Influence-Aware Attention Networks for Anomaly Detection in Surveillance Videos

Sijia Zhang, Maoguo Gong, *Senior Member, IEEE*, Yu Xie, A. K. Qin, *Senior Member, IEEE*, Hao Li, *Member, IEEE*, Yuan Gao, and Yew-Soon Ong, *Fellow, IEEE*

***Abstract*— Detecting anomalies in videos is a fundamental issue in public security. The majority of existing deep learning methods often perform anomaly detection based on the behavior or the trajectory of a single target. However, due to the overlaps of the crowd and the low-resolution of monitoring images, the segmentation of population is hard to implement and the features cannot be learned thoroughly, which make the methods be easily disturbed by visual elements and thus may lead to false detection sometimes. To tackle these problems, we propose the influence-aware attention to learn the representative attributes of the whole crowd. Walking pedestrians can be divided into numbers of flows, and in this paper, we aim to measure the consistency of movement patterns in the same stream and the interactions between different streams. Meanwhile, great importance is given to the relation between pedestrians and the circumstance for certain anomalies occur as a result of environmental issues. Specifically, the influence-aware attention module is composed of the motion attention and the location attention, which is designed to quantify the relations in the scene from spatial and temporal aspects. For the lack of abnormal samples, we utilize a dual generator-based framework to learn interactions among normal scenes. Experimental results on six benchmarks verify the effectiveness and robustness of our proposed method.**

***Index Terms*— Influence-aware attention, GAN, Anomaly detection.**

I. Introduction

With a rapid growth of video acquisition devices, there is an increasing need for detecting anomalies from a massive amount of video data [1]–[13] to protect citizens and their properties. At present, crowd abnormal detection is particularly important, especially in some festivals with gathering crowds. The application of anomaly detection not only guarantees the personal safety, but also benefits the social security. In our previous work [14], the behaviors of individuals are extracted and a binary classification module is assigned to recognize actions and detect anomalies jointly, and in this work, we attach great importance to the interactions in the crowd, including the individual-individual interaction and the individual-environment interaction.

The main research in anomaly detection can be divided into two parts, the individual-based methods and the crowd-based methods. The individual-based methods often conduct the action recognition and the target tracking of each pedestrian, and anomalies are then detected by the abnormal behaviors or the unexpected trajectories [1], [15], [16]. However, methods of this category are more suitable for the crowd with low density than that with high density. There are inevitable overlaps and occlusions accompanying with the crowd density increasing, which causes difficulties to obtain the clear segmentation of each single person and will seriously reduce the accuracy of abnormal event detection.

The crowd-based methods avoid these problems by extracting population characters [1], [2] and constructing the group mode [3], [5]. For extracting population characters, visual features of different aspects are extracted, such as histogram of oriented gradients, optical flow [17]–[19], etc. These features are integrated to present different events. For modeling the group mode, there have been a variety of event models according to different environments and requirements, which follow two main approaches based on the availability of data labels. The supervised learning methods use several different samples as training sets to build event models. Among them, a typical method is the Bag of Words (BoW) [15], which first gains the feature vectors and forms the vision dictionary containing different words, and then counts the frequency of occurrence of each word in the image, thus the image can be represented as a k-dimensional value vector. However, it is difficult to obtain labeled anomaly instances with various behaviors for their diversity and irregularity. Compared with the methods mentioned earlier, the unsupervised learning methods are easier to implement and do not require prior knowledge. Hidden Markov Model (HMM) [20] and Dynamic Bayesian Network (DBN) [2] are utilized to calculate the probabilities of the occurrence of different features and build models to represent

Manuscript received 19 October 2021; revised 7 January 2022; accepted 30 January 2022. Date of publication 1 February 2022; date of current version 4 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62036006 and Grant 61906146; in part by the Key Research and Development Program of Shaanxi Province under Grant 2018ZDXM-GY-045; in part by the Australian Research Council (ARC) under Grant LP180100114 and Grant DP200102611; in part by the A*STAR Centre for Frontier AI Research (CFAR); and in part by the Data Science Artificial Intelligence Research Center (DSAIR) at the School of Computer Science and Engineering, Nanyang Technological University. This article was recommended by Associate Editor H. Shi. *(Corresponding author: Maoguo Gong.)*

Sijia Zhang, Maoguo Gong, Yu Xie, Hao Li, and Yuan Gao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: zsj_stella@foxmail.com; gong@ieee.org; sxlljcxy@gmail.com; cn_gaoyuan@foxmail.com).

A. K. Qin is with the Department of Computer Science and Software Engineering, Swinburne University of Technology, Hawthorn, VIC 3122, Australia (e-mail: kqin@swin.edu.au).

Yew-Soon Ong is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: asysong@ntu.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3148392.

Digital Object Identifier 10.1109/TCSVT.2022.3148392

events according to the results of clustering algorithms, which establish probability density functions for different models. The attributes of test videos will be distinguished by the probability of each event to be measured.

However, a major challenge facing anomaly detection in video sequences is the lack of labeled abnormal samples. It is practically infeasible to enumerate the varieties of anomalies which are possible to occur. Besides, due to the intricate behavior patterns and different objects, unpredictable collisions occur more frequently under the dual effects of human factors and objective factors. In addition, there still exist some problems in surveillance video processing, such as the selection of appropriate features and accurate methods for discrimination, which make crowd anomaly detection more complicated. Therefore, if the above problems can be overcome, the anomaly in the video will be correctly identified and the abnormal situation will be timely warned, which is of great significance in the field of public safety.

To address these problems, we propose the influence-aware attention module to focus on the pattern of the whole crowd. In practice, due to the mutual influence of the crowd, it is often manifested as the simultaneous abnormality of multiple targets. Thus, the scope of our study is the interactions which take place in the street scene. It is easily to be observed that a crowd can be divided into several groups, and persons in the same group have similar moving patterns, while persons in different groups may exert influence on each other. Moreover, surrounding objects including vehicles, stairs, etc. can exert a certain influence on pedestrians and isolating only human behavior may affect the detection result. From the above two points, we propose the influence-aware attention module and synthesize it with a dual generator-based anomaly detection framework to evaluate interactions in the crowd. The influence-aware attention includes motion attention and location attention, where motion attention mainly corresponds to learning the pedestrians moving patterns, and location attention is proposed to learn the interactions between persons and objects. Anomalies can reflect both on appearance and motion, and the influence-aware attention learn them simultaneously. Our framework consists of two parts, including the generation part, which contains the influence-aware attention module, generator I and generator II, and the discrimination part. Firstly, in the training phase, to tackle the problem of lacking labeled abnormal samples, we employ the generative adversarial networks (GAN) with influence-aware attention to reconstruct normal frames. Then we construct the generator II through the residual square loss on the basis of the trained generator I, which can generate frames as idealized normal input. Meanwhile, the discrimination part inputs the reconstructed frames from generator I to ensure the learning ability of the latter. The experimental results on several benchmark datasets demonstrate the effectiveness of our approach.

The main contributions of our work are summarized as follows:

- We propose the influence-aware attention module to allow a deeper insight into characteristics of crowds. The motion attention part is used to capture the consistency between people in the same moving patterns and the interactions between people in different moving flows. The location attention part is utilized to measure the environment contextual influence, which ensures the generalization of our method.
- We utilize a dual generator framework to detect anomalies. On the basis of normal GAN structure, we add the influence-aware attention and set an extra generator with the residual square loss to produce the idealized normal data corresponding to each input data.
- The proposed framework is evaluated on six benchmarks of UCSD Ped1, UCSD Ped2, UMN, CUHK Avenue, ShanghaiTech and Street Scene datasets. Experimental results show that our method achieves competitive performance, especially in complex scenes with large objects.

II. Related Work

In recent years, along with the increased frequency of events such as the dense crowd stampede, detecting anomalies of the crowd has aroused extensive attention of the researchers from all over the world, and new deep learning models have made promising progress in this task. Specifically, we give a brief review of relevant works from anomaly detection and the attention mechanism two aspects.

A. Anomaly Detection in Videos

The methods of anomaly detection in videos can be mainly divided into two categories: individual-based methods and crowd-based methods. The former usually make the segmentation of the crowd and obtain the behavior or trajectory of each moving person. Anomalies can be detected through learning the potential information from the behavior or the trajectory of each pedestrian. Morais *et al.* [1] leveraged 2D human skeleton trajectories and separated the skeletal actions in two aspects: global and local. Message-Passing Encoder-Decoder Recurrent Network (MPED-RNN) is proposed to model the global and local components. Anomalies are detected by learning a regularity model of the dynamic skeleton features. Markovitz *et al.* [15] extracted poses in the frames and encoded them to a latent vector using the encoder part of a spatiotemporal graph autoencoder (ST-GCAE). The latent vector is soft-assigned to clusters and then a Dirichlet process-based mixture is utilized to handle the vectors and determine if an action is normal or not. Rodrigues *et al.* [16] developed a multi-timescale framework to predict pose trajectories in past and future. The multi-timescale predictions are combined to detect abnormal activities.

As for the crowd-based methods, features of the whole crowd are extracted.

Liu *et al.* [21] proposed a video prediction framework which adopts U-Net as a generator and constraints of appearance and motion. The difference between predicted frame and its ground truth is calculated as a score to predict whether a frame is normal or not. In [22], Vu *et al.* processed the data to extract low-level visual features like optical flow and high-level features from denoising autoencoders. The abnormal objects are detected at each representation level, and then united to obtain a detection result by the threshold. In [23], anomaly

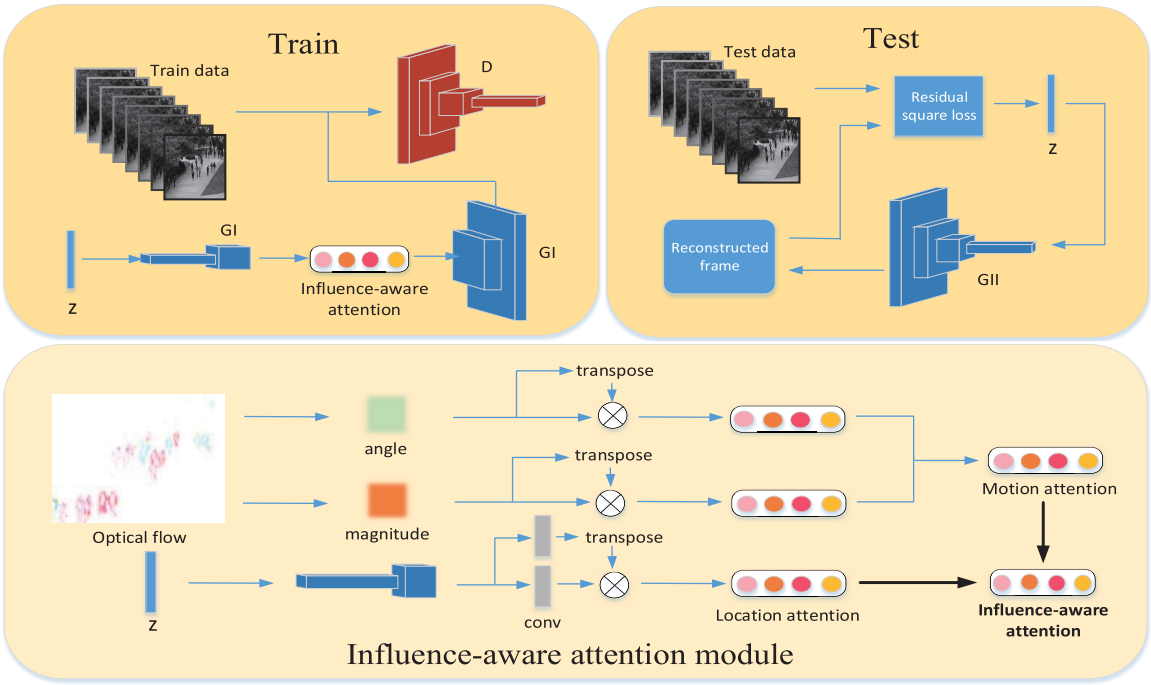


Fig. 1. The architecture of our main framework. The input is composed of a bunch of consecutive frames.

detection is conducted by learning a correspondence between common object appearances and their associated motions. The proposed framework is a combination of an appearance reconstruction model and a motion prediction model to learn the correspondence. In [24], deep multiple instance ranking framework is proposed by Sultani *et al.* to learn anomalies through treating normal and anomalous surveillance videos as bags and short segments of each video as instances in a bag. They used an anomaly ranking model to predict anomaly scores for each video segment in the testing phase. Due to the imbalance of positive and negative samples in anomaly detection, many works formulate abnormal events detection as an unsupervised learning problem. In [25], a memory module is proposed to augment the autoencoder. The module retrieves the most relevant items in the memory, which are delivered to the decoder for reconstruction. The reconstructed errors on anomalies will be strengthened for detection. The method proposed by [26] learns the generation of high-dimensional image space and the inference of latent space. Akcay *et al.* introduced an encoder-decoder-encoder architectural model and compared the distance between normal images and the test data. The work of [27] considers the diversity of normal patterns and a memory module with an update scheme is proposed. Meanwhile, the feature compactness and separateness losses are also presented to boost the discriminative and the feature learning capability of the model.

B. The Attention Mechanism

The attention mechanism has been widely used in natural language processing and object detection [28]–[31]. The main concept of attention mechanism derives from human visual attention. When people perceive things visually, instead of looking from the beginning to the end of a scene, they tend to observe and pay attention to certain parts according to their needs. Learning long-range dependencies is a key challenge in many sequence transduction tasks and anomaly detection tasks. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals having to traverse in the network [32]. The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies.

There are some applications of attention mechanism in anomaly detection. Zhou *et al.* [33] proposed two attention-driven loss, attention-driven content loss and attention-driven gradient loss, to alleviate the foreground-background imbalance problem in anomaly detection. Zheng *et al.* [34] proposed AddGraph and it is an extension of temporal Graph Convolutional Network (GCN). GCN is added with attention to detect anomalous edges in dynamic graph. The contextual attention-based model is applied to catch the short-term pattern of nodes.

III. Methodology

The influence-aware attention networks explore the potential information of interactions in the surveillance video. As shown in Fig. 1, we do this by processing the information in a two-stream architecture: motion attention and location attention. The motion attention captures the interaction among moving objects in time dimension. On the other hand, the location attention acquires the interactive information between moving objects and the stationary background in spatial dimension. The combination of two attention makes the learning of the normal state more detailed.

Fig. 2. An illustration of our primary points in terms of motion and location. The visualization of optical flow presents the movements of different streams of people in color. The following line of figures reflect the correlation between the anomaly and the location.

In this paper, we build our model based on the architecture of GAN, and the influence-aware attention is utilized in the generator of the training phase to retain the spatiotemporal information of videos. Generator I (GI) is used in the training, and as the clone of GI, generator II (GII) is applied in the testing, which is adopted by [35]. In the training phase, frames of normal scenes and noise are input to train GI and the discriminator. GI maps the noise z to the reconstructed output x'. In the testing phase, the test frames and noise are fed into GII, which shares weights with GI. Different from GI, GII employs the residual square loss to update the noise z, which can map the input x to noise z. Then the mapping of $x \rightarrow z \rightarrow x'$ is achieved. Thus the output will obey a distribution which is close to the normal state, and if the input is abnormal, the output of GII will be different from the input. The final result comes from the comparison between the input and the output of GII.

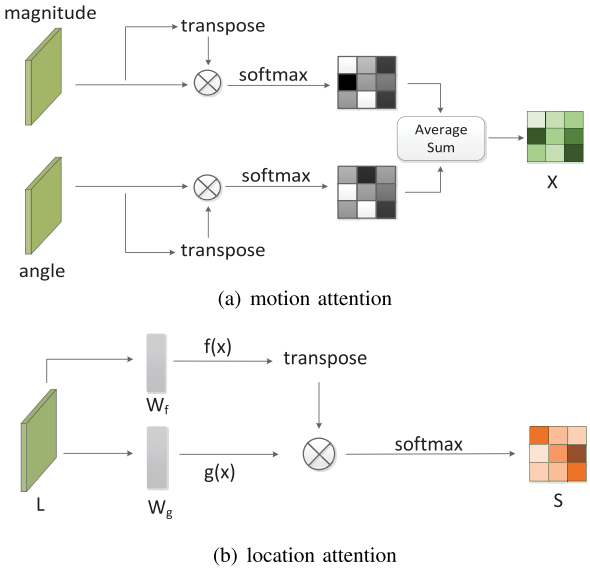
A. Problem Statement

Let V be a given test video, $\mathrm{V} = \{v_1, v_2, \ldots, v_n\}$, and v_i represents one of the consecutive frames in the video. Our purpose is to detect whether there are anomalies in $v_i \in \mathrm{V}$.

There exist interactions between the objects and pedestrians, and we aim to train an anomaly detection model containing the influence-aware attention module, which can extract such information by motion attention and location attention as $\mathbf{X} \in \mathbb{R}^{N\times N}$ and $\mathbf{S} \in \mathbb{R}^{N\times N}$. Using the learned normal characteristics, our approach outputs the generated frame v_i' from the corresponding input v_i, which is in the idealized normal state. For the normal behavior input, the output will be restored closely, and for the abnormal behavior, the output and the input will be far apart. The difference of each pair of input and output is compared and calculated as the anomaly score s_n, which indicates the extent of the anomaly.

B. Influence-Aware Attention Module

1) Motion Attention: In the videos of pedestrians, motion has been an essential feature for video understanding. Due to occlusions, illumination and the cluttered background, features of motion are challengeable to acquire. To promote the ability to recognize the pattern of normal behaviors, we build the motion attention module to explicitly model interdependencies between crowd movements.



(a) motion attention

(b) location attention

Fig. 3. The details of the influence-aware attention module.

The structure of the motion attention module is illustrated in Fig. 3(a). The optical flow field is calculated as $\mathbf{F} \in \mathbb{R}^{2\times C\times H\times W}$ on pixel level and the velocity of each region can be obtained. The velocity contains two attributes, and we separate it as two feature maps, the magnitude map $\mathbf{M} \in \mathbb{R}^{C\times H\times W}$ and the angle map $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$. Specifically, we reshape $\mathbf{M}$, $\mathbf{A}$ to $\mathbb{R}^{C\times N}$, where $N = H \times W$, and then separately perform a matrix multiplication between them and the transpose of themselves as in equation 1, 2. Finally, we apply a softmax layer and an average sum operation to obtain the motion attention map $\mathbf{X} \in \mathbb{R}^{N\times N}$, which is composed of the value of α of each pixel in a image.

$$\alpha 1_{j,i} = \frac{\exp(m(x_i)^T m(x_j))}{\sum_{i=1}^{N} \exp(m(x_i)^T m(x_j))},$$
$$\alpha 2_{j,i} = \frac{\exp(a(x_i)^T a(x_j))}{\sum_{i=1}^{N} \exp(a(x_i)^T a(x_j))}, \tag{1}$$

where $m(x)$ and $a(x)$ denote the magnitude feature and the angle feature, which are obtained from the calculation of optical flow. $\alpha 1$, $\alpha 2$ indicate the attention values of each pixel in the frame after performing the multiplication and the softmax operations.

2) Location Attention: As depicted in Fig. 3(b), $\mathbf{L} \in \mathbb{R}^{C\times H\times W}$ is the feature processed in the first half of hidden layers, and as the input of the influence-aware attention module, it is transformed into two feature spaces f, g after a convolution layer, where $f(x) = W_f x$, $g(x) = W_g x$. The location attention map aims to represent the influence of each pixel in appearance as a value. To calculate the corresponding weights, s_{ij} is introduced in equation 3, which is the matrix multiplication of $f(x)$ and $g(x)$. After the softmax calculation, $\beta_{j,i}$ indicates the extent to which the model attends to the

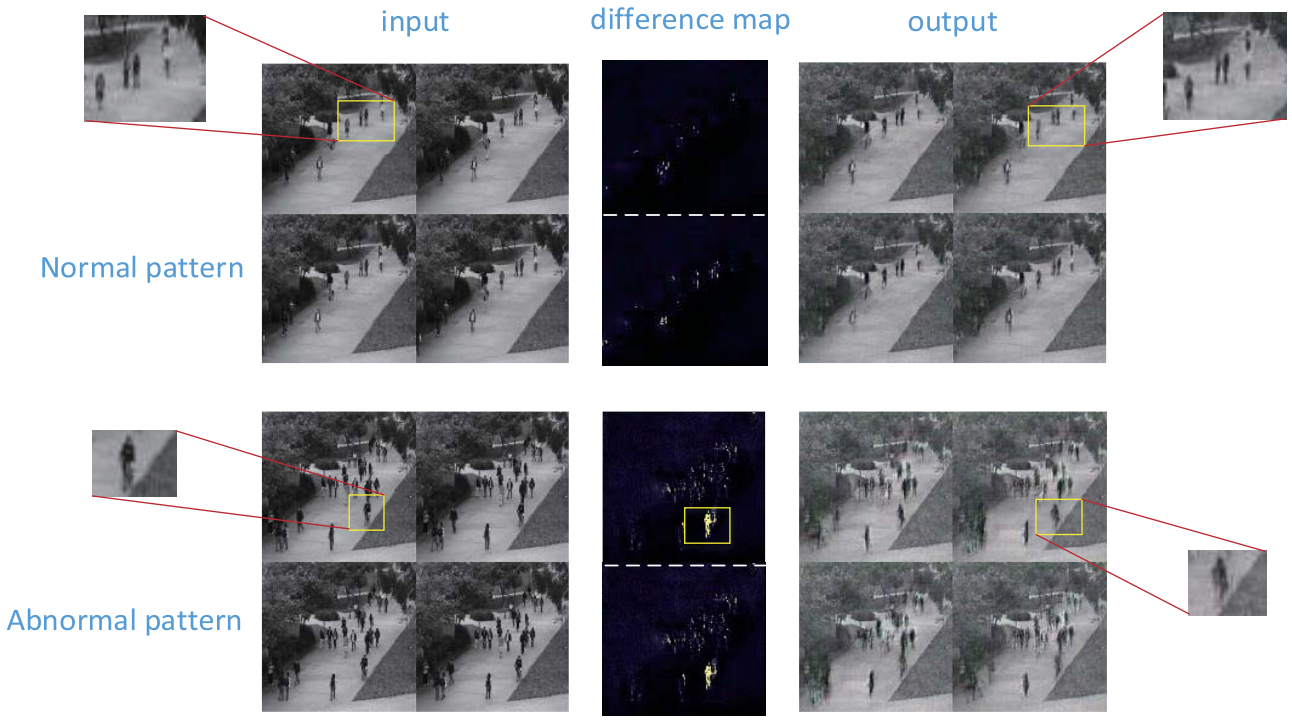


Fig. 4. The left image column is the input, and the right image column is the output of GII, which are of the normal pattern and the abnormal pattern respectively. The middle column is the difference map between input and output. We can see that compared with the input, the output of the normal pattern is clear and similar to the input, but the output of the abnormal pattern is blurred, especially in the abnormal regions of the frames where the cycling appears.

i^{th} region when synthesizing the j^{th} region.

$$s_{ij} = f(x_i)^T g(x_j), \tag{2}$$

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}. \tag{3}$$

In the above equations, W_g, W_f are the learned weight matrices, which are implemented as 1×1 convolutions.

C. Generator

1) Train-Generator I: GANs are adopted here to learn the normal pattern features. GI is trained with the discriminator in an adversarial way with the input of normal frames and the random noise z. The influence-aware attention module is added in the middle of the hidden layers of GI. After the first half hidden layers of processing, the processed data is sent into the influence-aware attention module. Armed with the influence-aware attention module, GI reconstructs each pixel of the input frame, taking into account the interactions with other pixels. The principle of GI is that the distinctive features can restore the normal behaviors well, meanwhile, the model will be perceptive to the occurrence of anomalies.

2) Test-Generator II : GI has been trained to generate images containing normal distributions. Based on the model of GI, the purpose of GII is to output the idealized normal data of the test data which are in normal or abnormal pattern. GI has mapped $z \rightarrow x'$, and to map $x \rightarrow z$, GII adds the residual square loss L_{RS} to update the noise z:

$$L_{RS}(z) = \sum |x - G(z)|^2. \tag{4}$$

The residual square loss measures the dissimilarity of the appearance between input test data and generated data of GII. The reflection of $x \rightarrow z \rightarrow x'$ has been established in the effect of the residual square loss. As shown in Fig. 4, if the test data is normal, the output data will be reconstructed well, and if the test data is abnormal, the output data will be blurred and far from the input data. The more significant difference between input and output exists, the more abnormal event happens.

D. Discriminator

The effect of the discriminator here is to lead the GI with better modeling performance. DRAGAN [36] is applied to the discriminator to stabilize the training process. Different from the general discriminator loss, DRAGAN has a gradient penalty scheme:

$$GP = \lambda \cdot E_{x \sim P_{real}, \delta \sim N_d(0,cI)}[||\nabla_x D_\theta(x+\delta)|| - k]^2. \tag{5}$$

In the equation 6, x is a real point, and $x + \delta$ is the noise. k is the hyperparameter for DRAGAN which depends on the corresponding architecture. In our experiments, k is chosen as 1. According to the penalty scheme, the discriminator assigns different probabilities of being real to training data and noisy samples, which constrains the norm of discriminator's gradients around real points to be steady and therefore can stabilize the training process. GI and the discriminator are simultaneously optimized through the following minimax

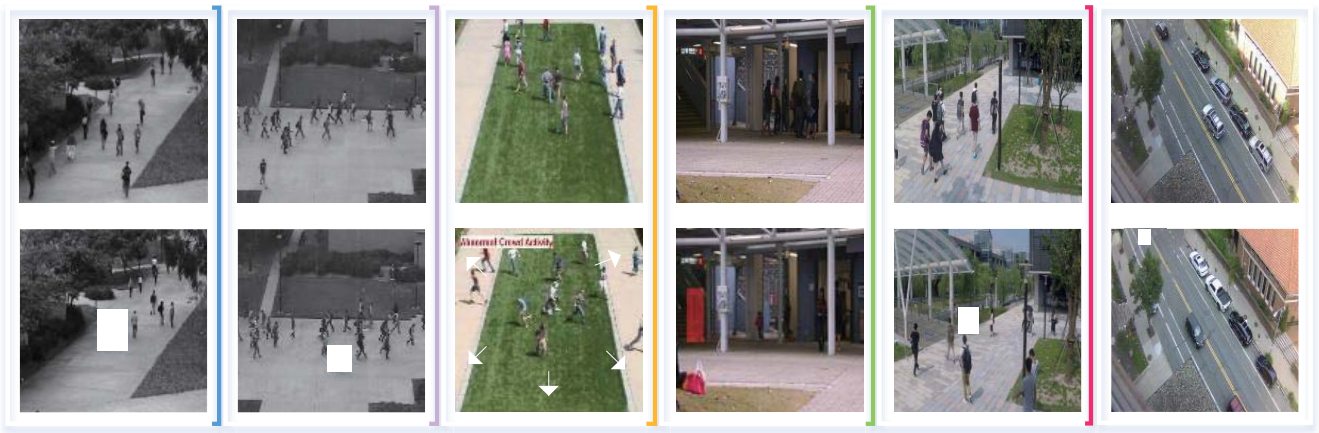


Fig. 5. The samples of different scenes in the UCSD Ped1, UCSD Ped2, UMN, Avenue, ShanghaiTech and Street Scene datasets. The anomalies are denoted by the red boxes or the red arrows.

value function $V(D, G)$:

$$\min_G \max_D V(D, G) = \sum_{i=1}^{3} V_i, \tag{6}$$

where

$$\begin{aligned} V_1 &= E_{x \sim P_{data}}[log D(x)], \\ V_2 &= E_{z \sim p_z}[\log(1 - D(G(z)))], \\ V_3 &= \lambda \cdot E_{x \sim P_{data}, \delta \sim N_d(0, cI)}[||\nabla_x D_\theta(x + \delta)|| - k]^2. \end{aligned} \tag{7}$$

After the adversarial training, the ability of GI to model the normal pattern improves, and on account of the influence-aware attention module, interactions can be retained in the model. With the shared weights of GI, GII will output the images with normal distributions.

E. Influence-Aware Attention Networks for Anomaly Detection

In our work, the influence-aware attention module acquires detailed interactive information from appearance and motion. The normal frames are input to extract features for the lack of abnormal samples. In the training phase, consecutive frames and the noise are fed into GI and the discriminator. GI has learned the normal patterns in an adversarial manner. In the testing phase, test frames and the noise are fed into GII to reconstruct the idealized normal frames. The anomalies are detected by comparing the reconstructed images with the input images. The metric of mean-square error (MSE) is introduced here to evaluate the degree of reconstruction, which indicates the degree of anomaly.

$$MSE(x') = E(x' - x)^2, \tag{8}$$

where x' is the output of GII, and x is the corresponding input. The dissimilarity determines whether the video frame is classified as an anomalous state.

IV. Experiments

In this section, we evaluate our proposed method on six public anomaly detection datasets, including the UCSD Ped1 and Ped2 datasets [37], the UMN dataset, the CUHK Avenue dataset [38], the ShanghaiTech Campus dataset [39] and the Street Scene dataset [40]. In the experiments our approach shows its effectiveness and robustness. A comparison between the count of abnormal frames detected and the ground truth is also provided, and an analysis of the result is given. The curve of anomaly score of samples is presented in this section to visually demonstrate the practical application of our method. We further validate the benefit of the influence-aware attention module on Ped1 dataset.

A. Datasets

Here we briefly introduce the datasets used in our experiments. The UCSD Ped1 and UCSD Ped2 datasets use fixed cameras to collect pedestrians on the sidewalk at a bird's eye angle. The density of pedestrians varies from sparse to crowded. The UMN dataset collects the abnormal behaviors of the video artificially arranged, which identifies the behaviors of the whole crowd. The videos of CUHK Avenue dataset are captured in CUHK campus avenue, which mainly detect the actions of pedestrians. The ShanghaiTech Campus dataset contains videos with complex light conditions and camera angles, and the anomalies caused by sudden motion such as chasing and brawling are also introduced. The videos of Street Scene dataset are collected on a scene of a two-lane street at various time during two consecutive summers, which contain changing shadows and moving background.

In the experiments, the definition of anomalies is same as that of the datasets, which is both about pedestrians and vehicles. Some samples of normal and abnormal states in these datasets are shown in Fig. 5 and we denote the anomalies with red boxes and red arrows, which also indicates that different types of anomalies are tested in our experiments.

TABLE I
COMPARISON OF AREA UNDER ROC CURVE (AUC) OF DIFFERENT METHODS ON SIX DATASETS

	UCSD Ped1	UCSD Ped2	UMN	CUHK Avenue	ShanghaiTech	Street Scene
MPPCA [41]	66.8%	69.3%	-	-	-	-
SF [3]	67.5%	55.6%	94.9%	-	-	-
MPPCA+SF [37]	74.2%	61.3%	-	-	-	-
Sparse Reconstruction [42]	86.0%	-	97.0%	-	-	-
ConvAE [43]	81.0%	90.0%	-	70.2%	-	-
ConvLSTM-AE [44]	81.5%	91.1%	-	77.0%	-	-
Motion Influence Map [17]	61.9%	77.3%	56.9%	-	-	-
Unmasking [45]	68.4%	82.2%	-	80.2%	-	-
Chong *et al.* [46]	89.9%	87.4%	-	80.3%	-	65.4%
AMDN [47]	92.1%	90.8%	-	-	-	-
GrowingGas [48]	93.8%	94.1%	-	-	-	-
Stacked RNN [39]	-	92.2%	-	81.7%	68.0%	-
Frame-Pred [21]	83.1%	95.4%	-	84.9%	72.8%	-
Plug and Play CNN [49]	**95.7**%	88.4%	-	-	-	-
Morais *et al.* [1]	-	-	-	-	73.4%	-
Nguyen *et al.* [23]	-	96.2%	-	-	-	-
Ionescu *et al.* [50]	-	97.8%	**99.6**%	**90.4**%	-	-
Vu *et al.* [22]	-	**99.2**%	-	71.5%	-	-
Deep Ordinal Regression [51]	71.7%	83.2%	97.4%	-	-	-
Ramachandra *et al.* [40]	77.3%	88.3%	-	72.0%	-	-
Ramachandra *et al.* [52]	86.0%	94.0%	-	87.2%	-	-
Georgescu *et al.* [53]	-	92.4%	-	86.9%	**83.5**%	-
Our method	94.2%	92.9%	98.8%	80.5%	80.3%	**73.0**%

B. Evaluation Metric

In the previous work [37], [38], the Receiver Operation Characteristic (ROC) is a popular evaluation metric, which is calculated by gradually changing the threshold of regular scores. The Area Under Curve (AUC) is calculated as a measurement for performance evaluation. Performance can also be summarized by the equal error rate (EER), which corresponds to the equal probability of misclassifying a positive or negative sample. These metrics are computed by the true-positive rate (TPR) and the false-positive rate (FPR):

$$\text{TPR} = \frac{\#\text{of true} - \text{positive frames}}{\#\text{of positive frames}}, \tag{9}$$

and

$$\text{FPR} = \frac{\#\text{of false} - \text{positive frames}}{\#\text{of negative frames}}. \tag{10}$$

The ROC curve is drawn using pairs of TPR and FPR, and EER is the ratio of misclassification at FPR = 1-TPR. Higher AUC and lower EER indicate better anomaly detection performance. In this paper, we leverage frame-level AUC and EER for performance evaluation [39].

C. Performance Comparison With Existing Methods

In this part, we compare our method with different handcraft feature-based methods and latest deep learning-based methods. The AUC of different methods on six datasets is listed in Table I. We can see that our method achieves competitive performance. The EERs of different methods are compared on UCSD Ped1 dataset in Fig. 6. The EER of the proposed method is lower than that of the second place by 10.0%. The comparison of these results reveals that our algorithm is effective.

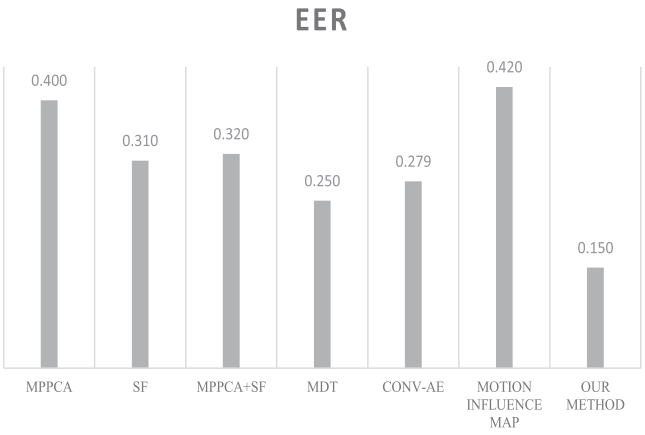


Fig. 6. Equal Error Rates for anomaly detection on UCSD Ped1 dataset of different methods.

In Table II, we provide the event count comparison on UCSD Ped1, Ped2 and UMN datasets. To demonstrate the detection and generalization ability of our method, the results of different scenes are listed separately. It can be seen that the proposed method works well in these scenes.

However, from the results of Table II, we can find that our method does not perform as well as other datasets on UCSD Ped2. It can be analyzed mainly from two aspects: 1) Most anomalies in Ped2 are about bicycles and skateboards,

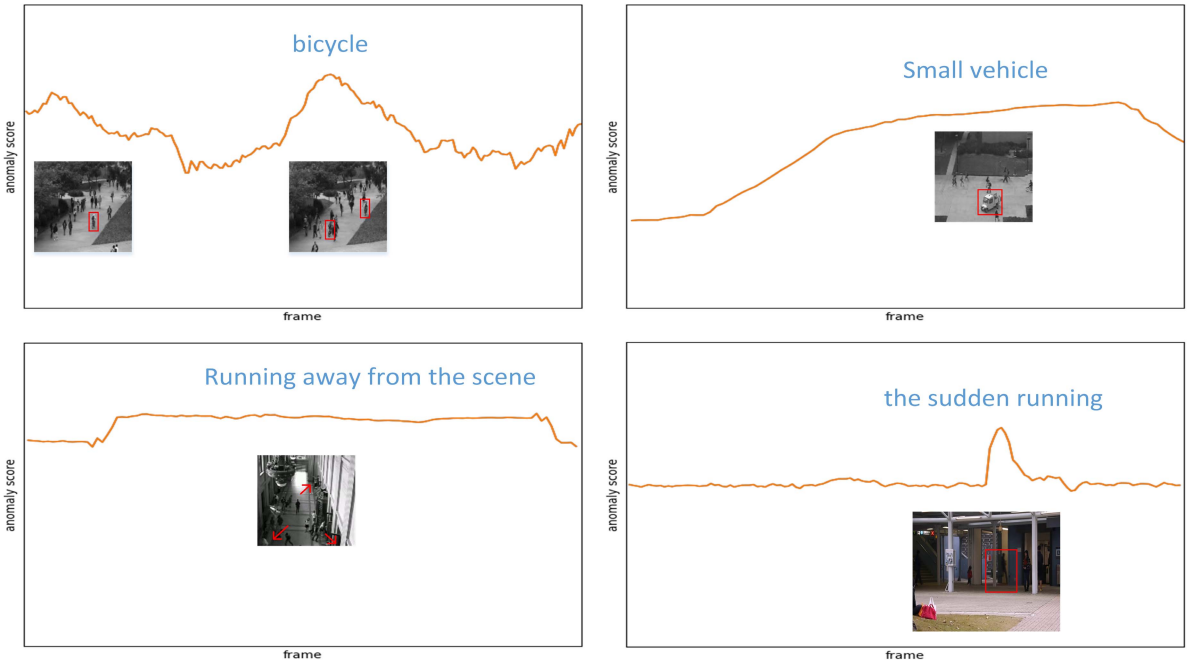


Fig. 7. The curves of anomaly scores on samples of UCSD Ped1, UCSD Ped2, UMN and Avenue datasets.

TABLE II
ANOMALOUS FRAMES AND ALARM COUNT DETECTED ON UCSD AND UMN DATASETS

	UCSD		UMN		
	Ped1	Ped2	Scene1	Scene2	Scene3
Our method	3839	1390	197	839	89
Groundtruth	4044	1648	200	843	90
Rate	94.9%	84.3%	98.5%	99.5%	98.9%

which appear in a small size in the whole surveillance footage; 2) Compared with the videos of Ped1, the vehicles including bicycles and skateboards move much more slowly and farther away from people in Ped2, therefore they exert less influence on the crowd. After analysis we find that a slow and small vehicle far away from people is a little harder to be detected as anomalous objects, but less likely to cause collisions.

D. Curve of Anomaly Score

The anomalies in UCSD Ped1 and Ped2 datasets are related to the vehicles especially on their appearance, while the anomalies in UMN dataset are about the crowd movement, which reflects on the motion information. Different anomalies are related to appearance and motion to different extent, and to check the ability of generalization of our algorithm, we draw the cures of anomaly scores in several scenes. In Fig. 7, the output anomaly scores of the proposed framework on samples of UCSD Ped1, Ped2, UMN and Avenue datasets are illustrated respectively. Our method detects anomalies correctly in these cases including in crowded scenes. Almost all of the anomalies show upward curves which indicate larger anomaly scores. Large anomaly scores will persist until the abnormal event passes. It can be seen that different kinds of events appear in the different form of upward curves. For the anomaly of the bicycle, the curve goes up with many zigzags, which corresponds to the scene that a bicycle is flexible when moving, and easy to mix in the crowd and cause uncertain anomalies. For the anomaly of small vehicles, the curve rises gradually, which is smoother than that of the bicycle, indicating that it is easier to be perceived because of its large size. As for the crowd running away from the scene, this kind of anomaly occurs quickly and continuously, thus the curve remains high and horizontal. For the sudden running, the curve shows a sharp rise and fall. In the score curves, we can find that different anomalies correspond to various curves, which depends on the characteristics of the abnormal events. In feature work, it may be feasible to predict the type of abnormal events by the changes in the score curve.

E. Ablation Study

Ablation studies are conducted to verify the contribution of our proposed sub-modules. We compared the ROC curves of the feature-branch ablation experiments on UCSD Ped1 dataset in Fig. 8. The influence-aware attention module is composed of the motion attention and the location attention, and to validate the benefits of them, we train models in an ordinary GAN way. First, the comparisons between GAN and two attention modules show that the separate learning of spatial and temporal feature boosts the robustness of feature representation. Second, the performance of complete influence-aware method is more advanced than either of the single branch baselines, which indicates that our algorithm learns the complementary representation information from the co-attention.

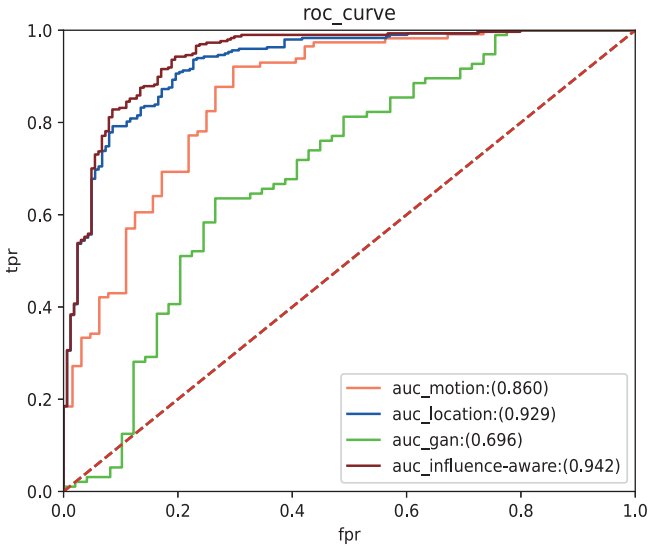


Fig. 8. Performance of variants of proposed method on UCSD Ped1 dataset. GAN refers to the network which is without the motion attention or the location attention.

The motion attention module extracts information from time series, and it is sensitive to the changes in movement. The location attention module extracts the appearance information, and it can obtain the anomalies of inappropriate behaviors or objects in a scene. Things tend to happen coherently in time series and with multiple connections on visual scenes, so it can be concluded that our method achieves accurate anomaly detection.

V. Conclusion

In this paper, we develop a novel framework with the influence-aware attention module for anomaly detection in surveillance videos, which adaptively integrates motion and appearance features. Specifically, the motion attention and the location attention are introduced to capture global dependencies in spatial and temporal dimensions respectively. The ablation experiments show that the influence-aware attention module obtains contextual information effectively and gives precise detection results. Additionally, we provide the anomaly score curves and discuss the characteristics of different anomalies. In future work, we will explore the following directions:

(1) In the experiments, it is found that the score curves present corresponding characteristics under different kinds of abnormal conditions. A step can be added into the future work to learn the rising law of the curve or the changing law of scores to obtain the types of anomalies.

(2) The real-time performance of our method still needs to be improved. In terms of being able to detect anomalies in time, including improving the detection speed, early warnings of potentially anomalous events should also be taken into consideration rather than only the fact detection of real-time anomalous events. Our future research will focus on speeding up the operations and conducting the precognition through features that may lead up to an anomalous event.

References

[1] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11996–12004.

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

[3] 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.

[4] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.

[5] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 673–682, Sep. 2021.

[6] B. Ramachandra, M. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 25, 2020, doi: 10.1109/TPAMI.2020.3040591.

[7] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 22, 2021, doi: 10.1109/TPAMI.2021.3053577.

[8] G. Wu *et al.*, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1993–2007, Apr. 2019.

[9] M. I. Georgescu, R. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "A background-agnostic framework with adversarial training for abnormal event detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 21, 2021, doi: 10.1109/TPAMI.2021.3074805.

[10] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Integrating part-object relationship and contrast for camouflaged object detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5154–5166, 2021.

[11] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 13588–13597.

[12] P. Ren *et al.*, "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–34, May 2022.

[13] B. Ramachandra, M. Jones, and R. R. Vatsavai, "Perceptual metric learning for video anomaly detection," *Mach. Vis. Appl.*, vol. 32, no. 3, pp. 1–17, May 2021.

[14] M. Gong, H. Zeng, Y. Xie, H. Li, and Z. Tang, "Local distinguishability aggrandizing network for human anomaly detection," *Neural Netw.*, vol. 122, pp. 364–373, Feb. 2020.

[15] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10539–10547.

[16] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2626–2634.

[17] D.-G. Lee, H.-I. Suk, S.-K. Park, and S.-W. Lee, "Motion influence map for unusual human activity detection and localization in crowded scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1612–1623, Oct. 2015.

[18] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7251–7259.

[19] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Adaptive unsupervised feature selection with structure regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 944–956, Apr. 2017.

[20] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[21] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[22] H. Vu, T. D. Nguyen, T. Le, W. Luo, and D. Phung, "Robust anomaly detection in videos using multilevel representations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33. Palo Alto, CA, USA: AAAI Press, 2019, pp. 5216–5223.

[23] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.
[24] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.
[25] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.
[26] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 622–637.
[27] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.
[28] J. Li, X. Liu, Z. Zong, W. Zhao, M. Zhang, and J. Song, "Graph attention based proposal 3D ConvNets for action detection," in *Proc. AAAI*, 2020, pp. 4626–4633.
[29] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio–temporal visual attention on skeleton image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2405–2415, Aug. 2019.
[30] K. Song, H. Yang, and Z. Yin, "Multi-scale attention deep neural network for fast accurate object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2972–2985, Oct. 2019.
[31] J. Wang, W. Wang, and W. Gao, "Fast and accurate action detection in videos with motion-centric attention model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 117–130, Jan. 2020.
[32] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2017, pp. 5998–6008.
[33] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, "Attention-driven loss for anomaly detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4639–4647, Dec. 2020.
[34] L. Zheng, Z. Li, J. Li, Z. Li, and J. Gao, "AddGraph: Anomaly detection in dynamic graph using attention-based temporal GCN," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4419–4425.
[35] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 146–157.
[36] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*.
[37] 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.
[38] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.
[39] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.
[40] B. Ramachandra and M. J. Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2569–2578.
[41] 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.
[42] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3449–3456.
[43] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.
[44] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.
[45] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2895–2903.
[46] 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.
[47] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2017.
[48] Q. Sun, H. Liu, and T. Harada, "Online growing neural gas for anomaly detection in changing surveillance scenes," *Pattern Recognit.*, vol. 64, pp. 187–201, Apr. 2017.
[49] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1689–1698.
[50] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7842–7851.
[51] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12173–12182.
[52] B. Ramachandra, M. J. Jones, and R. Raju Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2598–2607.
[53] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12742–12752.

Sijia Zhang was born in 1996. She is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an, China. Her research interests include deep learning and image processing.

Maoguo Gong (Senior Member, IEEE) received the B.S. degree (Hons.) in electronic engineering and the Ph.D. degree in electronic science and technology from Xidian University, Xi'an, China, in 2003 and 2009, respectively. Since 2006, he has been a Teacher with Xidian University. He was promoted as an Associate Professor in 2008 and a Full Professor in 2010, both with exceptive admission. His research interests are in the area of computational intelligence with applications to optimization, learning, data mining, and image understanding. He received the prestigious National Program for the support of Top-Notch Young Professionals from the Central Organization Department of China, the Excellent Young Scientist Foundation from the National Natural Science Foundation of China, and the New Century Excellent Talent in University from the Ministry of Education of China. He is also an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Yu Xie was born in 1993. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an, China. His research interests include deep learning, complex network analysis, and privacy preserving.

A. K. Qin (Senior Member, IEEE) received the B.Eng. degree from Southeast University, Nanjing, China, in 2001, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2007. From 2007 to 2017, he was with the University of Waterloo, Waterloo, ON, Canada; INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin, France; and RMIT University, Melbourne, VIC, Australia. He joined the Swinburne University of Technology, Hawthorn, VIC, Australia, in 2017, where he is currently a Professor. He is also the Director of the Swinburne Intelligent Data Analytics Laboratory, the Deputy Director of the Swinburne Space Technology and Industry Institute, and the Program Lead of the Swinburne Data Science Research Institute. His major research interests include machine learning, evolutionary computation, computer vision, remote sensing, services computing, and pervasive computing.

Hao Li (Member, IEEE) received the B.S. degree in electronic engineering and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2013 and 2018, respectively. He is currently an Associate Professor with the School of Electronic Engineering, Xidian University. His research interests include computational intelligence and machine learning.

Yuan Gao was born in 1996. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an, China. His research interests include knowledge discovery, privacy preserving, and user portrait.

Yew-Soon Ong (Fellow, IEEE) received the Ph.D. degree in artificial intelligence in complex design from the University of Southampton, U.K., in 2003. He is currently the President's Chair Professor in computer science with Nanyang Technological University (NTU), and holds the position of the Chief Artificial Intelligence Scientist at the Agency for Science, Technology and Research, Singapore. At NTU, he serves as the Co-Director of the Singtel-NTU Cognitive and Artificial Intelligence Joint Laboratory. His research interest is in artificial and computational intelligence. He has received several IEEE outstanding paper awards and was listed as a Thomson Reuters Highly Cited Researcher and among the World's Most Influential Scientific Minds. He is the founding Editor-in-Chief of the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, and an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE.