IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 19, 2019, accepted January 16, 2020, date of publication January 30, 2020, date of current version February 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970497

Spatiotemporal Representation Learning for Video Anomaly Detection

ZHAOYAN LI, YAOSHUN LI, AND ZHISHENG GAO
School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

Corresponding author: Zhisheng Gao (gzs_xihua@mail.xhu.edu.cn)

This work was supported in part by the Ministry of education Chunhui project under Grant Z2016149, in part by the Key scientific research fund of Xihua University under Grant Z17134, in part by the Xihua University Key Laboratory Development Program under Grant szjj2017-065, in part by the Fund of Sichuan Educational Committee under Grant 17ZA0360, and in part by the Sichuan science and technology program under Grant 2019YFG0108.

ABSTRACT Video-based anomalous human behavior detection is widely studied in many fields such as security, medical care, education, and energy. However, there are still some open problems in anomalous behavior detection, such as the large and complicated model is difficult to train, the accuracy of anomalous behavior detection is not high enough and the speed is not fast enough. A spatiotemporal representation learning model is proposed in this paper. Firstly, the spatial-temporal features of the video are extracted by the constructed multi-scale 3D convolutional neural network. Then the scene background is modeled by the high-dimensional mixed Gaussian model and used for anomaly detection. Finally, the accurate position of anomalous behavior in the video data is achieved by calculating the position of the last output feature, that is, the position of the receptive field. The proposed model does not require specific training. Moreover, the proposed method has the advantages of high versatility, fast calculation speed and high detection accuracy. We validated the proposed algorithm on two representative surveillance scene datasets, the Subway and the UCSDSped2. Results show that proposed algorithm has achieved the detection rate of 18 FPS under the condition of common computing resources, and meet the real-time requirements. Moreover, compared the similar methods, the proposed method has achieved the competitive results in both frame-level accuracy and pixel-level accuracy.

INDEX TERMS Spatiotemporal representation learning, anomaly detection, 3D convolutional neural network, mixed Gaussian model.

I. INTRODUCTION

The video-based anomalous human behavior detection aims to intelligently discriminate whether there is anomalous behavior in a video data captured by video sensors through a computer algorithm and if so, give a location marker for the anomalous behavior. Video-based anomalous behavior monitoring is extremely valuable in both public safety and defense security. In the past few decades, video-based anomalous behavior detection has been widely studied in both academia and industry, and it is still a hot and challenging research issue. Difficulties faced by video-based anomalous behavior monitoring are mainly as follows: (1) There is only a small part of data in the background, and the absolute dominance is non-anomalous data, that is asymmetry of data; (2) The definition of anomalous behavior is subjective or context-dependent, lacking uniform standards; (3) The diversity of monitored scenes has led to the diversification of noise. Therefore, the key to video-based anomalous behavior monitoring is whether the proposed algorithm can overcome noise interference in various video monitoring scenarios and robustly output various anomalous behaviors.

In video anomaly detection, behaviors that differ from the majority of behaviors in the scene are treated as anomalies, reflected as unusual object shapes, poses, and motions. The current popular anomaly detection method firstly takes the no-anomalous frames or the no-anomalous blocks in the training data as the no-anomalous modes and then extracts the features of these frames or blocks. Finally, in the test phase, the behavior of the object is determined by discriminating the difference between the test frame (or the the blocks) and the no-anomalous modes. In this way, anomaly

The associate editor coordinating the review of this manuscript and approving it for publication was Peng Liu.

detection is a special kind of identification, so the focus of research on these popular methods is the extraction of features and the discrimination of anomalies. There are many ways to describe regional features, such as histograms of oriented gradient (HOG) and histograms of optical flow (HOF). Motion trajectory-based approaches use such low-level features to construct a spatiotemporal model. However, they have extremely low detection accuracy in the presence of occlusion and will fail in highly complex environments. In recent years, deep learning has been extremely successful in the field of recognition, and it has also been used for anomaly detection. So in this paper, we divide the anomaly detection into two categories: traditional video and image processing based methods and deep learning based methods.

Traditional image processing mainly extracts features by artificially constructing feature operators, and performs anomaly discrimination. Representative methods include the construction of social force model methods [1], analysis of video behavior methods [2], Markov random field method [3], construction of mixed dynamic texture model methods [4], object motion trajectory modelings [5]–[16], motion trajectory and shape detection [17], [18], dense optical flow and space-time gradient [3], [19], structural acquaintance [20], sparse semi-negative matrix decomposition to learn local features [21], Dense trajectory algorithm [22], attitude and motion information recovery [23], point model [24], silhouette and contour [25]–[27], etc. These traditional methods have achieved good results in specific scenarios. The main disadvantage is that the model expression ability is weak, and the detection accuracy is significantly reduced in complex scenes. Moreover, some of them have high computational complexity and it is difficult to meet the real-time requirements of video detection.

Ddeep learning-based methods use the powerful representation learning ability to automatically extract the features of the video data and complete the anomalous behavior discrimination. Literature [28] used the deep neural network to identify whether a window image is anomalous or not. Literature [29] constructed a very complex cascaded neural network to extract features from cubic patches in video data and discriminate the anomalies. Literature [30] proposed a method by extracting features using 2D convolutional neural network and modeling anomalous behavior detection using a Gaussian model. Literature [31] used an unsupervised deep learning framework for anomaly detection and literature [32] used convolutional auto-encoders to perform anomaly detection.

The detection methods of anomalous behavior using the deep neural network directly have the problems such as low computational efficiency and difficulty in constructing training samples. Moreover, a method of feature extraction using 2D CNN has the problem of weak ability to express temporal behavior features. In order to overcome these shortcomings, this paper proposes a deep neural network model (STF-Net) for temporal and spatial representation learning, extracts video data features through the constructed network, and then uses hybrid Gaussian modeling to identify and locate anomalous behavior regions. In the proposed model, we introduce a multi-scale learning structure, which enriches receptive field of the feature in the original image, and makes the feature representation ability stronger. The main contributions of this paper are as follows:

(1) A deep learning model for anomalous behavior representation is proposed. This model extracts the spatiotemporal features of video data through a 3D convolutional neural network and enhances the feature receptive field through a multi-scale model. The extracted features have better adaptability to multiple viewing distances.

(2) The first five layers in the proposed model are directly transfer from C3D, and only the remaining parameters of the model need to be trained, which reduces the difficulty of model training. At the same time, the model parameters come from different behavioral data sets (which is similar to multi-task regularization), and it can improve the generalization ability of the model.

(3) The proposed model is trained using behavior recognition data, and it is not necessary to train the network model using the video of tested surveillance scene, so the proposed method is a completely unsupervised method.

(4) The proposed model and the corresponding training methods have the advantages of strong behavioral representation ability and high detection accuracy for anomalous behavior detection. At the same time, the proposed algorithm has high speed and can meet the requirements of real-time detection under the general computing resources.

The rest of this paper is organized as follows. Section II reviews related work. Section III introduces a novel method of spatiotemporal representation learning for video anomaly detection. Section IV describes the experiments conducted to verify the effectiveness of the proposed method. Finally, Section V concludes this work.

II. RELATED WORK

Accurately identifying the anomalous behavior in various behaviors with different scenes is a very challenging task. Anomalous behavior has varied and diverse characteristics. For example, the definitions of anomalous behavior in the financial field [33] and video surveillance [34] are not exactly the same. At the same time, in complex scenarios, anomalous behavior is often submerged in various no-anomalous behaviors. The in-depth research and extensive application of deep learning provide new ideas for this field. In recent years, anomaly detection using deep neural networks has been a hot-spot and new direction in the industry, and there already has some exploratory work in this area. The work of this paper is also based on deep learning, and we will introduce and analyze some typical closely related research work.

In 2015, a C3D network proposed by Du Tran proved that 3DCNNs are significantly superior to CNNs and traditional manual feature extraction methods in video feature extraction [35]. However, the method of directly using deep neural network to identify is currently faced with the problem of lack of training data, and the unsupervised method needs

to improve the accuracy of anomalous behavior detection in various scenarios. This requires a designed model that can adapt to different monitoring scenarios and does not require a large amount of anomalous behavior data for training.

In 2016, a method of using neural networks alternately between two-dimensional convolution and three-dimensional convolution for anomalous behavior prediction was proposed [28]. This method first extracts the regions of interest by using the optical flow in the video, then directly interpolates and scales the regions of interest to the same size, and then brings these blocks with the same size into the neural network for identification. This method requires the use of anomalous samples to train the network, so half of the standard data sets are used directly to train the network and the others are used for testing. However, the extreme lack of training samples has affected the effectiveness of the network to some extent. In addition, the method uses the optical flow to extract regions of interest which brings a large computational cost. A similar direct recognition algorithm is proposed in [29]. The algorithm adopts a cascaded network model to extract features for each cubic patch in the video, and then designs a set of Gaussian model to identify anomalous. Simple cubic pathes use a simple neural network and the remained cubic pathes (which are difficult to distinguish) use a more complex and deep neural network to extract features.

In 2017, based on the principle that the neural network obtained by training with no-anomalous data is difficult to reconstruct the anomalous data frame accurately, a 2D auto-encoder network was designed to extract the spatial features of the images, and the long short term memory (LSTM) was used to obtain the temporal evolution of spatial features [32]. This algorithm trains the network using the data that does not contain anomalies, and then inputs the image frames to be detected into the trained network for reconstruction. If the reconstruction accuracy is high, it is considered that there is no anomalies, and vice versa, there are data frames with anomalous behaviors. However, this method requires careful setting of thresholds to classify the reconstructed errors. Moreover, it only determines whether the entire image frame is anomalous and cannot locate the position of anomalous behavior in the frame.

In 2018, Sabokrou M [30] proposed a method that first compresses the mean of adjacent image frames into the three channels to obtain a sample data, and then uses the full convolutional networks (FCN) to obtain spatiotemporal features, and finally uses the convolutional auto-encoders to obtain more stable features. The Gaussian model is trained by the videos monitored in a no-anomalous scenario, and then the anomalous discrimination and positioning are completed by the distance metric. Sabokrou M also proved that the feature extraction method based on FCN can meet the requirements of anomaly detection such as accuracy and speed. The advantage of this type of method is that it is faster. However, this simple method of compressing a video into three-channel data would affect the accuracy of extracting time-dimensional features.

TABLE 1. Classification network model structure and configuration.

Layer name	Structure
Pool4	Pooling-2*2*2-0-2*2*2
Conv10	Conv3D-1*3*3-512-1-1*1*1
Conv11	Conv3D-1*3*3-2048-1-1*1*1
Conv12	Conv3D-1*1*1-1024-1-1*1*1
Dropout(0.5)	/
Conv13	Conv3D-1*1*1-101-0-1*1*1

Inspired by these efforts, this paper proposes an anomaly detection algorithm based on the three-dimensional fully convolutional networks (3DFCN). This method can solve the problems in the work of the literature [30], such as the lack of time dimension features, and the lack of spatial features to adapt to the scale.

III. OUR WORK

For the difficult problem of anomaly detection in video surveillance, this paper proposes a deep neural network model STF-Net for the representation of anomalous behavior based on the existing research results. The STF-Net network consists of a basic three-dimensional convolutional (3DC) and a three-dimensional pooling (3D Pooling) and activation units (as shown in Figure 1). It is noted that the first 4 layers of STF-Net are identical to the network structure C3D [35], which is used for the extraction of the primary features of the video data. Further, in order to improve the diversity of high-level features and enhance the receptive field corresponding to high-level features to better express the anomalous behavior of small and medium-sized targets, a multi-scale structure is added to the STF-Net. And the advanced spatiotemporal features of the final video data are output through the Conv9 layer for subsequent detection of anomalous behavior. It is noted that the spatiotemporal representation learning of STF-Net for anomalous behavior cannot be directly trained, so in practice, we add a 3DC layer and a classification layer after the network. The parameters in the added layers are shown in Table 1. When training, for the previous part of the network we use the C3D [35] network to get some of the features, the network parameters are obtained by training in the data set sport1M [36]. For the rest of the networks, we train them through the data set UCF101 [37]. After obtaining the high discriminative spatiotemporal features of the video data through STF-Net, a mixed Gaussian model is trained through no-anomalous video data. At the time of detection, whether the input video data is anomalous is determined by calculating the Mahalanobis distance between the extracted feature and the mixed Gaussian model. In this way, the algorithm does not need to use a special anomalous behavior dataset for training. Moreover, the trained network model can also be used in many different monitoring scenarios.

A. SPATIOTEMPORAL REPRESENTATION LEARNING

1) MODEL FOR SPATIOTEMPORAL REPRESENTATION LEARNING

In order to analyze the spatiotemporal information of the video, I_t is used to represent the video data at the current t

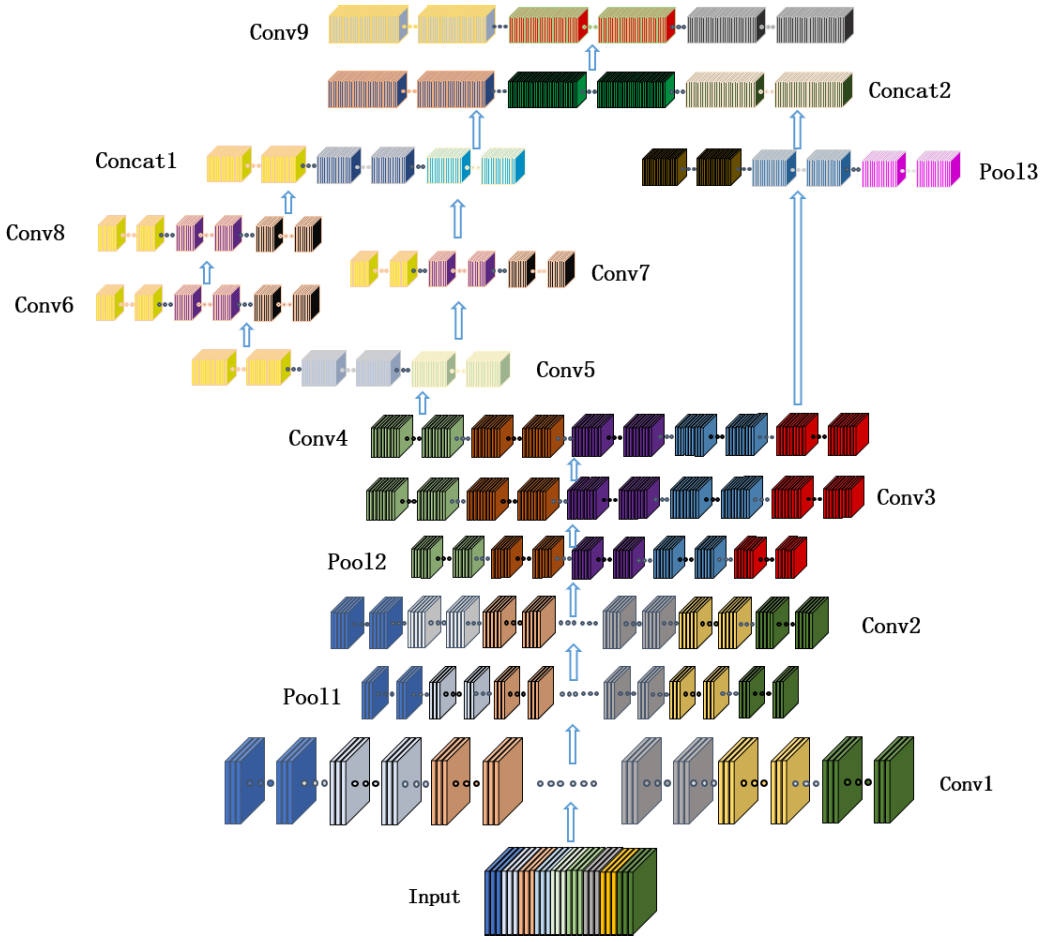


FIGURE 1. The Structure of STF-Net.

time, I_{t+1} is the video data at time $t+1$, and S_t is the video sequence at time t-8 to time t. The $D = \langle S_t, S_{t+1}, ..., S_{n-1}, S_n \rangle$ set contains all video sequence sets, where n represents the number of video sequences in the data set. In our proposed spatiotemporal feature model STF-Net, the data of the input layer is a video segment S_i. The learned features of the l^{th} layer is represented by $f_l = [f_l^1, f_l^2, \ldots, f_l^i, ..., f_l^d]$, which is composed by d dimension data vectors. And $f_l^i = [a_1^1, ..., a_1^k]$, where a_1^1 represents the first value of the l^{th} layer output feature of the neural network, $k = L_{feats} * H_{feats} * W_{feats}$ is the length of the current dimension feature vector, H_{feats} is the height of the spatiotemporal feature, W_{feats} is the width of the spatiotemporal feature, and L_{feats} is the dimension of the time. in the STF-Net, the value of l is set to 9, the value of d is 512, which is the number of feature maps of the l^{th} layer, the value of L_{feats} is set to 1, the value of H_{feats} and W_{feats} are both 14 when the size of the input frame is 112×112. The specific parameter settings of the STF-Net network are shown in Table 2, where the item "Struct" represents the information of configuration and structure of each layer. The convolutional layer structure consists of conv3D-kernel-num-padding-strides, and the pooling layer structure consists of pooling-kernel-padding-strides. It is well known that too short video sequences can contain insufficient temporal features, while too long video sequences are more susceptible to noise interference in complex scenes. Therefore, in practice, the number of video clip frames we selected is 9, so that the output of the Conv9 layer is extracted as a representation of anomalous behavior after multiple three-dimensional convolutions. There are a total of 512 feature maps, which contain both the spatial texture features of pixel locations and the time variation information at that location.

2) MODEL TRAINING

As mentioned earlier, the STF-Net network proposed in this paper uses a general motion behavior recognition database for training. At the same time, after adding the network structure part (shown in Table 1) to the STF-Net, the whole network constitutes a deep neural network model that can identify 101 kinds of motion behaviors. The sample data required by the input layer of the STF-Net network model is a tensor with the size $N * L * H * W * C$, where N is the number of batch samples, L is the time dimension of the sample, H and W are the height and width of the sample, respectively, and C is the number of channels in the sample. In the training phase,

TABLE 2. STF-Net model parameter configuration.

Layer name	Struct
Input	Input(None*9*112*112*3)
Conv1	Conv3D-3*3*3-64-1-1*1*1
Pool1	Pooling-1*2*2-0-1*2*2
Conv2	Conv3D-3*3*3-128-1-1*1*1
Pool2	Pooling-2*2*2-0-2*2*2
Conv3	Conv3D-3*3*3-256-1-1*1*1
Conv4	Conv3D-3*3*3-256-1-1*1*1
Pool3	Pooling-2*2*2-0-2*2*2
Conv5	Conv3D-3*3*3-256-1-2*2*2
Conv6	Conv3D-3*3*3-128-1-1*1*1
Conv7	Conv3D-3*3*3-128-1-1*1*1
Conv8	Conv3D-3*3*3-128-1-1*1*1
Concat1	Concatenate(256)
Concat2	Concatenate(512)
Conv9	Conv3D-1*1*1-512-0-1*1*1

the first 4 layers of STF-Net are initialized using the weights obtained by C3D [35], while the no-anomalous distribution $Y \backsim N(\mu, \sigma^2)$ is used to initialize the other layer network weights, where $\mu = 0$, $\sigma = 0.5$. The loss function used in the network model for classification is as follows:

$$Loss = \sum_{i=1}^{N} y^i log y'^{(i)} + (1 - y^i) log(1 - y'^{(i)}), \tag{1}$$

where $logy'^{(i)}$ is the predicted value of the classification task, y^i is the true value of the input sample, N is the total number of samples, and $Loss$ is the cross entropy of the overall sample.

The strategy of STF-Net training is to update the model parameter θ by the using mini-batch gradient descent (MBGD) on the training set D_{train}. And the condition of the stop of the STF-Net training is that the convergence is achieved in both the training set and the verification set and the model no longer converges after multiple trainings, as shown in Algorithm 1. D_{train} is the training set, S^i is the sample of the i^{th} video clip in the D_{train} (the sample is composed of consecutive 9 frames that do not overlap each other), N_t is the number of training samples, D_{valid} is the validation dataset, S^j is the j^{th} validation sample in the D_{valid} (which is also composed of 9 frames of non-overlapping images), and N_v is the number of validation samples. Function forward (S^j) represents a forward calculation in the training phase, $loss_t$ represents the loss value for each training, and derivation($loss$) represents the gradient of the calculated model parameters, update(θ,η) means to update the model parameter θ with the learning rate η, $step$ is the minimum number of trainings to start the verification model, min_thr represents the threshold of the loss caused by the early stop of the model, $epoch$ is the current number of training, $epochs$ is the predetermined total number of training the optimization model needed, $Count$ is the number of consecutive times the model parameters are not updated, and $batch\_size$ is the sample batch size. In practice, we use the following parameters: $batch\_size = 2$, $\eta = 0.001$, $epoch = 100{,}000$, $step = 150$, $min\_thr = 0.5$, and $Count = 100$.

Algorithm 1 Training Based on Multi-Scale 3D Convolutional Neural Network Model for Classification

Require:
training data $D_{train} = \{S^i, i = 0, 1, 2, \ldots, N_t\}$, test data $D_{valid} = \{S^j, j = 0, 1, 2, \ldots, N_v\}$, behavior recognition threshold φ, early stop parameters $Count$, $minn\_thr$, η, $batch\_size$.

Ensure:
The parameter of the classification model θ

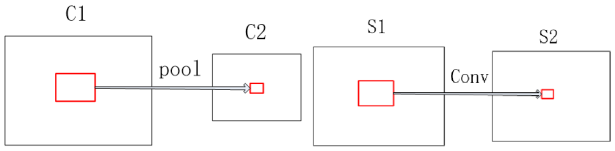
```
Data preprocessing of training and validation sets respectively
Index ← 0;
for epoch ← 0 to epochs do
  for i ← 0 to N_t / batch_size do
    loss ← 0;
    for b ← 0 to batch_size do
      loss_t ← forward(s^{i*batch_size+b});
      loss ← loss + loss_t;
    end for
    derivation (loss / batch_size)
    update(θ, η);
  end for
  if epoch == step then
    for j ← 0 to N_v do
      loss_v ← forward (S^j)
      if loss_v < min_thr then
        Index ← Index + 1
      else
        Index ← 0
      end if
    end for
  end if
  if index ≥ Count then
    stop training
  end if
end for
```

B. REVERSE CALCULATION OF ANOMALOUS POINT POSITION

In a convolutional neural network, the output of each layer of network features can be calculated by forward calculation. In STF-Net, the output features of each layer can be obtained by the following formula:

$$Out = \frac{W - F + 2P}{S} + 1, \tag{2}$$

where Out is the output of the convolutional layer, W is the width or height of the input data, F is the width or height of the convolution kernel, P is the padding performed during convolution, and S is the step size. As shown in Figure 2(a), assume that the input data of the $C1$ layer is a tensor with size $1\times 3 \times 100 \times 100$, the size of pooling kernel is 2×2, S is 2, and P is 0, then the output size is $1\times 3 \times 50 \times 50$. And as shown in Figure 2(b), assume that the input data of the $S1$



(a) receptive field of a pooling feature (b) receptive field of a convolutional feature

FIGURE 2. Calculation of the receptive field of a feature.

layer is a tensor with size $1 \times 3 \times 100 \times 100$, the number of convolutional kernels is 10, and the size of the convolution kernel is 3×3, S is 1, and P is 0, then the size of the $S2$ layer is $1 \times 10 \times 98 \times 98$.

The position and size of located anomalous feature area B obtained by discriminating the outputting map needs to be marked and measured on the input image frame. The whole step is actually a reverse process of forward convolution and pooling, including the calculation of the input receptive field and the calculation of the positional offset.

- Calculation of receptive field
  The receptive field from the hidden layer to the input layer is calculated by reversing from the hidden layer to the input layer. For example, to calculate the receptive field of the Conv2 layer relative to the input layer, the absolute receptive field of Conv2 layer is first calculated, then the receptive field relative to the Pool1 layer is calculated, and then the receptive field relative to the Conv1 layer is calculated, finally the receptive field relative to the input is calculated. The process of inverse calculation to obtain the receptive field is shown in Algorithm 2. Among them, *RFs* is the receptive field of each layer to the input layer, *reversed_layer* is the set of reverse subscripts of the network, *Stride* is the step size of a layer, K is the size of the convolution kernel or pooled size and the function ''getStrideAndKernelFromNet'' is to extract the step size and convolution kernel parameters of a layer.
- Offset calculation of the receptive field position
  After obtaining the receptive field size, the specific area information of the anomalous behavior can be calculated by the coordinates of the hidden layer, the size of the receptive field and the offset of the hidden layer with respect to the input layer. The offset of the hidden layer relative to the input layer is calculated as shown in Algorithm 3, where *layers* is the set of network forward subscripts, *Strides* is the offset of all layers relative to the input layer, and the function getStrideFromNet is used to extract the information of step size from a layer of the convolutional neural network, *Stride* is the step size of one layer.

C. CONSTRUCTING A NO-ANOMALOUS BEHAVIOR MODEL

The spatiotemporal features of the video clips can be extracted by the STF-Net model as f_l^d. In order to use the spatiotemporal feature to detect the anomalous behavior of video, this paper uses the no-anomalous behavior of video data to construct a mixed Gaussian model, which is used as the background scene modeling of anomalous behavior. The specific process is shown in Algorithm 4, where N is the number of no-anomalous video clips, S^n is the n^{th} video clip sample, $feats[n]$ is the spatiotemporal features obtained from the n^{th} sample, $f_l^{d\times N}(i,j)$ is the set of features at the (i,j) position for all samples of the l^{th} layer, d is the spatiotemporal feature dimension, and $G(\theta)$ is a mixed Gaussian model for no-anomalous behavior, $\Leftarrow$ adds features to the end of the array in turn, the function "predict(·)" represents the extraction of spatiotemporal features, and the function $F_G(\cdot)$ represents Gaussian modeling.

Algorithm 2 Calculation of Receptive Field in Convolutional Neural Networks

Require:
Convolutional neural network structure
Ensure:
Receptive field of each layer in the network relative to the input layer

```
RF = 1;
RFs = [ ];
for layer in reversed_layer do
    Stride, K ← getStrideAndKernelFromNet[layer];
    RF ← ((RF − 1) ∗ Stride) + K;
    RFs.append(RF);
end for
```

Algorithm 3 Calculation of the Offset of Each Layer Relative to the Input Layer in a Convolutional Neural Network

Require:
Convolutional neural network structure
Ensure:
Offset of each layer relative to the input layer in a convolutional neural network

```
tmpStride ← 1;
Strides ← [ ];
for each layer in layers do
    Stride ← getStrideFromNet[layer];
    tmpStride ← tmpStride ∗ Stride;
    Stride.append(tmpStride);
end for
```

D. ANOMALOUS BEHAVIOR DISCRIMINATION

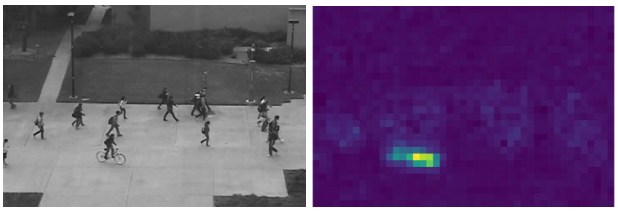
The trained algorithm model can be used to perform anomaly detection. First, the spatiotemporal features of the test sample S^n are calculated through the STF-Net and then the Mahalanobis distance $d(G(\theta), S^i)$ between the spatiotemporal features and the mean of the mixed Gaussian model is calculated. Finally, whether the sample has an anomaly is determined by whether $d(G(\theta), S^i)$ is greater than a given threshold φ.

Algorithm 4 Construction of No-Anomalous Model Based on Spatiotemporal Feature Learning

Require:
Training data D_{train}=$\{S^n, n = 0, 1, 2, \ldots, N\}$, model parameters of STF-Net

Ensure:
G(θ)

Data preprocessing on the training set
for n= 0 to N **do**
  $feats[n] \leftarrow$ predict(S^n)
end for
for $i = 0$ to W_{feats} **do**
  for $j \leftarrow 0$ to H_{feats} **do**
    Initialize vector $f_l^{d*N}(i,j)$ by using 0;
    for $n \leftarrow 0$ to N **do**
      $f_l^{d*n}(i,j) \leftarrow feats[n](i,j)$;
      $f_l^{d*N}(i,j) \Leftarrow f_l^{d*n}(i,j)$;
    end for
  end for
end for
$G(\theta_{i,j}) \leftarrow F_G(f_l^{d*N})$;



(a) A original frame (b) The corresponding Mahalanobis distance matrix

FIGURE 3. **Original frame and the corresponding Mahalanobis distance matrix.**

The discrimination of anomaly detection is shown in formula (3).

$$f(i,j) = \begin{cases} No-anomaly & if\ d(G(\theta), s^i) < \varphi \\ Anomaly & if\ d(G(\theta), s^i) \geq \varphi \end{cases} \quad (3)$$

where $f(i,j)$ indicates whether the (i,j) position is anomalous. The anomaly detection process is shown in Algorithm 5, where $Feats(i,j)^d$ is the eigenvector of the test data at the (i,j) position, $\sum^{-1}$ is the inverse of the covariance matrix, μ is the mean of the training dataset, $MahasDis(i,j)$ is the Mahalanobis distance of the (i,j) position, and φ is the threshold for determining the anomaly. The Mahalanobis distance matrix containing anomalous information is shown in Figure 3. It can be clearly seen that the values of the small portion are significantly larger than the values of other regions, and the threshold segmentation can be used to extract them and its positional information in the input data can be obtained by reverse calculation, so that the anomalous position in the sample can be determined.

Algorithm 5 Anomaly Detection Based on Spatiotemporal Feature Learning

Require:
$\sum^{-1}$, μ, test data $D_{test} = \{S^n, n = 0, 1, 2, \ldots, N\}$, threshold φ for anomaly detection.

Ensure:
The determination of whether the given sample is anomalous. If it is anomalous, the anomalous position will be output.

Preprocessing test data
for $n \leftarrow 0$ to N **do**
  $feats \leftarrow$ predict(S^n);
  for $i \leftarrow 0$ to W_{feats} **do**
    for $j \leftarrow 0$ to H_{feats} **do**
      $MahasDis(i,j) \leftarrow (feats(i,j)^d - \mu)^\top * \sum^{-1} * (feats(i,j)^d - \mu)$;
      if $MahasDis(i,j) > \varphi$ **then**
        Reverse calculation of the area represented by the (i,j) position in the sample S^n;
      end if
    end for
  end for
end for

IV. EXPERIMENTS AND ANALYSES

A. EVALUATION INDEXES

In this paper, two objective evaluation indexes are introduced to evaluate the accuracy of anomalous behavior detection. They are the frame-level standard and the pixel-level standard. Both of them are based on the true positive rate (*TPR*) and the false-positive rate (*FPR*). Anomalous events are positive and no-anomalous events are negative. The data containing the anomaly is positive, otherwise it is negative. The true and false definitions under the two standards are as follows:

(1) Frame level: the algorithm predicts the frames containing the anomalous event and compares them to the frame-level annotations of anomaly in the video clips to determine the number of true positive and false positive frames.

(2) Pixel level: The algorithm first predicts the pixels associated with the anomalous event and then compares it to the pixel-level annotations of anomaly to determine the number of true positive and false positive frames. If at least 40% of the anomalous pixels are identified, the frame is anomalous, otherwise, it is no-anomalous.

The calculations of *TPR* and *FPR* are given by equations (4) and (5), respectively, and the receiver operating characteristic curve (ROC) is plotted with *FPR* as the horizontal axis and *TPR* as the vertical axis. Next, a straight line (given by Equation 6) is used to intersect the curve, and the resulting intersection is the equal error rate (EER), which is the evaluation index used in this paper.

$$TPR = \frac{tp}{tp + fn}, \quad (4)$$

$$FPR = \frac{fp}{fp + tn}, \tag{5}$$

$$Y = -x, \tag{6}$$

where tp is the number of frames (or pixels) that are predicted to be positive and marked as positive, fn is the number of frames (or pixels) that are predicted to be negative and marked as positive, $tp + fn$ is the number of all positive samples, fp is the number of frames (or pixels) that are predicted to be positive and marked negative, tn is the number of frames(or pixels) that are predicted to be negative and marked negative, and $fp + tn$ is the number of all negative samples.

B. DATASETS

Mahadevan *et al.* [4] created no-anomalous behaviors based on the mixture of dynamic textures (MDT), and the outliers were anomalous behaviors. Two data sets (UCSDped1 and UCSDped2) for anomaly detection and the baseline detection rate of the corresponding data set are given in the reference [4]. Adam *et al.* [19] divided the scene into multiple monitoring areas to monitor anomalous events separately. Two data sets (the Subway and the Mall) for anomaly detection and the baseline detection rate of the corresponding data set are given in the reference [19]. The data sets published in the above two papers are the standard data sets with the highest frequency of reference in the current research work for anomalous behavior detection.

The Subway [19] data set contains two videos: the subway entrance video and the subway exit video. The subway entrance video contains 144,249 frames for 1 hour and 36 minutes. The exit video contains 64,900 frames for 43 minutes. Most behaviors in these two videos are no-anomalous, and the anomalous behavior is determined by the direction in which the pedestrian moves. For example, a pedestrian passes through the subway exit into the subway or passes a security check without a card.

The UCSDSped [4] dataset is obtained by mounting on a highly fixed camera overlooking the sidewalk. In this data set, the population density in the aisle varies from sparse to very crowded. Common anomalies come from cyclists, skaters, strollers, people walking across the sidewalk or around the grass, and some people in wheelchairs. All anomalies occur naturally, and the data is divided into two subsets, each of which corresponds to a different scene. The UCSDSped1 contains 34 training video samples and 36 test video samples, the UCSDSped2 contains 16 training video samples and 12 test video samples. There is a binary label on each frame in each segment indicating whether there is an anomaly in the frame. In addition, there are 10 segments in the UCSDSped1 and 12 segments in the UCSDSped2 that provide a manually generated pixel-level binary mask that indicates the anomalous regions. This helps to evaluate whether the algorithm can be accurately located.

The UCF101 [37] data set has 13320 video clips for a total of 24 hours. There are 101 behavioral categories, which are divided into five broad categories: interactions between people and objects, body movements, interactions between people and people, playing instruments and sports. The data image has a resolution of 320 × 240 and a sample frame rate of 25 FPS. In this paper, we use this data set to train the network model.

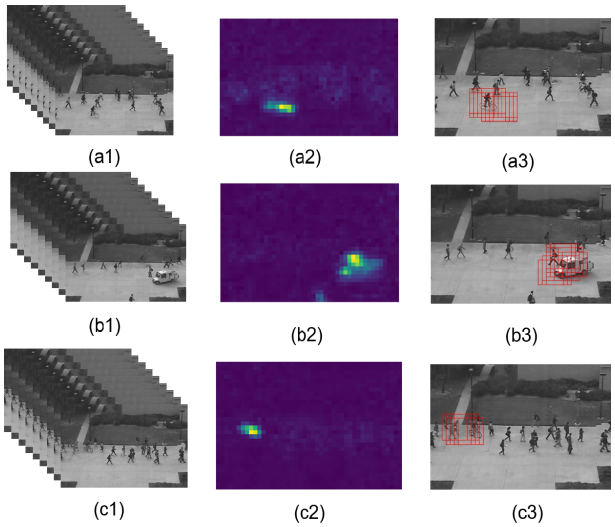


FIGURE 4. Part of the experimental results on the UCSDSped2.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) EXPERIMENTAL RESULTS ON THE UCSDSped2

Figure 4 is a partial test result of the proposed algorithm. The figure shows the test results of three samples, where (a1), (b1), (c1) are test samples, and (a2), (b2), (c2) are heat maps of the Mahalanobis distance matrix, (a3), (b3), (c3) are results of anomalous positioning. It can be seen from the results that even in the case of high population density, the algorithm can detect anomalous behavior better. Moreover, whether the anomaly is from a bicycle or a car can be reliably determined.

2) EXPERIMENTAL RESULTS ON SUBWAY

Some of the subway data test results are given in Figures 5 and 6. The scene view of the Subway is small and the anomalous behavior is mainly reverse walking, so the detection is difficult. It can be seen that our algorithm has obtained relatively reliable detection results both in the exit of the subway and at the entrance of the subway.

D. COMPARATIVE ANALYSIS OF EXPERIMENTAL RESULTS

The ROC curves for the UCSDSped2 test data set are shown in Figures 7 and 8. Fig. 7 is an ROC curve of accuracy of anomalous frame detection. It can be seen that the EER value of the proposed algorithm is the smallest, indicating that the method of this paper works best on this data set. The curve of Sabokrou 2015 in the figure is the result of method proposed in [20]. The curve of Sabokrou 2017 is the result of method proposed in [29]. The curve of MDT is the result of the method for anomalous behavior detection proposed in [4].

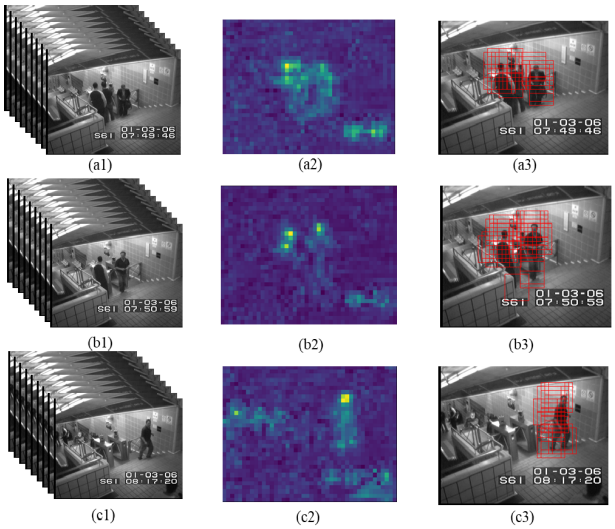


FIGURE 5. Part of the experimental results at the entrance to the subway.

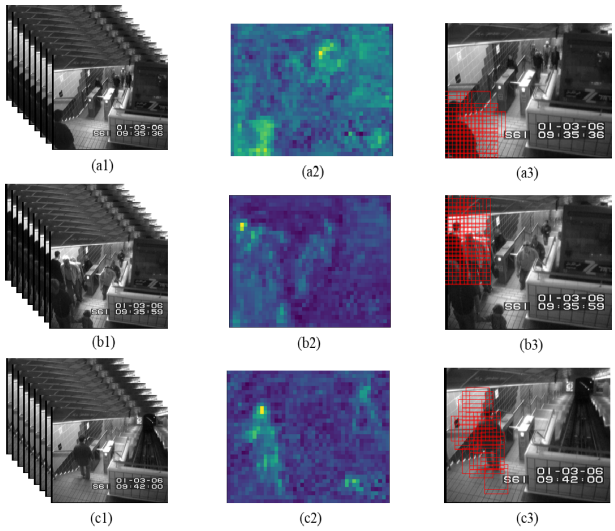


FIGURE 6. Part of the experimental results at the exit of the subway.

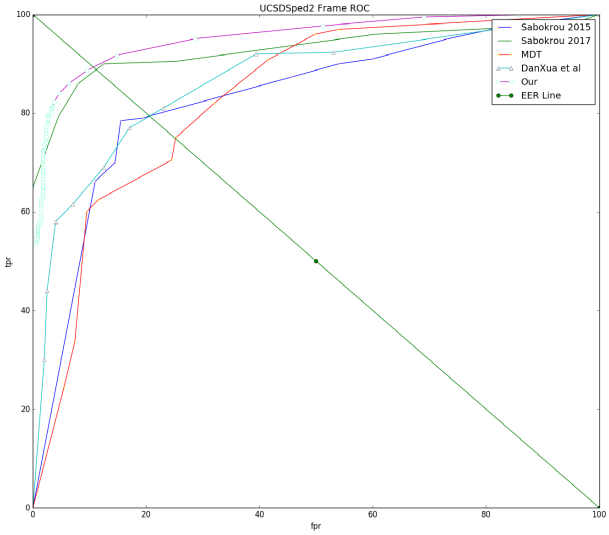


FIGURE 7. Frame-level based ROC curve on the UCSDSped2.

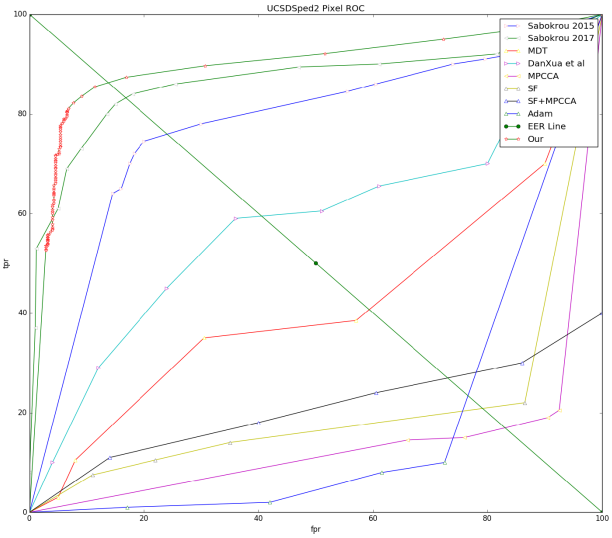


FIGURE 8. Pixel-level based ROC curve on the UCSDSped2.

The curve of Dan Xua et al is the result coming from the reference [29] directly, which is considered reasonable because it is a standard published data. Among them, Sabokrou 2017 is closest to the algorithm proposed in this paper, and its performance is even slightly higher than the algorithm when the FPR is close to zero. However, with the increase of the FPR, it gradually loses its competitiveness, indicating that the recognition-based algorithm is insufficient in versatility and robustness. In addition, it can be clearly seen from the figure that the ROC curve of this paper covers the largest area in two-dimensional coordinates. In general, the ROC curve of the proposed algorithm is not only optimal in the EER index, but also optimal in the AUC index. Fig. 9 is an ROC curve of pixel-level detection accuracy of anomalous behavior. It can be seen that at the pixel-level, the proposed algorithm is significantly better than the currently popular algorithms in the EER index. In the figure, MPCC is the result of the algorithm proposed in [3], SF is the result of the algorithm proposed in [1], SF+MPCCA is the result of the algorithm proposed in [4], and Adam is the result of the algorithm proposed in [19]. It can be seen that an pixel-level based detection method requires locating anomalies in the frame accurately which is more difficult than the method of frame-level based detection, so that most algorithms have lower detection accuracy on pixel-level based detection. Combining the EER index and the AUC index at the frame-level and pixel-level, the anomaly detection algorithm proposed in this paper extracts spatiotemporal features of video clips through the multi-scale 3D deep neural network, and has the best results and reaches the current leading level. Especially at the pixel-level, the algorithm proposed in this paper is obviously superior to the other algorithms.

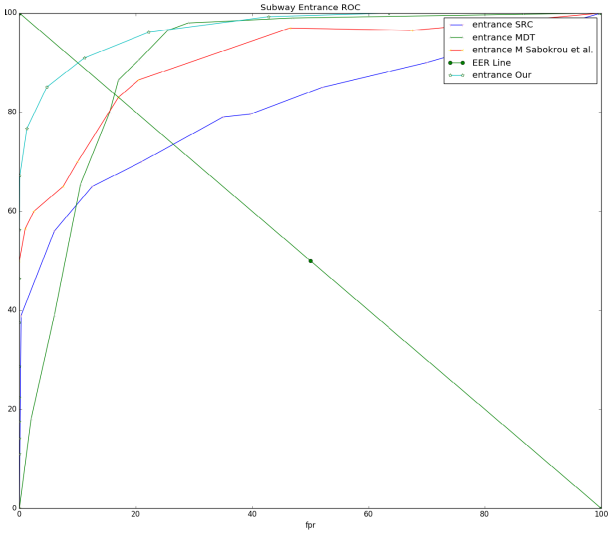


FIGURE 9. The ROC curve for anomaly detection at the entrance of the Subway.

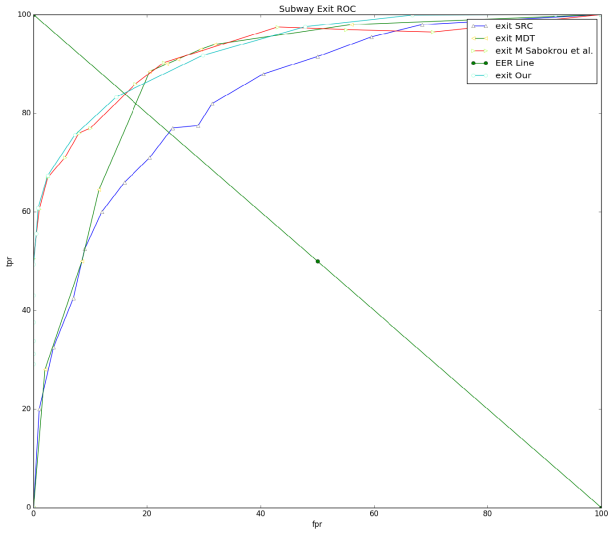


FIGURE 10. The ROC curve of anomaly detection at the exit of Subway.

We also conducted a comparative test on the Subway dataset to verify the effectiveness and performance of the anomaly detection algorithm proposed in this paper. Figures 9 and 10 are the ROC curves for the anomaly detection of the entrance and exit of the subway, respectively. Among them, entrance SRC and exit SRC are the results of the algorithm proposed in [19]. The entrance MDT and exit MDT are the results of the algorithm proposed in [4]. Entrance M Sabokrou et al and exit M Sabokrou et al are the results of the algorithm proposed in [29]. The MDT algorithm has poor performance when FPR is less than 20, and its detection accuracy is close to the algorithm when FPR is higher than 25. It can be seen from Fig. 10 and Fig. 11 that the EER index of this paper is obviously better than the other methods and the AUC index is obviously superior to other comparison methods. Compared with the currently popular

TABLE 3. The EER results of frame-level on the Subway.

Method	Exit (EER%)	Entrance (EER%)
SRC [38]	26.4	24.4
MDT [4]	16.4	16.7
Saligrama and Chen [39]	17.9	-
FCN [30]	16	17
Chong Y S et al. [32]	**9.5**	23.7
Ours	15.9	**10.1**

TABLE 4. The EER results of frame-level and pixel-level on the UCSDSped2 respectively.

Method	Frame-level (EER%)	Pixel-level (EER%))
IBC [17]	13	26
Adam [19]	42	76
SF [1]	42	80
MPCCA [3]	30	71
MPCCA+SF [4]	36	72
Zaharescu [40]	17	30
MDT [4]	24	54
Sabokrou [30]	11	15
Reddy [41]	20	-
Bertini [42]	30	-
Saligrama [39]	18	-
Xu [31]	20	42
Li [43]	18.5	29.9
Xiao [21]	10	17
Sabokrou [20]	19	24
Sabokrou [29]	**8.2**	19
Zhou [28]	21.5	19.5
Chong [32]	12	-
Ours	10.5	**13.8**

methods, the proposed method achieves a leading level in the anomaly detection of subway entrance and exit, and the effect is optimal on the anomaly detection of the entrance.

Through experiments on the UCSDSped2 and the Subway, we can see that the proposed algorithm has very good results of the detections in different scenarios, especially the pixel-level based anomaly detection. Experiments show that the proposed algorithm has better performance on the generalization ability and evaluation indexes. The two evaluation indexes of different algorithms on different datatsets (or different locations in the same dataset) are given in Tables 3 and 4. As can be seen from the EER values of frame-level and pixel-level indexes on different algorithms in Tables 3 and 4, the proposed algorithm in this paper has obtained better effects in different scenarios. In particular, the EER of pixel-level index of method on the UCSDSped2 is 1.2% higher than that of the currently second-ranked method, and the EER index on the entry in the Subway of our method is 6.6% higher than that of the second-ranked method. Methods based on direct recognition would obtain better values on frame-level EER by special training [21], [29]. The method proposed in reference [21] comprehensively uses algorithms with high complexity such as sparse non-negative matrix factorization, histogram feature construction, and probability model. The experimental results show that it takes 0.29 seconds to process an average frame on the UCSDSped2 dataset, which fails to meet the requirements

of real-time detection. The method in [29] adopts a cascade method to improve the detection speed, but it needs to divide the image into multiple cubic patches, and each cubic patch needs to be identified individually. This method uses local features of a region for identification, thus losing global and contextual information. Therefore, this method has higher accuracy on frame-level detection, while has lower accuracy on pixel-level detection.

V. CONCLUSION

Anomalous behavior detection in complex scenarios remains a challenging issue, such as crowd-intensive scenarios. At present, the existing algorithms still have insufficient robustness and are difficulty to train. The detection algorithm based on spatiotemporal feature extraction proposed in this paper learns the anomalous behavior features by a multi-scale 3D deep neural network. The background is modeled by a mixed Gaussian model, and the anomalous behavior is discriminated by the Mahalanobis distance of the feature. The algorithm proposed in this paper has the advantages such as simple training, needing no special anomalous behavior database, having fast speed and high accuracy of detection. The performance of the algorithm is verified on a variety of representative data sets. The experiments show that the proposed algorithm has achieved the currently best results, especially in the field of pixel-level accuracy. This shows that the proposed algorithm has stronger ability to express anomalous behaviors and has better versatility and robustness. The work worthy of further research in this paper is as follows: (1)The combination of 2D convolution and 1D convolution can be used instead of 3D convolution, which would reduce the numbers of learning parameters of the network and make the learned spatiotemporal features more descriptive and discriminative by deepening the network. (2) The STF-Net model proposed in this paper has adopted multi-scale to enhance the receptive field of features. The combination of more popular multi-scale algorithms (such as FPN [28], DCN [44], SSD [45]) can be designed to increase the multi-scale learning ability of spatiotemporal extraction models.

REFERENCES

[1] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.

[2] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 893–908, May 2008.

[3] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2921–2928.

[4] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.

[5] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[6] N. T. Siebel and S. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2002, pp. 373–387.

[7] T. Zhang, H. Lu, and S. Z. Li, "Learning semantic scene models by object classification and trajectory clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1940–1947.

[8] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Comput. Vis. Image Understand.*, vol. 115, no. 3, pp. 323–333, Mar. 2011.

[9] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2054–2060.

[10] C. Piciarelli and G. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognit. Lett.*, vol. 27, no. 15, pp. 1835–1842, Nov. 2006.

[11] C. Piciarelli, C. Micheloni, and G. Foresti, "Trajectory-based anomalous event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1544–1554, Nov. 2008.

[12] P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis, "Detecting abnormal human behaviour using multiple cameras," *Signal Process.*, vol. 89, no. 9, pp. 1723–1738, Sep. 2009.

[13] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby, "Detecting anomalies in people's trajectories using spectral graph analysis," *Comput. Vis. Image Understand.*, vol. 115, no. 8, pp. 1099–1111, 2011.

[14] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2287–2301, Nov. 2011.

[15] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, Sep. 2006.

[16] F. Tung, J. S. Zelek, and D. A. Clausi, "Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance," *Image Vis. Comput.*, vol. 29, no. 4, pp. 230–240, Mar. 2011.

[17] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, Apr. 2007.

[18] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1446–1453.

[19] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[20] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 56–62.

[21] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1477–1481, Sep. 2015.

[22] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.

[23] D. A. Forsyth, O. Arikan, L. Ikemoto, J. F. O'Brien, and D. Ramanan, "Computational sudies of human motion: Part1, tracking and motion systhesis," *Found. Trends Comput. Graph. Vis.*, vol. 1, nos. 2–3, pp. 77–254, 2005.

[24] G. Johansson, "Visual motion perception," *Sci. Amer.*, vol. 232, no. 6, pp. 76–89, 1975.

[25] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on r transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[26] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proc. 4th ACM Int. Workshop Video Surveill. Sensor Netw.*, 2006, pp. 171–178.

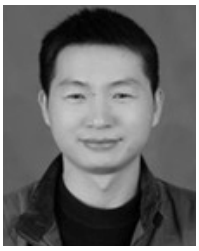
[27] L. Wang and D. Suter, "Informative shape representations for human action recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 2, Aug. 2006, pp. 1266–1269.

[28] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Process., Image Commun.*, vol. 47, pp. 358–368, 2016.

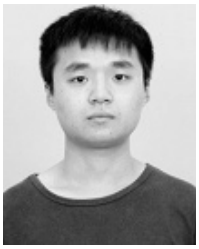
[29] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.

[30] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88–97, Jul. 2018.

[31] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," 2015, *arXiv:1510.01553*. [Online]. Available: https://arxiv.org/abs/1510.01553
[32] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 189–196.
[33] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Gener. Comput. Syst.*, vol. 55, pp. 278–288, Feb. 2016.
[34] B. Kiran, D. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *J. Imag.*, vol. 4, no. 2, p. 36, Feb. 2018.
[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
[36] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[37] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402
[38] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.
[39] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2112–2119.
[40] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 563–576.
[41] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *Proc. CVPR Workshops*, Jun. 2011, pp. 55–61.
[42] M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Comput. Vis. Image Understand.*, vol. 116, no. 3, pp. 320–329, Mar. 2012.
[43] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.
[44] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.
[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.



ZHAOYAN LI received the M.S. degree in computer science from Sichuan University, in 2007. He is currently an Associate Professor with Xihua University. He is the author of more than ten journal papers. His current research interests include event detection, image processing, and information fusion.



YAOSHUN LI received the B.S. degree from the School of Computer and Software Engineering, Xihua University, in 2017, where he is currently pursuing the master's degree. His current research interests include image restoration and image super-resolution.

ZHISHENG GAO received the Ph.D. degree in computer science from Sichuan University, in 2012. He is currently an Associate Professor with Xihua University. He is the author of more than 20 journal articles. His current research interests include machine learning, image processing, and information fusion.