


Article

Spatial–Temporal Attention Two-Stream Convolution Neural Network for Smoke Region Detection

Zhipeng Ding , Yaqin Zhao *, Ao Li and Zhaoxiang Zheng

College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; dingzhipeng@njfu.edu.cn (Z.D.); lostfate@njfu.edu.cn (A.L.); zhengzhaoxiang@njfu.edu.cn (Z.Z.)

* Correspondence: zhaoyaqin@njfu.edu.cn

Abstract: Smoke detection is of great significance for fire location and fire behavior analysis in a fire video surveillance system. Smoke image classification methods based on a deep convolution network have achieved high accuracy. However, the combustion of different types of fuel can produce smoke with different colors, such as black smoke, grey smoke, and white smoke. Additionally, the diffusion characteristic of smoke can lead to transparent smoke regions accompanied by colors and textures of background objects. Therefore, compared with smoke image classification, smoke region detection is a challenging task. This paper proposes a two-stream convolutional neural network based on spatio-temporal attention mechanism for smoke region segmentation (STCNNsmoke). The spatial stream extracts spatial features of foreground objects using the semi-supervised ranking model. The temporal stream uses optical flow characteristics to represent the dynamic characteristics of smoke such as diffusion and flutter features. Specifically, the spatio-temporal attention mechanism is presented to fuse the spatial and temporal characteristics of smoke and pay more attention to the moving regions with smoke colors and textures by predicting attention weights of channels. Furthermore, the spatio-temporal attention model improves the channel response of smoke-moving regions for the segmentation of complete smoke regions. The proposed method is evaluated and analyzed from multiple perspectives such as region detection accuracy and anti-interference. The experimental results showed that the proposed method significantly improved the ability of segmenting thin smoke and small smoke.

Keywords: smoke detection; convolutional neural network; two-stream; spatio-temporal attention



Citation: Ding, Z.; Zhao, Y.; Li, A.; Zheng, Z. Spatial–Temporal Attention Two-Stream Convolution Neural Network for Smoke Region Detection. *Fire* **2021**, *4*, 66. <https://doi.org/10.3390/fire4040066>

Academic Editor: James A. Lutz

Received: 21 August 2021

Accepted: 29 September 2021

Published: 3 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 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

The fire occurrence not only leads to the destruction of the natural ecological environment but also seriously threatens the safety of human life and property. With the rapid development of communication, image processing, and artificial intelligence technology, a fire video monitoring system has become important auxiliary means for fire alarms. As a main visual phenomenon in early fires, smoke is often used to give an alarm for the early fire or analyze the fire spread trend. Therefore, many efforts have been made in the studies on smoke image recognition and smoke region detection.

The traditional smoke detection algorithm is based on manually extracted features, such as color, shape, texture, and motion features. For example, the spectral probability density was represented to extract candidate smoke regions by comparing the color histogram models in HSI color space that describe colors by hue, saturation, and intensity [1]. The literature [2] utilized the low-chromaticity characteristic of smoke in the YUV color space; the Y component determines the brightness of the color, and the U and V components refer to the chrominance to detect the smoke region. Besides the color characteristics, some smoke region methods also identified the geometric features (shape, contour, and area) of smoke regions [3,4]. In order to improve the detection accuracy, some studies combined color and shape features [5]. Furthermore, the texture features such as LBP (local binary

pattern) histograms and LBPV (local binary pattern variance) pyramids were used for smoke region detection [6,7]. However, the above static features are unable to distinguish true smoke from smoke-like static objects. Thus, several attempts have been made to apply the motion characteristics, such as the optical flow [8] and cumulative motion direction features, to eliminate the interference of static camouflaged objects [9].

Although the above methods based on the manual feature extraction can identify smoke objects that are obviously different from backgrounds in visual, the anti-interference ability of complex smoke scenes and the detection accuracy of small smoke regions are not satisfactory. In recent years, with the rapid development of artificial intelligence and computer vision technology, the convolutional neural network (CNN) has been successfully applied in image classification, object detection, face recognition, semantic segmentation, and other fields [10,11]. CNN was also used to extract smoke features [12,13]. In order to further improve the computational efficiency, several studies tried to use a smaller convolution kernel in the CNN architecture and canceled the dense full connection layer so as to obtain a high smoke-classification accuracy [14]. Some deep convolution networks were used to determine the location of smoke. For example, the VGG-16 (Visual Geometry Group Network) architecture was applied to locate smoke so as to achieve a better balance between the detection accuracy and time efficiency [15]. An attempt was made to predict the coordinates of smoke objects based on Faster R-CNN (region-based CNN) and realize the end-to-end detection of smoke objects in forest fires [16]. A joint detection framework based on fast R-CNN and 3D-CNN (three-dimensional CNN) was developed to realize smoke target localization [17]. A video smoke detection method based on a deep saliency network is proposed by combining a depth feature map and a saliency map to realize smoke pixel-wise detection [18].

As mentioned above, the above research on CNN-based smoke image recognition or smoke target location has achieved great progress. Although most of the traditional methods based on the manual feature extraction have focused on pixel-wise smoke detection (that is, smoke region detection), CNN-based smoke region detection research is rarely reported. Sometimes, we need to focus on the smoke spreading or emerging smoke regions expected for locating smoke targets. Compared with smoke object location, the temporal changes of smoke regions or contours in video sequences can provide the more detailed and timely guiding data support for fire spread forecasting [19] and dangerous smoke classification [20], especially relatively small changes of smoke regions in a short time interval, which often occur in the early stage fire. In that case, pixel-wise smoke detection is needed to compute the geometry size changes of smoke-moving regions in video sequences [19,20]; however, smoke proposal boxes obtained by smoke object detection methods overlap almost entirely, which fails to achieve fire-spreading analysis. Sometimes, region-based smoke object location is unable to accurately grade the level of fire when the shapes of smoke regions are long and narrow, as shown in Figure 1. Therefore, smoke region detection has greater research value and practical significance for grading the level of fire or fire behaviors, including the fire spreading [21]. Therefore, in view of the above problems, this paper proposes a two-stream network based on a spatio-temporal attention model for smoke region detection.



Figure 1. Unsatisfactory examples of region-based smoke detection methods.

The spatial and channel attention mechanism was also introduced to recognize the smoke image [22]. However, the models cannot realize pixel-wise smoke detection. In [23], a new ranking attention module (RANet) was proposed to automatically rank and select foreground and background feature maps based on pixel-wise similarity. Although it can achieve high accuracy, the ranking model emphasizes the spatial similarity between the current frame and the template frame more and thus ignores the temporal information of continuous frames in the video. The ranking model [23] cannot achieve good results when the positions or the shapes of smoke regions change greatly in a continuous frame sequence due to smoke diffusion characteristics. Therefore, on the basis of the ranking attention module, we try to introduce the spatio-temporal attention mechanism to make the network pay more attention to the motion pixels in a smoke video in order to reflect the flutter or diffusion characteristics of smoke and improve the accuracy of smoke region segmentation.

The main contributions of this paper are as follows:

(1) We establish a spatio-temporal two-stream network to integrate the spatial and temporal features effectively. The spatial stream uses the semi-supervised ranking model to extract the appearance characteristics of smoke objects. The temporal stream takes optical flow features as the input to represent the dynamic characteristics, such as diffusion and flutter, and integrates the temporal and spatial features to realize smoke region detection.

(2) In order to remove the smoke-like interference and pay more attention to the dynamic characteristics of smoke regions, this paper utilizes a spatio-temporal attention mechanism to realize the fusion of temporal and spatial characteristics. The mechanism can predict the channel attention weights and then improve the response of the corresponding attributes or channels of smoke motion parts, which is helpful to segment complete smoke regions.

(3) Aiming to resolve the conflict between the large amount of fire monitoring data and real-time requirements, this paper uses the semi-supervised moving target ranking model to compute the feature correlations between the current frame and the template frame and selects the feature maps with higher similarity as the static feature maps of smoke regions instead of all the feature maps so as to achieve the balance between the detection speed and accuracy.

2. Materials and Methods

2.1. Network Structure

This paper proposes a smoke region detection network based on temporal and spatial characteristics. The overall structure of the network is shown in Figure 2; it contains a spatial stream and a temporal stream. The spatial stream uses the convolutional network to extract the feature maps of the current frame and the reference frame, respectively, and then the ranking attention module [23] calculates the correlation between the two convolutional feature maps. The temporal stream takes the optical flow images as input, and its convolution feature maps are fused with the correlation feature maps of the spatial stream through the spatio-temporal attention module to generate a spatio-temporal feature map. Finally, the size of the feature map is enlarged by the decoding module.

In the temporal stream branch, we first use the Siamese network (the backbone is ResNet-101 network) as the encoder to extract the feature maps of both the template frame and the current frame. The network structure for extracting the feature maps is shown in Table 1, where the convolution strides of the first residual modules in Conv1, Conv2_x, Conv3_x, and Conv4_x are 2. Therefore, the size of the feature map finally output by the encoder module is 1/16 of its corresponding original image.

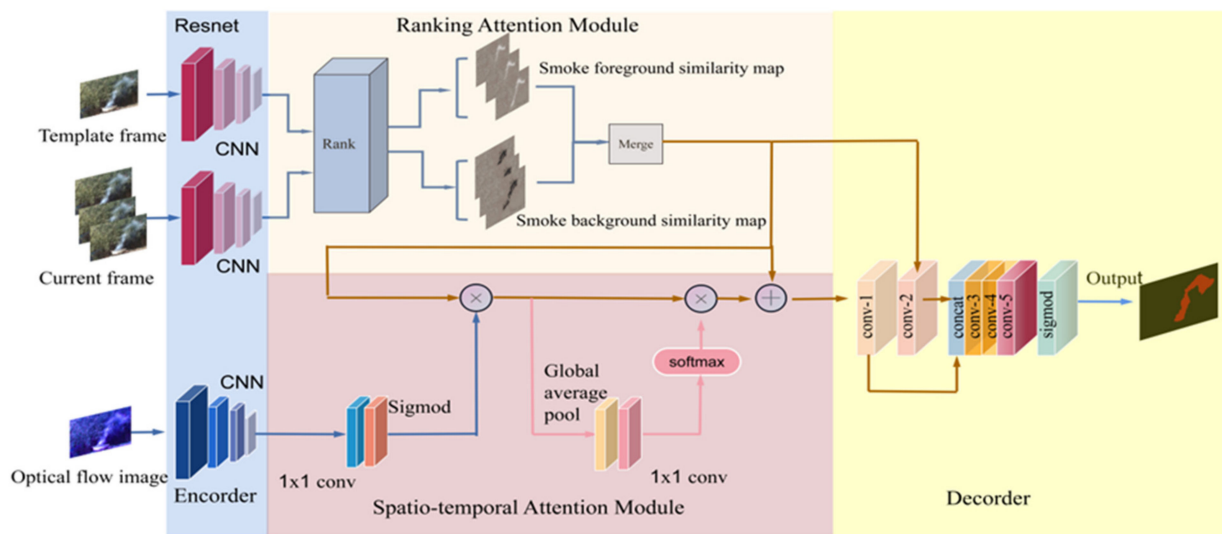


Figure 2. The complete structure of the proposed network.

Table 1. The structure of the feature extraction network.

Layer Name	Input		Output
Conv1	$W \times H \times 3$	$7 \times 7, 64$, stride 2	$W/2 \times H/2 \times 64$
Pool	$W/2 \times H/2 \times 64$	3×3 max pool, stride 2	$W/4 \times H/4 \times 64$
Conv2_x	$W/4 \times H/4 \times 64$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$W/4 \times H/4 \times 256$
Conv3_x	$W/4 \times H/4 \times 256$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$W/8 \times H/8 \times 512$
Conv4_x	$W/8 \times H/8 \times 512$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$W/16 \times H/16 \times 1024$

We treat each pixel of the template frame as a template to perform correlation calculation with the current frame, so each foreground and background pixel of the template frame can obtain a similarity map after matching with the current frame (detailed explanation of correlation calculation in Section 2.3). Since the number of the similarity maps dynamically changes with the number of the foreground or background pixels in different videos, which lead to change in the number of channels, we rank and select the similarity maps according to the correlation value between the current frame and the template frame.

2.2. Spatio-Temporal Attention Module

The encoder structure of extracting the temporal feature maps is the same as that of spatial stream branch. After obtaining spatial and temporal feature maps, we utilize the spatio-temporal attention mechanism to emphasize important positions or elements in smoke images. Firstly, a 1×1 convolution is used to align the shape of the motion feature with that of the spatial feature. Then, spatial feature maps are multiplied by an attention map of size $H \times W$ to generate the spatio-temporal feature map spatially highlighted by a motion feature. However, due to the lack of texture information and complex semantic in optical flow images, the features extracted by the above method may contain additional noise. Therefore, it is necessary to consider how to reasonably realize the channel-wise attention of smoke motion features to spatial features. We obtain a one-dimensional vector by global average pooling followed by a 1×1 convolution, which predicts the channel-

wise weights to realize global representation of spatio-temporal features. Further, smoke corresponding attributes or channel response can be improved by predicting the one-dimensional channel attention weights based on the spatio-temporal feature vectors of smoke moving regions, which helps to segment complete smoke regions.

The spatio-temporal attention model can be described by the following formulas:

$$S_k = S_e \otimes \text{Sigmoid}(f(S_m)) \quad (1)$$

$$S_c = S_k \otimes [\text{Softmax}(f'(GAP(S_k))) \bullet C] + S_e \quad (2)$$

where S_e , S_k and S_c are the tensors with the size of $C \times W \times H$. S_e and S_m represent the static appearance feature and motion feature of smoke, respectively, and S_k represents the spatio-temporal feature already spatially highlighted by a motion feature. S_c is the spatio-temporal feature of smoke after introducing channel attention.

In the above Formula (1), \otimes denotes the element multiplication. $f()$ and $f'()$ are 1×1 convolution (as shown in Figure 2), and the output channels are 1 and C , respectively. $GAP()$ represents the global average pool in the spatial dimension. C is a single scalar, and equals to the number of elements in the output of Softmax function. $GAP(S_k)$ obtains the global representation of S_k and outputs a single vector of C elements. Based on global representation, $f'()$ predicts the weight vector of C scalar weights for channels. The attention weights on these channels aim to select or strengthen the response of the basic attributes of smoke, such as edges, boundaries, colors, textures, and semantics. $\text{Softmax}() \bullet C$ normalizes the output of $f'()$ and makes the average value of attention weights equal to 1. $S_k \otimes []$ multiplies the characteristic column of each spatial position of S_k by the normalized attention vector. After obtaining the smoke appearance features, we send S_c to the decoder to predict the final smoke segmentation map with the output of the appearance branch.

2.3. Smoke Ranking Module

To evaluate the correlation between the template frame and the current frame, we find the matching relationship of pixels by calculating and ranking the similarity maps. The schematic diagram of calculating the correlation is shown in Figure 3.

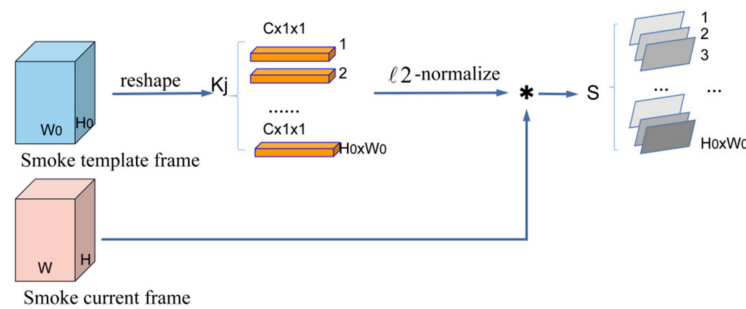


Figure 3. Correlation between the template frame and the current frame.

As shown in Figure 3, we define $I^1 \in \mathbb{R}^{C \times H_0 \times W_0}$ and $I^t \in \mathbb{R}^{C \times H \times W}$ as the features of the smoke template frame and current frame, respectively, which are extracted by the Siamese encoder, where C is the number of feature channels, and $H_0(W_0)$ and $H(W)$ represent the height (width) of the feature maps of the template frame and the current frame, respectively. We first reshape the template feature $I^1 \in \mathbb{R}^{C \times H_0 \times W_0}$ into $H_0 \times W_0 \times (C \times 1 \times 1)$ and represent the reconstructed template feature set as $K = \{K_j | j = 1, \dots, H_0 \times W_0\}$, which contains $H_0 \times W_0$ features with the size of $C \times 1 \times 1$. Then, by calculating the correlation between the $\downarrow 2$ normalized feature K_j in the template frame I^1 and the current frame I^t , we obtain the similarity map $S_j = K_j * I^t$. As shown in Formula (3). We

define $S \in \mathbb{R}^{H_0 W_0 \times H \times W}$ as the set of correlation maps, and each element in S represents a similar map.

$$S = \{S_j | S_j = K_j * I^t\}_{j \in \{1, \dots, H_0 \times W_0\}} \quad (3)$$

The structure of the smoke-ranking module is shown in Figure 4; the foreground and background masks of the template frame are used to filter the foreground and background similarity maps. Specifically, we exchange the space and channel dimensions of the pixel-wise similarity maps and then multiply them with the foreground or background mask, respectively, to obtain the foreground features and background features. After that, a ranking score is computed to indicate the importance of each similarity map. We calculate the sum of tensors after the channel-wise global maximum pool of tensors to obtain the ranking score. The higher the score, the higher the importance of the corresponding similarity map. The maximum value of the channels of each similarity map represents the probability of the corresponding pixel in the template frame to find a matching pixel in the current frame. Finally, we reshape the ranking score metric into a vector, and rank the maps according to the corresponding scores from largest to smallest. If the number of smoke foreground similarity maps is larger than the target channel size (set to 256), the redundant features are discarded. Otherwise, we need to use zero maps to pad the ranking feature in order that the channel size can be fixed. A similar scheme is also used to build the background similarity maps.

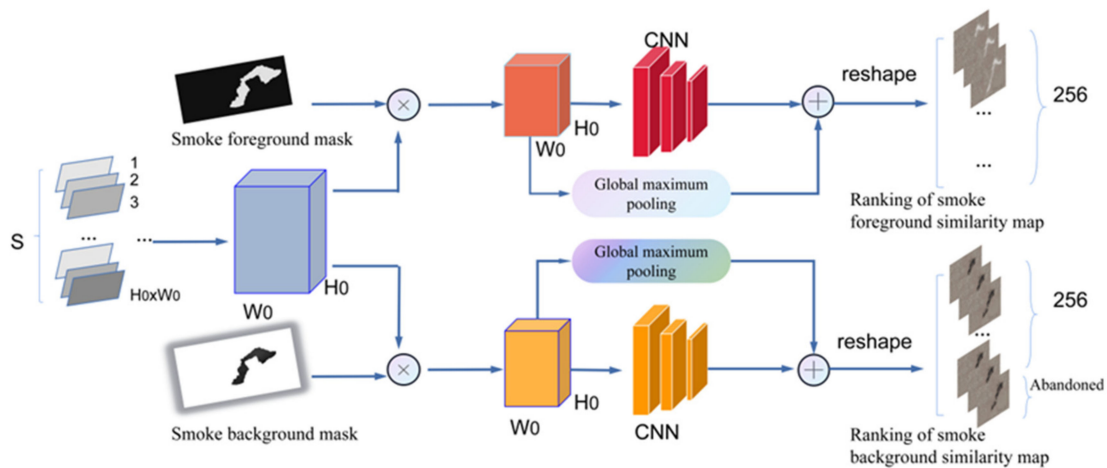


Figure 4. Smoke ranking module.

2.4. Datasets

We selected the public datasets widely used for evaluating smoke region detection methods. The public datasets were sourced from the websites <https://cvpr.kmu.ac.kr/> (accessed on 10 June 2020) and <http://smoke.ustc.edu.cn/datasets.htm> (accessed on 11 June 2020). We also recorded smoke videos from five different scenes and 10 smoke-like videos. As shown in Table 2, the datasets are divided into four categories: Dataset 1, Dataset 2, Dataset 3, and Dataset 4. Dataset 1 and Dataset 2 were recorded by us and used for positive samples and negative samples, respectively. All the samples in Dataset 3 and Dataset 4 are from public datasets. We divided each video into some small video sequences of the same amount of frames (30 frames each sequence) and randomly selected a total of about 800 smoke video sequences and 200 smoke-like video sequences to train the proposed network. The resting video sequences were used for testing sets. To better test the performance of our proposed method, we selected the public datasets from Dataset 3 and Dataset 4 as most of the testing sets that are composed of both the challenging smoke videos such as thin smoke, thick smoke, small smoke, and interference scenes including fog and clouds in the sky. Figure 5 shows several samples from the above four datasets. All the videos from the four datasets were captured with a fixed camera. Although the

proposed network supports any size input, the size of the images in the training sets was resized to 400×400 in order to support multiple batches of training.

Table 2. Details of the datasets in the experiments.

Dataset	Sample Type	Total Number of Images	Source
Dataset 1	Smoke samples	6000	Recorded by us
Dataset 2	Smoke-like samples	1500	Recorded by us
Dataset 3	Smoke samples	18,000	Public dataset
Dataset 4	Smoke-like samples	4500	Public dataset

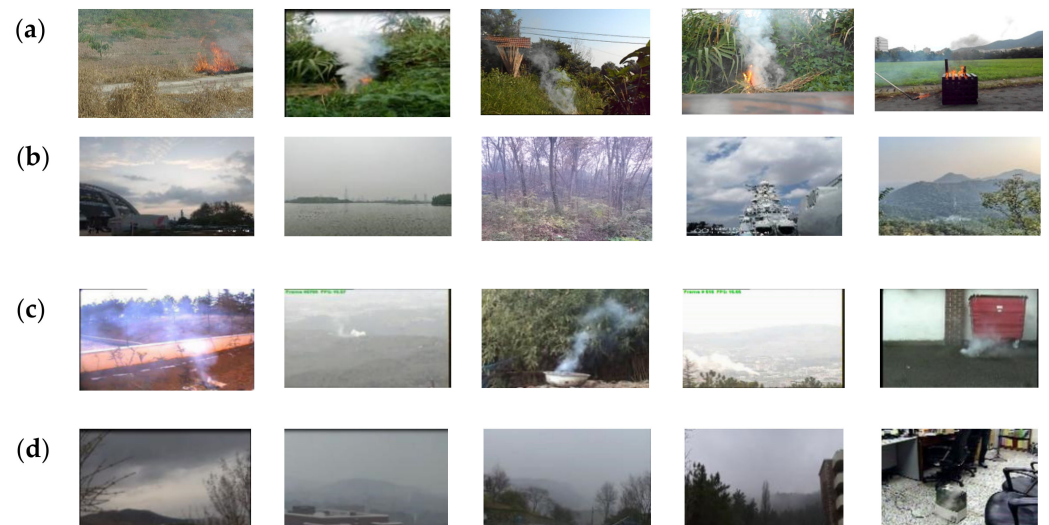


Figure 5. Several samples from the above four datasets. (a) smoke samples from Dataset 1; (b) smoke-like samples from Dataset 2; (c) smoke samples from Dataset 3; (d) smoke-like samples from Dataset 4.

2.5. Model Evaluation

Object detection algorithms are often evaluated by accuracy, precision, recall, *IoU*, etc. *IoU* is a common evaluation metric for target detection and object segmentation. The balance between recall and precision is very important for a model. It can be seen from Formula (2) that has a short-board effect compared to the method of obtaining the average value, which can better explain the quality of a model. Therefore, we use *IoU* [24] and *F1 – score* [24] to more objectively evaluate the smoke region detection method proposed in this paper.

$$IoU = \frac{groundTruth \cap prediction}{groundTruth \cup prediction} \quad (4)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (5)$$

where the metric *IoU* is defined as the overlap area between the detected flame region (named as *prediction* in Equation (4)) and the ground truth (named as *groundTruth* in Equation (4)) divided by the area of union between the detected flame region and the ground truth. *F1 – score* is the most used evaluation metric and defined as the harmonic mean of precision and recall, taking into account the detected regions with *TP*, *FP* and *FN*. In Equation (5), $precision = TP / (TP + FP)$ and $recall = TP / (TP + FN)$, where *TP* represents the number of smoke pixels correctly predicted, *FP* represents the number of the background pixels that are incorrectly predicted as smoke pixels, and *FN* represents the number of the smoke pixels that are incorrectly predicted as backgrounds.

3. Results

3.1. Training

In this paper, we used the PyTorch deep learning framework to implement the smoke region detection network model and used the stochastic gradient method (SGD) to update the parameters of each layer. The initial learning rate is 0.0001. The computer for training and testing is configured as an Intel Core i9 processor and a 2080 ti 11G GPU.

In the experiment, the three deep convolutional networks, FCN (fully convolutional networks), Deeplab V3+ (Deeplab version 3), and RANet [23], were compared with our proposed network to verify its performance. FCN and Deeplabv3+ are two kinds of deep learning networks that are recognized for image segmentation with better results. RANet, as a semi-supervised target detection network, is the spatial feature extraction model adopted by the spatial stream branch of this paper, and its application in experimental comparison can highlight the performance improvement of the proposed method.

3.2. Comparison

We used part of the data in Dataset 3 to evaluate recognition accuracy; several representative testing samples are shown in Figure 6. Dataset 3 contains different types of smoke regions, such as thin smoke regions caused by smoke diffusion (left image of the first row in Figure 6), distant smoke area (right image of the first row in Figure 6), small smoke area in the fog (left image of the second row in Figure 6), etc. These challenging videos bring great difficulty to smoke region detection.

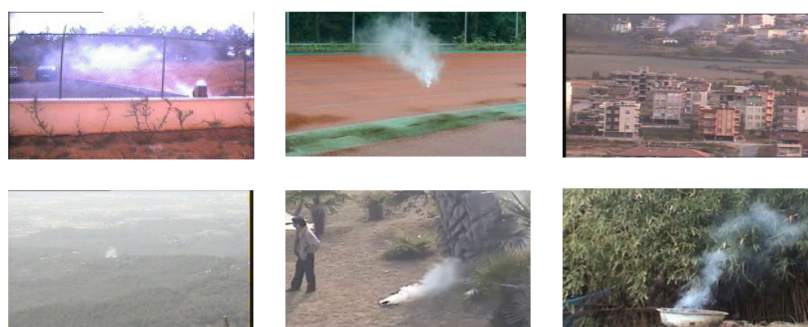


Figure 6. Test samples in Dataset 3. Top row: from Video 1 to Video 3; bottom row: from Video 4 to Video 6.

The *IoU* values and *F1 – score* values are listed in Tables 3 and 4 respectively. It can be seen that the proposed algorithm is superior to the other three methods (FCN, Deeplab V3+, and RANet) in both the average *IoU* value and *F1 – score*. FCN has the worst generalization. When a large number of interference pixels appear in the video, FCN generates more false-positive pixels. For the smoke videos (such as Video 1) with irregular moving edges and violent diffusion, both RANet and Deeplab V3+ algorithms are prone to false detection and missing detection, resulting in low *IoU* and *F1 – score*.

Table 3. *IoU* values of testing examples in Dataset 3.

Algorithm	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Mean
FCN	64.5%	82.6%	81.0%	72.8%	81.2%	69.4%	75.25%
Deeplab V3+	67.8%	87.7%	82.9%	73.1%	87.6%	73.9%	78.83%
RANet	70.4%	86.9%	83.8%	78.3%	86.1%	75.6%	80.18%
Ours	78.7%	87.2%	86.7%	78.7%	87.3%	82.5%	83.52%

Table 4. *F1 – score* of testing examples in Dataset 3.

Algorithm	Video1	Video2	Video3	Video4	Video5	Video6	Mean
FCN	70.2%	85.1%	85.2%	76.7%	85.7%	76.9%	79.97%
Deeplab V3+	72.5%	89.3%	86.3%	75.5%	89.6%	78.3%	81.92%
RANet	73.9%	88.3%	87.6%	80.1%	88.9%	79.4%	83.03%
Ours	79.8%	89.1%	88.5%	81.9%	89.3%	85.9%	85.75%

Due to the sparsity of label maps and the simple deconvolution of the upsampling process without considering the global context information, the pixel segmentation results based on FCN lack spatial consistency. Compared with FCN, the network proposed in this paper uses ResNet to extract feature maps and then uses the ranking attention module to capture the global context information. In addition, in the backbone network of Deeplab V3+, the edge information of the segmented objects is lost due to the multiple downsampling resulting in blurred edges. While the proposed network in this paper combines both the spatial stream and temporal stream and then decodes to restore the target boundary details. In the temporal branch, the network pays more attention to the dynamic characteristics of smoke regions. By predicting the channel attention weights, the channel response of the corresponding attributes of smoke moving parts is improved to help to segment complete smoke regions.

As shown in Figure 7, for Video 2 (the second row) and Video 5 (the fifth row), due to the simple background and clear smoke edge, the four comparison algorithms achieved good smoke segmentation results. For Video 4, FCN and Deeplab V3+ produce more false-positive pixels than the proposed network in this paper. Considering that the smoke area occupies a small proportion of the image, we used the spatio-temporal attention model to pay attention to moving pixels meeting smoke color characteristics. Therefore, the proposed network has a higher *F1 – score* than the other three networks.

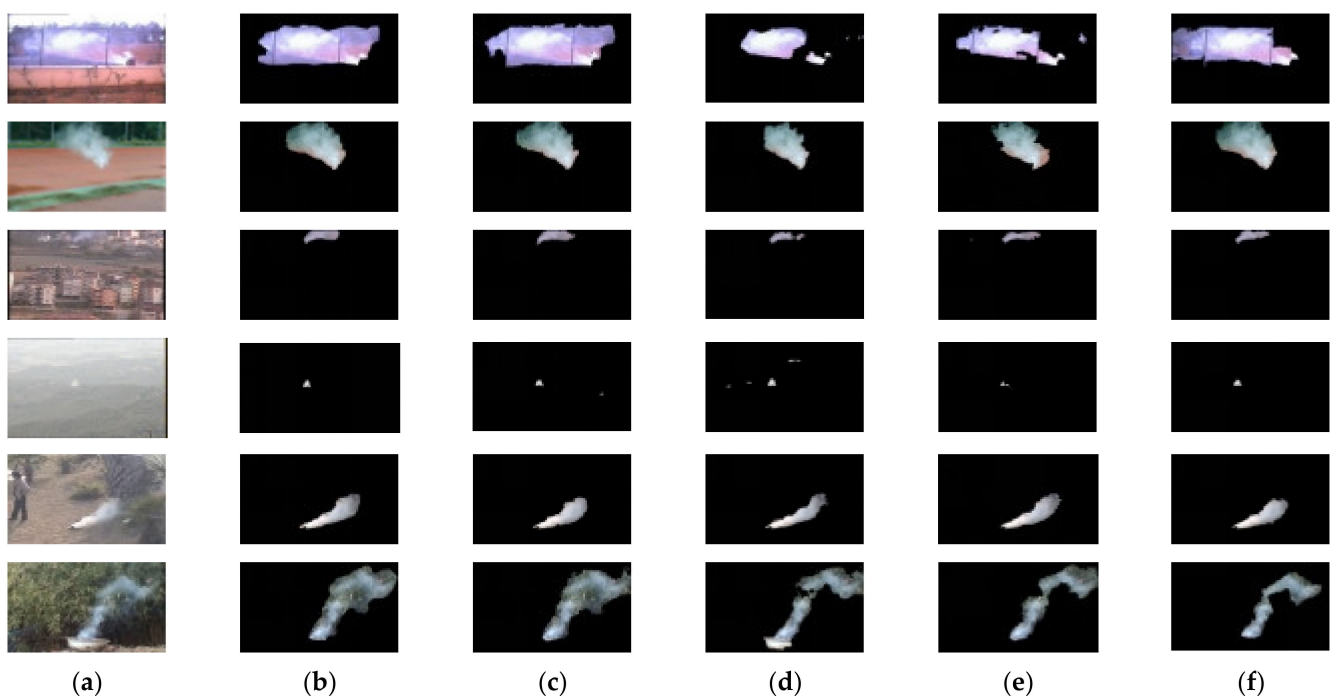


Figure 7. Output of each method from Dataset 3, from top row to bottom row, Video 1 to Video 6, respectively. (a) Input frame; (b) ground truth; (c) ours; (d) FCN; (e) RANet; (f) DeepLab V3+.

For Video 6, the edges of smoke regions are not clear, which results in missing pixels in the segmentation results of FCN and Deeplab V3+ to different degrees. The missing pixels of RANet are relatively low, but its segmented smoke region is not as complete as

that of the proposed network. The proposed network considers feature enhancement and fuses the spatio-temporal information of smoke to supplement the lack of saliency. We multiply the spatial and temporal feature maps to extract the common parts that meet both static characteristics and moving characteristics and then add the common parts to the saliency feature to supplement the motion information on the saliency features.

For Video 1, the proposed network considers that the continuous frames of smoke often have a similar diffusion state. Therefore, the segmentations result of the smoke edge is closer to the ground truth. However, in general, the results of the above four methods are not satisfactory. The main reason is that the smoke in Video 1 is affected by the strong wind, and the smoke area is erratic and spreads almost across the entire screen. The thin and light smoke on the periphery is reflected by the background, which brings great difficulty to the identification of the smoke area. In addition, the calibration of the ground truth of the video is also a dilemma.

3.3. Anti-Interference Test

In this section, we use the camouflaged videos to evaluate the anti-interference performance of our model. Figure 8 shows the results of two challenging videos. As shown in Figure 8, compared with the other three methods, the proposed method reduces the false detection rate for the complex background and has higher detection accuracy for smoke pixels. For the suspected smoke videos, these four methods all mistakenly detect the interfering pixels as smoke pixels to different degrees, but in contrast, the proposed algorithm has a lower false detection rate. This is because the temporal stream takes into account the spatial consistency and motion contrast in the smoke optical flow image. Furthermore, we integrated temporal and spatial features in order to eliminate static targets with camouflage colors.

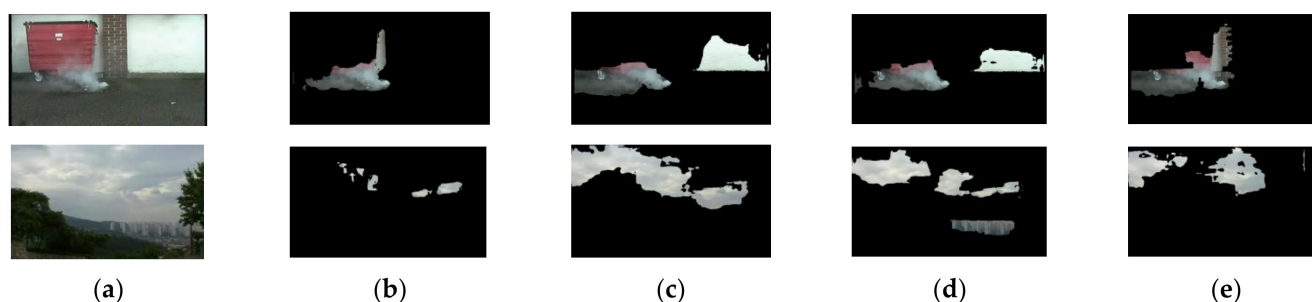


Figure 8. Anti-interference test results: (a) input frame; (b) ours; (c) FCN; (d) RANet; (e) DeepLab V3+.

The ROC (receiver operating characteristic) shows the relationship between the false-positive rate and true-positive rate. In this paper, we calculated the number of overlapping smoke pixels in the detection maps and ground truth images as true positives. Similarly, we also determined the number of non-overlapping flame pixels in the detection maps and take those as false positives. Figure 9 shows the ROC space for the four methods. It can be seen that the proposed method maintained a better balance between the true positive rate and the false positive rate than the other three methods.

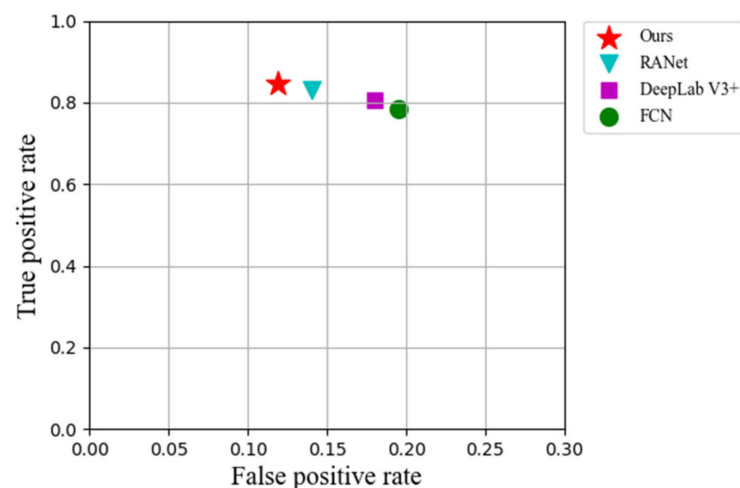


Figure 9. Receiver operating characteristic (ROC) space.

4. Conclusions and Future Works

In this paper, a new semi-supervised two-stream convolution network for smoke region detection is proposed. The semi-supervised ranking model can detect smoke pixels by calculating the similarity of smoke images and extracting the spatial characteristics of smoke. The spatio-temporal attention model is used to fuse the temporal and spatial characteristics and highlight the motion characteristics of smoke, which improves the detection accuracy of smoke region with obvious diffusion and drift characteristics. The experimental results show that the average *IoU* value of the proposed method is 83.52% and the average *F1 – score* is 85.75%, which are 3.34% and 2.72% higher than those of the three compared methods, respectively. At the same time, the proposed method has more robust anti-interference ability to camouflaged objects. For the videos containing a large number of light and thin smoke, the metric values *IoU* and *F1 – score* of the proposed method are slightly improved compared with other state-of-the-art algorithms; the two measurement values are still low. In future research, we will consider using convolutional LSTM (long short-term memory) to mine the drift direction and diffusion characteristics of smoke in order to improve the performance of smoke region detection. In addition, the architecture of the spatial stream branch will be considered to simplify using a weight-shared network. Inspired by the literature [25], we plan to utilize the multi-relation detector [25] to match the correlation between the temple frame and the current frame, which may improve the efficiency of evaluating the matching correlation.

Author Contributions: Conceptualization, Z.D. and Y.Z.; methodology, Z.D.; software, Z.D.; validation, Y.Z., A.L., and Z.Z.; formal analysis, A.L.; investigation, Z.D.; resources, Z.D.; data curation, Z.Z.; writing—original draft preparation, Z.D.; writing—review and editing, Y.Z.; visualization, Z.D.; supervision, Y.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ho, C.-C. Machine vision-based real-time early flame and smoke detection. *Meas. Sci. Technol.* **2009**, *20*, 045502. [\[CrossRef\]](#)
2. Genovese, A.; Labati, R.D.; Piuri, V.; Scotti, F. Wildfire Smoke Detection Using Computational Intelligence Techniques. In Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems & Applications, Ottawa, ON, Canada, 19–21 September 2011; pp. 1–6.
3. Xiong, Z.; Rodrigo Caballero, R.; Wang, H.; Finn, A.M.; Lelic, M.A.; Peng, P.Y. Video-based smoke detection: Possibilities, techniques, and challenges. *J. Hubei Radio Telev. Univ.* **2007**, 112–114. Available online: https://www.academia.edu/30284548/Video_Based_Smoke_Detection_Possibilities_Techniques_and_Challenges (accessed on 30 September 2021).
4. Töreyn, B.U.; Dedeolu, Y.; Çetin, A.E. Wavelet Based Real-Time Smoke Detection in Video. In Proceedings of the European Signal Processing Conference, Antalya, Turkey, 4–8 September 2005; pp. 4–8.
5. Filonenko, A.; Hernandez, D.C.; Jo, K.H. Fast Smoke Detection for Video Surveillance using CUDA. *IEEE Trans. Ind. Inform.* **2017**, *14*, 725–733. [\[CrossRef\]](#)
6. Maruta, H.; Nakamura, A.; Kurokawa, F. A New Approach for Smoke Detection with Texture Analysis and Support Vector Machine. In Proceedings of the IEEE International Symposium on Industrial Electronics, Bari, Italy, 4–7 July 2010; pp. 1550–1555.
7. Yuan, F. Video-based smoke detection with histogram sequence of LBP and LBPV pyramids. *Fire Saf. J.* **2011**, *46*, 132–139. [\[CrossRef\]](#)
8. Yu, C.; Fang, J.; Wang, J.; Zhang, Y. Video Fire Smoke Detection Using Motion and Color Features. *Fire Technol.* **2010**, *46*, 651–663.
9. Yuan, F. A fast accumulative motion orientation model based on integral image for video smoke detection. *Pattern Recognit. Lett.* **2008**, *29*, 925–932. [\[CrossRef\]](#)
10. He, T.; Liu, Y.; Xu, C.; Zhou, X.; Hu, Z.; Fan, J. A Fully Convolutional Neural Network for Wood Defect Location and Identification. *IEEE Access* **2019**, *7*, 123453–123462. [\[CrossRef\]](#)
11. Jin, X.; Che, J.; Chen, Y. Weed Identification Using Deep Learning and Image Processing in Vegetable Plantation. *IEEE Access* **2021**, *9*, 10940–10950. [\[CrossRef\]](#)
12. 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 Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23–26 October 2016; pp. 877–882.
13. Yuan, F.; Zhang, L.; Xia, X.; Wan, B.; Huang, Q.; Li, X. Deep smoke segmentation. *Neurocomputing* **2019**, *357*, 248–260. [\[CrossRef\]](#)
14. Muhammad, K.; Ahmad, J.; Lv, Z.; Bellavista, P.; Yang, P.; Baik, S.W. Efficient Deep CNN-Based Fire Detection and Localization in Video Surveillance Applications. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *49*, 1419–1434. [\[CrossRef\]](#)
15. Khan, S.; Muhammad, K.; Mumtaz, S.; Baik, S.W.; de Albuquerque, V.H.C. Energy-Efficient Deep CNN for Smoke Detection in Foggy IoT Environment. *IEEE Internet Things J.* **2019**, *6*, 9237–9245. [\[CrossRef\]](#)
16. 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. [\[CrossRef\]](#)
17. Lin, G.; Zhang, Y.; Guo, X.; Zhang, Q. Smoke Detection on Video Sequences Using 3D Convolutional Neural Networks. *Fire Technol.* **2019**, *55*, 1827–1847. [\[CrossRef\]](#)
18. Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, Z.; Jia, Y.; Wang, J. Video Smoke Detection Based on Deep Saliency Network. *Fire Saf. J.* **2019**, *105*, 277–285. [\[CrossRef\]](#)
19. Verstockt, S.; Beji, T.; Potter, P.D.; Hoecke, S.V.; Sette, B.; Merci, B.; Walle, R. Video Driven Fire Spread Forecasting Using Multi-modal LWIR and Visual Flame and Smoke Data. *Pattern Recognit. Lett.* **2013**, *34*, 62–69. [\[CrossRef\]](#)
20. Zen, R.I.M.; Widianto, M.R.; Kiswanto, G.; Dharsono, G.; Nugroho, Y.S. Dangerous Smoke Classification Using Mathematical Model of Meaning. *Procedia Eng.* **2013**, *62*, 963–971. [\[CrossRef\]](#)
21. Pan, J.; Ou, X.M.; Xu, L. A Collaborative Region Detection and Grading Framework for Forest Fire Smoke Using Weakly Supervised Fine Segmentation and Lightweight Faster-RCNN. *Forests* **2012**, *12*, 768. [\[CrossRef\]](#)
22. Ba, R.; Chen, C.; Yuan, J.; Song, W.; Lo, S. SmokeNet: Satellite Smoke Scene Detection Using Convolutional Neural Network with Spatial and Channel-Wise Attention. *Remote Sens.* **2019**, *11*, 1702. [\[CrossRef\]](#)
23. Wang, Z.; Xu, J.; Liu, L.; Zhu, F.; Shao, L. RANet Ranking Attention Network for Fast Video Object Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 3978–3987.
24. Barmpoutis, P.; Stathaki, T.; Dimitropoulos, K.; Grammalidis, N. Early Fire Detection Based on Aerial 360-Degree Sensors, Deep Convolution Neural Networks and Exploitation of Fire Dynamic Textures. *Remote Sens.* **2020**, *12*, 3177. [\[CrossRef\]](#)
25. Fan, Q.; Zhuo, W.; Tang, C.-K.; Tai, Y.-W. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 4012–4021.