forests

MDPI

Article

Novel Recursive BiFPN Combining with Swin Transformer for Wildland Fire Smoke Detection

Ao Li, Yaqin Zhao * and Zhaoxiang Zheng

College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China
* Correspondence: yaqinzhao@163.com

Abstract: The technologies and models based on machine vision are widely used for early wildfire detection. Due to the broadness of wild scene and the occlusion of the vegetation, smoke is more easily detected than flame. However, the shapes of the smoke blown by the wind change constantly and the smoke colors from different combustors vary greatly. Therefore, the existing target detection networks have limitations in detecting wildland fire smoke, such as low detection accuracy and high false alarm rate. This paper designs the attention model Recursive Bidirectional Feature Pyramid Network (RBiFPN for short) for the fusion and enhancement of smoke features. We introduce RBiFPN into the backbone network of YOLOV5 frame to better distinguish the subtle difference between clouds and smoke. In addition, we replace the classification head of YOLOV5 with Swin Transformer, which helps to change the receptive fields of the network with the size of smoke regions and enhance the capability of modeling local features and global features. We tested the proposed model on the dataset containing a large number of interference objects such as clouds and fog. The experimental results show that our model can detect wildfire smoke with a higher performance than the state-of-the-art methods.

Keywords: wildland fire smoke detection; Recursive BiFPN; Swin Transformer; feature enhancement

Citation: Li, A.; Zhao, Y.; Zheng, Z. Novel Recursive BiFPN Combining with Swin Transformer for Wildland Fire Smoke Detection. *Forests* **2022**, *13*, 2032. https://doi.org/10.3390/f13122032

Academic Editors: Rafael Coll Delgado and Rafael De Ávila Rodrigues

Received: 26 October 2022
Accepted: 25 November 2022
Published: 30 November 2022

1. Introduction

Wildfires not only seriously damage vegetation, ecology and environment, but also pose a huge threat to people's life and property safety [1]. In the early stage of wildfire, due to the broadness of wild scene and the occlusion of the vegetation, smoke is more easily detected than flame. Therefore, how to detect an early mountain fire and wildfire smoke in the field environment is particularly important.

Traditional sensor-based methods [2–4] are easily affected by some interference factors such as space size and airflow in wide fields, so the methods suffer from some limitations such as limited detection range and delayed alarm time due to the reduction of smoke concentration caused by its diffusion characteristics [5]. Due to the extensive use of image processing technology [6–8], computer vision methods were used to detect fire smoke. The methods artificially design smoke features such as color features and texture features [9], variance, mean, and gradient features extracted by wavelet transform [10], and high-frequency energy extracted by double-layer two-dimensional wavelet transform [11]. However, to improve the detection accuracy and reduce the false rate, the color texture and energy characteristics must be carefully designed for certain smoke, which reduces the generalization ability of the model [5].

Deep learning was also used for smoke image detection. For example, Yang et al. [12] and Wang et al. [5], respectively, applied the existing object detection models such as CNN (Convolutional Neural Networks) and YOLO series for fire smoke recognition. He et al. [13] introduced the attention mechanism combining spatial attention and channel attention into CNN and used a decision fusion module to distinguish smoke and fog. In [14–17], Faster RCNN, Efficientdet and SSD [14–17] were used to extract smoke static features that

were combined with the dynamic characteristics. Luo et al. [18] proposed a two-stage scheme, where the candidate smoke regions are first detected by a background update model and the prior knowledge of dark channels, and then CNN network is used to remove the non-smoke candidate regions.

Although these deep networks can obtain higher accuracy than the methods based on artificial features when detecting smoke images with complex visual characteristics, these methods suffered from the following problems in the face of outdoor environment. The thick smoke, especially white smoke, is very similar to the clouds in the sky and the fog in mountains, and thus these methods have a high false alarm rate. Moreover, for the fluctuating smoke blown by the wind and the smoke images with low resolution, the detecting accuracy of these methods is unsatisfactory.

To better extract the features of objects, FPN (Feature Pyramid Network) structure and its variants were proposed and widely used in deep learning networks [19–21]. In addition, several FPN variants were also presented, such as the simple two-way fusion structure PANet [22], the complex bidirectional fusion structures NAS-FPN [23], BiFPN [24], and Recursive-FPN [25]. BiFPN adds edges of contextual information to the original FPN structure in order to fuse more features and assigns different weights to each layer during feature fusion so that the network can pay more attention to important layers, thereby reducing unnecessary layer node connections. Recursive-FPN (RFPN for short) recursively inputs the features output by FPN back to the backbone network for a second cycle. After this operation, the generated feature representation is stronger and stronger, so it can more effectively obtain multi-scale information.

Aimed at the above problems, this paper constructs a Recursive BiFPN attention model and introduces it into the backbone network of YOLOV5 to extract and fuse the multi-scale features of smoke images in order to better distinguish the subtle difference between smoke from clouds and fog. Zhu et al. [26] proposed a Transformer Prediction Head (TPH for short) for small target detection. To better capture the features in different scales, Shifted Window Transformer [27] was presented to meet the needs of fine-grained forecasting. Inspired by [27], we introduce the Shifted Window Transformer (Swin-TPH for short) to replace prediction head of YOLOV5, which contributes to increase the detection accuracy of small smoke.

The main contributions of our method are as follows:

(1) We propose a novel Recursive BiFPN (RBiFPN) attention model, which combines the features extracted by the first BiFPN with the initial features of the backbone network, and then passes through the BiFPN for the second feature fusion. Therefore, the model pays more attention to the important features of the smoke and the interference objects such as clouds. In particular, it can more effectively enhance and fuse multi-scale features through the recursive operation, which is beneficial to detect fire smoke in outdoor complex backgrounds.

(2) We replace the prediction head of YOLOV5 with Swin-TPH. The hierarchical structure of Swin-TPH can change receptive fields with the size of smoke regions in fire images. In addition, non-overlapping local windows and overlapping cross-window operations enhance local features and global modeling capabilities. The above schemes contribute to detecting the different smoke targets with the large difference in smoke areas, especially small smoke. Since self-attention is calculated in the shifted window instead of the whole images, the fine-grained features in different scales can be captured, which is beneficial to the recognition of small smoke.

The rest of this article is presented as follows: Section 2 describes the content of the data source, and Section 3 details the proposed methodology, including the overall framework, structural design, etc.; Section 4 discusses the comparison of different design results; the last section ends with a summary.

2. Datasets

To our best knowledge, there are two public wildfire datasets (https://cvpr.kmu.ac.kr/ and http://smoke.ustc.edu.cn/datasets.htm accessed on 23 March 2022). Some of the fire images in the two datasets are captured in the wild, and others are from some urban fires. In order to fairly evaluate the performance of our model in detecting wildfire smoke, we collect and record a number of smoke images in some complex outdoor environments. In our experiment, the datasets include the above two public datasets, 9 videos recorded by us, and the wildfire smoke images collected from the video data website (https://www.vcg.com/creative-video-search/yanwu/ accessed on 5 April 2022). In order to verify the anti-interference performance of our model, we also collected a large number of interference images such as clouds and fog on the internet. Table 1 shows the details of our datasets, and Figure 1 shows some example images from our datasets. Our datasets contain white smoke, black smoke, thick smoke, and thin smoke from different scenes in the wild, such as mountains, forests, fields, plains, etc. Especially, we also collect some challenging images such as fogged smoke images and small smoke images.

Table 1. The details of our datasets.

Datasets	Fire Smoke Image	Interference Images
Total number	700	300
Training	630	270
Testing	70	30

Figure 1. Example images from the datasets, which include white smoke and black smoke images, thick smoke and thin smoke images, small smoke images, smoke images with clouds, and interference images. (**a**–**c**) contain different types of smoke dataset in the field, and (**d**) show cloud interference dataset.

3. Methods

The existing fire smoke methods often have high false alarm rates in the face of the interference objects such as clouds and fog. In addition, wildfire images are usually taken from a long-distance shot, so the proportion of smoke area in an image is very small. In this paper, a novel Recursive BiFPN structure (RBiFPN) is proposed, and on this basis, the

YOLOV5 network structure is improved to build the wildlife smoke detection model shown in Figure 2. The network is divided into backbone, neckbone and prediction head. We retain the backbone structure of YOLOV5, and mainly improve its neckbone and prediction head.

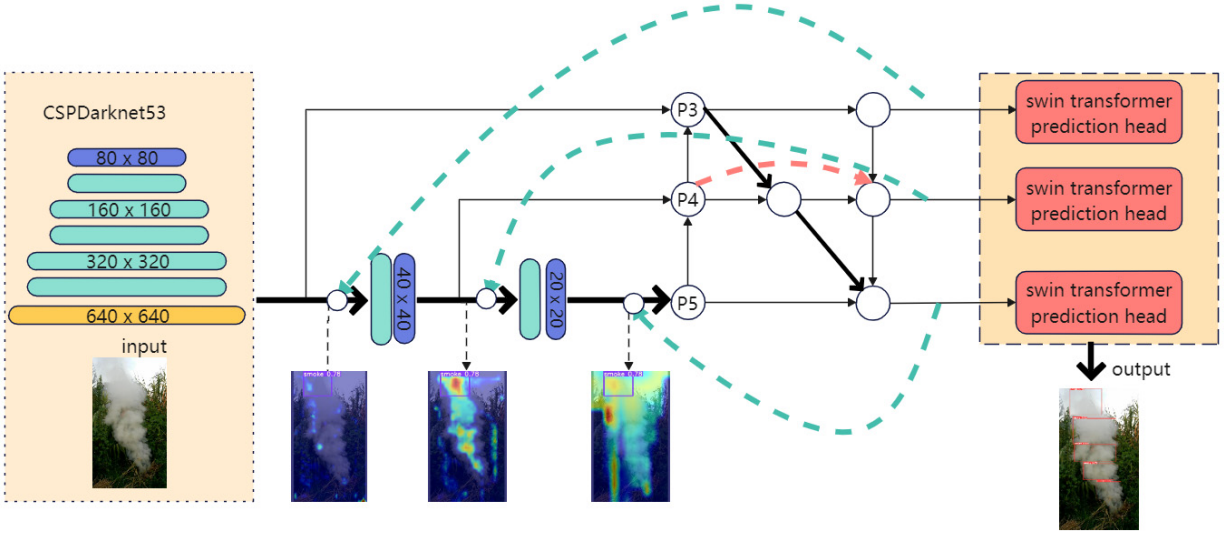


Figure 2. Structure of the proposed wildfire smoke detection network. The backbone network is CSPDarknet53 framework of YOLOV5, and Recursive BiFPN is designed for feature fusion and enhancement, and Swin Transformer for classification. The green dashed arrow indicates that the feature will be iterated back to the original backbone network once after being processed by BiFPN, and fused with the feature of the corresponding size, and then continue the subsequent processing. It is worth mentioning that the iteration is performed only once.

3.1. The Architecture of Wildfire Smoke Detection Network

The overall architecture of the wildfire smoke detection network is shown in Figure 2. As shown in Figure 2, the backbone network is CSPDarknet53 framework with the excellent feature extraction capability in YOLOV5.

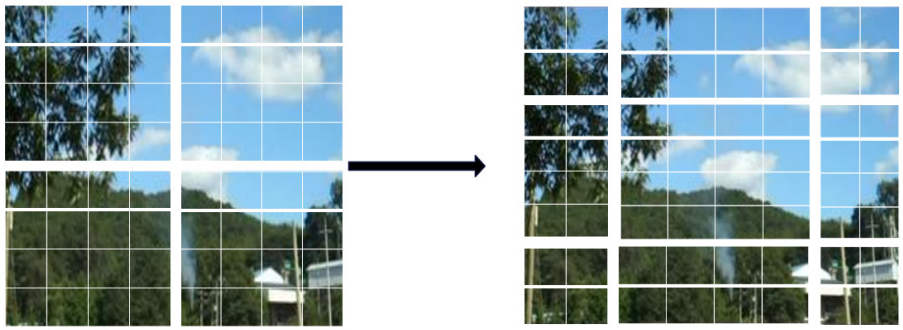
There are many factors that make it more difficult to locate the smoke objects. For instance, due to the influence of strong wind, smoke diffuses in the scene. In addition, the thick fog can also make the images blurred. Small smoke is generated by shooting from a long distance or early fire, and there are often interferences such as clouds in the background. In order to improve the feature extraction ability of the proposed network, we combine the advantages of BiFPN and RFPN and construct a novel Recursive BiFPN structure that replace the FPN of YOLOV5 in order to fuse and enhance features of different resolutions. The structure utilizes the feature enhancement function of the recursive operation, and at the same time, it can also filter out important feature layers, which is helpful for the multi-scale feature extraction of wildfire smoke.

Following the operation of backbone and neckbone, the prediction head identifies fire smoke according to the extracted features. In order to detect small smoke regions in images, inspired by [26,27], we replaced the prediction head of YOLOV5 with a Swin Transformer to improve the ability of detecting small smoke regions. Different from TPH, a Swin Transformer calculates self-attention only within non-overlapping windows, and thus the locality operation has a near-linear time complexity in the size of the input image. Moreover, as the network depth increases, image patches can be gradually merged to build a hierarchy structure.

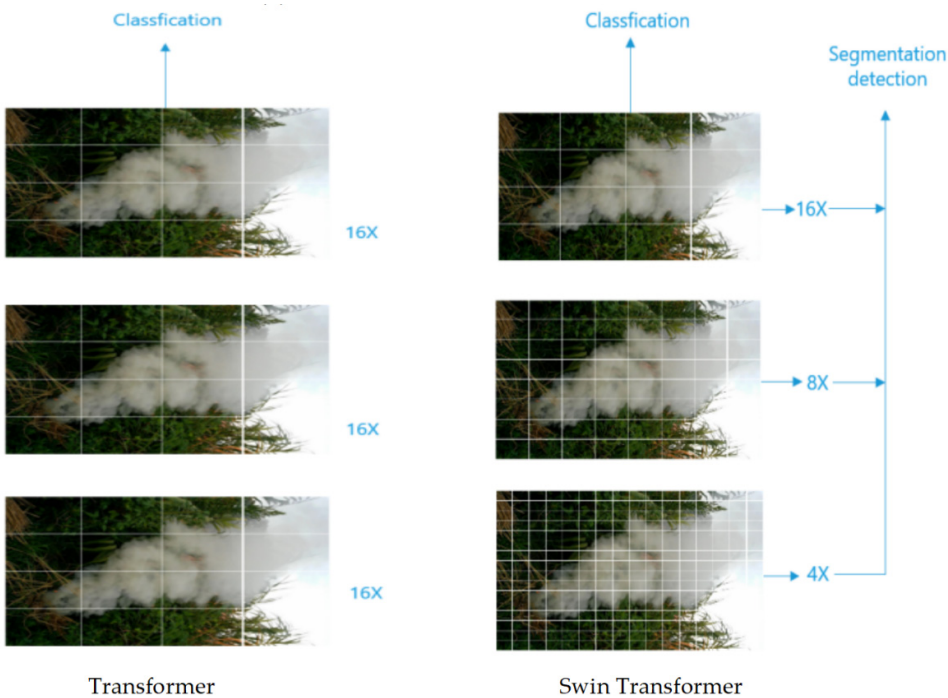
3.2. Workflow of Network Structure

As shown in Figure 2, we extract multi-scale features through the operations of several convolutional layers and the C3 layer, and obtain three feature layers with different sizes, P3, P4 and P5. The original channel of the input image is 64, after a series of convolution

processing, the channel number of the three feature layers P3, P4, and P5 is 256, 512, and 1024, respectively. Then, RBiFPN layer is constructed using the three effective feature layers [28,29]. RBiFPN performs feature fusion twice, that is, the output features of previous BiFPN are fused with the initial features obtained by the backbone network, and the new features generated are passed to the BiFPN structure again. The output feature map of the second BiFPN is input to the Swin Transformer prediction head. The Swin Transformer head has two Multi-Head Self Attention (MSA), window Multi-Head Self Attention (W-MSA for short) and Shifted window Multi-Head Self Attention (SW-MSA for short), followed by MLP. As shown in Figure 3, W-MSA extracts local information by computing self-attention within each local window. Compared with taking pixels as a sequence, W-MSA reduces the length of the sequence, [30]. SW-MSA uses the shifted windows to generate a cross-window connection in order to enable the interaction between two adjacent windows, thus achieving the ability of global modeling.



(**a**)Shifted windows of Swin Transformer



(**b**) Difference between Vit Transformer and Swin Transformer

Figure 3. Window operations of Swin Transformer (**a**) Through W-MSA, self-attention is calculated within each local window, and the shifted window enables the interaction between two adjacent windows and achieves the ability of global modeling. (**b**) The Vit Transformer generates a single low-resolution feature map, which has quadratic computational complexity in the input image size; Swin Transformer is based on the characteristics of W-MSA and SW-MSA, and can merge image patches in deeper layers, and the calculation of self-attention is performed in each local window, so it has linear computational complexity.

4. Experiments

In this section, we first introduce the experimental configuration. Then, we compare the impact of different neckbones and prediction heads on the recognition performance of our model. After that, we tested the anti-interference ability of the proposed model, and the ability to detect challenging images such as small smoke, black smoke and brown smoke images. Finally, we chose state-of-the-art smoke detection methods for comparison.

4.1. Experimental Configuration

The programs are written under the Pytorch framework and are trained and tested on a computer with an Intel i99900k CPU and an NVIDIA GeForce RTX 2080ti GPU. The ratio of the training subset to the testing subset is 9:1, and the initial learning rate is set to 0.001. We used the stochastic gradient method (SGD) to update the parameters of each layer and chose yolov5s.pt as the initial weight to start training. At the same time, considering that the backbone of the network has not been adjusted, its backbone part is frozen during training. The size of all images that are input into the network is fixed to 640×640. We train every network for 300 epochs, and the warmup_epochs is set to 3. We used average precision (AP) and average recall (AR) to evaluate the detection accuracy and false positive rate. In addition, we set false positive rate (FPR) to evaluate the false alarm rate. FPR represents the probability of false positive samples. We also used Map0.5 as an evaluation index, where Map0.5 denotes the value of Map (mean average precision) is 0.5. The detected smoke regions are considered valid if the value of the IoU (Intersection Over Union) is greater than 0.5. In order to evaluate the computation speed of our models, we chose the evaluation metric FPS (Frames Per Second) to compare more intuitively.

4.2. Evaluation of Feature Extraction

Since RBiFPN is an improvement on BiFPN, we compared the ability to the smoke feature enhancement and fusion of the two modules. The experiment adopts the backbone network of YOLOV5. Three prediction heads, YOLOV5 Head, TPH and Swin-TPH, are, respectively combined with the above two feature enhancement models. The experimental results of different network structures with different combinations are shown in Table 2. As seen from Table 2, after using RBiFPN to fuse and enhance the multi-scale features output by the backbone network, compared with BiFPN, the Precision and Map0.5 have a slight increase. Surprisingly, when using Transformer Prediction Head (whether TPH or Swin-TPH) to replace the original YOLOV5 Head, Precision and Map0.5 are significantly improved. The main reason is that white smoke is very similar to the cloud or fog. Although RBiFPN has been carried out to enhance features, the YOLOV5 Head may not be able to handle and judge the additional parameters obtained by RBiFPN well. However, as the depth deepens, RBiFPN can better merge the image patches to form a hierarchical structure, which makes the classification head more sensitive to the enhanced features. In addition, the features extracted by BiFPN are not enough to support the completely hierarchical processing of the Transformer Head, resulting in some noise interference.

Table 2. Comparision of neckbone.

Models	Precision	Map
BiFPN + GIOU_Loss	0.764	0.728
RBiFPN + GIOU_Loss	0.765	0.736
BiFPN + TPH	0.774	0.765
RBiFPN + TPH	0.821	0.805
BiFPN + Swin-TPH	0.764	0.751
RBiFPN + Swin-TPH	0.858	0.823

4.3. Anti-Interference

How to distinguish fire smoke from clouds in the sky has always been a challenging problem for wildfire smoke detection models. To solve the problems, we used RBiFPN

to fuse and enhance the multi-scale features, which contributes to assign high weights to smoke regions. As shown in Figure 4, compared with BiFPN, the feature maps generated by RBiFPN not only pays more attention to the smoke regions, but also can better distinguish clouds and smoke. We chose AP and FPR to evaluate the anti-interference ability, and the experimental results are shown in Section 4.5. As shown in Section 4.5, compared with other models, the proposed model has higher AP and lower FPR, so the proposed model has stronger anti-interference while detecting more smoke images.

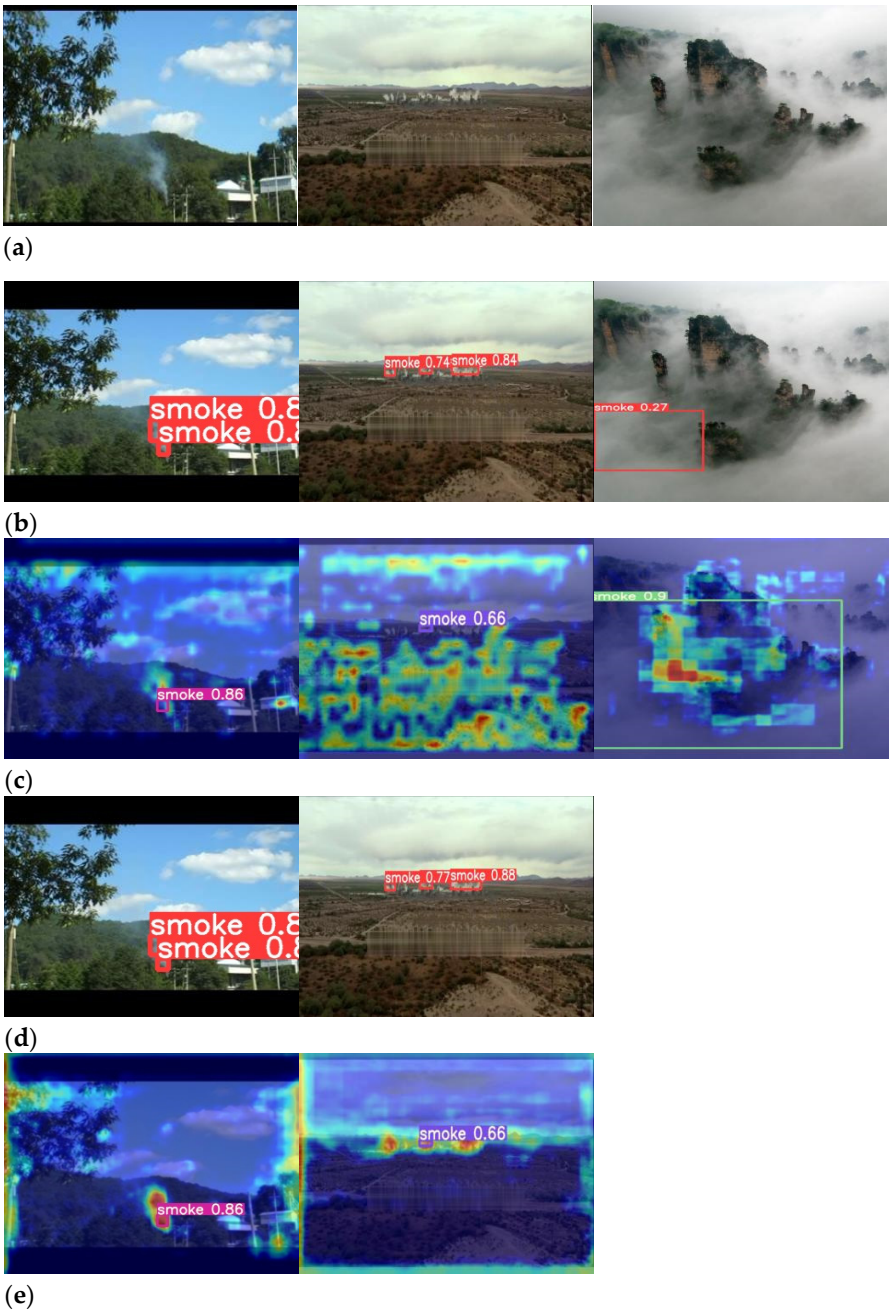


Figure 4. Comparison of anti-interference results. First row (**a**): the original images, Second row (**b**) the BiFPN detection results, Third row (**c**) the feature maps obtained by BiFPN, Fourth row (**d**) the RBiFPN detection results, Fifth row (**e**) the feature maps of obtained by RBiFPN. It is worth mentioning that the cloud and fog image (the third column) without fire do not have the attention map and the detected result to be shown in Fourth row (**d**) and Fifth row (**e**). However, in Second row (**b**) and Third row (**c**), the interference objects are falsely detected as fire smoke.

4.4. Identification of the Challenging Smoke Images

Swin-TPH assigns different weights for specific feature layers to detect small smoke in the images shot at a long distance or in an early fire, as shown in Figure 5. This is because the SW-MSA mechanism of Swin-TPH can form different receptive fields with different smoke areas by building a hierarchical structure. Furthermore, our model also has high detection accuracy for brown, black and other thick smoke except white smoke, because non-overlapping local windows and overlapping cross-window operations enhance the ability of local feature representation and global modeling.

Figure 5. The detection results of challenging smoke images. (**a**): small smoke, (**b**): black and brown smoke.

We also construct the dataset of small smoke and black smoke, and the dataset includes 100 small smoke images and 50 black smoke images for testing. As shown in Table 3, the proposed model can achieve better balance between accuracy and recall and has higher Map value than other competing models.

Table 3. Comparison of different smoke.

Models	Precision	Recall	Map
Faster RCNN	0.378	0.701	0.584
Efficientdet	0.728	0.406	0.497
SSD	0.780	0.131	0.411
BiFPN + GIOU_Loss	0.845	0.631	0.632
RBiFPN + GIOU_Loss	0.894	0.636	0.647
RBiFPN + TPH	0.819	0.671	0.663
Ours (RBiFPN + Swin-TPH)	0.847	0.674	0.674

4.5. Comparison of the Model

In this section, we compare the proposed network with Faster RCNN [14], Efficientdet [16], SSD [17] and YOLOV5 [5] that are widely used for smoke detection. We also designed different combination modes of the neckbone (BiFPN or RBiFPN) and the prediction head (YOLOV5 Head, TPH, or Swin-TPN) for comparison. As shown in Table 4, the proposed network has obvious advantages on the metric Map0.5, which means that our network can not only detect more wildfire smoke and have less missed rate than other comparative network structures. Faster RCNN or the original yolov5, still have a certain gap with our network in the detection of smoke. The neckbones of both Faster RCNN and YOLOV5 are inferior to our network, which proves that RBiFPN can better distinguish smoke from smoke-like objects by enhancing the features.

Table 4. Comparison of precision and recall of fire smoke detection.

Framework	Neckbone	Prediction Head	AP	AR	Map	FPR	Param(M)	FPS	FLOPs(G)
Faster RCNN	None	Sparse-Prediction	0.750	0.817	0.752	0.387	28.48	11.5	939.6
Efficientdet	BiFPN	Class + Box prediction net	0.729	0.469	0.611	-	3.874	25.9	5.1
SSD	None	None	0.820	0.278	0.599	-	26.3	93.8	62.8
YOLOV5	PANet	GIOU_Loss	0.760	0.781	0.717	0.203	46.14	10.6	107.8
YOLOV5	BiFPN	GIOU_Loss	0.764	0.783	0.728	0.093	46.86	10.4	114.9
YOLOV5	RBiFPN	GIOU_Loss	0.785	0.787	0.796	0.060	101.04	6.1	199.4
YOLOV5	BiFPN	TPH	0.774	0.803	0.765	0.058	51.94	5.5	111.3
YOLOV5	RBiFPN	TPH	0.821	0.812	0.805	0.054	106.56	4.0	202.9
Ours	RBiFPN	Swin-TPH	0.858	0.826	0.823	0.053	103.14	4.1	407.8

AP and AR are also important metrics for detection accuracy. FPR is also an important indicator which is used to evaluate the false alarm rate. As shown in Table 4, our network achieves the best performance according to the three metrics: AP, AR and Map. However, other competing models are unable to get a balance between AP and AR. Meantime, the Map value of our network also remains above 0.82. It can be also seen that the FPR value of our network is the lowest in comparison. It is worth mentioning that all the three metrics can be improved by replacing the prediction head with Swin-TPH. The Swin-TPH in the proposed network not only increases the smoke detection performance, but also the parameters and FPS are competitive with TPH. However, FLOPs of our model are double the TPH. To achieve the optimal metrics, we adopt the mode Swin-B of Swin Transformer [27]. Swin-B has more channels and layers, so it has computation complexity similar to computation complexity similar to ViT-B/DeiT-B, which leads to the problem of a large growth of FLOPs. We also depicted the training and validation loss curves for our network. As shown in Figure 6, the model converges gradually from 200 epochs. In addition, both Efficientdet and SSD miss many positive samples while carrying out fire smoke detection tasks, so their AR values are extremely low. Due to high rejection rate for positive samples, we do not calculate the metric FPR for Efficientdet and SSD.

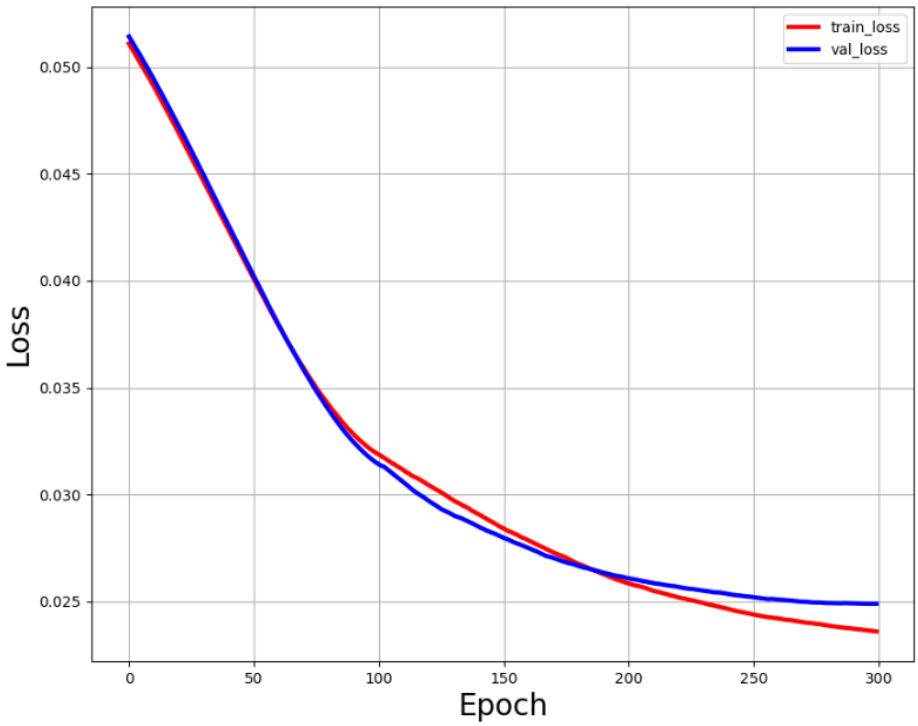


Figure 6. Training and validation loss curves for our network. The red line and the blue line represent the changes in the loss function generated during training and validation, respectively. The model gradually converges from 200 epochs.

5. Conclusions

This paper proposes a convolutional network for wildfire smoke detection based on RBiFPN and Swin-TPH. The RBiFPN structure can enhance and fuse multi-scale features, and the hierarchical Swin-TPH can improve the expression ability of local features and global information. In addition to being able to recognize common white smoke well, our network also shows good recognition performance in the face of small smoke shot from a long distance, or smoke with cloud interference in the background. These improvements ensure that our network is more suitable for wildfire smoke images.

However, our model suffers from a similar dilemma as the existing models, that is, the model needs to be further improved for different color wildfire smoke with different combustion materials and erratic wildfire smoke affected by strong winds. In the future, our work will extract the temporal features of fire smoke such as optical flow and integrate spatial-temporal features. We will also focus on the lightweight of the network, so that it is cost-efficient. We will try to prune the Swin Transformer using some state-of-the-art pruning methods [31,32], and adopt separable convolution to make the backbone network more lightweight. In particular, we hope to apply the design ideas about small target detection in Swin Transformer to other more lightweight Transformer network like mobile ViT [33]. Furthermore, since there are fewer images in the smoke dataset, we will try to develop wildfire smoke detection strategy for small sample datasets, in order to reduce the network's dependence on the number of training samples.

Author Contributions: Conceptualization, A.L. and Y.Z.; Methodology, A.L.; Software, A.L.; Validation, A.L. and Z.Z.; Formal Analysis, A.L.; Investigation, Z.Z.; Resources, Y.Z.; Data Curation, A.L.; Writing—Original Draft Preparation, A.L.; Writing—Review and Editing, Y.Z.; Visualization, A.L; Supervision, Y.Z.; Project Administration, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Guo, Y.; Chen, G.; Wang, Y.-N.; Zha, X.-M.; Xu, Z. Wildfire Identification Based on an Improved Two-Channel Convolutional Neural Network. *Forests* **2022**, *13*, 1302. [CrossRef]
2. Zhang, J.; Li, W.; Han, N.; Kan, J. Forest fire detection system based on a zigbee wireless sensor network. *Front. For. China* **2008**, *3*, 369–374. [CrossRef]
3. Aslan, Y.E.; Korpeoglu, I.; Ulusoy, Ö. A framework for use of wireless sensor networks in forest fire detection and monitoring. *Comput. Environ. Urban Syst.* **2012**, *36*, 614–625. [CrossRef]
4. Dener, M.; Özkök, Y.; Bostancıolu, C. Fire detection systems in wireless sensor networks. *Procedia-Soc. Behav. Sci.* **2015**, *195*, 1846–1850. [CrossRef]
5. Wang, Z.; Wu, L.; Li, T.; Shi, P. A Smoke Detection Model Based on Improved YOLOv5. *Mathematics* **2022**, *10*, 1190. [CrossRef]
6. Jiang, H.; Yuan, W.; Ru, Y.; Chen, Q.; Wang, J.; Zhou, H. Feasibility of identifying the authenticity of fresh and cooked mutton kebabs using visible and near-infrared hyperspectral imaging. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *282*, 121689. [CrossRef]
7. Shi, F.; Wang, J.; Shi, J.; Wu, Z.; Wang, Q.; Tang, Z.; He, K.; Shi, Y.; Shen, D. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* **2021**, *14*, 4–15. [CrossRef]
8. Zhu, Z.; Wei, H.; Hu, G.; Li, Y.; Qi, G.; Mazur, N. A Novel Fast Single Image Dehazing Algorithm Based on Artificial Multiexposure Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–23. [CrossRef]
9. Wang, T.; Shi, L.; Yuan, P.; Bu, L.; Hou, X. A new fire detection method based on flame color dispersion and similarity inconsecutive frames. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 151–156.
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. Gunay, O.; Toreyin, B.U.; Kose, K.; Cetin, A.E. Entropy-functional-based online adaptive decision fusion framework with application to wildfire detection in video. *IEEE Trans. Image Process.* **2012**, *21*, 2853–2865. [CrossRef]
12. Jia, Y.; Chen, W.; Yang, M.; Wang, L.; Liu, D.; Zhang, Q. Video smoke detection with domain knowledge and transfer learning from deep convolutional neural networks. *OPTIK* **2021**, *240*, 166947. [CrossRef]
13. He, L.; Gong, X.; Zhang, S.; Wang, L.; Li, F. Efficient attention based deep fusion CNN for smoke detection in fog environment. *Neurocomputing* **2021**, *434*, 224–238. [CrossRef]
14. 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]
15. Zhao, E.; Liu, Y.; Zhang, J.; Tian, Y. Forest Fire Smoke Recognition Based on Anchor Box Adaptive Generation Method. *Electronics* **2021**, *10*, 566. [CrossRef]
16. Li, F.S.; Yao, D.F.; Jiang, M.H.; Kang, X.C. Smoking behavior recognition based on a two-level attention fine-grained model and EfficientDet network. *J. Intell. Fuzzy Syst.* **2022**, *43*, 5733–5747. [CrossRef]
17. Wu, S.X.; Zhang, L.B. Using Popular Object Detection Methods for Real Time Forest Fire Detection. In Proceedings of the 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; pp. 280–284.
18. Luo, Y.; Zhao, L.; Liu, P.; Huang, D. Fire smoke detection algorithm based on motion characteristic and convolutional neural networks. *Multimed. Tools Appl.* **2018**, *77*, 15075–15092. [CrossRef]
19. Li, X.; Xu, Z.H.; Shen, X.; Zhou, Y.; Xiao, B.; Li, T.Q. Detection of Cervical Cancer Cells in Whole Slide Images Using Deformable and Global Context Aware Faster RCNN-FPN. *Curr. Oncol.* **2021**, *28*, 3585–3601. [CrossRef] [PubMed]
20. Wang, Y.; Zell, A. Yolo+FPN: 2D and 3D Fused Object Detection with an RGB-D Camera. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2020; pp. 4657–4664.
21. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
22. Yu, J.M.; Zhang, W. Face Mask Wearing Detection Algorithm Based on Improved YOLO-v4. *Sensors* **2021**, *21*, 3263. [CrossRef]
23. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038.
24. Wang, Y.F.; Hua, C.C.; Ding, W.L.; Wu, R. Real-time detection of flame and smoke using an improved YOLOv4 network. *Signal Image Video Process.* **2022**, *16*, 1109–1116. [CrossRef]
25. Qiao, S.Y.; Chen, L.C.; Yuile, A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2022; pp. 10208–10219.

26. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the 2021 IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
27. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.
28. Li, Z.; Zou, H.; Sun, X.; Zhu, T.; Ni, C. 3d expression-invariant face verification based on transfer learning and siamese network for small sample size. *Electronics* **2021**, *10*, 2128. [CrossRef]
29. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. *arXiv* **2021**, arXiv:1812.10025.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 5998–6008.
31. He, H.; Liu, J.; Pan, Z.; Cai, J.; Zhang, J.; Tao, D.; Zhuang, B. Pruning Self-attentions into Convolutional Layers in Single Path. *arXiv* **2021**, arXiv:2111.11802.
32. Zhu, M.; Han, K.; Tang, Y.; Wang, Y. Visual Transformer Pruning. *arXiv* **2021**, arXiv:2104.08500.
33. Mehta, S.; Rastegari, M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. *arXiv* **2021**, arXiv:2110.02178.