


## Article

# A Small-Target Forest Fire Smoke Detection Model Based on Deformable Transformer for End-to-End Object Detection

Jingwen Huang <sup>1</sup>, Jiashun Zhou <sup>1</sup>, Huizhou Yang <sup>1</sup>, Yunfei Liu <sup>1,\*</sup> and Han Liu <sup>2</sup><sup>1</sup> College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China<sup>2</sup> College of Letters and Science, University of Wisconsin-Madison, Madison, WI 53706, USA

\* Correspondence: lyf@njfu.com.cn; Tel.: +86-139-1389-5117

**Abstract:** Forest fires have continually endangered personal safety and social property. To reduce the occurrences of forest fires, it is essential to detect forest fire smoke accurately and quickly. Traditional forest fire smoke detection based on convolutional neural networks (CNNs) needs many hand-designed components and shows poor ability to detect small and inconspicuous smoke in complex forest scenes. Therefore, we propose an improved early forest fire smoke detection model based on deformable transformer for end-to-end object detection (deformable DETR). We use deformable DETR as a baseline containing the best sparse spatial sampling for smoke with deformable convolution and relation modeling capability of the transformer. We integrate a Multi-scale Context Contrast Local Feature module (MCCL) and a Dense Pyramid Pooling module (DPPM) into the feature extraction module for perceiving features of small or inconspicuous smoke. To improve detection accuracy and reduce false and missed detections, we propose an iterative bounding box combination method to generate precise bounding boxes which can cover the entire smoke object. In addition, we evaluate the proposed approach using a quantitative and qualitative self-made forest fire smoke dataset, which includes forest fire smoke images of different scales. Extensive experiments show that our improved model's forest fire smoke detection accuracy is significantly higher than that of the mainstream models. Compared with deformable DETR, our model shows better performance with improvement of *mAP* (mean average precision) by 4.2%, *AP<sub>S</sub>* (*AP* for small objects) by 5.1%, and other metrics by 2% to 3%. Our model is adequate for early forest fire smoke detection with high detection accuracy of different-scale smoke objects.

**Keywords:** forest fire smoke detection; computer vision; small smoke objects; transformer; multi-scale features



**Citation:** Huang, J.; Zhou, J.; Yang, H.; Liu, Y.; Liu, H. A Small-Target Forest Fire Smoke Detection Model Based on Deformable Transformer for End-to-End Object Detection. *Forests* **2023**, *14*, 162. <https://doi.org/10.3390/f14010162>

Academic Editor: Viacheslav I. Kharuk

Received: 13 December 2022

Revised: 29 December 2022

Accepted: 9 January 2023

Published: 16 January 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Forest resources are essential for the global environment and human society. In addition to improving the quality of the atmospheric environment, forests also play a crucial role in the global carbon cycle, soil properties, and climate regulation [1]. The increasing occurrence of forest fires is destroying the world's forest resources and impacting human society in terms of considerable losses in human lives and public properties [2,3]. Due to forest fires being too difficult to rapidly control and extinguish once they occur, effective detection of early forest fires is an urgent need. The characteristics of smoke are more obvious, always appearing earlier than fires when a forest fire breaks out. It is of great significance for fire detection if forest fire smoke can be detected quickly and accurately.

Traditional forest fire smoke monitoring is based on manual inspection and smoke sensor monitoring [4]. However, manual inspection consumes substantial human and material resources with low efficiency and unsatisfactory results. Various sensors have been used to detect fire and smoke in the last two decades. Point sensors [5,6] obtain remarkable results indoors, but the investment of a fire smoke wireless sensor network over an entire forest is too expensive, and sensors are easily interfered with and damaged. Smoke

sensors require close proximity to the forest fire because the alarm needs particles to trigger. However, when the particle concentration reaches the threshold, the forest fire might be too strong to be controlled. Unmanned Aerial Vehicles (UAVs) can collect important visual information on early forest fire smoke detection during patrols [7]. Satellite sensors [8] have been used widely in forest fire smoke detection, and are not affected by various environmental factors, but can only monitor large-scale fires. Due to infrequent periods and resolution limitations, satellite sensors cannot immediately detect forest fires. Currently, with the development of computer vision technology, video surveillance systems that can be installed in forests have become a suitable alternative to previous detection methods and have lower cost, convenient deployment, and high detection efficiency. Watchtowers [9] and UAVs [4] equipped with cameras are appropriate for automatically monitoring forest fire smoke. Previous forest fire smoke detection methods based on computer vision technology usually make use of color and motion characteristics of the pixels from surveillance video frames. They mostly adopt pattern recognition processes, including feature extraction and classification, which are human-designed. After the candidate areas are extracted, static and dynamic smoke features are used for smoke recognition. Gubbi et al. [10] used wavelets to extract smoke characteristics and then classified them by using SVM (Support Vector Machines). ByoungChul et al. [11] trained two random forests for wildfire smoke classification using RGB (Red Green Blue) color, wavelets coefficients, motion orientation and a histogram of oriented gradients as independent temporal and spatial feature vectors. Prema et al. [12] proposed an image-processing approach using YUV color space, wavelet energy, and correlation and contrast of smoke to detect smoke. However, such methods are heavily dependent on human prior knowledge and are limited in various scenarios due to complex changeable forest environments, small-target smoke, and low-contrast flame and smoke.

Deep learning methods have attracted more attention in recent years than traditional image processing methods. Compared with traditional fire smoke detection methods, the fire smoke detection methods based on deep learning could extract more abstract and high-level features, and have the advantages of fast speed, high accuracy and strong robustness in complex forest environments. Convolutional neural networks (CNNs) have become prevalent object detection methods due to their outstanding performance in image recognition [13]. Frizzi et al. [14] proposed a CNN for fire and smoke detection and classification by extracting features in video. Wu et al. [15] used classical object detection models to detect forest fires. The adopted models contained You Only Look Once (YOLO) [16], Single Shot multi-box Detector (SSD) [17] and Faster Region-CNN (R-CNN) [18,19], and the experiments showed that an improved YOLO model could detect early forest fires efficiently. Semantic segmentation is also a common method to detect smoke. The task of semantic segmentation is to classify the input image pixel by pixel and mark the pixel-level objects. Pan et al. [20] introduced a collaborative region detection and grading framework for forest fire and smoke using a weakly supervised fine segmentation and a lightweight Faster R-CNN. Frizzi et al. [21] showed the comparison of network performance on two smoke semantic segmentation databases. Semantic segmentation and object detection, which have similar task objectives, mark objects and specific classification information of objects. The difference is that the object marked by semantic segmentation is at the pixel level, while the object marked by target detection is its bounding box. When preparing the smoke dataset, there is no need for tedious pixel-level marking operation, nor to classify every pixel in the image during detection, which leads to great optimization in running speed.

There is a transformer [22] model which has become a preferred settlement for machine translation, text generation, etc. [23–25] with the development of natural language processing (NLP). The self-attention mechanism could gather global information and pay attention to important elements more quickly and efficiently. Inspired by the success of transformer model and self-attention mechanism, Dosovitskiy et al. [26] proposed Vision Transformer (ViT) for image recognition. The first end-to-end object detection method based on transformer (DETR) demonstrated higher accuracy and speed on par with the previous

well-established Faster R-CNN on COCO dataset [27]. DETR has a simple architecture with a CNN backbone and transformer encoder–decoders. However, DETR needs more epochs than Faster R-CNN to converge and shows low performance in detecting small targets. Deformable DETR [28], which is modified from DETR by using a deformable attention module, obtains satisfactory results in object detection tasks, especially in detecting small targets. Here, we set deformable DETR as our baseline for forest fire smoke detection and demonstrate its efficiency through experiments.

In previous studies of forest fire smoke detection, many detection models have been used and have obtained good results. However, there are many existing problems for early forest fire smoke detection in forest environments due to the complex background and the difficulty of extracting smoke features. Firstly, forest images usually contain not only smoke but other irrelevant background information with similar characteristics to smoke, such as clouds, lake surface, fog, etc. The light change in the natural environment will also cause interference, resulting in the change of some image features, affecting the subsequent feature extraction and recognition. Secondly, it is challenging to detect early smoke precisely with their dynamic characteristics and small fuzzy shape. Therefore, in this paper, we aim to address this critical issue by improving feature extraction and small object detecting abilities using a Multi-scale Context Contrast Local Feature module (MCCL), Dense Pyramid Pooling module (DPPM) [29–31], and iterative bounding box combination method.

The contributions of our paper are as follows:

- We propose an improved deformable DETR model to detect forest fire smoke which involves a Multi-scale Context Contrast Local Feature module (MCCL) and Dense Pyramid Pooling module (DPPM). The modules enhance low contrast smoke for detecting small and inconspicuous smoke by capturing locally discriminative features.
- An iterative bounding box combination method is proposed to obtain precise boxes for smoke objects to obtain more accurate localization and bounding boxes of semi-transparent blurred smoke.
- In order to evaluate our model, we build a forest fire smoke dataset from public resources, including various kinds of smoke and smoke-like objects in complex forest environments.

The rest of the paper is organized as follows. Section 2 describes our dataset and the details of the improved deformable DETR model, Section 3 presents the experimental results and performance analysis, the discussion is given in Section 4, and finally Section 5 concludes this paper.

## 2. Materials and Methods

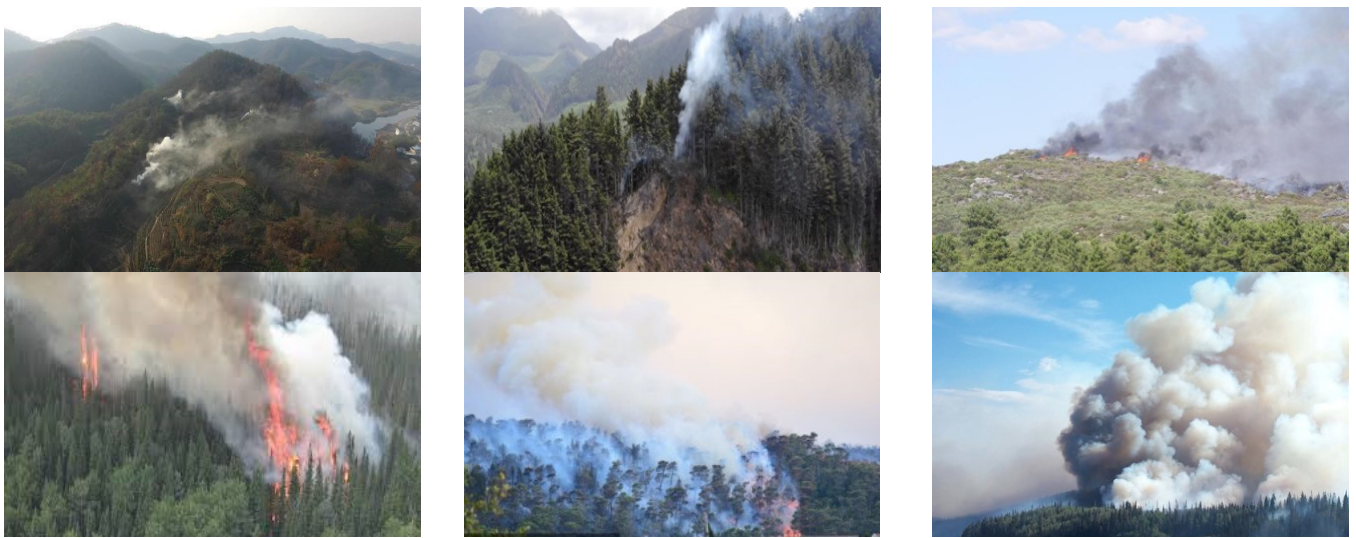
### 2.1. Dataset and Annotation

It is well-known that the quality and size of a dataset are essential for a deep learning model's performance. However, there are few public datasets about forest fire smoke or smoke datasets suitable for forest environments. Therefore, we proposed a forest fire smoke dataset (FFS dataset) by collecting forest fire smoke images (JPG format) from crawling open data on the Internet. Our self-built dataset contained different views and scales of forest fire smoke images. We manually labeled smoke areas in images and converted them to COCO [32] format. The dataset contained 10,250 images total, and we randomly divided them by 9:1; thus, 90% of the dataset was used as a training set and 10% as a validation set. Some sample images are shown in Figure 1.

### 2.2. Deformable DETR

Recently, DETR has demonstrated very competitive performance in the object detection field as a real end-to-end detector. In contrast to other modern object detect models, it does not need any hand-crafted components such as anchor generation and non-maximum suppression (NMS) and has a very simple architecture: a CNN backbone and an encoder–decoder transformer model. However, DETR has its own issues. Firstly, DETR need more

epochs to converge, which is mainly due to the difficulty of processing image features to train for the attention module. While the model is initializing, the cross-attention module gives average attention to the whole feature map. After training, the attention module gives attention to feature maps sparsely. Secondly, it is hard for DETR to detect small objects. The self-attention module in the encoder part of the transformer cannot handle high-resolution feature maps with unacceptable complexity. Zhu et al. [28] proposed a deformable DETR, which achieved satisfactory results in small object detection, and training epochs were reduced by almost a factor of 10. Authors provided the deformable attention module on each query to pay attention to the more meaningful locations that the network thought contained more local information, which were fewer in number, and a fixed number of locations as keys. This alleviated the problem of large computation requirements caused by high-resolution feature maps.



**Figure 1.** Samples of the FFS dataset (images show smoke objects at different scales, images of first row shows light smoke, the second row shows dense smoke).

The deformable attention feature was calculated by:

$$DeformAttn(z_q, p_q, x) = \sum_{m=1}^M W_m \left[ \sum_{k=1}^K A_{mqk} W'_m x(p_q + \Delta_{mqk}) \right] \quad (1)$$

In the formula,  $x$  is the input feature map,  $q$  represents the query element with content feature  $z_q$  and 2-d reference point  $p_q$ ,  $k$  indexes the sampled keys with the sampling offset  $\Delta_{mqk}$  and normalized attention weight  $A_{mqk}$  of the  $k$  sampling point in  $m$  attention head. In addition,  $W_m$  represents attention weights after linear transformation from different heads and  $K$  sampling offsets  $\Delta_{mqk}$  are calculated according to the linear layer, then  $K$  sampling offsets and  $p_q$  determine the selected points in the neighborhood.

Furthermore, a deformable attention module could be extended to a multi-scale deformable attention module in the deformable DETR's encoder part. Output feature maps of the encoder have the same resolution with the input feature maps. The input feature maps  $\{x^l\}_{l=1}^{L-1}$  ( $L = 4$ ) of the encoder are extracted from the backbone's output feature maps of stages  $C_3$  to  $C_5$  (such as ResNet [33], transformed by  $1 \times 1$  convolution). Every resolution of  $C_l$  is  $2^l$  lower than the input images. Authors proposed  $C_6$  stage, which was obtained by  $3 \times 3$  stride 2 convolution from  $C_5$  stage. For clarity of each query pixel's location, scale-level embedding is used for feature representation.

Then, the multi-scale deformable attention module is applied as:

$$MSDeformAttn(z_q, p_q, \{x^l\}_{l=1}^L) = \sum_{m=1}^M W_m \left[ \sum_{l=1}^L \sum_{k=1}^K A_{mqlk} W'_m x^l (\phi_l(p_q) + \Delta_{mqlk}) \right] \tag{2}$$

On the basis of the deformable attention module,  $l$  indexes input feature level,  $\{x^l\}_{l=1}^L$  are input feature maps which are divided by different levels,  $p_q$  are normalized coordinates which are not equivalent to reference points  $p_q$  in deformable attention module, function  $\phi_l$  rescales normalized coordinates at every feature layer to locate points that are sampled at different levels, and the remaining elements are similar to Equation (1) except for an additional  $l$  element.

The network structure of the deformable DETR is shown in Figure 2.

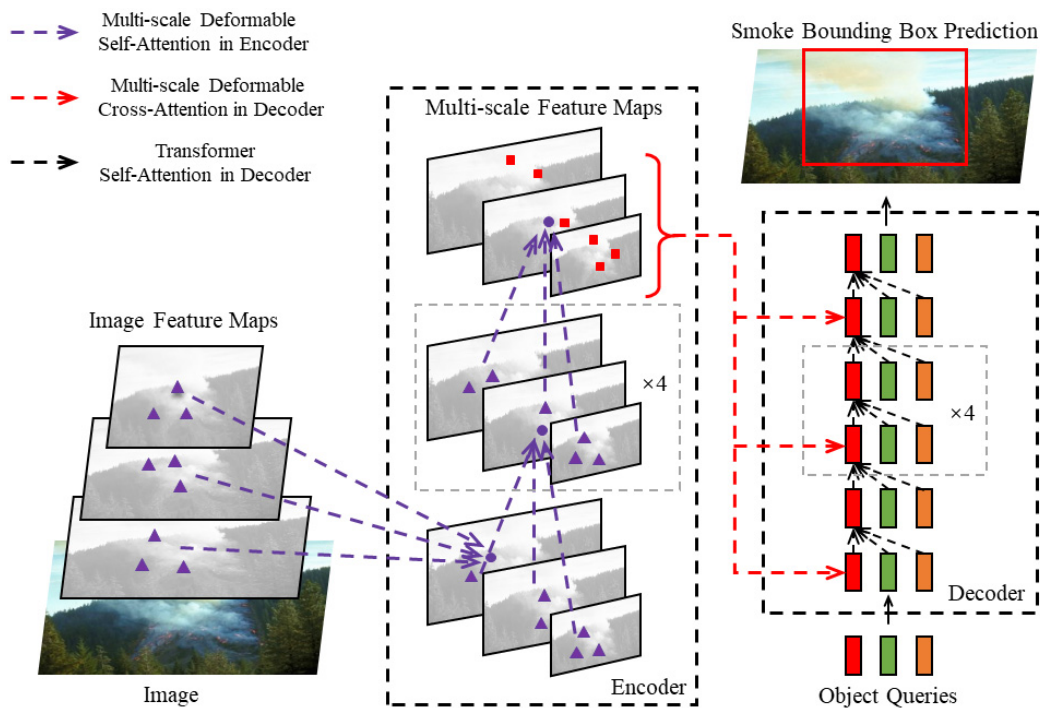


Figure 2. The network structure of deformable DETR.

By replacing the traditional transformer attention module, deformable DETR used a multi-scale deformable attention module to process feature maps which could be extended by aggregating multi-scale features naturally.

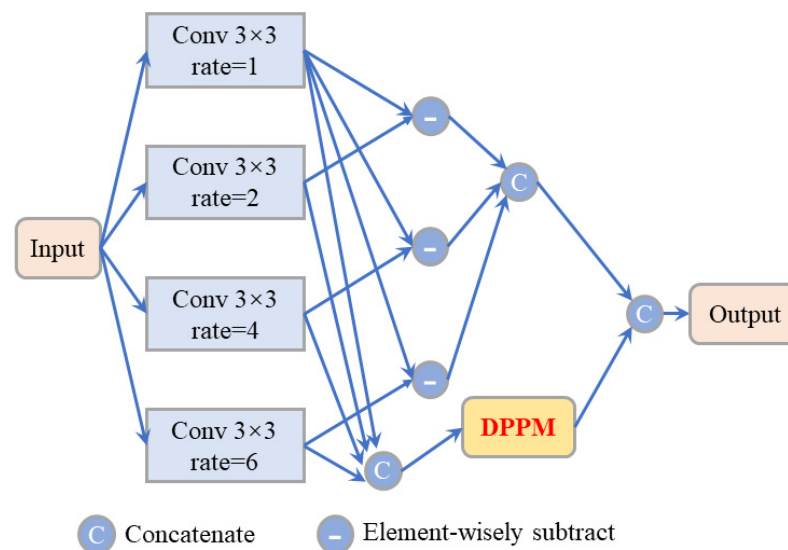
### 2.3. Multi-Scale Context Contrast Local Feature Module

As we know, context information can improve performance through scene labeling. CNN provides high-level context features which contain abstract and global information on the whole image for object recognition [31,33]. However, there are many inconspicuous targets in complex natural environments. Those context features from CNN usually focus on the dominated objects in the image and cannot make sure that they are useful for inconspicuous objects recognition. The Context Contrast Local Feature (CCL) module solves this problem well by computing the contrast of local context information, which not only makes full use of context but foregrounds the local information. This is an imitation of human behavior. Human beings concentrate on an object while we pay attention to its surrounding context. The contrast is computed by:

$$CL = C_l(F, \theta_l) - C_c(F, \theta_c) \tag{3}$$

where  $CL$  indexes the context contrasted local features,  $C_l$  and  $C_c$  are the local convolution block and context convolution block, respectively,  $F$  is the input features and  $\theta$  denotes respective parameters. The CCL module obtains context-local information from different levels by several chained context-local blocks. Each block contains dilated convolution blocks with dilation rate = 1 and rate = 5 to integrate multi-level context aware local features.

Early forest fire smoke can usually be considered as inconspicuous and blurred objects with low contrast and the CCL module cannot obtain satisfactory results for this task. To obtain more multi-scale features of inconspicuous smoke objects effectively, we use the Multi-scale Context Contrast Local Feature module in our model, which is modified from a previous module. The MCCL module is shown in Figure 3.



**Figure 3.** Illustration of Multi-scale Context Contrast Local Feature module.

To process subsequent high-level feature maps conveniently, we resize the input features as  $16 \times 16$  and restore them at the output block. This module contains 4 different levels of dilated convolution blocks with dilation rates = 1, 2, 4 and 6, respectively. Then we concatenate their output feature maps from each of the two blocks. We use a Dense Pyramid Pooling Module (DPPM) to extract more abstract information from the concatenated multi-scale feature maps. Confusion categories are a common problem in classification. It is an enormous challenge to distinguish between smoke and smoke-like objects such as clouds and haze. Zhao et al. [34] proposed a Pyramid Pooling Module (PPM) for global scene prior construction upon the high-level feature maps, and this obtains context information between sub-regions efficiently. Furthermore, the Dense Pyramid Pooling Module (DPPM) is used to process feature maps efficiently with fewer parameters and a larger size of the receptive field, as shown in Figure 4.

The module contains features under four different scales. We use four average pooling layers with different kernels and strides to generate feature maps (size  $1 \times 1$ ,  $2 \times 2$ ,  $4 \times 4$  and  $8 \times 8$ , respectively) into different sub-regions. After that, we use a  $1 \times 1$  convolution layer to reduce the dimension of features which could keep the weights of global feature consistent. Then we concatenate multi-scale feature maps from different pyramid levels and upsample several times directly to obtain the same size between input and output features via bilinear interpolation. Finally, we concatenate these feature maps as multi-scale features.

As we discussed in Section 2.2, the input feature maps  $\{x^l\}_{l=1}^{L-1}$  ( $L = 4$ ) of the encoder are extracted from the backbone's output feature maps of stages  $C_3$  to  $C_5$  (such as ResNet [33], transformed by  $1 \times 1$  convolution). The input multi-scale feature maps are obtained via  $1 \times 1$  stride 1 convolution on  $C_3$ ,  $C_4$  and  $C_5$  stage. In addition, we use the MCCL module to process the lowest-resolution feature maps on final  $C_5$  stage, then use  $3 \times 3$

stride 2 convolution to get the highest-dimensional feature maps as illustrated in Figure 5. The numbers below each layer represent the size and dimension of the feature maps.

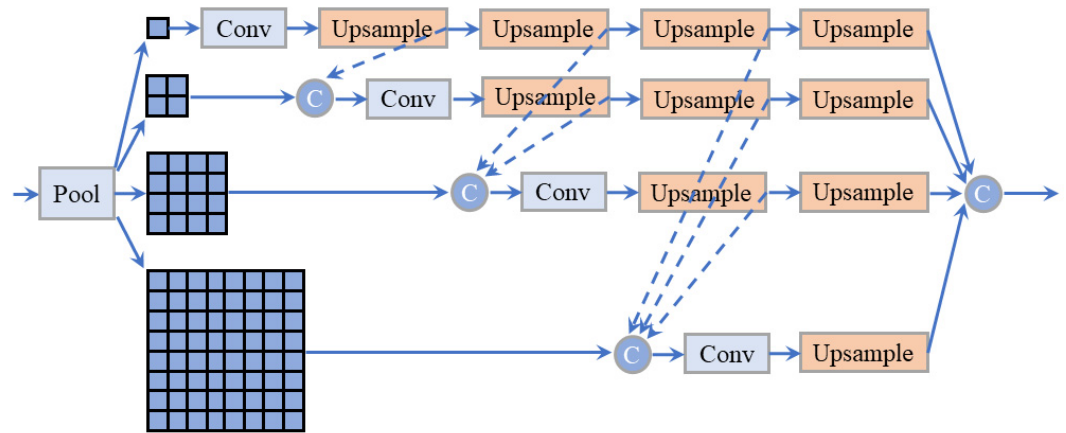


Figure 4. Illustration of the Dense Pyramid Pooling Module.

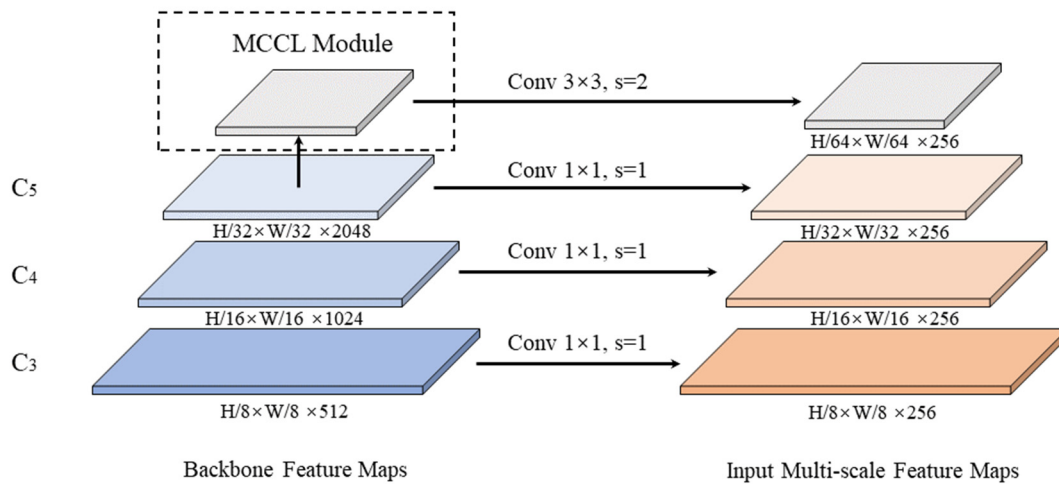
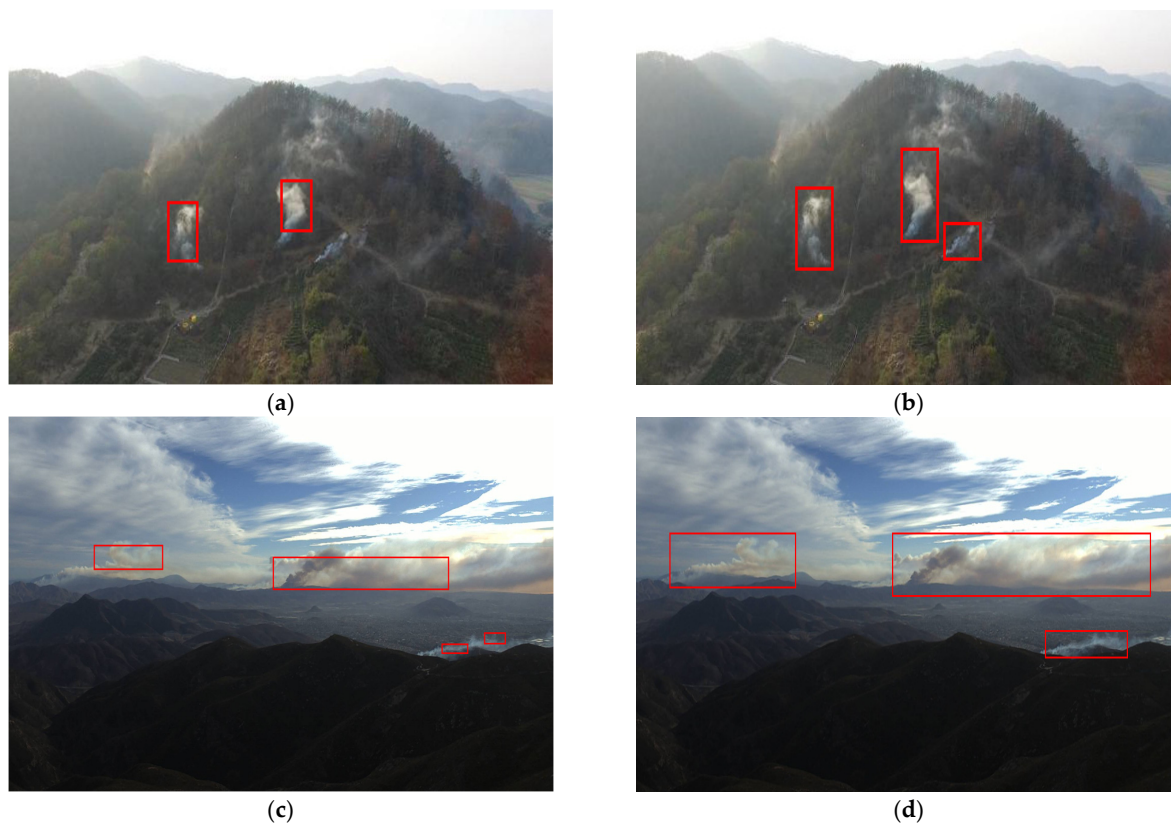


Figure 5. Location of the MCCL module in deformable DETR. MCCL module processes the lowest-resolution feature maps on C<sub>5</sub> stage.

#### 2.4. Iterative Bounding Box Combination Method

Forest fire smoke is easily affected by complex forest environments, and its characteristics change easily. Early smoke usually represents a semitransparent characteristic which leads to a blurred boundary. Unlike general object detection, it is difficult to obtain a precise bounding box for smoke. These uncertain elements inevitably lead to missed and false detections, as shown in Figure 6. In the previous object detection model, Non-Maximum Suppression (NMS) is proposed to obtain bounding boxes based on their scores. However, NMS is not necessary for DETR which lowers AP (Average Precision) in final layers and only improves AP<sub>50</sub> (AP at IoU = 0.5) slightly [27]. Deformable DETR uses iterative bounding box refinement to obtain precise bounding boxes based on predictions from each layer and different layers compute parameters independently [28]; each decoder layer predicts bounding boxes based on the predictions from the previous layer. As shown in Equation (4), for the boxes from the  $d$ -th decoder layer, the key elements are sampled to boxes predicted from the  $(d-1)$ -th decoder layer and the new reference points are set as  $(b_{jx}^{d-1}, b_{jy}^{d-1})$ . Additionally, these methods are not suitable for blurred smoke box proposals. Considering that our ideal goal is to detect early smoke rapidly and obtain an accurate position in images, we propose an iterative bounding box combination method based on NMS and iterative bounding box refinement to obtain satisfactory results and decrease the occurrence of missed and false detections. Our algorithm generates bounding boxes that

do not overlap with each other, and where the whole smoke objects are surrounded by bounding boxes. Ablation experimental results are shown in Figure 6.



**Figure 6.** Different detection samples before and after using iterative bounding box combination method. (a,c) Original detection results; (a) contains one missed detection; (c) contains one false detection. (b,d) The updated detection results where bounding boxes are generated by our method. The bounding boxes can cover the whole smoke accurately in both (b,d).

Firstly, we set  $D$  numbers of deformable DETR decoder layers (e.g.,  $D = 6$ ) and the predictions of bounding boxes  $box_j$  from every decoder layer are sorted by their confidences. The  $box_j$  is defined as:

$$box_j = \left\{ \sigma((\Delta b_{jx}^d) + \sigma^{-1}(\Delta b_{jx}^{d-1})), \sigma((\Delta b_{jy}^d) + \sigma^{-1}(\Delta b_{jy}^{d-1})), \right. \\ \left. \sigma((\Delta b_{jw}^d) + \sigma^{-1}(\Delta b_{jw}^{d-1})), \sigma((\Delta b_{jh}^d) + \sigma^{-1}(\Delta b_{jh}^{d-1})) \right\} \quad (4)$$

where  $d = \{1, 2, \dots, D\}$ ,  $b_{j\{x,y,w,h\}}^d$  are the predictions of the  $d$ -th decoder layer, and  $box_j$  is relevant to the predictions of  $d-1$ -th layer. The  $\sigma(\cdot)$  and  $\sigma^{-1}(\cdot)$  represent sigmoid function and inverse sigmoid function, respectively.

Secondly, we delete the  $box_j$  whose confidences are lower than 0.01. Then we calculate the Intersection over Union (IoU) between  $boundingbox_i$  and  $box_j$ :

$$IoU = \frac{boundingbox_i \cap box_j}{boundingbox_i \cup box_j} \quad (5)$$

We keep the  $box_j$  as a new bounding box if its IoU equals to zero.

Moreover, we combine  $boundingbox_i$  and  $box_j$  as a new  $boundingbox_{i+1}$  if the  $box_j$  only coincides with one bounding box and the IoU between two boxes is less than 0.7. We also need to keep the new boxes independent from other bounding boxes. Based on these, we improve the bounding box generation algorithm and our iterative bounding box combination algorithm is shown in Algorithm 1.



**Algorithm 1** Iterative Bounding Box Combination

---

**Input:**  $bbox = \{bbox_1, \dots, bbox_i\}$ ,  $box = \{box_1, \dots, box_j\}$ ,  $D = \{1, \dots, d, \dots, D\}$   
 $bbox$  is the bounding boxes.  
 $box$  is the box predictions from each decoder layers.  
 $D$  is a list of decoder layers.  
**Begin:**  
**For**  $d$  in  $D$  **do**  
    Rank  $box$  by confidence  
    **While** confidence of  $box_j < 0.01$  **do**  
        delete  $box_j$   
    **If**  $\sum_{i=1}^l IoU(bbox_i, box_j) = 0$  **do**  
         $bbox \leftarrow bbox_j$   
    **Else If**  $\sum_{i=1}^l IoU(bbox_i, box_j) \leq 0.7$  &&  $\sum_{k \neq i}^l IoU(bbox_k, bbox_i \cup bbox_j) = 0$  **do**  
         $bbox \leftarrow bbox_i \cup bbox_j$   
    **End**  
**End**  
**Return**  $bbox$   
**End**

---

### 2.5. Loss Function

In terms of loss function, our model follows the function of deformable DETR. Therefore, we totally set three components to the loss, classification loss, bounding box distance loss and GIoU loss [35]. The classification loss is necessary for the training model and classification task, which are represented as cross-entropy loss. The bounding box distance loss is set as L1 loss, which calculates the distance between prediction box and the ground truth then propagates gradients. Furthermore, we use *GIoU* loss to make the prediction box closer to the ground truth:

$$GIoU = IoU \cdot \frac{|C \setminus (A \cup B)|}{|C|} \quad (6)$$

$$L_{GIoU} = 1 - GIoU \quad (7)$$

where  $A$ ,  $B$  and  $C$  represent prediction box, ground truth and smallest closing box between  $A$  and  $B$ , respectively. Thus, our total loss is weighed sum of three loss:

$$Loss = L_{cls} + L_1 + L_{GIoU} \quad (8)$$

## 3. Results

### 3.1. Training

The details of our experimental environments are shown in Table 1. Training parameters of our model were designed based on the deformable DETR as shown in Table 2. Furthermore, we set  $M$  as the number of heads for multi-scale deformable attention module, which equals to 8, and  $K$  indexes the number of sample keys, which equals to 4.

**Table 1.** Experimental environments.

Experimental Environments	Details
Program Language	Python 3.7
Framework	Pytorch 1.5.1
Operating System	Windows 10
GPU Type	RTX 2080ti
Acceleration Tool	CUDA 10.2

**Table 2.** Training parameters.

Training Parameters	Details
Epochs	50
Batch Size	4
Learning Rate	$2 \times 10^{-5}$
Optimizer	SGD
Momentum	0.9
Weight Decay	$1 \times 10^{-4}$

### 3.2. Comparison and Evaluation

In order to analyze and demonstrate the early forest fire smoke detection performance of our improved deformable DETR model, we used Microsoft COCO evaluation metrics here, which are widely used to evaluate object detection tasks. Our model trains on the training set and evaluates on the validation set. The formulas of the two main metrics AP (Average Precision) and AR (Average Recall), which are calculated based on Precision (P) and Recall (R), are shown in Equations (9)–(12).

$$P = \frac{TP}{TP + FP} \quad (9)$$

$$R = \frac{TP}{TP + FN} \quad (10)$$

$$AP = \sum_{i=1}^{n-1} (R_{i+1} - R_i) P(R_{i+1}) \quad (11)$$

$$AR = 2 \int_{0.5}^1 R(o) do \quad (12)$$

$TP$ ,  $FP$  and  $FN$  represent the numbers of true positive samples, false positive samples and false negative samples, respectively. In Equation (12), the variable  $o$  indexes the IoU between the prediction box and the ground truth box.

Microsoft COCO evaluation metrics include various object detection accuracies of different area sizes. Therefore, we use AP and AR for comparison.  $mAP$  is mean Average Precision and  $mAR$  is mean Average Recall for all categories;  $AP_S$ ,  $AP_M$  and  $AP_L$  represent the AP for small objects (area size  $< 32^2$ ), medium objects ( $32^2 < \text{area size} < 96^2$ ) and large objects (area size  $> 96^2$ ), respectively.  $AP_{50}$  means average precision at IoU = 0.5 and AR indicators are similar to AP. Specifically, the units of AP and AR are percentages. We also added ablation experiments. The experimental results are shown in Table 3.

**Table 3.** Experimental results. Comparison of our improved model with other detection models on our FFS dataset.

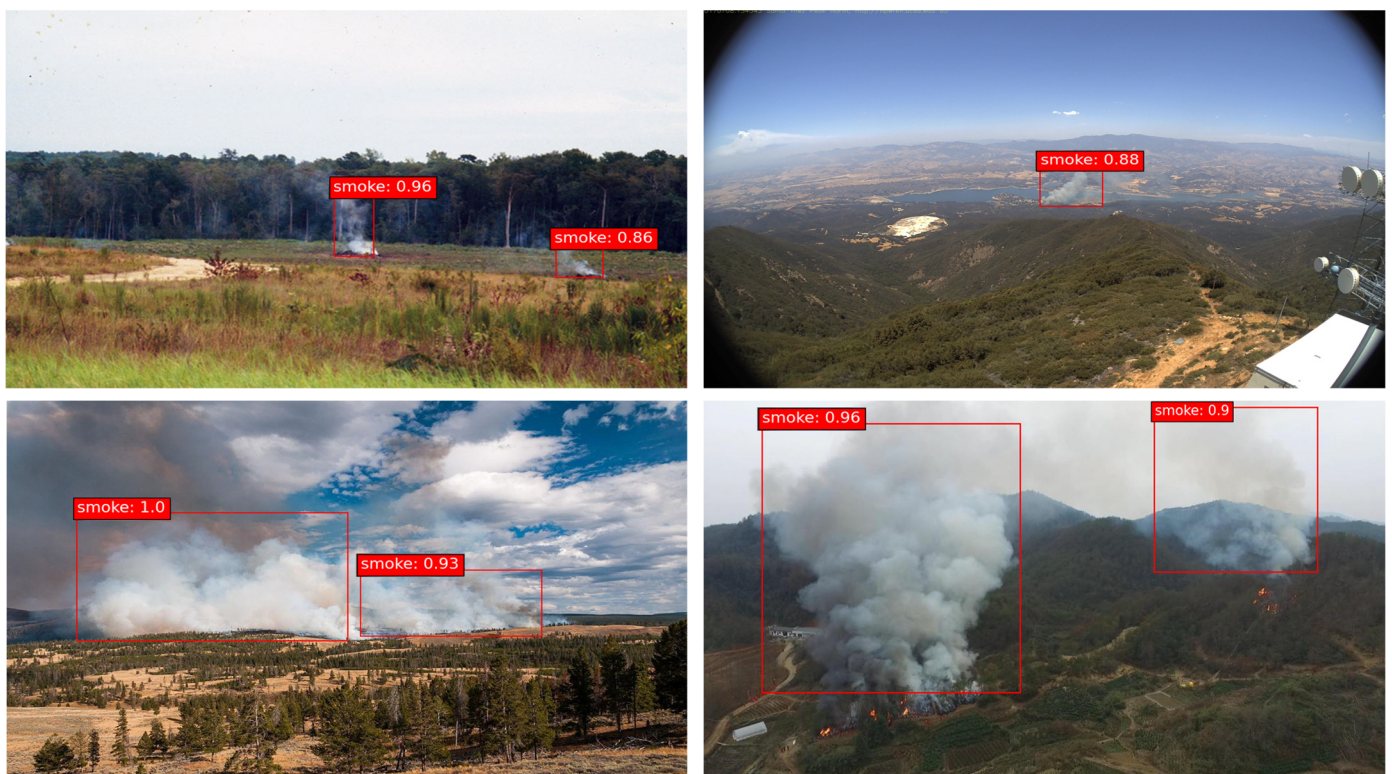
Model	Epoch	mAP	AP <sub>50</sub>	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>	mAR	AR <sub>S</sub>	AR <sub>M</sub>	AR <sub>L</sub>	Params	Speed
Faster R-CNN + FPN	100	37.4	80.0	24.2	34.3	49.7	47.2	28.3	43.3	54.0	42M	235 ms
YOLOv5s	100	42.7	82.2	29.6	41.1	56.0	48.7	34.9	56.1	62.3	7.2M	52 ms
DETR	500	44.2	84.8	27.4	40.8	60.2	53.6	33.8	51.0	62.4	40M	192 ms
DETR DC5	500	45.0	85.5	28.1	42.4	60.3	56.2	37.6	51.9	62.5	40M	441 ms
Deformable DETR (Baseline)	50	45.5	85.8	33.5	42.6	58.7	54.0	42.8	50.6	59.7	37M	245 ms
+ MCCL Module	50	48.4	86.9	<b>38.6</b>	46.1	60.2	57.7	44.0	<b>59.3</b>	62.8	37M	240 ms
++ iterative bounding box combination method	50	<b>49.7</b>	<b>88.4</b>	36.9	<b>48.7</b>	<b>62.3</b>	<b>60.1</b>	<b>44.2</b>	59.1	<b>65.3</b>	37M	240 ms

The backbone of DETR series is set to ResNet50, Faster R-CNN and YOLOv5s use ResNet101 and C3+SPPF as backbone, respectively. Training epochs are set to different values for the best training results of models. The bolded numbers indicate the best performance in the comparison. + Add ablation experiments are based on deformable DETR.

### 3.3. Detection Performance and Analysis

Compared with other remarkable detection models, extensive experiments indicated that our improved deformable DETR model with MCCL module and iterative bounding box combination method achieved satisfactory results in early forest fire smoke detection tasks. We also used YOLOv5s and DETR for comparison, which are widely used in object detection. Compared with Faster R-CNN + FPN, DETR shows higher accuracy of detection performance but needs much more training time to converge and delivers low accuracy in detecting small smoke. Our baseline, deformable DETR achieves more satisfactory performance with small targets with fewer training epochs. Compared with the baseline, the Multi-scale Context Contrasted Local Feature module improves the overall performance, especially with improvement at  $AP_5$  by 5.1%. After adding the iterative bounding box combination method, the detection of our model on forest fire smoke obtains higher accuracy with 4.2% in  $mAP$ , 2.6% in  $AP_{50}$  and 6.1% in  $mAR$  (compared with baseline), improving other metrics by 3%. Based on these experiments, we can conclude that our improved deformable DETR model is competent for small and inconspicuous smoke detection and the detection accuracy of smoke at different scales is higher than other common models. Some detection results are shown in Figures 7–11.

As shown in Figure 7, the detection results of the improved model show that there are no false and missed detections, and the bounding boxes cover the entire smoke objects with high accuracy. We also used YOLOv5s, DETR and baseline to detect ultra-small smoke targets in the wild with strong interference (such as strong direct sunlight interference in Figures 7 and 8); they all had a missed detection, but our model detected them accurately (as shown in Figures 8–11). A series of images on the left show that small gray smoke can be detected well by common models. As shown in the right images, ultra-small white smoke with strong light is too difficult to be detected by general models, but our model could detect it well.



**Figure 7.** Detection results of our improved deformable DETR model. The first row shows small-target smoke images, the second row shows large-target smoke images.

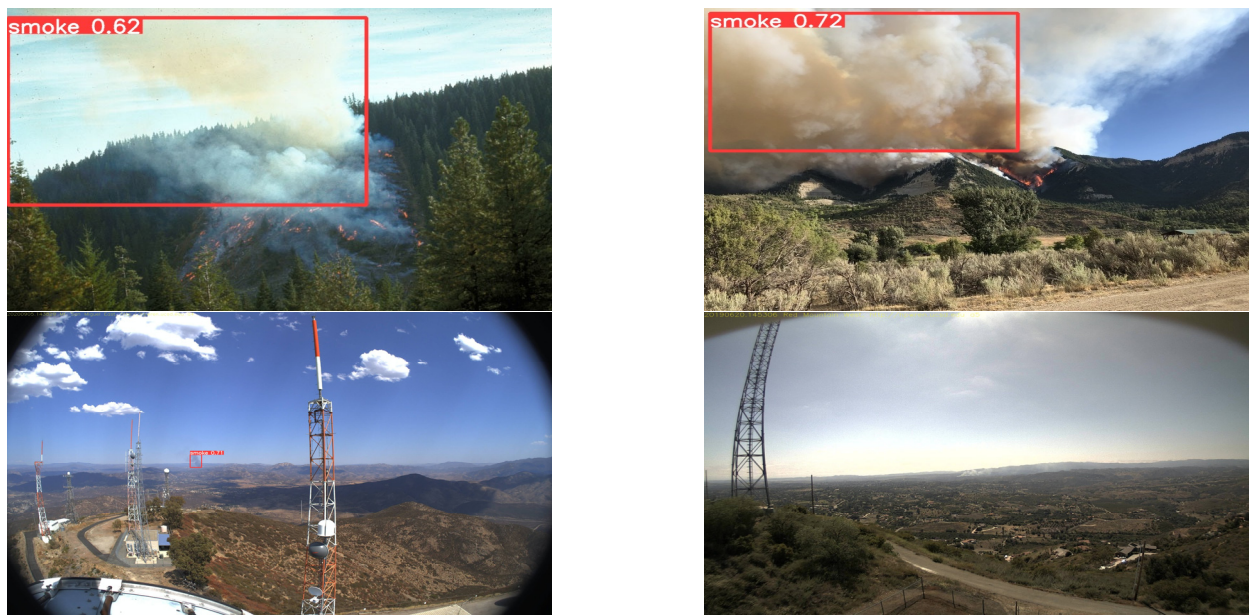


Figure 8. Detection results using YOLOv5s.

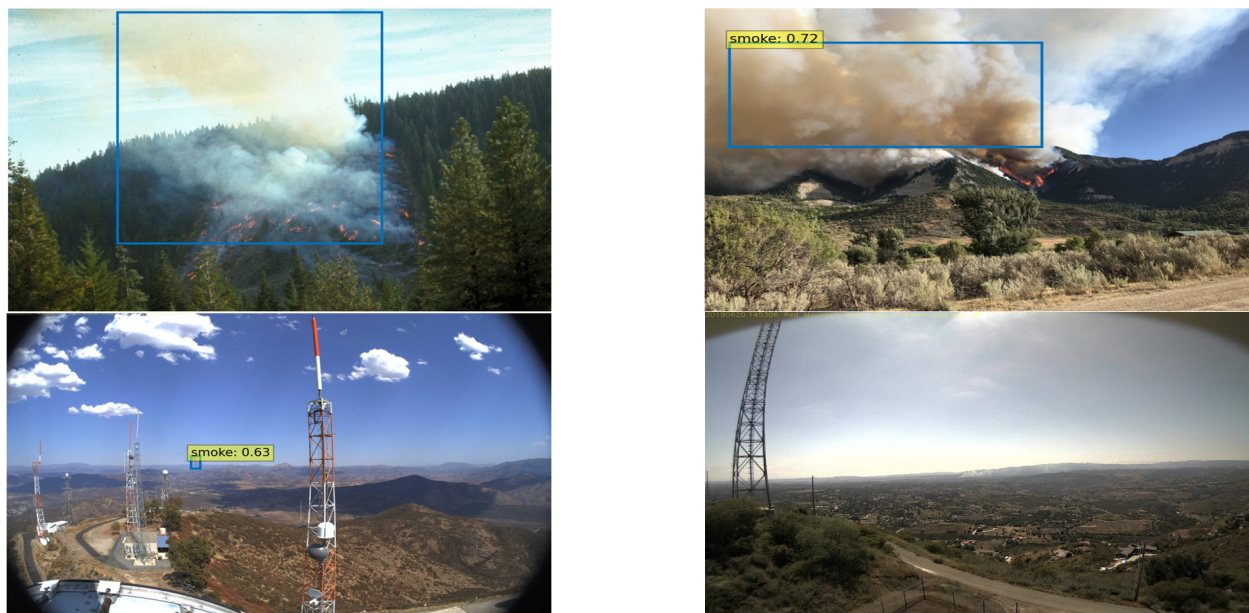


Figure 9. Detection results using DETR.

To investigate our improvement of feature extraction and understand the multi-scale attention module better, we visualize sampling points and attention weights of the last layer in the encoder. As shown in Figure 12, compared with the baseline, our improved model can focus more precisely on the inconspicuous smoke part by giving it larger attention weights, while the original model pays attention to the boundary of smoke roughly. The attention weights and the positions of sampling points lead to the difference in the subsequent learning and detection modules of the two models.



Figure 10. Detection results using deformable DETR.



Figure 11. Detection results using our model.

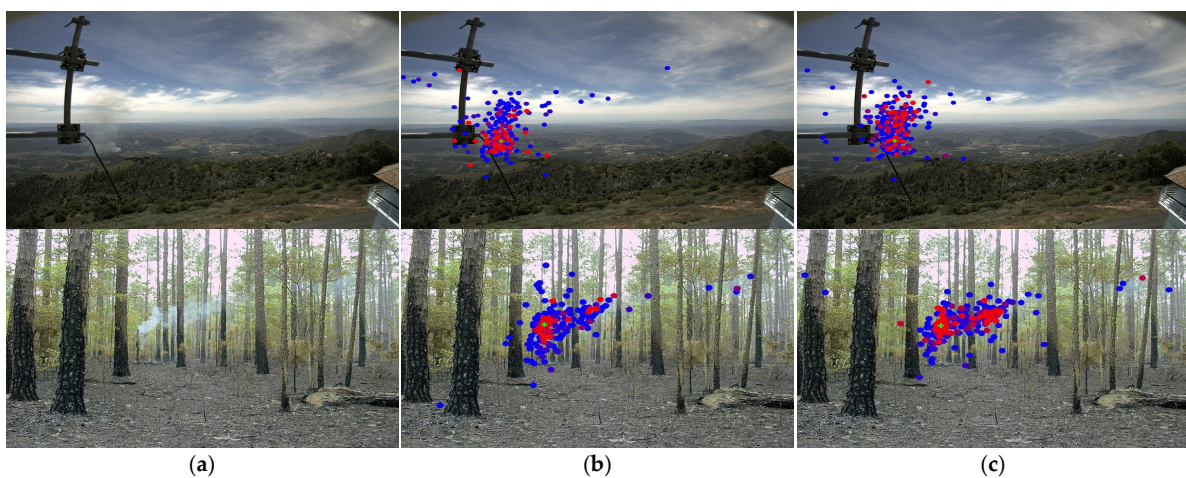


Figure 12. Visualization results of the multi-scale deformable attention in encoder. We draw the sampling points and attention weights from feature maps in one image. Each circle represents a sampling point and its color represents the attention weight. Color from blue to red indicates the weight from small to large. (a) Raw image; (b) deformable DETR; (c) our improved model.

#### 4. Discussion

It is very important to detect forest fires quickly and accurately. Smoke, as a significant feature of early fires, should be paid more attention to during detection. However, objects such as smoke and flames have irregular shapes and are easily disturbed by complex forest environments. Delayed or even missed detection of forest fire smoke can lead to the rapid spread of fire, which causes immeasurable losses. The development of computer vision has made it possible for high-precision automatic inspection to replace manual inspection in the last two decades. Because of the translucency and blurred boundary of smoke, it is easily influenced by other factors such as light and wind. Previous smoke detection methods based on deep learning have mainly studied the texture and spatio-temporal characteristics from smoke videos to achieve more accurate smoke detection results [36–38]. Smoke detection can also adopt another strategy, that is, paying attention to data re-processing such as dark channel prior, optical flow, and super-pixel segmentation of images [20,39].

Our improved deformable DETR model concentrates on feature extraction in order to obtain higher accuracy of smoke detection. Through these comparisons and ablation experiments, we found that our model is more suitable for early forest fire smoke detection tasks compared with other common models, as shown in Table 3. The MCCL module provides precise multi-scale features of small and inconspicuous smoke objects for high-level feature processing and the module has more dilated convolution blocks and fewer parameters than CCL. We used the DPPM module, which is expanded from the Pyramid Pooling Module to generate more features with fewer parameters than the Pyramid Pooling Module. As shown in Figure 4, our DPPM module computes multi-scale features naturally by upsampling at each stage. The module we used combines efficient feature extraction with fewer calculation parameters. In Figure 12, we visualize sampling points and attention weights of the last layer in the encoder, and our improved model can focus more precisely on smoke objects while the MCCL module extracts more useful features for subsequent feature learning and detection modules. Compared with the original model, more accurate sampling points and attention weights show the advantages of our method in feature extraction. Additionally, the detection performance of our model also demonstrates advantages in this task (as shown in Figures 7–11). Our detection samples contain ultra-small smoke objects with strong inference (such as strong direct sunlight and smoke-like clouds in Figure 11). Due to the further processing of high-dimensional feature maps by the MCCL module and DPPM greatly reducing the possibility of misclassification, this model can more accurately obtain inconspicuous smoke features and distinguish the smoke from smoke-like objects. In the field of vision-based target detection, small target detection has always been a difficult problem. Mis-detection of our model occurs when detecting small targets. Early small smoke targets tend to be easily covered by trees and dissipate quickly. Limited pixel representations of early smoke flow and the interference from smoke-like objects usually lead to the problem of mis-detection in the original model. In order to improve the detection performance of inconspicuous smoke targets, we propose using several dilated convolutions with different rates to obtain useful context information, and also pay attention to local information of inconspicuous targets. The proposed improvement strategy obtained satisfactory result in detecting early smoke targets and improved the  $AP_S$  metric by 5.1% (compared with the original model).

The previous bounding box generation method is obviously suitable for smoke in forest fire smoke detection tasks; the generated bounding box always has a smaller or larger offset to the ground truth, which leads to high training loss. Considering this situation, we used an iterative bounding box combination method to generate bounding boxes more consistently with ground truth which reduced the occurrences of false and missed detections and improved  $mAP$  by 4.2%. With the addition of our bounding box generation method, the detection results become more accurate than the baseline in Figure 6. Furthermore, we constructed a large forest fire smoke dataset to evaluate our method. Four common object detection models were obtained in the experiments with good performance

on forest fire smoke detection, which made it possible to detect the forest fire smoke in the wild.

However, our model still has some disadvantages to improve. Small object detection is not only the detection of forest fire smoke but also one of the difficulties of computer vision. We extracted features from high dimensions to detect small smoke, which will still be limited by the lack of small target pixel information. Complex environments, such as foggy weather, greatly affect the detection of our model, but smoke sensors still have high accuracy in detecting smoke. Therefore, combining computer vision with traditional smoke sensor networks may make smoke detection more accurate.

## 5. Conclusions

In this paper, we propose an improved end-to-end deformable DETR model for forest fire smoke detection. Firstly, in order to capture the information of small and inconspicuous smoke, a feature extraction module with Multi-scale Context Contrasted Local Feature module and Dense Pyramid Pooling module is used. Several dilated convolutions with different rates make full use of context information and local information of inconspicuous objects, which improves the performance of early forest fire smoke detection. Secondly, we propose an iterative bounding box combination method to reduce the occurrences of false and missed detections and generate a bounding box for forest fire smoke more accurately to the ground truth. Lastly, due to the lack of relevant public datasets, we established a quantitative and qualitative forest fire smoke dataset to verify the performance of our model. Ablation experiments show that our improved model for detecting forest fire smoke is superior to the mainstream detection model in most metrics. Our model not only achieves high detection accuracy of smoke but can detect early forest fire smoke which is too small and inconspicuous to be detected by common models.

In the next stage, we plan to conduct joint detection of early fire and smoke, then prune and distill the knowledge for our improved model so that it can be deployed to edge devices such as UAVs and watchtowers for real-time detection with fewer parameters and higher processing speed.

**Author Contributions:** Conceptualization, J.H.; data curation, J.H., J.Z. and H.Y.; methodology, J.H.; resources, J.H., J.Z. and H.Y.; software, J.H. and J.Z.; validation, Y.L.; funding acquisition, Y.L.; writing—original draft preparation, J.H.; writing—review and editing, Y.L. and H.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number KYCX22\_1056) and National Key R&D Program of China (grant number 2017YFD0600904).

**Data Availability Statement:** The data in this study are available from the authors upon request.

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

## References

1. Yuan, C.; Zhang, Y.M.; Liu, Z.X. A survey on technologies for automatic forest fire monitoring, detection, and fighting using unmanned aerial vehicles and remote sensing techniques. *Can. J. For. Res.* **2015**, *45*, 783–792. [[CrossRef](#)]
2. Eugenio, F.C.; dos Santos, A.R.; Fiedler, N.C.; Ribeiro, G.A.; da Silva, A.G.; dos Santos, Á.B.; Paneto, G.G.; Schettino, V.R. Applying GIS to develop a model for forest fire risk: A case study in Espírito Santo, Brazil. *J. Environ. Manag.* **2016**, *173*, 65–71. [[CrossRef](#)] [[PubMed](#)]
3. Tang, X.; Machimura, T.; Li, J.; Liu, W.; Hong, H. A novel optimized repeatedly random undersampling for selecting negative samples: A case study in an SVM-based forest fire susceptibility assessment. *J. Environ. Manag.* **2020**, *271*, 111014. [[CrossRef](#)]
4. Yang, X.; Tang, L.; Wang, H.; He, X. Early Detection of Forest Fire Based on Unmanned Aerial Vehicle Platform. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–4.
5. 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](#)]

6. Qiu, X.; Wei, Y.; Li, N.; Guo, A.; Zhang, E.; Li, C.; Peng, Y.; Wei, J.; Zang, Z. Development of an early warning fire detection system based on a laser spectroscopic carbon monoxide sensor using a 32-bit system-on-chip. *Infrared Phys. Technol.* **2019**, *96*, 44–51. [[CrossRef](#)]
7. Sudhakar, S.; Vijayakumar, V.; Kumar, C.S.; Priya, V.; Ravi, L.; Subramaniaswamy, V. Unmanned Aerial Vehicle (UAV) based Forest Fire Detection and monitoring for reducing false alarms in forest-fires. *Comput. Commun.* **2020**, *149*, 1–16. [[CrossRef](#)]
8. Guo, C.H.; Qi, X.Y.; Gong, Y.L. Study on the Technology and Method of Forest Fire Monitoring by Using HJ Satellite Images. *Remote Sens. Inf.* **2010**, *4*, 85–99.
9. Zhang, F.; Zhao, P.; Xu, S.; Wu, Y.; Yang, X.; Zhang, Y. Integrating multiple factors to optimize watchtower deployment for wildfire detection. *Sci. Total Environ.* **2020**, *737*, 139561. [[CrossRef](#)] [[PubMed](#)]
10. Gubbi, J.; Marusic, S.; Palaniswami, M. Smoke detection in video using wavelets and support vector machines. *Fire Saf. J.* **2009**, *44*, 1110–1115. [[CrossRef](#)]
11. Ko, B.; Kwak, J.-Y.; Nam, J.-Y. Wildfire smoke detection using temporospatial features and random forest classifiers. *Opt. Eng.* **2012**, *51*, 017208-1–017208-10. [[CrossRef](#)]
12. Prema, C.E.; Vinsley, S.S.; Suresh, S. Multi Feature Analysis of Smoke in YUV Color Space for Early Forest Fire Detection. *Fire Technol.* **2016**, *52*, 1319–1342. [[CrossRef](#)]
13. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2012; Volume 25.
14. Frizzi, S.; Kaabi, R.; Bouchouicha, M.; Ginoux, J.M.; Moreau, E.; Fnaiech, F. Convolutional neural network for video fire and smoke detection. In Proceedings of the IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23–26 October 2016; pp. 877–882.
15. Wu, S.; Zhang, L. Using popular object detection methods for real time forest fire detection. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; pp. 280–284.
16. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**. [[CrossRef](#)]
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
19. Lin, G.; Zhang, Y.; Xu, G.; Zhang, Q. Smoke Detection on Video Sequences Using 3D Convolutional Neural Networks. *Fire Technol.* **2019**, *55*, 1827–1847. [[CrossRef](#)]
20. Pan, J.; Ou, X.; Xu, L. A Collaborative Region Detection and Grading Framework for Forest Fire Smoke Using Weakly Supervised Fine Segmentation and Lightweight Faster-RCNN. *Forests* **2021**, *12*, 768. [[CrossRef](#)]
21. Frizzi, S.; Bouchouicha, M.; Moreau, E. Comparison of two semantic segmentation databases for smoke detection. In Proceedings of the IEEE Conference on Industrial Technology (ICIT), Virtual Event, 10–12 March 2021; pp. 856–863.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**. [[CrossRef](#)]
23. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**. [[CrossRef](#)]
24. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**. [[CrossRef](#)]
25. Zhang, X.; Wei, F.; Zhou, M. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. *arXiv* **2019**. [[CrossRef](#)]
26. 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 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
27. 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.
28. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**. [[CrossRef](#)]
29. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Context Contrasted Feature and Gated Multi-scale Aggregation for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–21 June 2018; pp. 2393–2402.
30. Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A Gated Recurrent Network With Dual Classification Assistance for Smoke Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 4409–4422. [[CrossRef](#)] [[PubMed](#)]
31. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-Free Local Feature Matching with Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021; pp. 8918–8927.
32. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.



33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
35. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
36. Liu, T.; Cheng, J.; Du, X.; Luo, X.; Zhang, L.; Wang, Y. Video Smoke Detection Method Based on Change-Cumulative Image and Fusion Deep Network. *Sensors* **2019**, *19*, 5060. [[CrossRef](#)]
37. Cao, Y.; Tang, Q.; Lu, X. STCNet: Spatiotemporal cross network for industrial smoke detection. *Multimed. Tools Appl.* **2022**, *81*, 10261–10277. [[CrossRef](#)]
38. Li, X.; Song, W.; Lian, L.; Wei, X. Forest Fire Smoke Detection Using Back-Propagation Neural Network Based on MODIS Data. *Remote Sens.* **2015**, *7*, 4473–4498. [[CrossRef](#)]
39. Ryu, J.; Kwak, D. A Study on a Complex Flame and Smoke Detection Method Using Computer Vision Detection and Convolutional Neural Network. *Fire* **2022**, *5*, 108. [[CrossRef](#)]

**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.