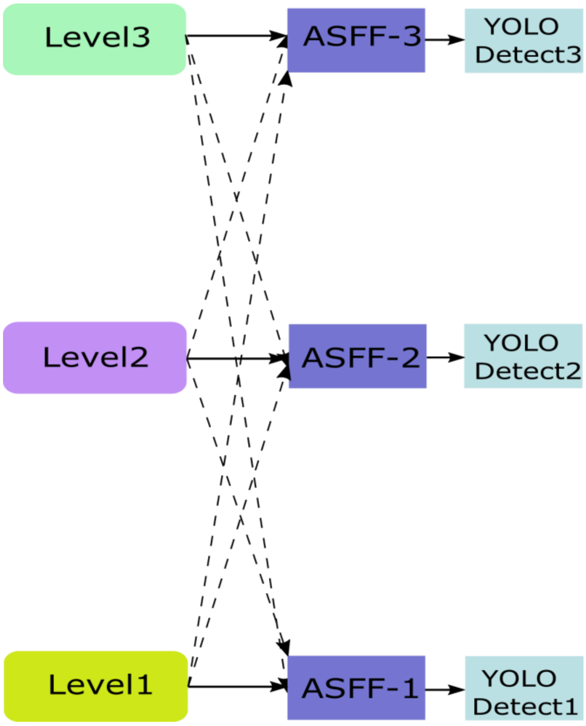


Figure 6. Schematic diagram of adaptively spatial feature fusion.

α_{ij}^{1} satisfies Equation (10):

$$a_{ij}^{1} = \frac{e^{\lambda_{aij}^{1}}}{e^{\lambda_{aij}^{1}} + e^{\lambda_{\beta ij}^{1}} + e^{\lambda_{\gamma ij}^{1}}} \tag{10}$$

a_{ij}^{l}, β_{ij}^{l}, and r_{ij}^{l} range from 0 to 1. $x_{ij}^{1\to l}$, $x_{ij}^{2\to l}$, and $x_{ij}^{3\to l}$ denote the features of layers 1, 2, and 3, respectively.

Since the summation operation is performed, it is necessary to ensure that Level 1, Level 2, and Level 3 have the same features and the same number of channels in each layer, so up-sampling or down-sampling is needed to adjust them. Taking ASFF-3 as an example, two rounds of upsampling are needed to make the results of Level 1 and Level 2 have the same size as level 3. First, Level 1 and Level 2 need to be compressed by a 1 × 1 convolution to the same number of channels as Level 3, and then quadruple and double upsamplings are performed using interpolation to obtain the same dimension as Level 3. Finally, the summation is performed.

In this paper, the integration of the ASFF technique into the forest fire detection model can improve the multi-scale feature fusion of forest fire targets, which can more fully utilize the underlying fine-grained features as well as the semantic information of high-level features. It enhances the ability of the model to represent the forest fire target features in the complex environment of the forest area and effectively suppresses the interference of invalid features in the complex background of the forest area for forest fire detection, thus improving detection accuracy.

2.3.5. The General Framework of the Proposed Forest Fire Target Detection Model TCA-YOLO

In this paper, a forest fire target detection model, TCA-YOLO, is proposed with the following improvements using the framework of YOLOv5. Figure 7 shows the overall framework of TCA-YOLO. Firstly, the Transformer module (the module obtained by embedding the Transformer encoder in the original CSP module of YOLOv5) was integrated into the original feature extraction network of YOLOv5 to form a feature extraction network with CNN+ Transformer architecture. The self-attention mechanism of Transformer enables the model to obtain a larger receptive field and more contextual information, which enhances the model's global feature extraction of forest fire targets. Secondly, in order to make the model imitate human vision to selectively focus on the forest fire target and ignore other invalid features, this paper inserted the Coordinate Attention mechanism module before the Head of YOLOv5 so that the CSP module of each one before the head of YOLOv5 performs an attention mechanism calculation to adjust the weight of the target location information in the feature map and enhance the model's attention to forest fire targets. Finally, in order to make up for the deficiencies in the original multi-scale feature fusion of YOLOv5, this paper uses the ASFF technique in the head part of the model to improve the multi-scale feature fusion of forest fire targets. It suppresses the interference of invalid features in the complex environment of forest areas for forest fire detection and improves the accuracy of TCA-YOLO in detecting forest fire targets in the complex environment of forest areas.

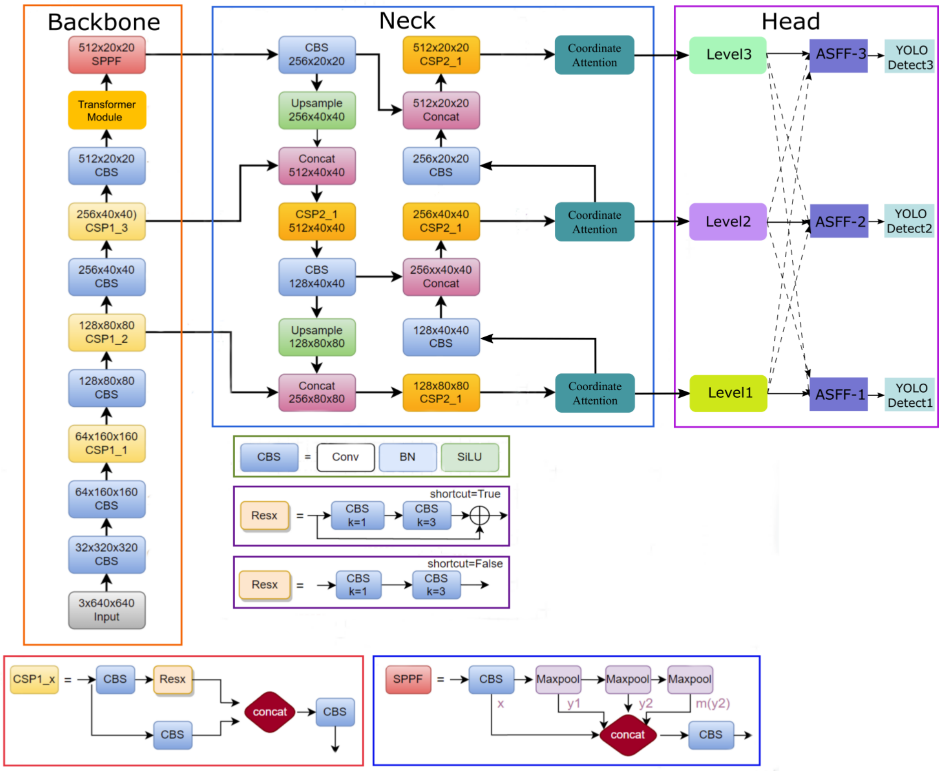


Figure 7. The general framework diagram of the proposed forest fire target detection model, TCA-YOLO.

2.4. The Proposed Semi-Supervised Learning Method for Training TCA-YOLO

Although the proposed TCA-YOLO has greatly improved YOLOv5, its performance also depends on the number of forest fire samples used for training. Manually labeling a large number of forest fire images takes a lot of time and requires professionals to do the labeling. In addition, although a certain number of labeled forest fire samples exist on the Internet, there are almost hundreds of times more unlabeled forest fire samples than labeled ones, and various forest fire videos can also be framed into unlabeled forest fire samples for exploitation. Making full use of the unlabeled samples can further improve the accuracy of the model. Therefore, this paper proposes a semi-supervised strategy to train the proposed forest fire target detection model. Only a small number of manually labeled forest fire samples are needed, and the remaining unlabeled samples are automatically labeled using the proposed method, saving a lot of manual labeling work and improving the accuracy of TCA-YOLO through a reasonable training strategy.

We found through preliminary experiments that using different numbers of manually labeled forest fire datasets has a certain impact on the accuracy of TCA-YOLO. We trained TCA-YOLO with 50 to 1000 manually labeled forest fire datasets, respectively, and found that the accuracy of the model improves with the increase in the number of samples. When the number of manually labeled forest fire samples reaches 700, the average accuracy of TCA-YOLO reaches 75%, but it starts to converge slowly thereafter. Therefore, using 700 manually labeled forest fire samples can stabilize the accuracy of the model at more than 75% and minimize the manual labeling effort.

As shown in Figure 8, 700 unlabeled forest fire samples are first selected from a large number of unlabeled forest fire samples with high training values for the model to be manually and accurately labeled. This small number of manually labeled forest fire samples is used to train TCA-YOLO to derive an initial model weight, M. Forest fire detection with weight M can achieve certain accuracy, but the accuracy needs to be further improved. The supervised model weight M is used to predict the unlabeled forest fire dataset. If the forest fire target can be detected, the prediction probability (confidence), target category and boundary box for locating the forest fire target will be output. $\mathrm{b_x}$, $\mathrm{b_y}$, $\mathrm{b_h}$, and $\mathrm{b_w}$ are parameterized representations of the boundary box of the detected forest fire target. The label file obtained from the manually labeled forest fire dataset contains only the positioning coordinates of the forest fire target bounding box and the judgment result of whether it is a forest fire target or not, which can also be obtained when the unlabeled forest fire dataset is predicted using weight M. Therefore, we wrote a program to automatically save the fire target bounding box information obtained when the unlabeled fire dataset is predicted by the weight M and save the information in the YOLO dataset format. Automatic labeling of forest fire datasets is thus achieved. Then, the labels with high confidence are screened and added to the training set as pseudo-labels to train TCA-YOLO to obtain the new model weights M′. We replace M with M′ and repeat the above steps until the model effect does not appear to be boosted, or the unlabeled dataset is empty. If new unlabeled forest fire datasets become available in the future, the strategy described above can also be used to automatically label and add labeled datasets to the training to continuously improve the accuracy of TCA-YOLO. The specific algorithm flow for training TCA-YOLO using the proposed strategy is shown in Algorithm 1.

Algorithm 1: The semi-supervised algorithm for training TCA-YOLO

```
The total number of forest fire samples is $I + J$, where the total number of manually labeled
    samples is $I$, the number of unlabeled samples is $J$, and $J >> I$. $D_L = \left\{\left(x^i, y^i\right)\right\}_{i=1}^{I}$ is the set
    of labeled forest fire datasets, $D_U = \left\{\left(x^i\right)\right\}_{j=1}^{J}$ is the set of unlabeled forest fire datasets,
    and x and $y$ are the true and predicted values of forest fire datasets, respectively.
Pre-training: Using a small number of manually labeled forest fire datasets $D_L$ to train the
    proposed forest fire target detection model TCA-YOLO in this paper yields an initial
    supervised model weight $M$.
Input: $D_L$, $D_U$, and the initial weight $M$ of TCA-YOLO.
Output: The final weight file $F$ for the forest fire target detection model TCA-YOLO.
Repeat:
    1. Predictions are made on the unlabeled forest fire dataset $D_U$ using $M$ to obtain the set
       of pseudo-labeled dataset $D_P = \left\{\left(x^j, y^j\right)\right\}_{j=1}^{J}$ and the predicted probability
       (confidence) $p^j$ corresponding to each label. $J$ is the total number of all pseudo-labeled
       samples.
    2. The samples with predicted probability $p^j >$ predetermined probability $P$ in $D_P$ are
       filtered as high-confidence samples to be used as pseudo-labeled samples for training.
       The set of pseudo-labeled samples for training is denoted as $D_{PT}= \{(x^e, y^e)\}_{e=1}^{E}$. $E$ is
       the total number of samples with confidence higher than the predetermined value $P$.
    3. Add $D_{PT}$ to $D_L$. The original dataset of $D_{PT}$ (unlabeled) is noted as $D_{UP} =\{(x^e)\}_{e=1}^{E}$.
       Remove $D_{UP}$ from $D_U$. Update $D_L$ and $D_U$. Clear $D_{P'}$, $D_{PT}$, and $D_{UP}$.
    4. Train TCA-YOLO with the new $D_L$ to obtain the new model weight $M'$ and replace $M$
       with $M'$.
Until: No forest fire samples are available in $D_U$, or the accuracy of TCA-YOLO is no longer
    improved.
Return: $F$
End.
```

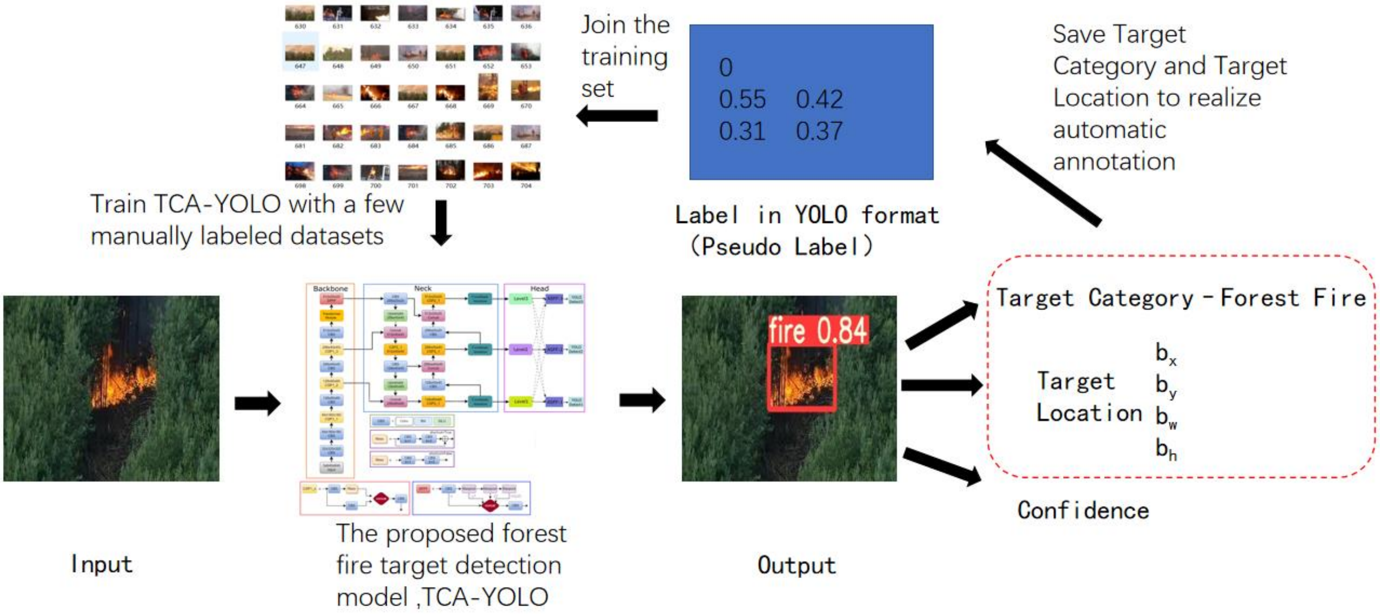

Join the training set
0
0.55 0.42
0.31 0.37
Save Target Category and Target Location to realize automatic annotation
Train TCA-YOLO with a few manually labeled datasets
Label in YOLO format (Pseudo Label)
Backbone
Neck
Head
fire 0.84
Target Category - Forest Fire
Target Location
b_x
b_y
b_w
b_h
Confidence
Input
The proposed forest fire target detection model ,TCA-YOLO
Output

Figure 8. Schematic diagram of automatic annotation of forest fire datasets.

2.5. Experimental Environment and Parameters

In this paper, model training and testing were performed on a Windows 64-bit operating system. The CPU of the computer was an AMD R7 5800H with 32G of running memory, and the GPU was an NVIDIA GeForce RTX 3060 with 6G of video memory. The model was programmed in Python 3.8 and built and improved using the deep learning framework Pytorch, with an AMD R7 5800H CPU, 32G of running memory, and an NVIDIA GeForce RTX 3060 GPU with 6G of video memory. The CUDA version is 11.1, and the CUDNN version is 8.0.5.

The input images during training were uniformly adjusted to 640 × 640 pixels, and the batch size was 8.

3. Results

3.1. Model Evaluation and Ablation Experiment

The detection effectiveness of the model was evaluated by the precision, recall, and average precision, using FPS to evaluate the detection speed. Average precision (AP) is the integral of the P-R curve constructed from precision (P) as the vertical axis and recall (R) as the horizontal axis. As shown in Figure 9, the value of AP is the area under the P-R curve, and the closer the P-R curve is to the upper right, the better the model performance is.

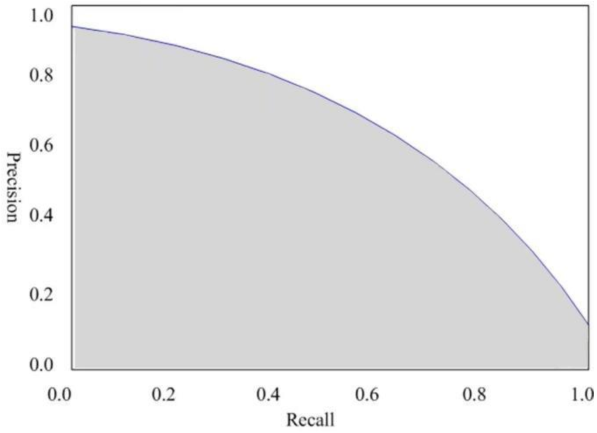


Figure 9. P-R curve diagram.

Recall reflects the ability of the forest fire detection model to find positive sample targets; precision reflects the ability of the model to classify samples, and mean average precision reflects the overall performance of the model to detect and classify targets. The mean average precision (mAP) represents the average of the average precision of all categories. Its calculation formulas are as follows (Equations (11)–(14)).

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$AP = \int_0^1 P(R)dR \tag{13}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{14}$$

TP represents the number of positive samples detected correctly. FP represents the number of positive samples detected incorrectly. FN represents the number of negative samples detected incorrectly. N represents the number of categories of data.

In this paper, IoU (intersection over union) is introduced to calculate the average accuracy of the model in accordance with the evaluation requirements of the target detection model. IoU is used to calculate the ratio of intersection over the union between the predicted bounding box and the true bounding box.

$mAP_{0.5}$ is the mean mAP obtained by evaluating the forest target detection model at an IoU threshold of 0.5. $mAP_{0.5:0.95}$ is the mean mAP obtained by evaluating the model at different IoU thresholds (0.5 to 0.95, step 0.05), which is a more stringent indicator of model accuracy.

To verify whether TCA-YOLO can detect forest fire targets in real time, the detection speed of TCA-YOLO was evaluated using FPS (frames per second), i.e., the number of forest fire images that can be processed within each second.

We used the proposed semi-supervised strategy to train the unmodified YOLOv5 and TCA-YOLO. We used the same test set to verify the detection accuracy of YOLOv5 and TCA-YOLO, and we used the above indicators to perform the evaluation. An ablation experiment was performed to verify the effectiveness of each module of TCA-YOLO. The model evaluation results and ablation experiment results are shown in Table 1.

Table 1. Results of the experiment.

Model	$mAP_{0.5}$ (%)	$mAP_{0.5:0.95}$ (%)	P (%)	R (%)	FPS
YOLOv 5	79.26	54.38	78.98	79.19	55.3
YOLO v5 + Transformer	81.56	56.01	82.45	81.23	54.5
YOLO v5 + Transformer + CA	82.78	56.45	84.37	83.01	54.0
YOLO v5 + Transformer + CA + ASFF (TCA-YOLO, ours)	84.56	57.38	85.26	83.37	53.7

The evaluation shows that the detection accuracy of TCA-YOLO for forest fire targets is significantly improved after the improvement; in particular, $mAP_{0.5}$ is 5.3 higher than YOLOv5. Although the FPS (the test set contains images of different resolutions) is slightly decreased, it still reaches 53.7, i.e., TCA-YOLO can detect 53.7 forest fire images per second. The video of real-time surveillance is usually 25 frames per second to 30 frames per second, so the detection speed of TCA-YOLO is much higher than the requirements for real-time detection.

Since the pixel sizes of the images in the test set vary, we also tested the detection speed of TCA-YOLO for forest fire images of different resolutions, as shown in Table 2 below. It can be seen that the detection speed of TCA-YOLO meets the requirement of real-time detection even for high-definition images.

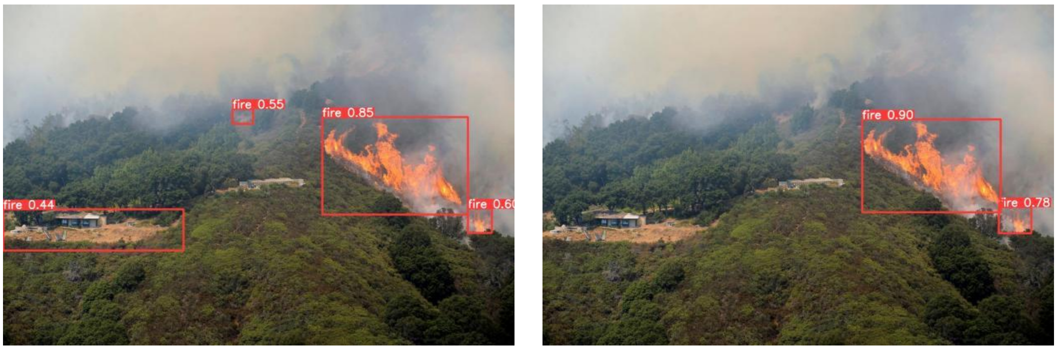
Table 2. The detection speed of TCA-YOLO for images of different resolutions.

Resolution	FPS
256 × 400	73.93
334 × 500	69.43
720 × 1280	55.82
2160 × 3840	46.51

3.2. Forest Fire Target Detection Performance and Comparative Analysis

By comparing the forest fire target detection results of the proposed model, TCA-YOLO, with the unimproved YOLOv5, we find that the detection effect of TCA-YOLO is greatly improved over the original unimproved YOLOv5. In particular, the resistance to complex background interference and the ability to extract the global information of forest fire targets are much improved. The number of missed and false detections of forest fire targets is also less. The focus on forest fire targets is higher, and the detection of small target forest fires is better. This further validates the effectiveness of each module of TCA-YOLO. Some of the identification results are shown below.

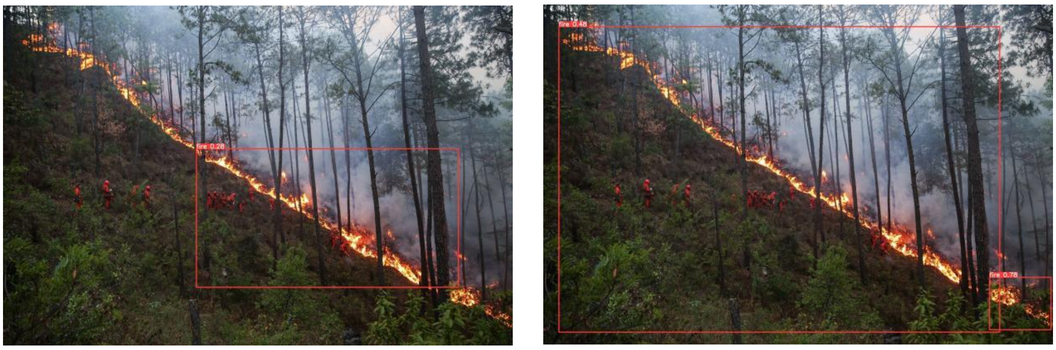
As shown in Figure 10, when disturbed by similar forest fire targets, YOLOv5 incorrectly treats forest fire-like targets as forest fire targets, while TCA-YOLO is able to distinguish them.



(**a**) YOLOv5 (**b**) TCA-YOLO (ours)

Figure 10. Comparison of anti-interference capability. (**a**) Two targets were misdetected by YOLOv5. (**b**) Our model has no misdetected targets.

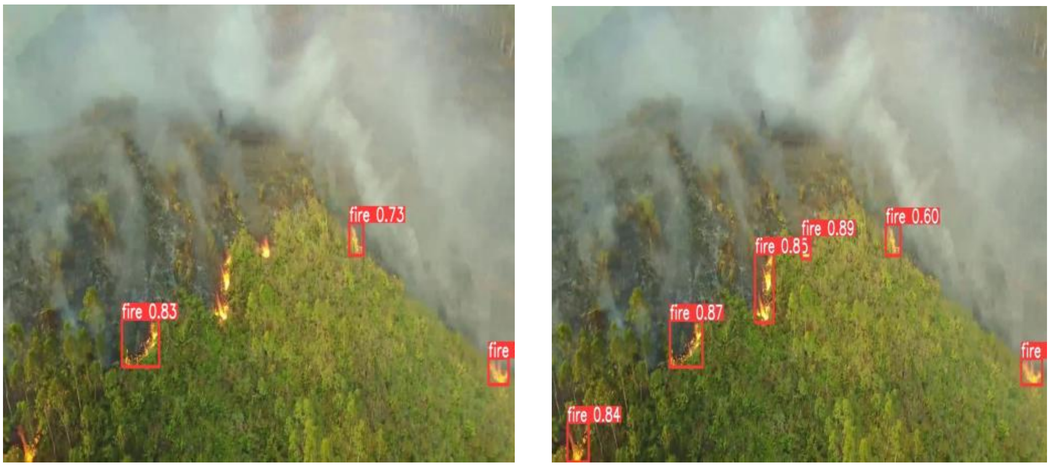
As shown in Figure 11, YOLOv5 fails to effectively extract the global information of the forest fire target in the figure; only the local area of the forest fire target is framed. It cannot effectively localize the forest fire target. In contrast. TCA-YOLO can effectively extract the global information of the forest fire target, and the receptive field is much larger.



(**a**) YOLOv5 (**b**) TCA-YOLO (ours)

Figure 11. Comparison of global information extraction ability for forest fires. (**a**) YOLOv5 locates only a local area of the fire target. (**b**) Our model locates the global area of the fire target.

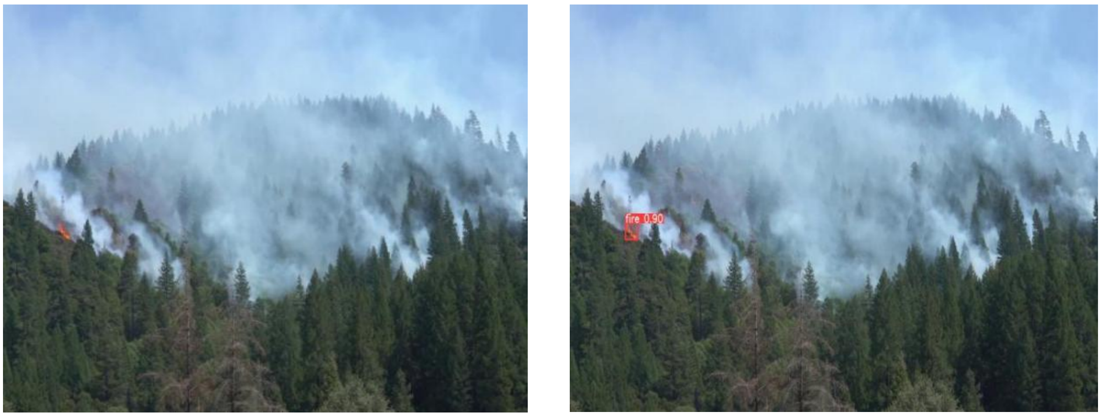
As shown in Figure 12, YOLOv5 has multiple missed targets when facing forest fire targets at different scales, while TCA-YOLO has no missed targets.



(**a**) YOLOv5 (**b**) TCA-YOLO (ours)

Figure 12. Comparison of whether forest fire targets will be missed. (**a**) YOLOv5 misses multiple fire targets. (**b**) Our model does not have missed detections.

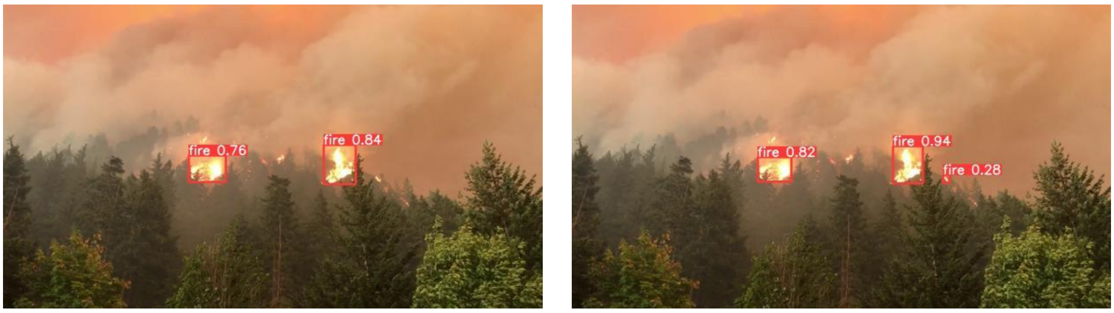
The timely detection of initial small target forest fires is crucial for forest fire detection. As shown in Figure 13 below, YOLOv5 fails to detect a small target forest fire at a long distance, while TCA-YOLO can accurately detect it in the image and has a confidence level of 0.90.



(**a**) YOLOv5 (**b**) TCA-YOLO (ours)

Figure 13. Comparison of small target forest fire detection. (**a**) YOLOv5 fails to detect the small target forest fire at a distance. (**b**) Our model can detect the distant small target forest fire.

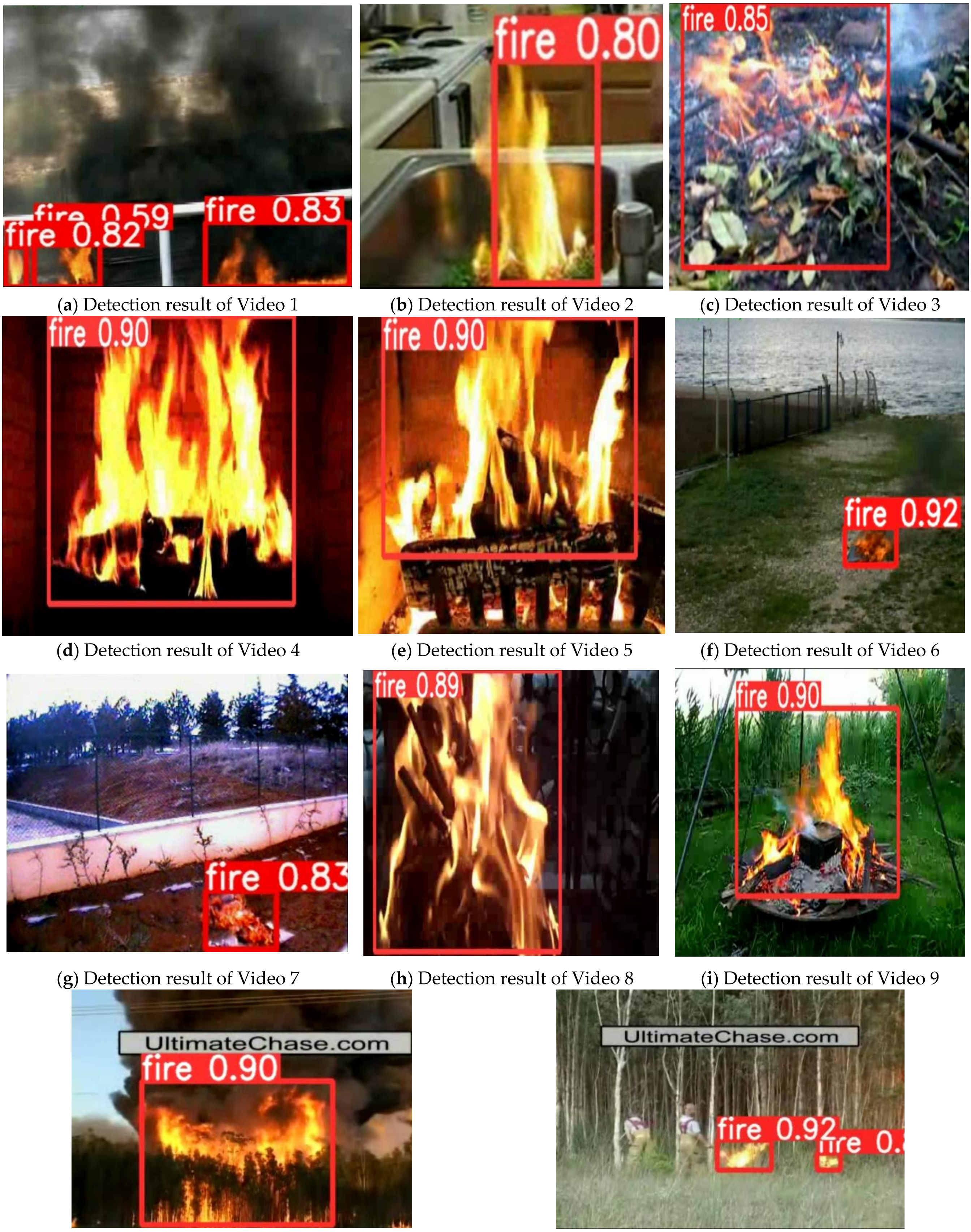
As shown in Figure 14 below, TCA-YOLO detects better than YOLOv5 in most cases. However, similar to YOLOv5, TCA-YOLO also has some very small fire points that cannot be detected because the features of these small fire points are not very similar to the flame features in most forest fire datasets (including our self-built dataset and most publicly available datasets), which is something we need to improve in the future.



(**a**) YOLOv5 (**b**) TCA-YOLO (ours)

Figure 14. Comparison of the detection effect of YOLOv5 and TCA-YOLO. (**a**) YOLOv5's performance in detecting small fire points. (**b**) Our model's performance in detecting small fire points.

In this paper, we also tested the detection effect of TCA-YOLO on the publicly available fire video dataset VisiFire [38] and compared it with the mainstream target detection models. A total of 11 positive sample videos of fire in forest scenes and other scenes are included in VisiFire. The slices of the proposed TCA-YOLO detection results for the 11 videos are shown in Figure 15a–k.



(**a**) Detection result of Video 1 (**b**) Detection result of Video 2 (**c**) Detection result of Video 3

(**d**) Detection result of Video 4 (**e**) Detection result of Video 5 (**f**) Detection result of Video 6

(**g**) Detection result of Video 7 (**h**) Detection result of Video 8 (**i**) Detection result of Video 9

(**j**) Detection result of Video 10 (**k**) Detection result of Video 11

Figure 15. Slicing of the public fire video dataset VisiFire detection results. (**a**–**k**) are slices of the results of detecting the 1st to 11th fire videos in VisiFire, respectively.

We evaluate the detection accuracy of TCA-YOLO and other mainstream target detection models for each frame of the open fire video dataset VisiFire using TPR (true positive rate) and FNR (false negative rate), and the results are shown in Table 3 below. It can be seen that the detection accuracy of TCA-YOLO is ahead of other mainstream target detection models.

Table 3. Performance when detecting open forest fire video dataset VisiFire.

Video	Total Frames	TCA-YOLO (Ours)		YOLO		Fast R-CNN [39]		EffcientDet [40]	
		TPR (%)	FNR (%)	TPR (%)	FNR (%)	TPR (%)	FNR (%)	TPR (%)	FNR (%)
Video 1	293	97.43	2.57	95.23	4.77	94.32	5.68	94.75	5.25
Video 2	510	97.98	2.02	96.99	3.01	93.66	6.34	94.01	5.99
Video 3	318	98.78	1.22	98.43	1.57	96.13	3.87	95.78	4.22
Video 4	1655	97.79	2.21	95.49	4.51	94.99	5.01	95.01	4.99
Video 5	2406	97.21	2.79	96.23	3.77	93.87	6.13	94.02	5.98
Video 6	258	98.23	1.77	97.31	2.69	95.95	4.05	94.76	5.24
Video 7	547	97.35	2.65	96.01	3.99	94.18	5.82	94.01	5.99
Video 8	513	98.19	1.81	96.63	3.37	95.11	4.89	95.01	4.99
Video 9	663	98.81	1.19	98.79	1.21	95.33	4.67	94.84	5.16
Video 10	235	98.35	1.65	97.01	2.99	96.01	3.99	95.87	4.13
Video 11	178	98.24	1.76	96.11	3.89	95.03	4.97	94.88	5.12
Average	728.3	98.03	1.97	96.75	3.25	94.96	5.04	94.81	5.19

4. Discussion and Conclusions

Forest fires occur frequently around the world, causing serious economic losses and human casualties. Therefore, forest fire detection is particularly important. The use of images for forest fire detection has become a mainstream trend. The intelligent recognition of forest fire images is mainly based on the deep learning technique of CNNs. Although the deep learning techniques based on CNNs have solved the problem of the intelligent recognition of forest fires to a certain extent, the extracted features of forest fire images are only limited to local regions due to the limitations of convolutional operations and their lack of global modeling capabilities. Therefore, CNNs have difficulty capturing global and contextual information on forest fire targets, which greatly affects the performance of forest fire detection. CNN-based forest fire target models also do not pay enough attention to forest fire targets and are susceptible to the interference of complex backgrounds in forest areas, treating invalid forest fire-like features as forest fire targets. In addition, CNN-based forest fire detection techniques require a large number of manually labeled forest fire datasets. This requires a lot of manual labeling work by professionals and is very time-consuming.

Therefore, this paper proposes a solution to these problems and designs an improved forest fire target detection model, TCA-YOLO, using the convolutional neural network-based target detector YOLOv5 as the basic framework and a series of improvements. Firstly, to make up for the deficiency of CNNs in the global feature extraction of forest fires, we use a Transformer encoder with a self-attention mechanism and CNN junction as the feature extraction network, which enhances the extraction of the global information of forest fire targets and has a more powerful receptive field to obtain more contextual information. Secondly, in order to enhance the model's focus on forest fire targets, we integrate the Coordinate Attention (CA) mechanism in the neck part of YOLOv5. CA not only obtains inter-channel information but also considers the information of direction-related location, which helps the model to better locate and identify forest fire targets. In addition, we also integrate the adaptively spatial feature fusion (ASFF) technique. ASFF improves the multi-scale fusion of forest fire features by adaptive methods to adjust the fusion ratio between different feature layers. It thus effectively suppresses the interference of invalid features in the complex background of forest areas for forest fire detection and further improves the detection accuracy of TCA-YOLO. Finally, this paper adopts semi-supervised learning to train the proposed forest fire target detection model, TCA-YOLO, using only a small number of manually labeled forest fire datasets to obtain the initial model weights, and the rest of the unlabeled datasets are automatically labeled using an automatic labeling method to filter out those with high confidence labels as pseudo-labels to join the training set to retrain TCA-YOLO. The model then continuously iterates to improve the accuracy step by step, saving a large amount of manual annotation time.

The average accuracy of forest fire target detection ($mAP_{0.5}$) reaches 84.56, which is 5.3 higher than the unimproved YOLOv5. The FPS reaches 53.7, which means TCA-YOLO can quickly detect forest fire targets in real-time. By comparing and analyzing the detection results, TCA-YOLO also outperformed the unimproved YOLOv5 in detecting forest fire targets in different scenarios. The ability of TCA-YOLO to extract global information on forest fire targets was much improved, and it could locate forest fire targets more accurately. TCA-YOLO misses fewer forest fire targets and is less likely to be interfered with by forest fire-like targets. The focus on forest fire targets is higher, and the detection of small target forest fires is better. If new unlabeled forest fire datasets are available, the unlabeled datasets can also be automatically labeled to join the training using the strategy proposed in this paper to continuously improve the detection accuracy of TCA-YOLO.

The proposed forest fire target detection model is mainly used for the real-time automatic detection of forest fire targets on watchtowers, forest video surveillance equipment and UAVs deployed in place of manual inspection. The model can mark the bounding box of the fire target in the image to determine the location of the fire target in the image. Since these video surveillance deployments capture small forest fire targets at long distances and take images with low numbers of pixels, it is not possible to segment the fire targets pixel-by-pixel with flame edges (the function of the semantic segmentation model) in most cases, so we only design a forest fire target detection model instead of a semantic segmentation model in this stage of work. Although it outputs a rectangular bounding box, it is as capable of localizing forest fire targets as the semantic segmentation model. However, both the semantic segmentation and target detection models require excellent feature extraction networks, and usually, the semantic segmentation model also has a target detection branch and shares the same feature extraction network with the target detection branch. Although this model is only a target detection model at this stage, the improved feature extraction network has good performance. In future work, we will continue to use the feature extraction network of the proposed model and add a semantic segmentation branch to share the same excellent feature extraction network with the proposed model to further improve the functionality of the model.

Author Contributions: J.L. devised the programs and drafted the initial manuscript. H.L. and F.W. designed the project and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development plan of Jiangsu Province (Grant No. BE2021716), the Jiangsu Modern Agricultural Machinery Equipment and Technology Demonstration and Promotion Project (NJ2021-19), Nanjing modern agricultural machinery equipment and technological innovation demonstration projects (NJ [2022]09), the National Natural Science Foundation of China (32101535), and the Jiangsu Postdoctoral Research Foundation (2021K112B).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chen, S.J.; Hovde, D.C.; Peterson, K.A.; Marshall, A.W. Fire detection using smoke and gas sensors. *Fire Saf. J.* **2007**, *42*, 507–515. [CrossRef]
2. Yu, L.; Wang, N.; Meng, X. Real-time forest fire detection with wireless sensor networks. In Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing, Zhangjiajie, China, 2–4 August 2005; pp. 1214–1217.
3. Zhang, J.; Li, W.; Yin, Z.; Liu, S.; Guo, X. Forest fire detection system based on wireless sensor network. In Proceedings of the 4th IEEE Conference on Industrial Electronics and Applications, Xi'an, China, 25–27 May 2009; pp. 520–523.
4. Lee, B.; Kwon, O.; Jung, C.; Park, S. The development of UV/IR combination flame detector. *J. KIIS* **2001**, *16*, 1–8.
5. Fernandes, A.M.; Utkin, A.B.; Lavrov, A.V.; Vilar, R.M. Development of neural network committee machines for automatic forest fire detection using lidar. *Pattern Recognit.* **2004**, *37*, 2039–2047. [CrossRef]
6. Celik, T.; Demirel, H.; Ozkaramanli, H.; Uyguroglu, M. Fire detection using statistical color model in video sequences. *J. Vis. Commun. Image Represent.* **2007**, *18*, 176–185. [CrossRef]
7. Habiboglu, Y.H.; Guenay, O.; Cetin, A.E. Covariance matrix-based fire and flame detection method in video. *Mach. Vis. Appl.* **2012**, *23*, 1103–1113. [CrossRef]
8. Kong, S.G.; Jin, D.; Li, S.; Kim, H. Fast fire flame detection in surveillance video using logistic regression and temporal smoothing. *Fire Saf. J.* **2016**, *79*, 37–43. [CrossRef]
9. Dimitropoulos, K.; Barmpoutis, P.; Grammalidis, N. Spatio-Temporal Flame Modeling and Dynamic Texture Analysis for Automatic Video-Based Fire Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 339–351. [CrossRef]
10. Yin, Z.; Wan, B.; Yuan, F.; Xia, X.; Shi, J. A Deep Normalization and Convolutional Neural Network for Image Smoke Detection. *IEEE Access* **2017**, *5*, 18429–18438. [CrossRef]
11. Zhang, Q.X.; Lin, G.H.; Zhang, Y.M.; Xu, G.; Wang, J.J. Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images. *Procedia Eng.* **2018**, *211*, 441–446. [CrossRef]
12. Avula, S.B.; Badri, S.J.; Gokul, R.P. A Novel Forest Fire Detection System Using Fuzzy Entropy Optimized Thresholding and STN-based CNN. In Proceedings of the 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, 7–11 January 2020.
13. Wang, G.; Zhang, Y.; Qu, Y.; Chen, Y.; Maqsood, H. Early Forest Fire Region Segmentation Based on Deep Learning. In Proceedings of the 2019 Chinese Control and Decision Conference (CCDC), Nanchang, China, 3–5 June 2019.
14. Jiao, Z.; Zhang, Y.; Mu, L.; Xin, J.; Jiao, S.; Liu, H.; Liu, D. A YOLOv3-based Learning Strategy for Real-time UAV-based Forest Fire Detection. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, 22–24 August 2020.
15. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
16. 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]
17. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Cham, Switzerland, 2015.
18. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 815–825.
19. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 213–229.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
21. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-End Object Detection with Dynamic Attention. In Proceedings of the International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; Available online: https://openaccess.thecvf.com/content/ICCV2021/papers/Dai_Dynamic_DETR_End-to-End_Object_Detection_With_Dynamic_Attention_ICCV_2021_paper.pdf (accessed on 1 December 2022).
22. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor DETR: Query Design for Transformer-Based Object Detection. *arXiv* **2021**, arXiv:2109.07107.
23. Ultralytics-Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 1 October 2022).
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
25. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
26. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
27. Du, X.; Cai, Y.; Wang, S.; Zhang, L. Overview of deep learning. In Proceedings of the 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, Wuhan, China, 11–13 November 2016.

28. Lu, K.; Huang, J.; Li, J.; Zhou, J.; Chen, X.; Liu, Y. MTL-FFDET: A Multi-Task Learning-Based Model for Forest Fire Detection. *Forests* **2022**, *13*, 1448. [CrossRef]
29. Guan, Z.; Miao, X.; Mu, Y.; Sun, Q.; Ye, Q.; Gao, D. Forest Fire Segmentation from Aerial Imagery Data Using an Improved Instance Segmentation Model. *Remote. Sens.* **2022**, *14*, 3159. [CrossRef]
30. Lu, K.; Xu, R.; Li, J.; Lv, Y.; Lin, H.; Liu, Y. A Vision-Based Detection and Spatial Localization Scheme for Forest Fire Inspection from UAV. *Forests* **2022**, *13*, 383. [CrossRef]
31. Lee, D.H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. 2013. Available online: https://www.researchgate.net/publication/280581078_Pseudo-Label_The_Simple_and_Efficient_Semi-Supervised_Learning_Method_for_Deep_Neural_Networks (accessed on 1 May 2022).
32. BoWFire Dataset. Available online: https://bitbucket.org/gbdi/bowfifire-dataset/downloads/ (accessed on 1 May 2022).
33. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
35. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
36. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *Proc. IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.
38. Çetin, A.E. Computer Vision Based Fire Detection Dataset. 2014. Available online: http://signal.ee.bilkent.edu.tr/VisiFire/ (accessed on 1 December 2022).
39. Girshick, R. Fast R-CNN. In Proceedings of the International Conference on Computer Vision, IEEE Computer Society, Santiago, Chile, 7–13 December 2015.
40. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.