Deep social force network for anomaly event detection

Xingming Yang¹,²,³ | Zhiming Wang³ | Kewei Wu¹,²,³ | Zhao Xie¹,²,³ | Jinkui Hou⁴

¹ Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei, China
² Anhui Province Key Laboratory of Affective Computing & Advanced Intelligent Machine, Hefei University of Technology, Hefei, China
³ School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
⁴ School of Computer Engineering, Weifang University, Weifang, China

Correspondence
Zhao Xie, Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei 230601, China.
Email: xiezhaoc@hfut.edu.cn
Jinkui Hou, School of Computer Engineering, Weifang University, Weifang 261061, China.
Email: jkhou@163.com

Abstract
Anomaly event detection is vital in surveillance video analysis. However, how to learn the discriminative motion in the crowd scene is still not tackled. Here, a deep social force network by exploiting both social force extracting and deep motion coding is proposed. Given a grid of particles with velocity provided by the optical flow, the interaction force in the crowd scene is investigated and a social force module is embedded in a deep network. A deep motion convolution was further designed with a 3D (DMC-3D) module. The DMC-3D not only eliminates the noise motion in the crowd scene with a spatial encoder-decoder but also learns the 3D feature with a spatio-temporal encoder. The deep social force coding is modelled with multiple features, in which each feature can describe specific anomaly motion. The experiments on UCF-Crime and ShanghaiTech datasets demonstrate that our method can predict the temporal localization of anomaly events and outperform the state-of-the-art methods.

1 | INTRODUCTION

Anomaly detection in the surveillance video is to monitor unusual events, which are not involved in general events. Anomaly detection is an important application in the surveillance video, for example, intelligent surveillance, violence alerting, evidence investigation. For crowd scenes, the anomaly events have many variations, including irregular motion speed, irregular motion direction, irregular motion flow, and even irregular motion patterns. Therefore, the difficulty of learning motion features for anomaly events is how to describe these motion variations.

As the success of a convolutional neural network of learning features, Hasan et al. learning anomaly features a fully convolutional autoencoder from HOG and HOF [1], and consider that the reconstruction error of regular motion is low and that of irregular motion is high. Sultani et al. use multiple instances learning to optimize the 3D ConvNet features [2]. Zhong et al. consider the anomaly features that suffer from the label noise and design a graph convolutional noise cleaner to denoise the anomaly features [3]. Zaheer et al. further discuss the label noise of batch with a normalcy suppression module [4]. However, the motion features of the above methods are learned from video frames. And when the frame label is noisy, these motion features can be degraded.

The social force model describes the motion feature of the crowd as the physical force. Mohammadi et al. define three
behavioural rules for anomaly events [5]. The obstacle force is the interaction of the presence of an obstacle. The contact force is the interaction from surrounding persons. And the aggression force is the interaction towards the opponents, which is mostly like violent actions. Li et al. use a region proposal to guide the social force feature extraction [6]. The deep model can optimize the feature from the social force model [7].

In this work, we propose a deep social force network for anomaly event detection, in which we focus on both social force coding and spatio-temporal motion coding. As shown in Figure 1, first, to describe the motion variations, we introduce three modules in the deep network for the interaction of obstacle force, contact force, and aggression forces. Then we design a deep motion convolution with a 3D (DMC-3D) module to learn spatio-temporal motion features. The DMC-3D module integrates three blocks. Especially, the spatial encoder–decoder denoises spatial motion feature. The spatio-temporal encoder joint optimizes features for spatio-temporal patterns. And the motion classifier predicts the anomaly score.

The contributions are summarized as follows. We propose a deep social force network for anomaly event detection, which targets discovering the interaction force of particles and learn the deep social force features. (1) We build a deep motion convolution with a 3D (DMC-3D) module for anomaly feature learning. The DMC-3D not only eliminates the noise motion in the crowd scene with a spatial encoder–decoder but also learns the 3D feature with a spatio-temporal encoder. (2) To describe the specific anomaly motion, we introduce three social force blocks in the deep network. And the features of the three forces, including obstacle force, contact force, and aggression force, can provide the force location and imply the violent action. (3) We design an FC pooling to learn the weight of multiple deep social force features, which can adaptively select the interesting feature for anomaly events. The experiment on the UCF-Crime anomaly detection dataset shows that our method can predict the temporal localization of anomaly events and outperform the state-of-the-art methods.

2 | RELATED WORK

2.1 | Anomaly event detection

The key to video anomaly detection is anomaly feature learning. Early attempts extract hand-crafted features, such as 3D gradient feature [8], local binary pattern, gray level co-occurrence matrix [9], HOG, HOF, and STIP [10]. However, the hand-crafted features cannot select the interesting feature for an event.

The deep feature can use backpropagation to learn the kernel of convolution, which is more successful for anomaly detection [11, 12]. The skeleton feature has been used to describe normal patterns of human action [13–15]. And the acoustic features enhance the concept-level representation of micro-videos [16]. The above methods consider anomaly detection as action recognition, and cannot solve the temporal localization in untrimmed videos.

As the weak label for temporal localization, weakly supervised learning has been introduced to reduce the noise labels. Sultani et al. adopts multi-instance learning for the weakly supervised label [2]. The label cleaner can use the original prediction score and even clean the noise iteratively [3, 17, 18]. They consider when the noise is clean; the rest frames will be left as anomaly events.

The anomaly spatial localization can be solved by comparing the ground truth next frame and the prediction result. An autoencoder can reconstruct the prediction image. The autoencoders have many variations, such as residual learning model [19], self-attention model [20] and feature memory model [21], super-resolution model [22], temporally coherent sparse coding RNN [23], stacked RNN auto-encoder [24]. Motion feature also has been introduced to feature reconstruction [25, 26].

2.2 | Action recognition

As the success of deep CNN, the majority of existing methods extract deep features. Two main CNN architectures of action
FIGURE 2  The framework of our deep social force network. For various motion features, we introduce a social force module to describe the interaction of individuals. For the deep motion feature, we design a DMC-3D module, which integrates the spatial encoder–decoder, the spatio-temporal encoder, and the motion classifier.

recognition are 2D CNN [27–31] and 3D CNN [32–36]. 2D CNN has higher efficiency because of fewer convolution operations than 3D CNN. 3D CNN can learn spatio-temporal features from 3D volume due to 3D convolution and 3D pooling operations.

Unlike appearance features in image classification, video understanding needs to learn additional motion features from several successive frames. Some focus on the spatio-temporal and motion encoding method [37] and other learn optical-flow-like features from data [38–41].

Another type of deep network is RNN, especially LSTM, which has greater capabilities of modelling dynamics among frames. RNN does not focus on visual patterns learning from the spatial-temporal cube. Therefore, the CNN-LSTM, as a cascade architecture, attracts widespread attention because RNN and CNN are complementary to each other. The most recent RNN focus on searching the convolutional network for temporal structure [42–45].

2.3 Social force model

Unlike appearance features, the social force model and its variations use physical models, such as repulsive and attractive forces, motion equations, and interaction energy to describe motion [46]. Mohammadi et al. define a set of simple behavioural heuristics to describe people’s behaviours in the crowd [5]. They implement these heuristics into physical equations and finally classify the anomaly event. Li et al. use social force to proposal region [6] and Sumon et al. add SFM into CNN and LSTM for violent crowd flow detection [7]. SFM analyses the particle trajectory and can be applied for multi-pedestrian interaction [47], separating crowd behaviour [48], alighting and boarding behaviours [49], and evacuation assistant [50]. Qi Wang et al. learn pixel-wise features with spatial CNN for crowd understanding [51]. Qi Wang et al. provide a structural crowd descriptor with motion and context similarities [52].

In our work, we imbed a social force module into the deep network for anomaly detection. We focus on the architecture search of the encoder–decoder to reduce the noise motion feature. To further explain what feature can support the anomaly event, our work aggregates multi-social force as a deep social force network.

3 DEEP SOCIAL FORCE NETWORK

The overview of our model is shown in Figure 2. Our model introduces two new modules for deep social force coding. First, because the optical flow provides the pixel-level velocity, we use these velocities to model the interaction force of individuals and design a social force module for three interaction forces. Second, we design the DMC-3D module, in which we denoise the spatial motion noise with spatial encoder–decoder, learn motion features with spatio-temporal encoder, and predict the anomaly score with motion classifier. This network solves the deep motion feature jointly for the social force model and spatio-temporal model.

3.1 Social force module

We design the social force module for motion dynamics by estimating the interaction force from the velocity. In this module, we consider individualistic goals or environmental constraints. When the individual \( p \) with mass \( m(p) \) changes his/her velocity \( v(p) \) the actual force \( F^a(p) \) is

\[
m(p) \cdot \frac{dv(p)}{dt} = F^a(p) = F^d(p) + F^{int}(p)
\]  

(1)
This actual force consists of two parts: personal desire force $F^d(p)$ and interaction force $F^{int}(p)$. Figure 3 shows the three interaction forces, including obstacle force, contact force, and aggression force.

The optical flow of video frame $t$ contains pixel-level velocity $v(t) = [v_x(t), v_y(t)]$, where $v_x(t)$ is the velocity in the horizontal direction and $v_y(t)$ is the velocity in the vertical direction. We consider each pixel as a moving particle. We design three blocks for the interaction force as

$$
\begin{align*}
F^{ob}(t) &= \text{obstacle}(v(t)) \\
F^{co}(t) &= \text{contact}(v(t)) \\
F^{agg}(t) &= \text{aggression}(v(t))
\end{align*}
$$

The obstacle force describes the force of the present obstacle. People often choose the most direct path to their destination, but the abnormal environment of an unexpected obstacle changes their desired velocity $v(t)$ to $v(t+1)$. If the velocity is constant over time, this means that the individual is approaching his/her target destination without encountering any obstacle. Otherwise, we can use the change of velocity to detect the abnormal obstacle. We consider each individual has the unit mass, and estimate the obstacle force $F^{ob} = [F^{obx}, F^{oby}]$ in the horizontal direction and the vertical direction according to the decomposition of the velocity. The formulation of the obstacle force of individual $p$ at time $t$ is

$$
\begin{align*}
F^{obx}(p, t) &= v_x(p, t+1) - v_x(p, t) \\
F^{oby}(p, t) &= v_y(p, t+1) - v_y(p, t)
\end{align*}
$$

The contact force describes the force of surrounding persons. The contacts of the unintentional surrounding physical body may strongly affect the motion of the individual. Similar to the obstacle, the contact can stop the surrounding person from $v(t)$ to 0. Thus, the surrounding obstacle force implies the contact force. We estimate the contact force $F^{co} = [F^{cox}, F^{coy}]$ in the horizontal direction and the vertical direction as

$$
\begin{align*}
F^{cox}(p, t) &= \sum_q v_x(q, t) \left| \vec{p}_q \right| < \tau \\
F^{coy}(p, t) &= \sum_q v_y(q, t) \left| \vec{p}_q \right| < \tau
\end{align*}
$$

where $p$ is the centre individual and $q$ is the surrounding individual, $|\vec{p}_q|$ is the distance between two individuals, and $\tau$ is the distance threshold to indicate the individuals who are close enough to have a body contact.

The aggression force describes the violent motion which is mainly toward the opponent. To estimate the effect on the opponent, we consider the component of the obstacle force as the aggression force, which is in the direction of the movement of the opponent. And the intersection angle between the force of two individuals is calculated as

$$
\cos \theta_{p,q} = \frac{\vec{p}_q \cdot \vec{q}_p}{|\vec{p}_q||\vec{q}_p|}
$$

Then the two components of the aggression force in the horizontal direction and vertical direction $F^{agg} = [F^{aggx}, F^{aggy}]$ are formulated as

$$
\begin{align*}
F^{aggx}(p, t) &= \sum_q F^{obx}(q, t) \cdot \cos \theta_{p,q} |\vec{p}_q| < \tau \\
F^{aggy}(p, t) &= \sum_q F^{oby}(q, t) \cdot \cos \theta_{p,q} |\vec{p}_q| < \tau
\end{align*}
$$

3.2 Deep motion coding with 3D module

The motion of particles in the scene describes the anomaly pattern. This pattern is a spatio-temporal feature, which can be learned with 3D convolution. Our DMC-3D module contains three blocks. The DMC module focuses on removing the
FIGURE 4 The variation of DMC block. (a) DMC with max-pooling. (b) DMC with interpretation. (c) DMC with deconvolution. (d) DMC with residual deconvolution

noisy motion with spatial encoder–decoder. The 3D convolution learns the spatio-temporal feature. And the FC layer predicts the anomaly score. The operation of DMC-3D is formulated as

\[
\begin{align*}
    x_{\text{DMC}}(t) &= DMC(x(t)) \\
    x_{3D} &= 3D\text{Conv}\{\text{concat}\{x_{\text{DMC}}(t)\}\} \\
    y(t) &= FC(x_{3D}(t))
\end{align*}
\]  

where \(x(t)\) is the feature of time \(t\), the feature can be the RGB, optical flow (OF), obstacle force, contact force, and aggression force. \(x_{\text{DMC}}(t)\) is the output result of the DMC module and \(x_{3D}\) is the output feature of 3D-Conv, \(y(t)\) is the final output of full connection operation.

Figure 4 shows the variation of the DMC block as a spatial encoder–decoder to smooth the noisy motion. The DMC-max-pooling (Figure 4a) has two downsampling layers with \(2 \times 2\) max-pooling, which can reduce the motion noise by pooling the high-resolution feature into the low-resolution feature. The DMC with downsampling is formulated as

\[
\begin{align*}
    x_{\text{DMC}}^{\text{maxpooling}}(t) &= DX(DY(x(t))))
\end{align*}
\]  

where \(DX(.)\) is the downsampling operation.

The DMC-downsampling block has only an encoder. This block decreases the resolution simultaneously and may decrease the anomaly pattern consequently. Therefore, we need another decoder to get a high resolution. The DMC-interpretation (Figure 4b) has two additional upsampling layers by bilinear interpolation with scale factor 2. This reduces the noise and keeps the resolution simultaneously. And the formulation of DMC-upsampling is

\[
\begin{align*}
    x_{\text{DMC}}^{\text{interpretation}}(t) &= UP(UP(DS(DS(x(t))))))
\end{align*}
\]  

where \(UP(.)\) is the upsampling operation.

The interpolations in downsampling and upsampling do not have any parameter, while the kernel in convolution will help to learn the anomaly pattern. Therefore, we use convolution and deconvolution to replace the downsampling and upsampling operation. The DMC-DeConv (Figure 4c) has two convolutions with \(2 \times 2\) max-pooling and two deconvolution layers with stride 2. The kernel in DMC-DeConv is \(3 \times 3\). And the formulation of DMC-DeConv is

\[
\begin{align*}
    x_{\text{DMC}}^{\text{DeConv}}(t) &= De(De(Conv(Conv(x(t))))))
\end{align*}
\]  

where \(Conv(.)\) means convolution operation and \(De(.)\) means deconvolution operation.

As the success of Resnet, the DMC-ResDeConv (Figure 4d) introduces identity mapping after the deconvolution layer to increase the gradient in backpropagation. This will quicken the convergence and alleviate the degradation of the deep model. And the formulation of DMC-ResDeConv is

\[
\begin{align*}
    x_{\text{DMC}}^{\text{ResDeConv}}(t) &= De(De(Conv(Conv(x(t)))))) + x(t)
\end{align*}
\]  

Our spatio-temporal encoder is a 3DConv block, which learns motion representation from the spatial-temporal cube.
This cube contains multiple features of successive frames. Therefore, we concatenate the features of 16 frames as the cube. The C3D network [28] is used as the 3DCnov block and the FC block. The C3D network has 8 3D convolutions, 4 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. All pooling kernels are $2 \times 2 \times 2$, except for pool1 with $1 \times 2 \times 2$. Each FC layer has 4096 output units. The softmax output layer predicts the anomaly score.

### 3.3 Aggregation of multiple features

To avoid the bias of the social force for anomaly detection, we design three multiple feature pooling strategies to learn the anomaly pattern. Given the prediction score $Z_s$, whereas include the feature of RGB, optical flow, obstacle force, contact force, and aggression forces. The mean pooling can reduce the effect of the noise score. If the mean feature prediction is the anomaly event, then the final prediction is also an anomaly. The mean pooling is

$$Z_{\text{mean}} = \text{mean}(\{Z_i\})$$

(12)

The max-pooling can select the interesting feature, and find what force supports the anomaly event. If one of the forces considers the event as an anomaly, then the final prediction is also an anomaly. The max-pooling is

$$Z_{\text{max}} = \max(\{Z_i\})$$

(13)

The selection in the max-pooling is binary processing, that is, we select one feature or none. This can be modified with weighted processing because the weight can describe the rule with multiple social forces. And the weight can be as an FC layer. The FC pooling is

$$Z_{\text{FC}} = \sum_i w_i Z_i$$

(14)

### 3.4 Multiple-instance ranking loss

We train our model with video-level annotation. Due to the lack of frame-level annotation, we use the multiple-instance ranking loss for this weakly labelled training. Given the bag of clips in anomaly action $B_a$ and clips in normal action $B_n$, the ranking loss considers the score of the anomaly bag to be larger than that of the normal bag. The score of the bag is estimated as the max score of the bag. According to [2], the ranking function is

$$\max_{i \in B_a} z(i) > \max_{i \in B_n} z(i)$$

(15)

where $i$ is the clip in the anomaly bag, and $i'$ is the clip in the normal bag. The input video sample into clips and each clip has 16 frames to feed into 3D Convolution to get the prediction $z$.

The anomaly loss contains three parts, the rank loss, the sparse loss, and the smooth loss. The ranking loss is the hinge-loss according to ranking function, which denote the normal clip in anomaly video. The sparse loss describes that the score in the anomaly video should be zero for the most time. This is because the anomaly action occurs only for a short time in real-world scenarios. The smooth loss constraints that in the short anomaly clip, the score should be stable. Therefore, the formulation of anomaly loss is

$$\begin{align*}
L_{\text{loss}}(B_a, B_n) &= L_{\text{loss}} \text{rank}(B_a, B_n) + L_{\text{loss}} \text{sparse}(B_a) + L_{\text{loss}} \text{smooth}(B_a) \\
L_{\text{loss}} \text{rank}(B_a, B_n) &= \max(0, 1 - \max_{i \in B_a} z(i') + \max_{i \in B_n} z(i')) \\
L_{\text{loss}} \text{sparse}(B_a) &= \lambda_1 \sum_i |B_a| \min(z(i), \lambda) \\
L_{\text{loss}} \text{smooth}(B_a) &= \lambda_2 \sum_i |B_a|^{-1} (z(i') - \lambda)^2
\end{align*}$$

We use the anomaly loss to train each single stream model. After training the single stream, the FC pooling uses each stream score as input and uses cross-entropy to train another stage. The first stage provides the parameter in each stream, and the second stage provides the parameter in the FC pooling layer. The mean pooling and the max-pooling for multiple streams do not need additional training.

### 4 EXPERIMENTS

#### 4.1 Dataset

We evaluate our method on UCF-Crime and ShanghaiTech. UCF-Crime has 14 types of anomalies with 1900 untrimmed videos, which are all captured in real-world scenarios. It has 290 videos with frame-level temporal annotation. We use video with only video-level annotation for training and test on the video with frame-level temporal annotation. Following the same split provided in [2], the training set consists of 800 normal and 810 anomalous videos, and the testing set including 150 normal and 140 anomalous videos. The length of the videos varies from 1 to 40 minutes. Each video contains 1–2 anomaly clips.

ShanghaiTech has 437 videos, including 130 abnormal events on 13 scenes. It has the pixel-level ground truth, which indicates the frame-level abnormal events. In the protocol [3], the training set consists of 175 normal videos and 63 anomalous videos, and the testing set includes 150 normal and 44 anomalous videos. The video clips ranges from 15 seconds to over a minute long.

#### 4.2 Implementation details

We use PyTorch to implement the whole framework. We employ two Nvidia GeForce GTX 1080Ti graphics cards to conduct our experiments. Before computing features, we re-size each video frame to 240 $\times$ 320 pixels as RGB input and fix the
frame rate to 30 fps. We obtain optical flow (OF) by dense flow to analyze the impact of our attention on OF stream. We obtained three social forces, including obstacle force (OBF), contact force (COF), and aggression force (AGF), by our social force module. We divided each video into 32 non-overlapping clips, and each clip contained 16 frames. We input the feature of each frame for the spatial encoder–decoder block. And we concatenated the features of 16 frames for the 3DConv block. The output of the FC layer is the prediction of anomaly action.

We train out the DMC module of every single feature on the UCF-Crime training set. We randomly selected 30 positive and 30 negative clips as a batch. We considered this batch as the bag for multiple instance learning. We use an Adagrad optimizer with an initial learning rate of 0.001. The whole training process stopped at 50 epochs. Then we learned the FC pooling of multiple features with the same batch size and same optimizer.

4.3 Evaluation metric

We use two metrics to evaluate anomaly detection. One is the area under the curve (AUC) of the frame-based receiver operating characteristic (ROC) curve. The ROC curve is a graphical plot that illustrates the diagnostic ability of a classifier system. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR is also known as recall, which defines how many correct positive results occur among all positive samples available during the test. The FPR, on the other hand, defines how many incorrect positive results occur among all negative samples. Given the number of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN), the formulation of TPR is $TPR = TP / (TP + FN)$, and the formulation of FPR is $FPR = FP / (FP + TN)$. Given the ROC curve, the AUC is equal to the probability of a classifier ranking a randomly chosen positive instance higher than a randomly chosen negative one. A robust model should have a larger AUC value.

The other is the false alarm rate (FAR). As the FPR is also known as the probability of false alarm, we use FPR at 50% threshold as the false alarm rate. As the major part of the real-world surveillance video is normal, a robust model should have a low false alarm rate on normal clips.

4.4 Experiments on UCF-Crime

4.4.1 Ablation experiments

The ablation studies contain three parts: (1) evaluate various spatial encoder–decoder in deep motion coding; (2) discuss various aggregations of multiple social forces; (3) verify the effect of various social forces.

**Effect of various deep motion coding**

Table 1 shows the result of various DMC with RGB feature. As the 3DConv block and motion classifier block is the same as C3D. We use C3D as a baseline and further add the spatial encoder–decoder block. The DMC-maxpooling gets higher performance than the C3D because the max-pooling reduces the noise of the feature. The DMC-interpretation increases the performance because the interpretation provides more features in the large resolution. Beyond the DMC without parameters, the DMC-DeConv uses convolution and deconvolution to replace the max-pooling and interpretation layer. The kernels of convolution and deconvolution are used to learn robust features. The DMC-ResDeConv further solves the vanishing gradient problem in backpropagation and gets the highest performance in Table 1.

We further showed the temporal anomaly detection (a) and the ROC curve (b) of a test video of the fighting action in Figure 5. The C3D model predicts a high anomaly score of normal clips before the second fighting. This is probably because of the irregular pattern from the noisy feature. And our DMC-ResDeConv can alleviate the noise feature. The ROC curve also shows the DMC-ResDeConv to have a larger AUC value and predicts a low false alarm rate on normal clips.

**Effect of multiple feature aggregation**

Table 2 shows the result of various aggregations with five-stream features, including RGB, optical flow, obstacle force, contact force, and aggression forces. As known, the mean pooling average score of each stream the max-pooling highlights the high score of each stream. The max-pooling outperforms the mean pooling because the max-pooling retains the maximum response of the discriminative feature. And the FC pooling gets the highest performance in Table 2 because it further learns the weight of multiple streams.

**Effect of various features**

Table 3 shows the result of various social forces and their aggregation. The optical flow outperforms RGB, which suggests...
Figure 5 (a) The temporal anomaly score and (b) the ROC curve of a test video on UCF-Crime. We show the comparison of C3D and DMC-ResDeConv with RGB feature.

Table 3 The performance of various features and their aggregation on UCF-Crime

| Method               | AUC (%) | False alarm (%) |
|----------------------|---------|-----------------|
| RGB                  | 77.50   | 0.98            |
| Optical flow         | 78.71   | 0.74            |
| Obstacle force       | 80.29   | 0.42            |
| Contact force        | 81.33   | 0.35            |
| Aggression force     | 80.98   | 0.32            |
| Aggregation with FC pooling | 84.90 | 0.27 |

that the motion feature is more important than the appearance features for anomaly action. Three social forces describe the interaction of individuals and outperform the optical flow. The performance from low to high is obstacle force, aggression force, and contact force, which implies that the contact force occurs in abnormal videos most frequently.

As a single feature focus on a specific pattern, its detection will miss some other pattern. Therefore, the aggregation of the five-stream feature describes various anomaly patterns for each anomaly action. And the aggregation model is more robust than a single feature. We also show the ROC curve in Figure 6. The curve shows that the false alarm rate of aggregation is lower than that of a single feature.

4.4.2 Comparison with state-of-the-art methods

Our comparison method includes sparse coding [8], fully convolutional autoencoder [1], multiple instance learning [2], graph convolutional label noise cleaner [3], and clustering assisted weakly supervised learning [4]. Table 4 shows the result of the above state-of-the-art methods.

Comparison with sparse coding

Our model outperforms sparse coding [8] because we use a C3D network with an additional spatial encoder-decoder, and the sparse coding uses 3D gradient features, which is a handcrafted feature. As shown in Table 3, our model with only the RGB stream has 77.50 AUC, which still outperforms the model with sparse coding. This suggests the deep feature is more robust than the 3D gradient features.

Figure 6 The ROC curve on UCF-Crime. We show the comparison of various features and their aggregation.

Table 4 The performance comparison with the state-of-art methods on UCF-Crime

| Method                | Backbone | AUC (%) | False alarm (%) |
|-----------------------|----------|---------|-----------------|
| Lu et al. 2013 [8]    | NA       | 50.60   | 27.20           |
| Hasan et al. 2016 [1] | Autoencoder | 65.51 | 3.10            |
| Sultani et al. 2018 [2]| C3D      | 75.41   | 1.90            |
| Zhong et al. 2019 [3] | C3D      | 82.12   | 0.10            |
| Zaheer et al. 2020 [4]| C3D      | 83.03   | NA              |
| DSFN [Ours]           | C3D      | 84.90   | 0.27            |
Comparison with fully convolutional autoencoder

Our model outperforms fully convolutional autoencoder [1]. This is because we use 3D convolution to learn features with spatial-temporal constraints. While the feature in a fully convolutional autoencoder only considers the spatial constraint it suggests that the 3D convolution is important for anomaly feature learning.

Comparison with label noise cleaner

The three methods, multiple instance learning [2], graph convolutional label noise cleaner [3], and clustering assisted weakly supervised learning [4], focus on cleaning the temporal label noise. The multiple instance learning [2] uses multiple instance bags to smoothen the temporal noise. The graph convolutional label noise cleaner [3] uses graph inference to estimate temporal noise. The clustering assisted weakly supervised learning [4] designs a normal suppression module to denoise the normal frames. Our model not only adopts the multiple instance bag to smoothen the temporal noise but also uses an encoder-decoder to remove spatial noise. Therefore, our model outperforms existing methods with label noise cleaner.

4.4.3 Visualization of social force

We further visualize the social force of video frames to indicate what social force supports anomaly action. Figure 7 shows the anomaly score of single features in (a) fighting, (b) assault, (c) abuse, (d) burglary, (e) explosion, (f) road accident. We label anomaly points in Figure 7, and further show their features in Figure 8, including RGB, optical flow, obstacle force, contact force, and aggression forces. We notice that (1) aggression force is most discriminative for the anomaly action, and has the highest score. This is probably because the interaction of aggression is the conflict at the pixel-level, which is more like a violent action than other interactions. (2) The localization of the interaction force for violent action is more accurate than that of optical flow. This is because optical flow only describes the velocity of pixels, while the interaction of force captures the local pattern of the velocity. (3) In the explosion, the motion feature predicts later than the RGB feature. This is because before the explosion, the appearance feature changes, while the motion feature is not evident at the beginning. (4) The responses of the interaction forces change on various anomaly actions. The obstacle force describes the irregular acceleration and has a high response for the anomaly action of the abuse and burglary. (5) The contact force mainly focuses on the action with limited range and can detect the abnormal motion in the assault. The road accident has both high responses of the obstacle and contact force. (6) The aggression force finds the directional motion toward the opponent. The directional motion most exists in the action of fighting and explosion.

4.5 Experiments on ShanghaiTech

4.5.1 Ablation experiments

For ShanghaiTech dataset, we provide the AUC and false alarm in Table 5, and the ROC curve in Figure 9. The force feature describes complex motion in the anomaly event and outperforms the RGB/Optical flow feature. The aggregation model
The visualization of various feature on UCF-Crime, including (a) RGB, (b) optical flow, (c) obstacle force, (d) contact force, and (e) aggression force. Each frame is the labelled frame in Figure 7.

The ROC curve of the testing set on ShanghaiTech. We show the comparison of various features and their aggregation improves because the FC pooling adaptively learns the impact of each feature.

| Method               | AUC (%) | False alarm (%) |
|----------------------|---------|-----------------|
| RGB                  | 74.89   | 1.32            |
| Optical flow         | 76.20   | 1.17            |
| Obstacle force       | 78.55   | 0.84            |
| Contact force        | 79.36   | 0.79            |
| Aggression force     | 80.55   | 0.77            |
| Aggregation with FC pooling | 82.14 | 0.74 |

4.5.2 Comparison with state-of-the-art methods

Our comparison method includes unsupervised methods and supervised methods. The unsupervised methods only use normal video for training, including memory-guided auto-encoder [20], spatial and temporal constrained frame prediction [23], temporally coherent sparse coding RNN [23], stacked RNN auto-encoder [24], skeleton GRU [13], skeleton graph
Comparison with unsupervised methods

The unsupervised methods learn the normal pattern with normal video and find abnormal frames that are different from those defined as normal. One can set a threshold and if the score of the frame is smaller than the threshold, the frame can be categorized as an abnormal case. By changing the threshold gradually, we can arrive at a ROC curve. The above methods learn normal feature pattern [23, 24], normal frame prediction [21, 26], and normal skeleton pattern [13, 14]. As abnormal video has not been seen in training, therefore, the unsupervised methods get lower performance than the supervised methods.

4.5.3 Visualization of social force

We visualize the social force scores of video frames in Figure 10 and their features in Figure 11. In the first column, the anomaly score of optical flow can describe the persons in chasing because of the change of speed. The chasing event generates contact force and aggregation force, and these two forces both provide a higher score than the optical flow. The aggregation force detects the chasing earlier than the contact force because the motion toward the opponent occurs before the contacting surrounding persons. In the second column, the RGB feature cannot describe the motion and predict confusion results of robbery. After frame 300, the speed is similar to running, and the optical flow mistakes the motion as normal. But the social force still occurs, specifically motioning toward the opponent, therefore, the aggregation force captures the robbery reliably.

5 CONCLUSION

In this work, we propose a deep social force network for anomaly event detection, which targets in discovering the interaction force of particles and learning the deep social force features. We build a deep motion convolution with a 3D (DMC-3D) module, which not only eliminates the noise motion in the crowd scene with a spatial encoder–decoder but also learns the 3D feature with a spatio-temporal encoder. To describe the specific anomaly motion, we introduce three social

| Method                          | Backbone | AUC (%) | False alarm (%) |
|--------------------------------|----------|---------|-----------------|
| Normal video training          |          |         |                 |
| Luo et al. 2017 [23]           | RNN      | 68.00   | NA              |
| Luo et al. 2021 [24]           | RNN      | 69.63   | NA              |
| Park et al. 2020 [21]          | U-Net    | 70.50   | NA              |
| Liu et al. 2018 [26]           | U-Net    | 72.80   | NA              |
| Morais et al. 2019 [13]        | GRU      | 73.40   | NA              |
| Normal/anomalous video training|          |         |                 |
| Markovitz et al. 2020 [14]     | GCN      | 76.10   | NA              |
| Zhong et al. 2019 [3]          | C3D      | 76.44   | NA              |
| DSFN [Ours]                    | C3D      | 82.14   | 0.74            |

FIGURE 10 The anomaly score of each feature on ShanghaiTech

```
force blocks in the deep network. We introduce pooling to fuse the prediction of multiple deep social force features. The ablation studies on the UCF-Crime dataset show that spatial convolution can improve the social force feature learning, and FC pooling can adaptively select the interesting feature for anomaly events. The visualization also shows that our method can predict the temporal localization of anomaly events and find what feature is important for anomaly events.

ACKNOWLEDGEMENTS
This research was supported by the National Key Research and Development Program of China (2017YFB1002203), the Fundamental Research Funds for the Central Universities of China (PA2020GDSK0059), National Nature Science Foundation of China (61503111, 61876058), and Anhui Natural Science Foundation Funds for the Central Universities of China (PA2020GDSK0059), National Nature Science Foundation of China (1808085MF168).

REFERENCES
1. Hasan, M., et al.: Learning temporal regularity in video sequences. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–742. IEEE (2016)
2. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488. IEEE (2018)
3. Zhong, J., et al.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1237–1246. IEEE (2019)
4. Zaheer, M.Z., et al.: CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: European Conference on Computer Vision, 22, 358–376 (2020)
5. Mohammadi, S., et al.: Angry crowds: Detecting violent events in videos. In: European Conference on Computer Vision, 7, pp. 3–18 (2016)
6. Li, X., et al.: Simulation of evacuating crowd based on deep learning and social force model. IEEE Access 7, 155361–155371 (2019)
7. Sumon, S.A., et al.: Violent crowd flow detection using deep learning. In: Asian Conference on Intelligent Information and Database Systems, vol. 1, pp. 613–625 (2019)
8. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in MATLAB. In: IEEE International Conference Computer Vision, pp. 2720–2727 (2013)
9. Lohithashva, B.H., Manjunath Aradhya, V.N., Guru, D.S.: Violent video event detection based on integrated LBP and GLCM texture features. Rev. d'Intelligence Artific. 34(2), 179–187 (2020)
10. Song, D., Kim, C., Park, S.: A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance. Inf. Sci. 447, 83–103 (2018)
11. Wei, X., et al.: A very deep two-stream network for crowd type recognition. Neurocomputing 396, 522–533 (2020)
12. Li, Q., et al.: Recurrent prediction with spatio-temporal attention for crowd attribute recognition. IEEE Trans. Circuits Syst. Video Technol. 30(7), 2167–2177 (2020)
13. Morais, R., et al.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11996–12004 (2019)
14. Markovitz, A., et al.: Graph embedded pose clustering for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10536–10544 (2020)
15. Su, Y., et al.: Human interaction learning on 3D skeleton point clouds for video violence recognition. In: European Conference on Computer Vision, vol. 4, pp. 74–90 (2020)
16. Nie, L., et al.: Enhancing Micro-video Understanding by Harnessing External Sounds. ACM Multimedia (2017)
17. Pang, G., et al.: Self-trained deep ordinal regression for end-to-end video anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 12170–12179. IEEE (2020)
18. Shi, B., et al.: Weakly-supervised action localization by generative attention modeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1006–1016. IEEE (2020)
19. Abati, D., et al.: Latent space autoregression for novelty detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 481–490. IEEE (2019)
20. Gong, D., et al.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: IEEE International Conference on Computer Vision, pp. 1705–1714. IEEE (2019)
21. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 14360–14369. IEEE (2020)
22. Sulun, S., Tekalp, A.M.: Can learned frame-prediction compete with block-motion compensation for video coding? (2020). https://arxiv.org/abs/2007.08922
23. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: IEEE International Conference on Computer Vision, pp. 341–349. IEEE (2017)
24. Luo, W., et al.: Video anomaly detection with sparse coding inspired deep neural networks. IEEE Trans. Pattern Anal. Mach. Intell. 43(5), 1070–1084 (2020)
25. Nguyen, T., Meunier, J.: Anomaly detection in video sequence with appearance-motion correspondence. In: IEEE International Conference on Computer Vision, pp. 1273–1283. IEEE (2019)
26. Liu, W., et al.: Future frame prediction for anomaly detection—A new baseline. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536–6545. IEEE (2018)

FIGURE 11 The visualization of various features, including (a) RGB, (b) optical flow, (c) obstacle force, (d) contact force, and (e) aggression force. Each frame is the labelled frame in Figure 10
27. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Conference on Neural Information Processing Systems, pp. 568–576 (2014)
28. Tran, D., et al.: Learning spatiotemporal features with 3D convolutional networks. In: IEEE International Conference on Computer Vision, pp. 4489–4497. IEEE (2015)
29. Chen, Y., et al.: Multi-filter networks for video recognition. In: European Conference on Computer Vision, vol. 1, pp. 364–380 (2018)
30. Lin, J., Gan, C., Han, S.: TSM: Temporal shift module for efficient video understanding. In: IEEE International Conference on Computer Vision, pp. 7082–7092. IEEE (2019)
31. Sudhakaran, S., Escalera, S., Lanz, O.: Gate-shift networks for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1430–1439. IEEE (2018)
32. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? A new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1099–1108. IEEE (2020)
33. Zhu, Y., et al.: Hidden two-stream convolutional networks for action recognition. In: Asian Conference on Computer Vision, vol. 3, pp. 363–378 (2018)
34. Ng, J.Y., et al.: Beyond short snippets: Deep networks for video classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702. IEEE (2015)
35. Li, Q., et al.: Action recognition by learning deep multi-granular spatiotemporal video representation. In: International Conference on Multimedia Retrieval, pp. 159–166 (2016)
36. Yang, D., Özgüner, U., Redmill, K.A.: A social force based pedestrian motion model considering multi-pedestrian interaction with a vehicle. ACM Trans. Spatial Algorithms Syst. 6(2), 11:1–11:27 (2020)
37. Yang, X., Wang, Q., Liu, T.: A fuzzy-theory-based social force model for studying the impact of alighting area width on alighting and boarding behaviors. IEEE Access 8, 28595–28606 (2020)
38. Zheng, Z., et al.: Improved social force model based on emotional contagion and evacuation assistant. IEEE Access 8, 195989–196001 (2020)
39. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vision 129(1), 225–245 (2021)
40. Diba, A., Pazandeh, A.M., Gool, L.V.: Efficient two-stream motion and appearance 3D CNNs for video classification (2016). https://arxiv.org/abs/1608.08851
How to cite this article: Yang, X., et al.: Deep social force network for anomaly event detection. IET Image Process. 2021; 1–13. https://doi.org/10.1049/ipr2.12299