STCNet: spatiotemporal cross network for industrial smoke detection

Yichao Cao¹,² · Qingfei Tang³ · Xiaobo Lu¹,²

Received: 20 May 2021 / Revised: 5 September 2021 / Accepted: 25 November 2021 / Published online: 14 February 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Industrial smoke emissions present a serious threat to natural ecosystems and human health. Prior works have shown that using computer vision techniques to identify smoke is a low-cost and convenient method. However, translucent smoke detection is a challenging task because of the irregular contours and complex motion state. To overcome these problems, we propose a novel spatiotemporal cross network (STCNet) to recognize industrial smoke emissions. The proposed STCNet involves a spatial pathway to extract appearance features and a temporal pathway to capture smoke motion information. Our STCNet is more targeted and goal oriented for dealing with translucent, nonrigid smoke objects. The spatial path can easily recognize obvious nonsmoking objects such as trees and buildings, and the temporal path can highlight the obscure traces of motion smoke. Our STCNet achieves the mutual guidance of multilevel spatiotemporal information by bidirectional feature fusion on multilevel feature maps. Extensive experiments on public datasets show that our STCNet achieves clear improvements against the best competitors by 6.2%. We also perform in-depth ablation studies on STCNet to explore the impacts of different feature fusion methods for the entire model. The code will be available at https://github.com/Caoyichao/STCNet.

Keywords Smoke detection · Video understanding · Deep learning
1 Introduction

Industrial smoke emissions may cause adverse effects on human health and the ecological environment. Large amounts of air pollutants may cause or contribute to an increase in mortality or serious illness. Smoke detection technology based on computer vision can help regulators obtain visual evidence and enterprises implement self-monitoring.

In industrial smoke detection tasks, plants usually not only emit smoke but also emit much steam. Steam and smoke have very similar appearances, which creates a great challenge to smoke detection in this scene. Different from some smoke datasets (in which steam and smoke are not deliberately distinguished), a large-scale RISE dataset [8] makes a clear distinction between steam and smoke. A more realistic industrial smoke dataset makes practical smoke detection methods possible.

Currently, there are many references about the recognition of specific smoke features. According to the dimension difference of input data, existing methods can be divided into image-based and video-based methods. Image-based smoke detection methods tend to detect smoke areas from single-frame images. Video-based methods usually not only learn spatial features from a single frame but also learn temporal information in the temporal domain from videos.

In some cases, the image-based method is a good choice when stable and reliable image sequences are not available. Tian et al. [24] proposed detecting and separating smoke from a single image frame by convex optimization. Yuan et al. [31] proposed combining local binary pattern (LBP)-like features, kernel principal component analysis (KPCA), and Gaussian process regression (GPR) for smoke detection. Some studies also applied convolutional neural networks (CNNs) to smoke detection and recognition. Yin et al. [29] proposed a deep normalization and convolutional neural network (DNCNN) with 14 layers for smoke recognition, in which batch normalization is used to speed up the training process and boost the smoke recognition accuracy. These methods recognize the specific static texture of smoke in a single frame image in order to distinguish smoke from background.

However, the dynamic characteristics of smoke often play an important role in the recognition process. When the human eye distinguishes smoke in a video, dynamic features are often used as key reference information. If the recognition model can learn context information from sequence data, its recognition accuracy will be improved theoretically. From a motion point of view, a higher-order linear dynamical system (h-LDS) descriptor was proposed as a dynamic texture descriptor for video smoke identification [4]. Some researchers have attempted to use deep learning methods for smoke detection. Lin et al. [13] proposed a joint detection framework for video smoke detection, in which faster RCNN is employed to generate suspected smoke boxes and 3D CNN is used to extract spatiotemporal features of the clip. However, this method is limited by the considerable computational cost. There are also some researchers who improve the object detection performance by utilizing depth information [21] or improve the control method of the visual system [15].

Despite these efforts, video-based smoke recognition is still a challenging task. Early smoke objects are usually small in size, and the variance in smoke color, texture, and interference is large, which makes smoke detection very difficult. While there is extensive literature regarding video understanding, research on the spatiotemporal modeling of smoke features is still necessary. Especially for large-scale smoke datasets, the deep learning based methods are far less studied. We hope to study more general and efficient methods to improve the recognition performance of CNN for translucent smoke objects.

Springer
Given the aforementioned concerns, we propose a novel industrial smoke detection framework, denoted as the spatiotemporal cross network (STCNet). Inspired by the two-stream methods [19, 20], this framework improves smoke spatiotemporal modeling capability by fully fusing multilevel features between spatial and temporal pathways. The main contributions of this paper are as follows:

- We propose a novel video smoke detection architecture, termed STCNet, which is good at spatiotemporal modeling for translucent smoke objects. Our STCNet can effectively highlight subtle motion smoke and suppress disturbing moving targets.
- Our STCNet integrates multilevel spatial and temporal feature representation through a dual pyramid and achieves collaborative promotion of two pathways.
- Experimental results on the challenging RISE dataset demonstrate that our proposed method achieves state-of-the-art results on video smoke detection tasks. We also perform in-depth ablation studies to explore the impacts of multiple fusion strategies at different levels and directions.

The rest of this paper is organized as follows. Section 2 summarizes related works. The proposed architecture for smoke representation is described in Section 3. Detailed performance studies and analysis are conducted in Section 4. Finally, conclusions and discussions are drawn in Section 5. Our model code will be released if the paper is accepted.

2 Related work

In this section, some representative smoke detection methods are reviewed. Although few studies have investigated video smoke detection, there is substantial literature on video understanding. It is a natural idea to use video understanding methods for smoke recognition. Therefore, we use a subsection to introduce and discuss video understanding methods in detail.

2.1 Smoke detection

The success of existing approaches for smoke detection relies on robust smoke feature descriptions. Both smoke feature descriptors and deep learning methods can be divided into two categories according to the dimension of smoke features: image-based methods [14, 16, 22–24, 30, 32] and video-based methods [4, 13]. To motivate the rationale for the proposed methods, some representative works are reviewed from three aspects: image-based, video-based and deep learning-based.

Image-based methods usually focus on the smoke texture, color, shape and edge. Yuan [30] proposed a double mapping framework for smoke detection, in which the first mapping calculates edge orientation histograms, edge magnitude and edge magnitude local binary pattern (LBP) bit and densities, LBP bit, color intensity and saturation. The second mapping computes the statistical characteristics of mean, variance, skewness, kurtosis and Hu moments. Some researchers formulated the smoke detection task as a sparse representation and convex optimization problem [22–24]. Tian et al. [24] proposed separating quasismoke and quasibackground components by dual overcomplete dictionaries, in which the respective sparse coefficients are concatenated for smoke detection. Based on the airlight
albedo ambiguity model, Long et al. [16] proposed detecting smoke and predicting the thickness distribution through transmission.

In recent years, deep learning methods have achieved competitive results on various tasks, such as visual recognition [10], speech recognition [3] and natural language processing [6]. Researchers designed a deep normalization and convolutional neural network (DNCNN) for smoke detection, which is a superior alternative to traditional handcrafted methods [29]. Zhao et al. [33] demonstrated the effectiveness of using saliency detection and deep convolutional neural networks in the localization and recognition of wildfires in unmanned aerial vehicle (UAV) images. Given the smoke candidate patch, the dark channel was reported to have more elaborate smoke information, and the detailed features of dark channel images were used as cues to perform smoke detection [14]. One of the difficulties in smoke recognition is the limitation of the number of smoke samples for training. To ease this limitation, Xu et al. [28] proposed a framework-based fast detector SSD and MSCNN for smoke detection using synthetic smoke image samples.

Although image-based methods have achieved impressive results, these methods have difficulty meeting the requirements of practical applications since they ignore the motion information of smoke. However, the dynamic characteristics of smoke often play an important role in the recognition process. Recently, there have been some video-based smoke detection methods using deep learning [4]. Lin et al. [13] proposed a joint smoke detection framework to locate and recognize smoke from videos, in which a faster RCNN is used to generate suspected smoke region proposals and a 3D CNN is used to extract temporal information. Although this method achieved better smoke detection performance than image-based methods, the large computational cost limited its practical applications.

2.2 Video understanding

Although smoke detection from video is still a challenging task, great breakthroughs have been made in video understanding [5]-[2].

Long-term recurrent convolutional networks (LRCNs) [5] combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were proposed for activity recognition and video description. A 2D CNN was applied to process individual frames and output representations of image features to stack LSTMs. Zhou et al. [34] proposed a temporal relation network (TRN) to learn and reason about temporal dependencies between video frames at multiple time scales. TRNs aim to describe the temporal relations between the spatial features extracted by 2D CNNs. The temporal shift module (TSM) [12] was designed to solve video understanding from another side, which shifts part of the channels along the temporal dimension, both forward and backward, to exchange information among neighboring frames. The above methods were explored to extract spatial features through a 2D CNN and then relied on different temporal model methods in the middle or output layer of the 2D CNN.

Another research direction of video understanding is 3D CNN methods and their (2+1)D CNN variants. Tran et al. [25] proposed using 3D CNN to model appearance and motion information simultaneously for videos. Carreira et al. [2] introduced a two-stream inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation for video and proved that after pretraining on a large video dataset, the performance of I3D models can be considerably improved. In this work, I3D models and their variants are considered baseline methods on video smoke detection datasets.
However, most of these methods focus on relatively obvious motion information, and few studies have studied translucent objects such as smoke. Here, we explore smoke detection via video classification approaches. The goal of this work is to explore whether the spatial and temporal features in the smoke detection model can work together to improve the modeling ability of the model for smoke characteristics. Our key motivation for STC-Net is that a) subtle smoke motion features may be highlighted by a simple and efficient residual frame calculation method, b) a multiscale spatiotemporal dual pyramid may be able to better integrate spatial and temporal information in the middle layers of a two-path network, and c) a coherent two-path network can consider both temporal and spatial characteristics in the video smoke detection process.

3 Spatiotemporal cross networks

In this section, we give a detailed description of the proposed video smoke detection method. The intuition behind the proposed methods is introduced first. Second, we present the architectures of the spatiotemporal cross network. After that, the multilevel spatiotemporal representation structure details are reported, which are very important to the decomposition of spatial and temporal components for smoke videos. Finally, we describe the spatiotemporal feature cross operation.

3.1 Intuition

For video smoke detection tasks, it is natural to apply existing video understanding methods directly to detect smoke from videos. However, experiments suggest that general video understanding frameworks are not good at dealing with light smoke objects. These methods seem to focus on obvious motion information. For the RISE dataset studied in this paper, industrial smoke objects usually have no obvious contours or texture features. Some smoke samples are hard to recognize from a single image, even for human eyes. In addition, due to the lack of effective smoke feature descriptors, it is difficult to generate the smoke optical flow representation. Detailed experimental results are shown in Section 4.3.

Some input RGB frames in the RISE dataset and corresponding residual frames are shown in Fig. 1. As can be seen from the Fig. 1, subtle smoke objects are well highlighted after frame difference calculation. We hope video smoke detection model to be better than

![Fig. 1 Input RGB frames (the top row) in the RISE dataset and corresponding residual frames (the bottom row). For clarity, red and blue arrows are used to mark smoke and steam (nonsmoke), respectively](image)
the plain CNN by giving the model more temporal guidance information. The intuition behind STCNet is that the temporal pathway can highlight smoke motion information, and spatiotemporal pathways can guide each to further improve smoke detection performance.

3.2 Overview of methodology

Figure 2 depicts the framework architecture of the proposed STCNet. The proposed method is a two-path architecture, which consists of a spatial path using a CNN to extract smoke appearance features and a temporal path using the same structure CNN (non-weight sharing) to capture motion features. For the smoke video detection task, an input video is split into $T + 1$ subsections of the same size, and one RGB frame is sampled in each subsection. Then, $T$ residual frames can be generated from $T + 1$ RGB frames. By jointly processing a small number of frames sampled from a whole video, the most relevant information of smoke objects can be captured. Additionally, processing fewer frames can speed up the inference of the model.

Let $Frame_i \in \mathbb{R}^{C \times T \times H \times W}$ and $ResFrame_i \in \mathbb{R}^{C \times T \times H \times W}$ be the $i_{th}$ RGB frame and $i_{th}$ residual frame, respectively, with $C$, $T$, $H$, and $W$ being the number of channels, temporal length, height and width of the image. The channel number $C$ is 3 for both RGB and residual frames. The RGB and residual frames are processed by spatial and temporal networks, respectively.

3.2.1 Spatial path

The backbone design of a neural network is very important in the whole framework design. In recent years, many famous network structures have been designed for image classification, such as ResNet [7] and ResNeXt [27]. Here, we adapt them to design our spatiotemporal cross network for smoke detection following these successful network structures in image classification.

For the spatial path, the dimension of the stacked RGB frames is $C \times T \times H \times W$. However, it should be noted that there is no temporal interaction between the stacked frames in

![Image of network architecture](image-url)
the inference of the spatial branch. The design goal of the space path is to focus on the texture, color, contours and other appearance features of smoke objects. Therefore, each input frame is processed as an independent individual. A multiscale spatial feature pyramid is generated by the spatial CNN backbone. The details of spatiotemporal cross operation are reported in the following subsections.

### 3.2.2 Residual frames

For smoke detection tasks, one of the most challenging problems is that smoke regions are not as intuitive as general rigid objects. Therefore, we obtain residual frames by subtracting adjacent frames to highlight the changing regions between frames. To preserve color and long-term dependence information, stacked residual frames of the RGB channel are set as temporal inputs.

Assuming the $i_{th}$ RGB frame is formulated as $\text{Frame}_i$, the $i_{th}$ residual frame can be defined as:

$$\text{ResFrame}_i = \left\{ \begin{array}{ll} \alpha \ast \mid \text{Frame}_i - \text{Frame}_{i+1} \mid, & \text{if } \beta < \beta \\ \beta, & \text{if } \beta \geq \beta \end{array} \right.$$  

where $\alpha$ is an expanding coefficient, which amplifies the frame differences. Then, the maximum residual frame pixel value is limited with the parameter $\beta$ ($\beta = 255$ in experiments), which is designed to prevent numerical overflow. Some RGB and corresponding residual frames are shown in Fig. 1. The first row shows the normal RGB images, in which the red arrows are used to indicate the smoke regions. Figure 1 shows that the subtle smoke area in RGB frames usually contains obvious light cyan components in residual frames. These visualization results verify the highlight effect of residual frames on smoke regions. Compared with RGB frames, the information in residual images is relatively sparse and mainly focused on moving objects. The motion regions are enhanced by the expanding coefficient $\alpha$ and then limited to 255 by parameter $\beta$. These operations can not only highlight the characteristics of smoke but also suppress the more obvious motion objects such as steam.

For each RGB frame, we only need to calculate the difference with the next frame. The computational cost for residual frames can even be ignored when compared with the convolutional neural network latency or optical flow calculation.

### 3.2.3 Residual frames

In parallel to the spatial path, the temporal pathway is another branch designed to capture smoke motion features. After obtaining the residual frames, the temporal pathway can extract motion features from the frame differences. To reduce the complexity of the network structure design, the architecture of the temporal pathway is the same as that of the spatial pathway. To capture different spatial and temporal information, the weights of the spatial and temporal branches are different. The spatial pathway focuses on the smoke appearance and background information, while the temporal branch focuses on the moving region.

Our spatiotemporal cross-network architecture is generic, and it can be instantiated with different backbones, such as Se-ResNeXt [27] and MobileNetv2 [17]. Generally, the complex backbone can obtain better recognition performance, while the advantage of the lightweight model lies in the computational speed. An STCNet example using SE-ResNeXt-50 [27] as the backbone is specified in Table 1.
3.2.4 Multilevel structure

Our STCNet generates feature maps layer by layer, and the feature maps in different stages have different sizes and depths. The high-resolution feature maps have low-level features that contain detailed spatial information. However, the low-resolution feature maps have high-level semantic features that have a stronger representational capacity for object recognition. The feature pyramid network (FPN) [11] is a common technique in object detection and image classification tasks to improve the feature representation ability of CNNs.

In STCNet, to effectively fuse spatial and temporal features at multiple levels, a dual pyramid fusion strategy is designed (as shown in Fig. 3). Our STCNet takes T RGB and residual frames as input and generates spatial and temporal feature maps at several scales with a scaling step of 2. There are many residual blocks generating feature maps of the same size, and we combine these blocks into the same stage (as shown in Table 1). For our STCNet, the output of the last layer of each residual block in the spatial and temporal paths is chosen to build the dual feature pyramid.

Specifically, each residual stage of SE-ResNext-50 [27] is denoted as {res1, res2, res3, res4}. Their output has a stride of {4, 8, 16, 32} with respect to the input frame. The first three pairs of feature maps are selected to build the dual pyramid. The last $7 \times 7$ size feature map is added as the top layer of the pyramid.

In addition, we also design a dual pyramid variant in which connections only fuse from the temporal to the spatial path. The comparison results are reported in Section 4.4.

3.3 Cross operation

As shown in Fig. 2, a spatiotemporal dual pyramid sample is constructed to generate multilevel feature representations. In this structure, the spatial and temporal feature maps of the same level are summed to fuse information. Specifically, the sum fusion operation of two kinds of feature maps can be defined as:

![Fig. 3 Framework of the proposed spatiotemporal cross network (STCNet) for smoke detection](image)
Table 1  An example instantiation of the STCNet

| Stage | Spatial path | Temporal path | Output sizes |
|-------|-------------|---------------|--------------|
| Data  | –           | –             | Spatial: T × 3 × 224² | Temporal: T × 3 × 224² |
| Conv1 | 1 × 7², 64 stride 2² | 1 × 7², 64 stride 2² | Spatial: T × 3 × 224² | Temporal: T × 3 × 224² |
| Pool1 | 1 × 3², max stride 2² | 1 × 3², max stride 2² | Spatial: T × 64 × 56² | Temporal: T × 64 × 56² |
| Res1  | 1 × 1², 128 \times 3 | 1 × 1², 128 \times 3 | Spatial: T × 256 × 56² | Temporal: T × 256 × 56² |
| Res2  | 1 × 1², 256 \times 4 | 1 × 1², 256 \times 4 | Spatial: T × 512 × 28² | Temporal: T × 512 × 28² |
| Res3  | 1 × 1², 128 \times 6 | 1 × 1², 128 \times 6 | Spatial: T × 1024 × 14² | Temporal: T × 1024 × 14² |
| Res4  | 1 × 1², 1024 \times 3 | 1 × 1², 1024 \times 3 | Spatial: T × 2048 × 7² | Temporal: T × 2048 × 7² |
| Fuse  | conv, 1 × 1², 256 | 1 × 1², 256 adaptive average pool | T × 256 × 7² |
| Cls   | conv, 1 × 1², 256 adaptive average pool | T × 256 × 1² |
| Out   | fully connected layer | # classes |

\[
SF_{i,j,c}^{\text{sum}} = TF_{i,j,c}^{\text{sum}} = SF_{i,j,c} + TF_{i,j,c}
\]

where \(1 \leq i \leq H, 1 \leq j \leq W, 1 \leq c \leq C\) and \(SF, TF \in \mathbb{R}^{C \times H \times W}\). \(SF\) and \(TF\) are the spatial and temporal feature maps, respectively. The sum fusion is elementwise addition between two kinds of feature maps. The feature maps to be fused are from the same location of neural networks with the same architecture. Although the added features in the two pathways are the same, the focuses of the two branches are different. In Section 4.5, we show the activated regions on 7 \times 7 size feature maps of two branches by Grad-CAM [18], which helps us to analyze whether two branches work as expected at the inference stage.

To make the research more convincing, we discuss different fusion methods between the two pathways. Conv fusion is another common feature fusion approach. We also explore changing the addition fusion into a convolution operation:

\[
SF_{i,j,c}^{\text{conv}} = \text{Conv}_s\left(\text{Concat}(SF_{i,j,c}, TF_{i,j,c})\right)
\]

\[
TF_{i,j,c}^{\text{conv}} = \text{Conv}_t\left(\text{Concat}(SF_{i,j,c}, TF_{i,j,c})\right)
\]

where \(\text{Conv}_s\) and \(\text{Conv}_t\) denotes convolutional fusion operation in spatial and temporal pathway, respectively. We use two convolution layers to integrate the two kind of feature maps respectively, and hope to obtain better recognition performance. However, two problems are exposed when this convolution fusion method is adopted: one is the slow convergence speed, and the other is the reduction in accuracy. The cross fusion model using the convolution operation is named STCNet-D, and the comparison results are shown in Section 4.4.
4 Experiments

In this section, we evaluate the proposed methods on the RISE industrial smoke dataset. Experiments were conducted on a GPU server with an Intel Core i7-9700 CPU and an NVIDIA RTX2080TI GPU using the PyTorch framework.

4.1 RISE dataset

The RISE video smoke dataset [8] is the first large-scale video dataset for recognizing industrial smoke emissions. The RISE dataset contains 12,567 clips with 19 distinct views from cameras on three sites that monitored three different industrial facilities. The clips are from 30 days that span four seasons in two years in the daytime. RISE is a challenging video classification dataset, as it covers various characteristics of smoke emissions, including opacity and color, under diverse weather (e.g., haze, fog, snow, cloud) and lighting conditions. Moreover, RISE involves distractions of various types of steam, which can be similar to smoke and challenging to distinguish.

4.2 Implementation details

Stochastic gradient descent (SGD) with a minibatch size of 3 is used to optimize the model weights. The weight decay is 0.0005, and the momentum is 0.9. We start from a learning rate of 0.001 and a pretrained SE-ResNeXt-50 [27] on ImageNet because of its balance between accuracy and efficiency. In addition, the proposed framework based on a lightweight backbone was also tested, such as MobileNetv2 [17].

4.3 Comparisons on RISE

We compare the proposed method with primary methods on video understanding and video smoke recognition [1, 2, 5, 9, 12, 26]. Abbreviations for these methods and operations are listed in Table 2. Abbreviations ND and FP indicate no data augmentation and frame perturbation, respectively. Abbreviations TSM, LSTM, NL and TC indicate the temporal shift module [12], long short-term memory layers [5], nonlocal module [26], and timeception layers [9], respectively.

| Abbreviation | Operation & Method                                      |
|--------------|---------------------------------------------------------|
| I3D          | Inflated 3D ConvNet [2]                                  |
| ND           | No Data Augmentation                                     |
| FP           | Frame Perturbation                                       |
| TSM          | Temporal Shift module [12]                               |
| LSTM         | Long Short-term Memory layers [5]                        |
| NL           | Non-local module [26]                                    |
| TC           | Timeception layers [9]                                   |
| EFFNet       | Enhanced Feature Foreground Network [1]                  |
| STC          | Spatiotemporal Cross operation                           |
The results in Table 3 show that the plain CNN model based on SE-ResNeXt-50 [27] achieves a better performance than the I3D-based methods, which shows that a stronger backbone is helpful for smoke recognition. When adopting SE-ResNeXt-50 as the backbone, our STCNet gains a 4.8% higher F-score than the previous best model EFFNet [1], which confirms the remarkable ability of STCNet for video smoke detection. Data augmentation technology can further improve the performance of STCNet. In the training stage, the data augmentation methods (horizontal flipping, random resizing and cropping, perspective transformation, area erasing, and color jittering) are the same as in [8] and are used by default.

Moreover, the parameters, FLOPs, latency and throughput characterization of each method are also reported in Table 4. Generally, our proposed STCNet with MobileNetv2 achieves 0.868 F-score on RISE, which is 4.5% better than the RGB-I3D-TC model, while being 3.3x smaller and 3.3x faster on inference than the RGB-I3D-TC model. STCNet with MobileNetv2 backbone can achieve 109.7 videos per second processing.
speed, as shown in Table 4. Using SE-ResNeXt-50 as the backbone of STCNet leads to another 1.7% improvement on the F-scores by using MobileNetv2 as backbone.

### 4.4 Comparison with other variants

At the beginning of architecture design, we also try some test schemes to gradually explore the optimal detection architecture. Here, four STCNet variants, which are denoted as STCNet-A (Fig. 4), STCNet-B (Fig. 5), STCNet-C (Fig. 6) and STCNet-D, are designed for comparison. Among them, STCNet-A is a typical two-stream network that sums the output feature maps of two pathways to predict probability. The difference is that there is no optical flow input, but the residual frames are calculated as described in Section 3.2.2. STCNet-B is equipped with a unidirectional feature fusion method, which only fuses temporal to spatial features. The aim of designing STCNet-B is to verify the efficiency of the fusion operation from the spatial to temporal path. In addition, STCNet-C performs spatiotemporal feature fusion only after the first residual block without our scale feature fusion operation. Finally, the designed STCNet-D architecture is similar to STCNet; the only difference is that the cross operation uses convolution instead of addition.

Four variant models are trained and tested on the RISE dataset. The same hyperparameters are adopted for the four models. The results are reported in Table 5.

**Table 5** F-scores for comparing different variants on RISE dataset

| Model   | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | Average |
|---------|-------|-------|-------|-------|-------|-------|---------|
| STCNet-A | 0.87  | 0.88  | 0.89  | 0.89  | 0.84  | 0.86  | 0.872   |
| STCNet-B | 0.88  | 0.89  | 0.90  | 0.90  | 0.85  | 0.87  | 0.882   |
| STCNet-C | 0.88  | 0.89  | 0.90  | 0.90  | 0.85  | 0.86  | 0.877   |
| STCNet-D | 0.87  | 0.87  | 0.89  | 0.89  | 0.85  | 0.86  | 0.870   |
| STCNet   | 0.88  | 0.89  | 0.90  | 0.90  | 0.86  | 0.88  | 0.885   |

Fig. 4 STCNet-A only fuses spatiotemporal information in deep features, which is similar to a two-stream network [19]
multiscale fusion architecture, STCNet increases by 1.3% compared with STCNet-A. After adding the multiscale fusion operation from the temporal path to the spatial path in STCNet-A, the F-score of STCNet-B is improved to 0.882. However, compared with the STCNet of bidirectional fusion, it is 0.3% behind. Finally, the F-score of STCNet-C is reduced by 0.7% using only shallow feature fusion. These ablation experiments prove the effectiveness of the multiscale spatiotemporal feature fusion operation in STCNet.

4.5 Visualization of two paths

To diagnose the attention of two pathways in STCNet, we apply gradient-weighted class activation mapping (Grad-CAM) [18] to visualize the active regions in input frames. The
output of the last residual blocks of two branches is visually analyzed, as shown in Fig. 7. For most of the input frames, the proposed method can focus on the smoke region in two paths while ignoring the interference of steam (nonsmoke). For example, in the second column, even if it is difficult for human eyes to find the smoke, the temporal path can focus well on the smoke movement region. It is difficult to detect this subtle smoke information only by spatial feature extraction. These visualization results show that the two paths in STCNet can achieve mutual guidance in the inference process.

Finally, we show some false positive and false negative cases in Figs. 8 and 9, respectively. We analyze the deficiencies of the proposed method through these cases. In Fig. 8, red arrows are used to mark smoke regions that cannot be detected. The smoke in the first and third rows is almost completely obscured by steam. The images in the second and third rows have a small amount of smoke for a short duration, these reasons may be why the model failed to detect smoke. Some false positive cases are shown in Fig. 9. We use red dotted arrows to mark the regions that cause false positives. In the first two rows, light steam leads to false detection results. The third cause of false
alarms is the large moving fog, which may mislead the model to make incorrect predictions. In fact, small areas and short-duration steam or smoke object detection are still challenging tasks.

5 Conclusion

In this work, a novel spatiotemporal cross network (STCNet) was proposed and verified for video smoke detection. In STCNet, a spatial pathway is designed to extract appearance features, and a temporal pathway is designed to capture smoke motion information. Our STCNet can integrate multilevel spatial and temporal feature maps to generate a more efficient smoke representation. Extensive experimental results on a challenging smoke dataset also demonstrated the remarkable ability of STCNet in video smoke detection. We hope this efficient spatiotemporal cross strategy can be a standard component for video smoke detection.

However, our method can only process fixed-length video sequences (e.g., 8 frames). And downsampling for variable-length sequences cannot solve the fundamental problem. For further research, we will study how to achieve reliable smoke recognition performance from variable-length video sequences.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.61871123), Key Research and Development Program in Jiangsu Province (No.BE2016739) and is a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions. We thank the Big Data Center of Southeast University for providing facility support for the numerical calculations in this paper.

Declarations

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

1. Cao Y, Tang Q, Wu X, Lu X (2021) EFFNet: enhanced feature foreground network for video smoke source prediction and detection. IEEE Trans Circ Syst Video Technol https://doi.org/10.1109/TCSVT.2021.3083112
2. Carreira J, Zisserman A (2017) Quo Vadis, action recognition? A new model and the kinetics dataset, presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
3. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28. Curran Associates, Inc., pp 577–585
4. Dimitropoulos K, Barmpoutis P, Grammalidis N (2017) Higher order linear dynamical systems for smoke detection in video surveillance applications. IEEE Transactions on Circuits and Systems for Video Technology 27(5):1143–1154. https://doi.org/10.1109/TCSVT.2016.2527340
5. Donahue J, et al (2015) Long-term recurrent convolutional networks for visual recognition and description, presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
6. Goldberg Y (2017) Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies 10(1):1–309
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition, presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
8. Hsu Y-C, et al (2020) RISE video dataset: Recognizing industrial smoke emissions. arXiv:2005.06111. Accessed 09 July 2020
9. Hussein N, Gavves E, Smeulders AWM (2019) Timeception for complex action recognition, presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 254–263
10. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Curran Associates, Inc., pp 1097–1105
11. Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection, presented at the Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2117–2125
12. Lin J, Gan C, Han S (2019) TSM: Temporal shift module for efficient video understanding. In: 2019 IEEE/CVF international conference on computer vision (ICCV). Seoul, Korea (South), pp 7082–7092 https://doi.org/10.1109/ICCV.2019.00718
13. Lin G, Zhang Y, Xu G, Zhang Q (2019) Smoke detection on video sequences using 3D convolutional neural networks. Fire Technol https://doi.org/10.1007/s10694-019-00832-w
14. Liu Y, Qin W, Liu K, Zhang F, Xiao Z (2019) A dual convolution network using dark channel prior for image smoke classification. IEEE Access 7:60697–60706. https://doi.org/10.1109/ACCESS.2019.2915599
15. Liu P, Yu H, Cang S (2019) Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances. Nonlinear Dyn 98(2):1447–1464. https://doi.org/10.1007/s11071-019-05170-8
16. Long C et al (2010) Transmission: A new feature for computer vision based smoke detection. In: Wang FL, Deng H, Gao Y, Lei J (eds) Artificial intelligence and computational intelligence, vol 6319. Springer, Berlin Heidelberg, pp 389–396
17. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2019) MobileNetV2: inverted residuals and linear bottlenecks. arXiv:1801.04381. Accessed 03 Sep 2020
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV), pp 618–626 https://doi.org/10.1109/ICCV.2017.74
19. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, pp 568–576
20. Sobral A, Vacavant A (2014) A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Computer Vision and Image Understanding 122:4–21. https://doi.org/10.1016/j.cviu.2013.12.005
21. Sun L, Zhao C, Yan Z, Liu P, Duckett T, Stolkin R (2019) A novel weakly-supervised approach for RGB-D-based nuclear waste object detection. IEEE Sensors Journal 19(9):3487–3500. https://doi.org/10.1109/JSEN.2018.2888815
22. Tian H, Li W, Wang L, Ogunbona P (2014) Smoke detection in video: An image separation approach. International Journal of Computer Vision 106(2):192–209
23. Tian H, Li W, Ogunbona P, Wang L (2015) Single image smoke detection. Computer vision - ACCV 2014:87–101
24. Tian H, Li W, Ogunbona PO, Wang L (2018) Detection and separation of smoke from single image frames. IEEE Transactions on Image Processing 27(3):1164–1177. https://doi.org/10.1109/TIP.2017.2771499
25. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE international conference on computer vision (ICCV), pp 4489–4497 https://doi.org/10.1109/ICCV.2015.510
26. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City, UT, USA, pp. 7794–7803 https://doi.org/10.1109/CVPR.2018.00813
27. Xie S, Girshick R, Dollar P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, pp 5987–5995 https://doi.org/10.1109/CVPR.2017.634
28. Xu G, Zhang Q, Liu D, Lin G, Wang J, Zhang Y (2019) Adversarial adaptation from synthesis to reality in fast detector for smoke detection. IEEE Access 7:29471–29483
29. Yin Z, Wan B, Yuan F, Xia X, Shi J (2017) A deep normalization and convolutional neural network for image smoke detection. IEEE Access 5:18429–18438
30. Yuan FN (2012) A double mapping framework for extraction of shape-invariant features based on multi-scale partitions with AdaBoost for video smoke detection. Pattern Recognition 45(12):4326–4336
31. Yuan F, Xia X, Shi J, Li H, Li G (2017) Non-linear dimensionality reduction and gaussian process based classification method for smoke detection. IEEE Access 5(99):6833–6841
32. Yuan F, Zhang L, Xia X, Huang Q, Li X (2019) A wave-shaped deep neural network for smoke density estimation. IEEE Trans Image Process 1 https://doi.org/10.1109/TIP.2019.2946126
33. Zhao Y, Ma J, Li X, Zhang J (2018) Saliency detection and deep learning-based wildfire identification in UAV imagery. Sensors (Basel) 18(3)
34. Zhou B, Andonian A, Oliva A, Torralba A (2018) Temporal relational reasoning in videos. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision-ECCV 2018, vol 11205. Springer International Publishing, Cham, pp 831–846
35. Zolfaghari M, Singh K, Brox T (2018) ECO: Efficient convolutional network for online video understanding. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision-ECCV 2018, vol 11206. Springer International Publishing, Cham, pp 713–730

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.