Computer Vision and Image Understanding 219 (2022) 103416

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Detecting abnormality with separated foreground and background: Mutual Generative Adversarial Networks for video abnormal event detection

Zhi Zhang[a], Sheng-hua Zhong[a,*], Ahmed Fares[a], Yan Liu[b]

[a] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China*
[b] *Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China*

ARTICLE INFO

Keywords:
Foreground–background separation mutual generative adversarial network
Video-level abnormal event detection

ABSTRACT

As one of the most important tasks in intelligent video analysis, video abnormal event detection has been extensively studied. Prior arts have made a great process in designing frameworks to capture spatio-temporal features of video frames. However, video frames usually consist of various objects. It is challenging to grasp the nuances of anomalies against noisy backgrounds. To tackle the bottleneck, we propose a novel Foreground–Background Separation Mutual Generative Adversarial Network (FSM-GAN) framework. The FSM-GAN permits the separation of video frames into the foreground and background. The separated foreground and background are utilized as the input of mutual generative adversarial networks, which transform raw-pixel images in optical-flow representations and vice versa. In the networks, the background is regarded as known conditions and the model focuses on learning the high-level spatio-temporal foreground features to represent the event with the given conditions during the mutual adversarial training. In the test stage, these high-level features instead of low-level visual primitives are utilized to measure the abnormality in the semantic level. Compared with state-of-the-art methods and other abnormal event detection approaches, the proposed framework demonstrates its effectiveness and reliability across various scenes and events.

1. Introduction

Video-level abnormal event detection refers to the identification of events that do not conform to expected behavior. It is a challenging problem due to the complexity of "anomaly" as well as the cluttered backgrounds, objects, and motions in the real-world video scenes (Zhao et al., 2017). In this task, we are given a set of normal training videos samples, and determine whether or not a test video contains an anomaly on these samples (Saligrama and Chen, 2012). In general, such anomalies can include unusual motion patterns and unusual objects on usual/unusual locations (Saligrama and Chen, 2012). In the past couple of years, the anomaly detection task has drawn much attention as a core problem of video modeling, and related technologies have widely been used in public places, e.g., streets, squares, and shopping centers, etc., to increase public safety (Sultani et al., 2018). Recently, deep learning-based methods provide state-of-the-art results for the task of video abnormal event detection. Popular frameworks include auto-encoders (Xu et al., 2015; Hasan et al., 2016; Luo et al., 2017a; Hinami et al., 2017; Ionescu et al., 2019) and generative adversarial networks (Ravanbakhsh et al., 2017; Liu et al., 2018; Lee et al., 2018).

Though lots of efforts have been made, the problem is still open. It is considered that human operators are more robust to scene changes, precisely locate abnormal events, and work well even in cases where given scenes are different from those in the training set. We have launched a survey on the Amazon Mechanical Turk (MTurk) to study how human operators perform abnormal event detection (Section 2). The survey results support that the detection performance gap may be caused by the different detection processes between human operators and existing methods: (i) Human operators tend to focus on moving objects instead of the static background. (ii) Human operators measure anomaly by high-level features instead of the low-level visual primitives.

Based on the observations, each video frame is separated into the foreground and background in this paper. Dynamic objects are directly related to the events. Thus, these objects are considered as the foreground. Stationary objects or other surroundings are not involved in events but they often provide conditions for the happening events. Naturally, they are regarded as background. Then, generative adversarial networks are designed to learn normal foreground patterns under the condition of background on the training dataset. To model the motion and appearance of the foreground, two mutual generative adversarial networks are proposed to transform raw-pixel images in optical-flow representations and vice versa. For an abnormal scene in the test stage,

[*] Corresponding author.
E-mail addresses: zhangzhi2018@email.szu.edu.cn, cszhizhang@comp.polyu.edu.hk (Z. Zhang), csshzhong@szu.edu.cn (S.-h. Zhong), ahmed.fares@szu.edu.cn (A. Fares), csyliu@comp.polyu.edu.hk (Y. Liu).

https://doi.org/10.1016/j.cviu.2022.103416
Received 10 November 2020; Received in revised form 30 October 2021; Accepted 20 March 2022
Available online 29 March 2022

Table 1
Seven tasks on the process of abnormal event detection required to be finished by workers.

Index	Task
T1	Determine whether the given video contains abnormal behavior.
T2	Mark on the frame where the abnormal events occur.
T3	Select the basis of the detection results. (“moving objects” or “static background”)
T4	Select the description on the detection process. (“I first noticed an unusual event, and then decided that this video was abnormal.” or “I first realized that the scene was abnormal, and then I found the abnormal event in the scene.”)
T5	Determine whether the bounding boxes contain abnormal events. (anomalies with motion blur)
T6	Determine whether the bounding boxes contain abnormal events. (normal events with motion blur)
T7	Determine whether the bounding boxes contain abnormal events. (normal events with distortion)

since the training set does not contain any abnormal samples, the networks are supposed to produce distorted reconstructions. By measuring the distortion with extracted high-level features instead of low-level visual primitives, the semantic anomalies can be easily captured.

In summary, the contributions of this paper can be highlighted as: (i) The proposed Foreground–Background Separation Mutual Generative Adversarial Network (FSM-GAN) permits the separation of video frames into foreground and background. By regarding the background as known conditions, the proposed FSM-GAN is able to focus on the foreground to detect abnormal events. (ii) The high-level features are proposed to learn to represent the foreground events. In the test stage, the feature-based anomaly metrics can take the place of low-level visual primitives to measure abnormality, forcing the model to capture abnormal semantics. (iii) we carry out extensive experiments to demonstrate the good generalization ability of our proposed FSM-GAN across three benchmarking datasets.

2. Human feedback in abnormal event detection

To support the insight of our proposed methods qualitatively, we investigate on how human operators perform abnormal event detection and explore the different detection processes between human operators and existing methods that may cause detection performance gap. Here, a survey has been launched based on the MTurk crowdsourcing platform, in which workers must read the instructions of video abnormal event detection and then finish seven tasks, as shown in Table 1:

The qualification requirement for this task is the number of HITs approved should be greater than or equal to 30. We allowed each HIT to be completed by a unique worker, and 20 HITs were created. The monetary reward was based on an effective hourly wage of $1. In total, 20 HITs were created with $29. For the actual experiment, 17 workers were granted a qualification to access the HITs. Furthermore, all of the 17 HITS completed by those qualified workers were approved without rejections.

In these HITs, all workers could identify abnormal events and completed T1 and T2 correctly. It shows that these workers are able to complete the anomaly detection task, and their answers are valid. For T3, while asked to select the basis for the detection results, all workers chose moving objects instead of static backgrounds. It supports that (i) human operators tend to focus on moving objects instead of the entire scene or all the objects. Inspired by the conclusion, each video frame is separated into the foreground and background in this paper. Especially, dynamic objects are directly related to the events. These objects are considered as the foreground. Stationary objects or other surroundings are not involved in events. They are regarded as background. Regarding the background as known conditions, the proposed FSM-GAN can focus on the foreground to detect abnormal events.

About the tasks of T4, one worker thought that he first realized that the scene was abnormal, and then he found the abnormal event in the scene. The remaining 16 workers believed that they first noticed an unusual event, and then decided that this video was abnormal. It means that (ii) human operators commonly detect abnormalities and identifying the abnormal frame in a bottom-up fashion. Inspired by such observations, we propose to use max-pooling to aggregate local evidences to represent the abnormality of the entire video frame.

The last three tasks are related to low-level visual primitives. In T5, one worker was affected by motion blur and detected the abnormality as a normal event. The remaining 16 workers correctly detected it as abnormal. In T6, two workers were affected by motion blur and detected normal events as anomalies. The remaining 15 workers made the correct judgment. In T7, two workers were affected by distortion and detected the abnormality as a normal event, and the detection results of the remaining 15 workers are correct. It shows that (iii) human operators have better robustness to pixel-level distortion. They measure anomaly by high-level features instead of the low-level visual primitives. Thus, in this paper, we propose to learn high-level features to represent the foreground events. In the test stage, the feature-based anomaly metrics can take the place of low-level visual primitives to measure abnormality, forcing the model to capture abnormal semantics.

3. Related work

In recent years, abnormal event detection in the video has gained attention from computer vision researchers and artificial intelligence application developers. Boosted by the recent success of neural networks (Jiang et al., 2011; Calderara et al., 2011; Dan et al., 2017; Fan et al., 2020), deep learning-based methods have outperformed hand-crafted feature engineering (Zaharescu and Wildes, 2010; Rota et al., 2012; Roshtkhari and Levine, 2013; Wiliem et al., 2012; Jeong et al., 2011; Zhu et al., 2014) in the field of abnormal event detection, as well as related tasks, such as crowd counting (Wan et al., 2019), human pose estimation (Zhao et al., 2019; Sengupta et al., 2020), object tracking (Luo et al., 2018), and action recognition (Hong et al., 2019; Idrees et al., 2017).

A popular approach adopted by researchers is to learn a deep neural network in an auto-encoder fashion, and they enforce it to reconstruct normal events with small reconstruction errors. Then, abnormal events would correspond to larger reconstruction errors. There are also some studies utilizing such learned auto-encoders to extract the features of events and perform outlier detection. Xu et al. (2015) first designed a multi-layer auto-encoder to reconstruct scenes. Hasan et al. (2016) further proposed two methods that are built upon the autoencoders. They first utilized the conventional spatio-temporal local features and learned a fully-connected autoencoder. Then, they built a fully convolutional autoencoder to learn both the local features and the classifiers in a single learning framework. Luo et al. further boosted the performance by the proposed Convolutional LSTMs auto-encoder (ConvLSTM-AE) (Luo et al., 2017a) to model normal appearance and motion patterns at the same time. All these approaches extracted features without explicitly taking into account the objects of interest.

To explicitly focus only on the objects that present in the scene, researchers proposed to train auto-encoder for object-level reconstruction. Hinami et al. (2017) used geodesic and object proposals to detect objects from scenes and then fine-tuned the classification branch of the

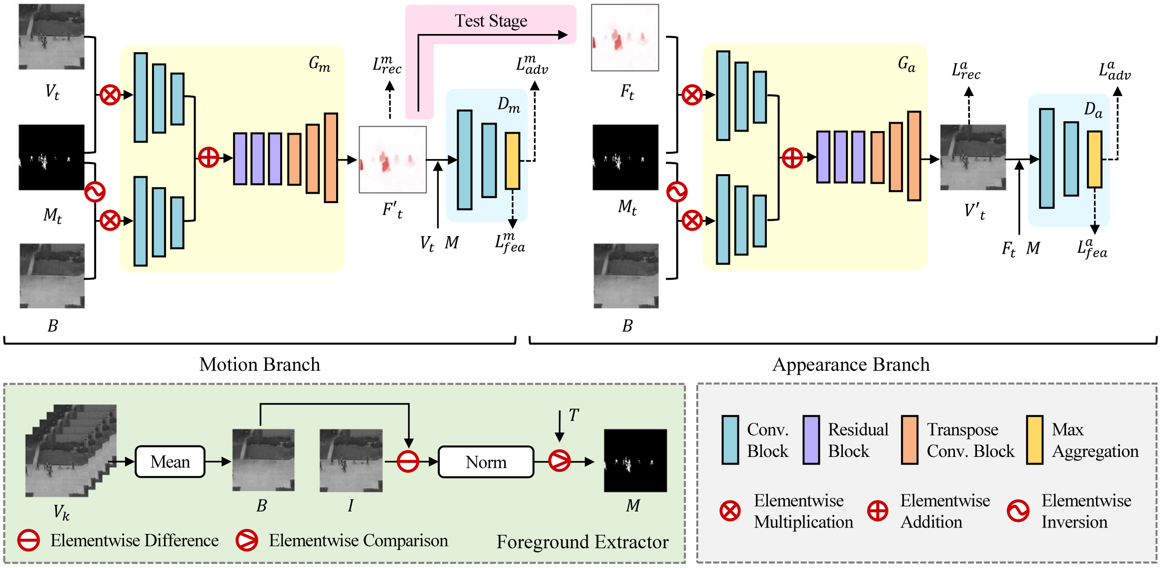


Fig. 1. Foreground–Background Separation Mutual Generative Adversarial Network (FSM-GAN).

Fast R-CNN model to exploit semantic information that is useful for detecting and recounting abnormal events. Differently, Ionescu et al. (2019) employed a single-shot detector based on Feature Pyramid Networks (FPN) and learned features of objects by training auto-encoders to reconstruct detected objects. Though these methods have opened the way for dividing objects and scenes, new problems have emerged. On the one hand, object proposals contain stationary objects not involved in events. These objects are paid unnecessary attention. On the other hand, only objects are modeled while the scenes are discarded. It leads to the fact that the location and the context of objects are missing. Given these observations, the question arises: how should we define objects of interest and how should we handle the separated objects and scenes.

Recently, generative adversarial networks (GAN) based methods are proposed to exploit the adversary game between generating and discriminating abnormal events to approximate the normal data distribution and train the final classifier. Ravanbakhsh et al. (2017, 2019) proposed to train GAN using normal frames and corresponding optical-flow images to obtain the appearance and motion representations. Liu et al. (2018) adopted U-Net as generator to predict a future frame. Lee et al. (2018) proposed a novel abnormal event detection method with spatio-temporal adversarial networks.

In this paper, we further improve GAN-based abnormal event detection methods and design the FSM-GAN. Regarding the background as a known condition to learn the foreground patterns, we are the very first to succeed in introducing the separation of foreground and background to GAN-based abnormal event detection.

4. Proposed method

4.1. Overall framework

In this paper, we propose a FSM-GAN to detect anomalies in videos. Fig. 1 describes the architecture of FSM-GAN, which contains three parts, i.e., the foreground extractor, the motion branch, and the appearance branch. The foreground extractor attempts to decouple foreground and background from given scenes producing the foreground mask M_t and background B. Then, the motion branch and the appearance branch are proposed based on the architecture of generative adversarial networks to reconstruct the optical flow F_t and video frame V_t corresponding to the tth frame under the guidance of the background, respectively. Each branch involves a generator and a discriminator. For normal scenes, the generator learns normal patterns from training datasets and reconstruct optical flows or video frames as much as possible. The discriminator learns to distinguish what is reconstructed from what is real. For an abnormal video frame in the test stage, since the training set does not contain any abnormal samples, the generator is supposed to generate distorted optical flows or video frames, which would be captured by the discriminator as the anomaly.

In the remainder of this section, we will detail the pipeline of the proposed method.

4.2. Pre-processing

Given a video clip, we first convert frames into gray-level images and resize images to a fixed size of 256 × 256. Then, data normalization technologies are utilized to divide pixel values by 255, subtract $mean = 0.5$ and divide by $std = 0.5$ per channel. Then, the normalized image tensors are provided as the input of FSM-GAN, denoted by V_t for each time step t. For the next step, DeepFlow (Weinzaepfel et al., 2013) is applied to the given frame V_t and the adjacent frame V_{t-3} at time step $(t-3)$ to estimate two-channel optical flow vectors. Due to the limited relationship between the movement direction of the behavior and its abnormality, we neglect the optical flow direction and focus on the magnitude of the optical flow vectors to simplify the modeling of events. Thus, we take the absolute value of the obtained optical flow as FSM-GAN's input, denoted by F_t.

4.3. Foreground extractor

As we described before, in video frames, only moving objects produce meaningful events. These objects are of interest for abnormal event detection tasks. To enable accurate detection of moving objects, the first task is to extract a reliable background image, which has to be robust to environmental changes and sensitive to detect objects of interest.

Since the position of the camera is relatively fixed in the abnormal event detection tasks, a simple and easy way to solve these problems is to generate the background image as an average of frames within a video clip. In this paper, given a video clip, each frame is first converted to a gray-level image, and then the background of the given video can be approximated as follows:

$$B(i,j) = \frac{1}{n} \sum^{t \in (0,n]} V_t(i,j) \tag{1}$$

where $B(i,j)$ denotes the pixel of background at row i and column j, the corresponding pixel of the tth frame is denoted by $V_t(i,j)$. The length of the given video is n. Then, the difference between the background B and tth video frame can be obtained:

$$D_t(i,j) = \|V_t(i,j) - B(i,j)\|_1 \tag{2}$$

Due to the diverse background images and foreground objects, the range of $D_t(i,j)$ may vary widely. Thus, we normalize the $D_t(i,j)$ and obtained:

$$D'_t(i,j) = \frac{D_t(i,j) - min(D_t)}{max(D_t) - min(D_t)} \tag{3}$$

Here, $min(D_t)$ denotes the minimized value of pixel difference D_t and $max(D_t)$ denotes the maximized value of pixel difference D_t. Based on the normalized difference between the background B and tth video frame, the mask of the foreground in tth frame can be obtained:

$$M_t(i,j) = \begin{cases} 1, D'_t(i,j) > T \\ 0, D'_t(i,j) \le T \end{cases} \tag{4}$$

Here, $M_t(i,j)$ denotes the mask of foreground at row i and column j of tth video frame. T is a hyperparameter. In this way, the foreground parts and background parts of each video frame are marked separately.

4.4. Motion branch and appearance branch

In this subsection, we introduce the motion branch and the appearance branch in our framework, which are both designed based on generative adversarial networks to model the motion and appearance of the foreground under the guidance of background conditions.

As shown in Fig. 1, the generator of the motion branch G_m takes the tth video frame V_t, the corresponding foreground mask M_t and the background B of the video clip as input to reconstruct the optical flow F_t. For the first step, we utilize $V_t \circ M_t$ to extract foreground pixels from V_t and provide the results as the input of three convolutional blocks to extract low-resolution features from the foreground. Here, $\circ$ is the elementwise multiplication. Meanwhile, we apply elementwise inversion on M_t to obtain the mask of background $(1 - M_t)$, and $B \circ (1 - M_t)$ is provided as the input of another three convolutional blocks to extract low-resolution features of background. Then, the low-resolution features of the foreground and background are summed elementwisely to serve as the input of five residual blocks, followed by three transpose convolutional blocks to predict the optical flow. The reconstructed optical flow is denoted by F'_t.

The generator of the appearance branch G_a has the same architecture as G_m. The difference is that instead of the tth video frame V_t, the appearance branch takes the optical flow F_t, the corresponding foreground masks M_t and the background of the video clip B as input to reconstruct the tth video frame V_t. Here, $F_t \circ M_t$ is provided as the input to extract low-resolution features of foreground, while background features are extracted using $B \circ (1 - M_t)$.

To enable the adversarial training against generators, the discriminator of the motion branch D_m learns to distinguish reconstructed optical flows F'_t from real ones F_t. The given optical flow is first concatenated with the corresponding video frame V_t at index t and the mask of the foreground M_t. To be specific, D_m consists of a feature extractor $\phi_m(.)$ and a classifier. Here, the $\phi_m(.)$ is formed by several convolutional blocks and a max-pooling layer. After extract low-resolution features from provided inputs, a max-pooling is applied to the obtained feature maps to aggregate all local evidence for representing the abnormality of the entire video frame. The pooling results are utilized as the outputs of $\phi_m(.)$. For the next step, the $\phi_m(F'_t|V_t, M_t)$ or $\phi_m(F_t|V_t, M_t)$ is provided as the input of the classifier to predict the probability of the video frame being abnormal.

The discriminator of the appearance branch D_a has the same network architecture as D_m. The difference lies in that D_a learns to distinguish the reconstructed video frame V'_t from real ones V_t. The given video frame is concatenated with the corresponding optical flow F_t and the mask of the foreground M_t to form the input. Moreover, in the discriminator D_a, the feature extractor $\phi_a(.)$ is utilized to extract high-level appearance features from the foreground.

4.5. Training strategy

This section introduces an adversarial training strategy to adapt the adversary game between the generators and discriminators on separated foreground and background and learn distinguishable feature-level representation for abnormal events.

First, we improve the WGAN-GP (Gulrajani et al., 2017) for optimizing the generators and discriminators to reconstruct optical flows or video frames and distinguish the reconstructed samples under the guidance of the background. Take the motion branch as an example, the game between the generator G_m and the discriminator D_m can be operated by optimizing the adversarial loss L^m_{adv}:

$$L^m_{adv} = \mathop{\mathbb{E}}_{F_t \sim \mathbb{P}_r}[D_m(F_t|V_t, M_t)] - \mathop{\mathbb{E}}_{F'_t \sim \mathbb{P}_g}[D_m(F'_t|V_t, M_t)] + \mathcal{P} \tag{5}$$

where $\mathbb{P}_r$ is the real data distribution and $\mathbb{P}_g$ is the model distribution implicitly defined by $F'_t = G_m(V_t, M_t, B)$. $\mathcal{P}$ is the gradient penalty proposed by WGAN-GP (Gulrajani et al., 2017).

Then, to improve the reconstruction ability of generators, the generators are tasked to not only fool the discriminator but also to be near the ground-truth output by optimizing pixel-level errors. Previous works have found it beneficial to mix the GAN objective with L2 distance (Pathak et al., 2016). We also explore this option and optimize L2 distance between reconstructed optical flows and real ones:

$$L^m_{rec} = \frac{1}{N}\|F_t - F'_t\|_2 \tag{6}$$

where L^m_{rec} denotes the reconstruction loss of the motion branch, and N is the number of samples.

As we described before, it can be found that human operators tend to recognize high-level semantic anomaly instead of low-level visual primitives during the process of anomaly detection. Such a strategy is able to enhance the interpretability and stability of abnormal event detection. Inspired by these observations, we propose to permit an adversary game to generate realistic high-level features and distinguish the generation against high-level features extracted from real samples:

$$L^m_{fea} = \|\phi_m(F'_t|V_t, M_t) - \phi_m(F_t|V_t, M_t)\|_1 \tag{7}$$

With the adversarial feature loss formulated as Eq. (7), in the motion branch, the discriminator learns to maximize the difference of high-level features extracted from reconstructed optical flow and real ones, and the generator tries to fool the discriminator by minimizing the difference. In this way, the discriminator is able to extract distinguishable high-level features to represent the motion of the foreground. Meanwhile, existing work (Salimans et al., 2016) also demonstrates that training the generator to match the expected value of the features on an intermediate layer of the discriminator is effective in situations where regular GAN becomes unstable.

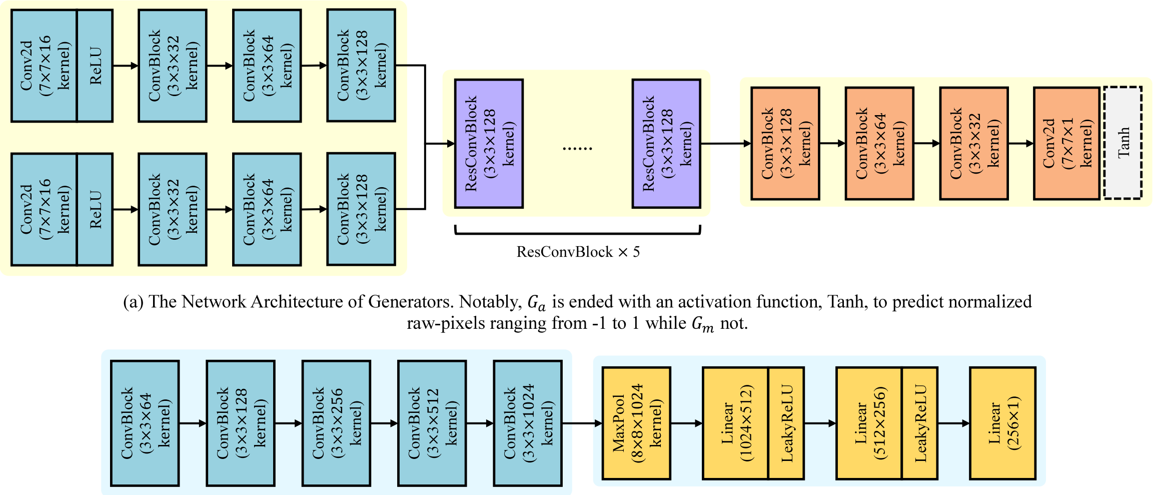
Similarly, the appearance branch is optimized by:

$$\begin{aligned} L^a_{adv} &= \mathop{\mathbb{E}}_{V_t \sim \mathbb{P}_r}[D_a(V_t|F_t, M_t)] - \mathop{\mathbb{E}}_{V'_t \sim \mathbb{P}_g}[D_a(V'_t|F_t, M_t)] + \mathcal{P} \\ L^a_{rec} &= \frac{1}{N}\|V_t - V'_t\|_2 \\ L^a_{fea} &= \|\phi_a(V'_t|F_t, M_t) - \phi_a(V_t|F_t, M_t)\|_1 \end{aligned} \tag{8}$$

where $\mathbb{P}_r$ is the data distribution and $\mathbb{P}_g$ is the model distribution implicitly defined by $V'_t = G_a(F_t, M_t, B)$. ϕ_a denotes the feature extractor part of discriminator.

4.6. Abnormality detection

In the test stage, a feature-based anomaly metric is applied to detect abnormal events based on these high-level features.


Conv2d (7×7×16 kernel)
ReLU
ConvBlock (3×3×32 kernel)
ConvBlock (3×3×64 kernel)
ConvBlock (3×3×128 kernel)
Conv2d (7×7×16 kernel)
ReLU
ConvBlock (3×3×32 kernel)
ConvBlock (3×3×64 kernel)
ConvBlock (3×3×128 kernel)
ResConvBlock (3×3×128 kernel)
......
ResConvBlock (3×3×128 kernel)
ResConvBlock × 5
ConvBlock (3×3×128 kernel)
ConvBlock (3×3×64 kernel)
ConvBlock (3×3×32 kernel)
Conv2d (7×7×1 kernel)
Tanh
ConvBlock (3×3×64 kernel)
ConvBlock (3×3×128 kernel)
ConvBlock (3×3×256 kernel)
ConvBlock (3×3×512 kernel)
ConvBlock (3×3×1024 kernel)
MaxPool (8×8×1024 kernel)
Linear (1024×512)
LeakyReLU
Linear (512×256)
LeakyReLU
Linear (256×1)

(a) The Network Architecture of Generators. Notably, G_a is ended with an activation function, Tanh, to predict normalized raw-pixels ranging from -1 to 1 while G_m not.

(b) The Network Architecture of Discriminators.

Fig. 2. Network architectures of the Foreground–Background Separation Mutual Generative Adversarial Network (FSM-GAN).

For a given test video frame V_t, the foreground extractor is utilized to calculate the mask M_t and background B. Then, the generators of the motion branch and the appearance branch are applied to reconstruct the optical flow F_t and video frame V_t corresponding to the tth frame under the guidance of the background. Especially, instead of the ground truth optical flow F_t, the reconstructed optical flow F'_t by the motion branch is taken as the input of the appearance branch. Since the training set does not contain any abnormal samples, the generator will reconstruct a distorted optical flow or video frame. In contrast, benefited from the learned normal patterns, the foreground not containing anomalies can be reconstructed with slight distortion. The difference caused by reconstruction can be utilized to distinguish anomaly.

Following the survey result of T5, T6, and T7 in Table 1, we find human operators tend to recognize high-level semantic anomalies and are robust to the outlier of low-level visual primitives. However, directly constructing a perceptual metric to simulate human judgments is challenging, which needs to capture high-order image structure and understand the context in video frames. Fortunately, related studies (Zhang et al., 2018) show that networks trained to solve challenging visual prediction and modeling tasks end up learning a representation of the world that correlates well with perceptual judgments. Inspired by related work, we propose to utilize the high-level features extracted by ϕ_m and ϕ_a to identify the abnormality of frame t to provide an alternative of widely-used per-pixel measures, such as pixel-based anomaly measures.

In detail, the frame-level anomaly score E_t of tth frame can be formulated as:

$$\begin{aligned}E_t =&S(\|\phi_m(V'_t|F'_t, M_t) - \phi_m(V_t|F_t, M_t)\|_1)^{\frac{1}{2}}\\ &+S(\|\phi_a(V'_t|F'_t, M_t) - \phi_a(V_t|F_t, M_t)\|_1)^{\frac{1}{2}}\end{aligned} \tag{9}$$

where $F'_t = G(V_t, M_t, B)$ and $V'_t = G(F'_t, M_t, B)$. By measuring the distance of high-level features before and after reconstruction, E_t ignores outliers of low-level visual primitives, such as motion blur and noise, which should not be defined as abnormal events, and only focus on the anomaly in semantics. Here, following previous works (Ionescu et al., 2019), the min–max normalization is applied over video frames to normalize abnormality prediction based on ϕ_m and ϕ_a between 0.0 and 1.0:

$$S(e_t) = \frac{e_t - min(e)}{max(e) - min(e)} \tag{10}$$

where $min(e)$ denotes the minimized value of e_i at each frame i of the given video and $max(e)$ denotes the maximized value of e_i at each frame i of the given video. Inspired by Berthelot et al. (2019), we utilize the normalized predictions to the power of $\frac{1}{2}$ to sharpen the distribution's temperature and encourages the model to produce lower-entropy predictions. Finally, a Gaussian filter is applied to smooth the frame-level anomaly scores E_t.

4.7. Implementation details

In this subsection, we further detail the implementation of network architectures in the proposed FSM-GAN.

As shown in Fig. 2, the generator network G_m of the motion branch takes the video frame $V_t \in \mathbb{R}^{1\times h\times w}$, the foreground mask $M_t \in \mathbb{R}^{1\times h\times w}$ of the corresponding video frame, and the background $B \in \mathbb{R}^{1\times h\times w}$ of the video clip as input. Here, h denotes the height of the video frame, and w is the width of the video frame. First, $V_t \circ M_t$ is provided as the input of three convolutional blocks to extract low-resolution features from the foreground. Each convolutional block uses a stride-2 convolutional layer and a stride-1 convolutional layer to downsample. Meanwhile, we apply elementwise inversion on M_t to obtain the mask $(1 - M_t)$ of the background, and $V_t \circ (1 - M_t)$ is provided as the input of another three convolutional blocks with the same architecture to extract low-resolution features of the background. Then, the low-resolution features are summed elementwisely to serve as the input of five residual blocks. Each of the residual blocks contains two convolutional layers with residual connections. Finally, three transpose convolutional blocks are utilized to predict the optical flow $F'_t \in \mathbb{R}^{2\times h\times w}$. Other than the first and the last layers, which use 7 × 7 kernels, all convolutional layers use 3 × 3 kernels.

For the generator network G_a, the input is the optical flow $F_t \in \mathbb{R}^{2\times h\times w}$, the foreground mask M_t, and the background B of the video clip. Unlike G_m, G_a takes $F_t \circ M_t$ as the input of three convolutional blocks to extract low-resolution features from the motion of the foreground. Meanwhile, the mask $(1 - M_t)$ of the background is obtained by the elementwise inversion applied to M_t, and $F_t \circ (1 - M_t)$ is provided as the input of another three convolutional blocks to extract low-resolution motion features of background. Then, most of the network architecture of G_a is consistent with that of G_m except one difference: G_a is ended with an activation function, Tanh, to predict normalized raw pixels ranging from −1 to 1.

The discriminators D_m and D_a share the same network architecture too, taking the given two-dimensional optical flow F_t (or F'_t), video frame V_t (or V'_t) and the corresponding mask M_t as input. For the first step, the given inputs are first concatenated along the channel dimension. Then, they are provided as the input of five convolutional

Table 2
AUC (%) of different methods on the UCSD Ped2, Avenue, and ShanghaiTech datasets.

Method		Avenue	UCSD Ped2	ShanghaiTech
GAN-based	AED-GAN (Ravanbakhsh et al., 2017)	–	93.5	–
	STAN (Lee et al., 2018)	87.2	96.5	–
	FramePred (Liu et al., 2018)	84.9	95.4	72.8
	CCAED-GAN (Ravanbakhsh et al., 2019)	–	95.5	–
Other SOTA	Conv-AE (Hasan et al., 2016)	70.2	–	60.9
	ConvLSTM-AE (Luo et al., 2017a)	77	–	–
	Stacked-RNN (Luo et al., 2017b)	81.7	92.2	68
	Mem-AE (Gong et al., 2019)	83.3	94.1	71.2
	MNAD (Park et al., 2020)	**88.5**	97.0	70.5
Proposed	**FSM-GAN**	80.1	**98.1**	**73.5**

blocks to extract low-resolution feature maps. Each convolutional block uses a stride-2 convolutional layer and a stride-1 convolutional layer to down-sample, where all convolutional layers use 3 × 3 kernels. A global max-pooling is then applied to the obtained feature maps to aggregate all local evidences for representing the abnormality of the entire video frame. Finally, the obtained feature maps are utilized as the input of three linear layers, each of which is followed by LeakyReLU nonlinearities except for the output layer.

5. Experiments

5.1. Overall performance

In this section, we validate the proposed FSM-GAN for anomaly detection and conduct experiments on the three most commonly-used benchmark datasets: Avenue (Lu et al., 2013), UCSD Ped2 (Mahadevan et al., 2010), and ShanghaiTech (Luo et al., 2017b). Each dataset is composed of two subsets: the training set and the test set. Training videos contain only normal events, while test videos have both normal and abnormal events.

Avenue dataset is one of the most commonly used datasets and is usually considered a challenging dataset, which contains various types of abnormal events on avenues. In detail, the Avenue dataset consists of 16 training videos and 21 test ones. The resolution of each video frame is 360 × 640 pixels. Notably, Hinami et al. (2017) argue that there are five test videos (01, 02, 08, 09, and 10) containing static abnormal objects that are not correctly labeled. Thus, we follow recent state-of-the-art researchers to carry out experiments on a subset excluding the respective five videos.

UCSD dataset is another benchmarking dataset with low frame-resolution, which is composed of two parts: UCSD Ped1 and UCSD Ped2. UCSD Ped1 includes 34 training videos and 36 test ones, while UCSD Ped2 contains 16 training videos and 12 test videos. Generally, different methods are evaluated on these two parts separately. Notably, a single frame possibly contains multiple abnormal events for the UCSD dataset.

ShanghaiTech dataset (Luo et al., 2017b) is one of the largest datasets for abnormal event detection, which contains 330 training videos and 107 test videos recorded in 13 different scenes. The resolution of each video frame is 480 × 856 pixels.

In the experiments, video frames are fed into the proposed framework to calculate the anomaly scores. Then, following the experimental protocol in Liu et al. (2018), we measure the Area Under the Curve (AUC) (Cohen et al., 1998) by computing the Receiver Operating Characteristic (ROC) curve with varying threshold values for abnormality scores.

As shown in Table 2, the AUC performances of different models are reported. It is shown that our approach is comparable or even better than the recent state-of-the-art methods across three benchmarking datasets, demonstrating the effectiveness of our approach.

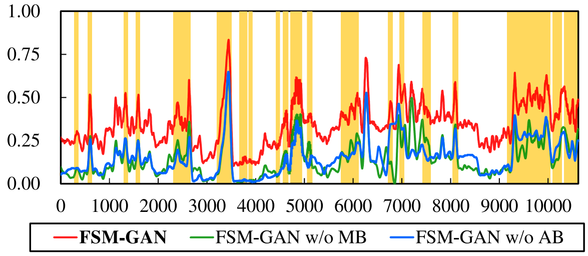


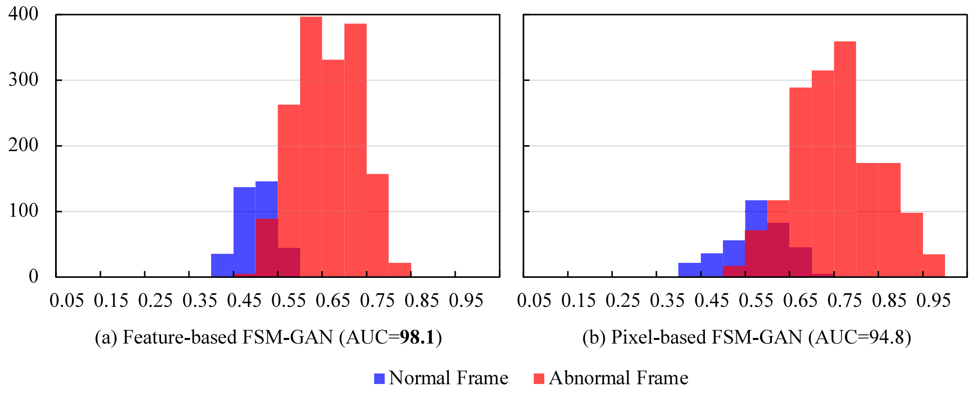
Fig. 3. Anomaly scores of ablation studies on the Avenue dataset.

5.2. Results on the avenue dataset

In the Avenue dataset, there are various types of abnormal events on avenues, which are widely considered to be challenging for precise abnormal event detection. The performances of the different methods on Avenue are reported in Table 2. It can be found that MNAD (Park et al., 2020) achieves the best performance of 88.5%. Our proposed method is comparable with the state-of-the-art result and attains a frame-level AUC score over 80% as well.

In this section, besides the overall anomaly scores calculated by Eq. (9), the scores obtained by the motion branch and the appearance branch are recorded separably to analyze the different roles of the two branches. By removing the motion branch, the given frame's abnormality can be measured only based on the perspective of appearance. Mathematically, the anomaly score can be formulated as $E_t = S(\|\phi_m(V_t'|F_t', M_t) - \phi_m(V_t|F_t, M_t)\|_1)$ as a contrast to Eq. (9). We denote the ablated framework as the FSM-GAN w/o MB for the convenience of the description. Similarly, removing the appearance branch, the anomaly score can be calculated $E_t = S(\|\phi_a(V_t'|F_t', M_t) - \phi_a(V_t|F_t, M_t)\|_1)$ from the perspective of motion. The ablated framework is denoted by FSM-GAN w/o AB.

The predicted anomaly scores and the ground truth of each frame are illustrated in Fig. 3. The yellow area represents the abnormal events labeled according to the ground truth. The green line indicates the anomaly scores obtained by FSM-GAN w/o MB. The blue line denotes the anomaly calculated by FSM-GAN w/o AB. The red line represents the overall score based on Eq. (9), which mixes the motion branch and the appearance branch. We can find that both FSM-GAN w/o MB and FSM-GAN w/o AB are able to keep the lowest score in normal scenarios. For the abnormal scenarios, sometimes FSM-GAN w/o MB gives higher anomaly scores, and sometimes FSM-GAN w/o AB does. It shows that the motion branch and appearance branch are sensitive to different anomalies. By fusing the FSM-GAN w/o MB and FSM-GAN w/o AB, the proposed FSM-GAN is able to take into consideration both the motion and appearance of the given frame to detect the anomaly. It can be found that in most cases, the anomaly scores correlate well to the ground-truth labels; i.e., strong upward spikes are produced by abnormal events. It means that abnormal events are rated as high scores, and normal events are the opposite.



(a) Feature-based FSM-GAN (AUC=**98.1**)

(b) Pixel-based FSM-GAN (AUC=94.8)

Fig. 4. The distribution of anomaly scores for normal and abnormal frames under different anomaly metrics on the UCSD Ped2 datase.

Table 3
The AUC (%) of ablation studies on the UCSD Ped2 dataset.

V_t/F_t	B	M_t	AUC(%)
✓			96.4
✓	✓		97.8
✓	✓	✓	**98.1**

5.3. Results on the UCSD dataset

It is generally considered that low frame resolution and multiple anomalies might affect the performance of abnormal event detection. The UCSD dataset, where the scenes have low frame resolutions and have multiple anomalies, is widely compared to evaluate the performance of abnormal event detection systems.

As shown in Table 2, the existing methods have made significant progress in the past ten years. Especially, Lee et al. (2018) proposed the best GAN-based framework, STAN, reaching a frame-level AUC of 96.5%. Recently, Park et al. (2020) introduced a memory-guided method and achieved state-of-the-art results with a frame-level AUC of 97.0%. Our proposed method further achieves a frame-level AUC of 98.1%, demonstrating the effectiveness of our proposed method. To show the performance gain brought by the designed module in FSM-GAN, we further conduct two ablation experiments.

On the one hand, one of the core ideas of the FSM-GAN is dividing the foreground and background to explicitly model foreground events under the conditions of background. To evaluate the performance gained by such design, we remove the foreground division and provide the ablation results for two stripped-down versions of the FSM-GAN. In detail, the foreground extractor and foreground division mask M_t are first removed. In this version of the ablated model, we provide V_t and B as the input of convolution blocks to extract low-resolution features and reconstruct F_t in the motion branch and mutually utilize F_t' and B as the input of convolution blocks to extract low-resolution features to synthesize V_t in the appearance branch. Then, we further ablate background image B for both branches without background prior B. In detail, we design a model taking the video frame V_t as the input of convolution blocks to extract low-resolution features and reconstruct the optical flow F_t, while mutually F_t' is provided as the input of convolution blocks to extract low-resolution features and reconstruct the video frame V_t. We remain the other setting of ablated models the same as FSM-GAN and report the performance of different stride-down versions. The ablation results are compared using the indicator of AUC (Cohen et al., 1998).

As shown in Table 3, we can find compared with the model without background image B as prior conditions, the ablated model remaining B achieves a performance improvement of near 1.4%. Meanwhile, the FSM-GAN demonstrates a performance gain with the help of foreground division based on M_t of near 0.3%. Based on these observations, we believe vanilla background images can provide references for the model recognizing the motion and appearance pattern of events. Furthermore, foreground division based on the mask M_t can also benefit the regional synthesizing of mutual transformation regarding appearance and optical flow vectors under the precise guidance of divided background conditions.

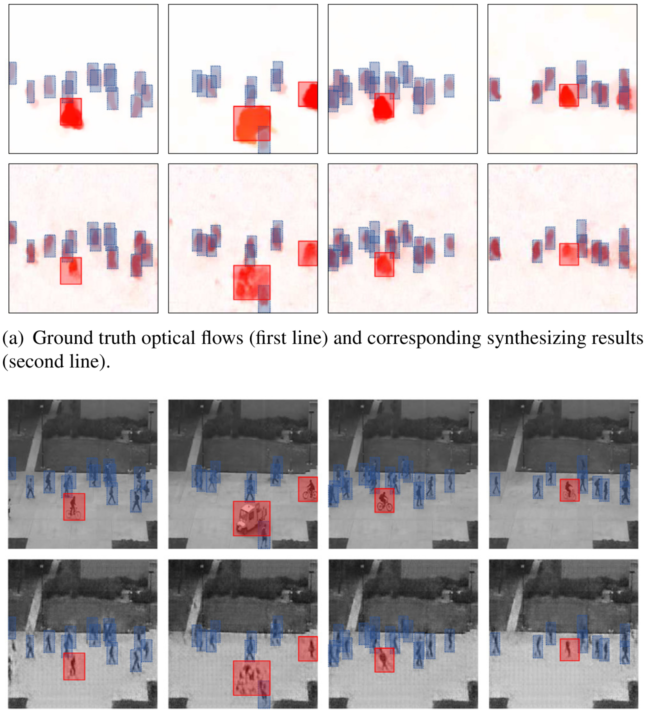
On the other hand, one of the most important improvements of this paper is that a novel anomaly metric is proposed to focus on the semantic anomalies in the foreground. While popular approaches reconstruct events at the pixel level to capture larger pixel-level reconstruction errors of anomaly, this paper takes feature-level reconstruction error of events into consideration and measure the shift of reconstructed feature to detect the anomaly. For comparison, in this paper, we modify the proposed FSM-GAN to an anomaly detection fashion based on measuring the reconstruction of visual primitives at the pixel level (hereinafter called pixel-based FSM-GAN). The training stage of pixel-based FSM-GAN remains the same as the FSM-GAN. In the test stage, the generators of trained FSM-GAN are utilized to reconstruct the optical flow F_t and video frame V_t under the foreground's guidance to obtain the reconstruction results F_t' and V_t'. Instead of Eq. (9), the pixel-level reconstruction error $E_t = S(\|F_t - F_t'\|_2)^{\frac{1}{2}} + S(\|V_t - V_t'\|_2)^{\frac{1}{2}}$ is utilized to measure the anomaly. The abnormal scores of video frames on the UCSD Ped2 dataset are counted, and the score distributions of two different anomaly metrics are provided in Fig. 4.

In Fig. 4, the histograms indicate how many anomaly scores are distributed in the interval corresponding to the abscissa. The red histograms correspond to the anomaly scores of the abnormal frames, while the blue histograms represent the anomaly scores of the normal frames. It can be found, no matter which anomaly metric method is applied, the blue histograms are always distributed to the left of the red histograms, demonstrating the model learns the normal patterns from the training dataset and can produce lower anomaly scores to normal scenes. It is also obvious that there is an overlap between the score distributions of normal frames and abnormal ones. In these cases, the anomaly scores of normal scenes and abnormal scenes are close. There is a high potential to confuse normal frames and anomalies, and the system will still make mistakes. The overlap of FSM-GAN is smaller than pixel-based FSM-GAN, showing better distinguishability of the proposed feature-based metric.

To further demonstrate how the proposed anomaly metric is robust to the outliers of low-level visual primitives, we conduct a series of simulation experiments to evaluate different anomaly metrics under different levels of motion blurring and multiplicative noise. In this experiment, the previously learned FSM-GAN is utilized in the test stage, and we apply motion blurring and multiplicative noise on the input video frames of all scenes, respectively, to simulate the outliers of low-level visual primitives. Here, motion blurring and multiplicative noise are implemented by the APIs, MotionBlur and MultiplicativeNoise, of Albumentations (Buslaev et al., 2020). We then report the AUC of ROC

Table 4
The comparison experiment of feature-based and pixel-based anomaly metrics on the UCSD Ped2 dataset under motion blurring and multiplicative noise. Here, AUC (%) of different methods are reported.

Transformation		Anomaly Metric	
		Feature-based metric	Pixel-based metric
None		**98.1**	**94.8**
Motion Blurring	*blur_limit=(3,7)*	96.4	87.4
	blur_limit=(6,10)	95.4	81.4
Multiplicative Noise	*multiplier=(0.8,1.2)*	97.9	81.7
	multiplier=(0.7,1.3)	97.0	75.5


(a) Ground truth optical flows (first line) and corresponding synthesizing results (second line).

(b) Ground truth video frames (first line) and corresponding synthesizing results (second line).

Fig. 5. Qualitative results are showing the ground truth optical flows and input video frames, besides corresponding synthesized results by the proposed FSM-GAN, on the UCSD Ped2 dataset. Especially, the regions where abnormal events occur are marked in red and normal behaviors are marked in blue.

curves of FSM-GAN based on feature-based and pixel-based anomaly metrics on the UCSD Ped2 dataset.

As shown in Table 4, *blur_limit* refers to the range of kernel size for blurring the input image. Furthermore, *multiplier* denotes the range of numbers to be multiplied. Larger kernel size and multiplier can lead to larger differences after transformation. We can find the performance drop down after introducing additional simulated motion blurring and multiplicative noise, no matter what anomaly detection metric is utilized. This phenomenon is related to the fact that motion blurring and multiplicative noise lead to the low-level visual primitives of test video frames distinct from training samples, affected by which FSM-GAN cannot recognize some patterns of normal behaviors. Especially, we can also find the performance dropdown more sharply with pixel-based metric compared with feature-based metric. Meanwhile, after the level of motion blurring and multiplicative noise rising up, the performance gap becomes more considerable. It demonstrates that feature-based anomaly score is less affected by low-level visual primitive variance, and compared with pixel-based metric, the feature-based metric can neglect parts of motion blurring and noises to achieve better performance in detecting abnormal events.

Besides the above quantitative analysis, we also provide a visualization experiment on the UCSD Ped2 dataset and report a few qualitative results showing the synthesized optical flows and video frames of normal events and anomalies, demonstrating the performance of generative networks in distinguishing abnormal events. In detail, after training is completed, the optimized generative networks on the UCSD Ped2 dataset, i.e., G_m in motion branch and G_a in appearance branch, are utilized to synthesize the optical flows and video frames based on given foreground prior information and background conditions. A few scenes are randomly selected, and the generated results are reported.

As shown in Fig. 5, the ground truth optical flows are presented on the first line of the sub-figure Fig. 5(a), while the second line of the sub-figure Fig. 5(a) corresponds to the reconstructed optical flows synthesized by the proposed FSM-GAN. Here, HSV optical flow visualization is applied, where the value (brightness) corresponds to the magnitude of optical flow vectors. We can find that the color distribution of generated optical flow images where events occur differs from ground truth, demonstrating that the estimated motion is distorted to different extents. Compared with the synthesizing results marked in blue, which represents normal events, the abnormal regions marked in red are distorted heavily: the high-brightness regions are severely deformed from that of ground truth optical flow. In Fig. 5(b), we further show the ground truth video frames and corresponding synthesized frames, and the ground truth video frames are presented on the first line of Fig. 5(b), while the second line corresponds to the reconstructed video frames. The phenomenon is consistent with that of optical flow: reconstructed frames are distorted where high-resolution details are smoothed compared with ground truth. Especially, the regions where abnormal events occur in reconstructed video frames are distorted heavily: the human body is approximated with a blob-like structure. The distortion of anomalies shows that without training on abnormal samples, the generator cannot recognize motion and appearance patterns of anomalies so that the discriminator can detect the out-of-distribution pattern and capture distorted abnormal samples.

5.4. Comparison experiments on mask guidance

In this subsection, we further analyze the strategies for generating a mask for foreground and background division and compare the designed strategy with other reasonable choices.

In detail, besides the designed strategy based on the average background algorithm, the optical flow can also guide the foreground and background division. That is, it is also a natural choice to threshold the optical flow image and get the centroid or area of the moving object. To evaluate and compare the performance gain of these two kinds of foreground division methods, we further carry out a comparison experiment. Here, we utilize average background subtraction algorithm (as Section 4.3) and an optical flow-based division to extract the background image, respectively. In detail, for the flow-based division, DeepFlow is utilized to estimate the optical flow F_t between the $(t-3)$th video frame and the tth video frame. Then, F_t is normalized via min–max normalization and the obtained result F'_t can be formulated as:

$$F'_t(i,j) = \frac{F_t(i,j) - min(F_t)}{max(F_t) - min(F_t)} \tag{11}$$

Table 5
AUC (%) of FSM-GAN-o and FSM-GAN on the Avenue and UCSD Ped2 dataset.

Method	Dataset	
	Avenue	UCSD Ped2
FSM-GAN-o	74.8	98.0
FSM-GAN	**80.1**	**98.1**

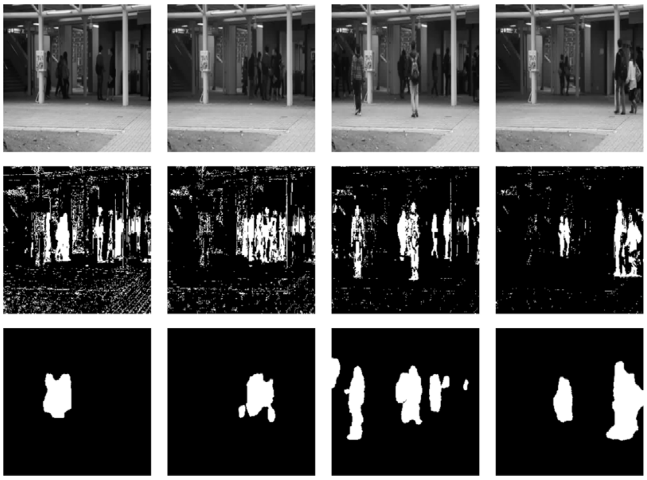


Fig. 6. Qualitative results are showing the video frames (first line), mask images generated by the proposed FSM-GAN (second line), and the FSM-GAN-o (third line) on the Avenue dataset. Here, the white color corresponds to the foreground, and the black color denotes the background.

Here, $min(F_t)$ denotes the minimized value of the optical flow image F_t, while $max(F_t)$ represents the maximized value of the optical flow image F_t. Next, we threshold the F'_t to obtain mask M_t:

$$M_t(i,j) = \begin{cases} 1, F'_t(i,j) > T \\ 0, F'_t(i,j) \leq T \end{cases} \tag{12}$$

Based on M_t, we can divide the dynamic area (foreground) and stationary (background) area over input frame to extract foreground and background. In the subsequence modeling of FSM-GAN, $V_t \circ M_t$ and $V_t \circ (1 - M_t)$ are provided as the input of motion branch, representing foreground appearance and background conditions, respectively. Meanwhile, $F_t \circ M_t$ and $V_t \circ (1 - M_t)$ are provided as the input of appearance branch, denoting foreground motion and background conditions, respectively. Here, $\circ$ is the elementwise multiplication.

Table 5 reports the performance of the alternative version of FSM-GAN with optical flow-based foreground division (hereinafter referred to as FSM-GAN-o) on two benchmark datasets, respectively. It can be found that FSM-GAN-o brings performance degradation to different extents on benchmark datasets. Especially, the performance margin on the Avenue dataset is the largest, up to over 5%.

For further exploration, we randomly select a few video frames on the Avenue dataset and provide the visualization results of mask generated by optical flow-based foreground division (from FSM-GAN-o) compared with the average background subtraction algorithm (from FSM-GAN). As shown in Fig. 6, although optical flow-based foreground division is robust to background noise, such a method only performs well on locating a rough area for behaviors. On the contrary, average background subtraction results include a few noises caused by camera movement, but contain extensive details of behaviors to infer the events. These details provide more accurate foreground information and contribute to modeling normal events, so that rarely happened anomalies are distinguished. The experimental results demonstrate the effectiveness of our designed foreground division strategy on the abnormal event detection task.

Table 6
AUC (%) of FSM-GAN-p and FSM-GAN on the Avenue, and UCSD Ped2 dataset.

Method	Dataset	
	Avenue	UCSD Ped2
FSM-GAN-p	77.0	97.7
FSM-GAN	**80.1**	**98.1**

5.5. Comparison experiments on motion modeling

As a task modeling pattern of motions, the optical flow estimation task has existed for a long time and is widely researched. Inspired by the practice in the optical flow estimation task field, we discuss a few reasonable alternatives to model motion patterns of events in this subsection. The main goal is to explore the models with the best applicability to abnormal event detection tasks.

First, considering that the optical flow represents the difference between the $(t-3)$th and the tth video frame, using $(t-3)$th video frame as the input is also a reasonable choice for motion branch. To evaluate the possible performance difference, we replace the input, V_t, M_t and B, of motion branch in the proposed FSM-GAN with the input, V_{t-3}, M_t and B, and conduct comparison experiments on the Avenue dataset to compare the performance of the revised method and the proposed model. For convenience, the modified version of FSM-GAN will be referred to as FSM-GAN-p in the following.

Table 6 reports the performance of the FSM-GAN and FSM-GAN-p on the Avenue dataset and UCSD Ped2 dataset, respectively. It can be found that the performance of FSM-GAN-p slightly dropdown by a margin of about 3% compared with the designed FSM-GAN on the Avenue dataset, and 1.3% on the UCSD Ped2 dataset. In this paper, we believe the performance gap may result from the mismatch between the provided mask and input video frame V_{t-3}. In detail, in both motion branch and appearance branch, the mask M_t is provided as guidance to divide the foreground and background, which is calculated based on V_t as shown in Eq. (2), (3), (4). If we replace V_t in the motion branch's input with V_{t-3}, the shift between V_t and V_{t-3} can lead to the ground truth foreground division different from the guidance mask M_t. Finally, inaccuracy foreground division leads to performance degradation of events modeling.

Then, we also consider the design of providing both $(t-3)$th and tth video frames as the input of the motion branch. Compared with such a design, the original FSM-GAN formulates an ill-posed optical flow estimation task in the motion branch for abnormal event detection; that is, the FSM-GAN is designed to estimate the optical flow between $(t-3)$th and tth timestep given only the information at the tth timestep. Though the ill-posed formulation can be adverse to precise optical flow estimation due to the lack of $(t-3)$th timestep information, the final goal of this paper is not to estimate optical flow but to distinguish abnormal events. The ill-posed formulation contributes to different levels of estimated distortion of anomalies and makes abnormal events distinct from normal behaviors. It relies on the hypotheses of abnormal event detection: *normal events always happen and are well collected in the training stage, while abnormal behaviors hardly appear and are omitted in the training dataset*. In ideal situations, the model learns the pattern of normal behavior in the training stage, which makes the near future of normal behaviors predictable. As a result, optical flow in the near temporal-neighborhood is assessable, producing low anomaly scores. In contrast, the model never sees the pattern of anomalies and cannot predict what will happen when anomalies appear. Due to the lack of the following frame information, the model fails to predict precise optical flow for abnormal events, and the model gives high scores.

To evaluate if it is necessary to tackle a well-posed optical flow estimation problem in the abnormal event detection task, we conduct a comparison experiment as follows. For the first step, we revise the motion branch of the proposed FSM-GAN to provide complete information, i.e., the video frame V_t and adjacency video frame V_{t-3}

Table 7
AUC (%) of FSM-GAN-q and FSM-GAN on the Avenue, UCSD Ped2, and ShanghaiTech dataset.

Method	Dataset		
	Avenue	UCSD Ped2	ShanghaiTech
FSM-GAN-q	**80.3**	97.9	71.4
FSM-GAN	80.1	**98.1**	**73.5**

as input for optical flow estimation. Then, similar to the network architecture of FSM-GAN, we utilize three convolutional blocks to extract low-resolution features from the video frame V_t, and another three convolutional blocks to extract low-resolution features from the adjacency video frame V_{t-3}. The other implementation details are also the same as FSM-GAN. For convenience, the modified version of FSM-GAN will be referred to as FSM-GAN-q in the following. In this way, given inputs, the excepted prediction of FSM-GAN-q is unique. We record the performance of the FSM-GAN-q on the Avenue, UCSD Ped2, and ShanghaiTech dataset.

As shown in Table 7, the proposed method outperforms the well-posed optical flow estimation method FSM-GAN-q, demonstrating the effectiveness of the ill-posed optical flow estimation on abnormal events detection on the UCSD Ped2 and ShanghaiTech dataset. Especially, we can observe near 2% performance gain on the ShanghaiTech dataset with the ill-posed formulation. On this dataset, enormous training videos are collected. With the help of large training samples, the generation ability of a deep learning network can lead to precise optical flow estimation for abnormal events, with a well-posed optical flow estimation, and make it hard to distinguish the distortion of normal events and anomalies. Thus, the performance of FSM-GAN-q drops down.

It is also worth noting that the proposed model is outperformed by the well-posed optical flow estimation method FSM-GAN-q slightly on the Avenue dataset. We find the failure cases related to the fact that the hypotheses of abnormal event detection are not always perfectly met in real-world scenarios. And parts of normal behaviors cannot be covered in the training dataset for some reason. In the Avenue dataset, training samples include pedestrians moving horizontally far away from the camera (from left to right or from right to left), which is considered normal behavior. However, in the test dataset, parts of pedestrians stay close to the camera or walk from far to near. In common sense, these pedestrians should be considered normal regardless of movement direction changes, since walking patterns appear in the training samples as normal events. However, from the computer vision perspective, these movements are quite different from known patterns due to the lack of samples with similar optical flow distributions, and thus the pedestrians are falsely recognized as abnormal events. In the future, we will explore the potential application of generative models to generalize on scale variance and fuse human common sense to produce low false-positive scores for normal behaviors.

5.6. Results on the ShanghaiTech dataset

To the best of our knowledge, the ShanghaiTech dataset is one of the most challenging dataset published in recent years. The large scale and various scenarios lead to the poor performance of many existing works. To further verify the robustness of our model for different scenarios, the designed framework is also applied to the ShanghaiTech dataset.

As shown in Table 2, our proposed method shows great abnormal event detection performance of over 70% and considerably outperforms the second-best method by a margin of 0.7%. Notably, the ShanghaiTech dataset contains 107 test videos and 40791 frames of scenes to be tested. Each percentage increase corresponds to the detection result of a large number of frames. The proposed model demonstrates good detection performance on this large scale anomaly detection dataset.

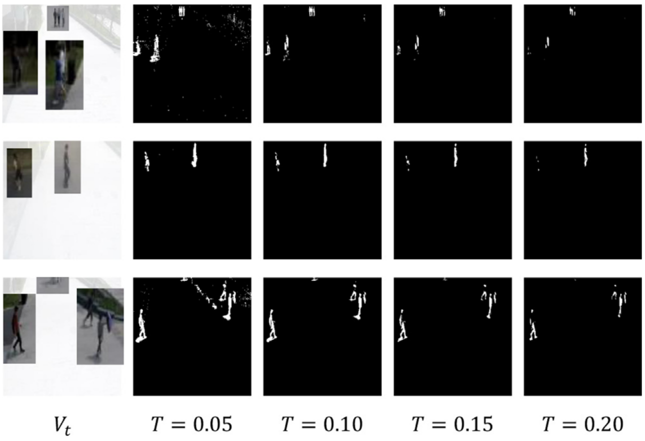


Fig. 7. A few qualitative results are showing the generated masks when T is 0.05, 0.10, 0.15, and 0.20, respectively. To facilitate observation, in the video frames shown in the first column, we zoomed in on the regions where events occur.

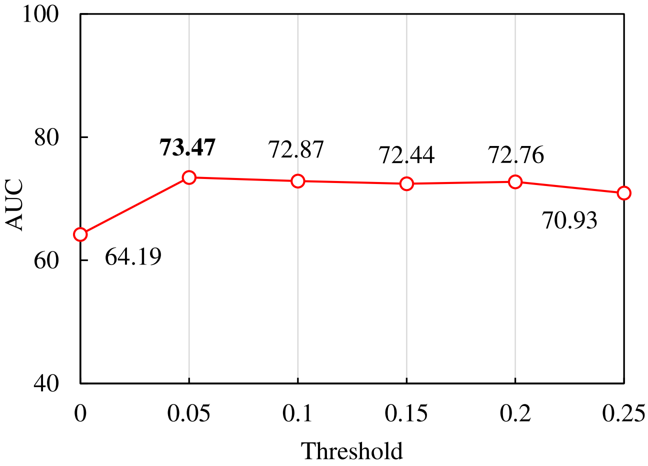


Fig. 8. AUC (%) of the sensitivity experiment on the ShanghaiTech dataset.

In this experiment, we further conduct the sensitivity experiment to evaluate our model under different hyperparameter settings. In this paper, one of the most important hyper-parameters is T in Eq. (4). When the normalized difference $d_t'(i,j)$ between a pixel of a given frame and a corresponding pixel in the background exceeds the threshold T, the pixel will be regarded as the foreground. As shown in Fig. 7, the masks of an example video frame under different threshold T are provided. It is obvious that smaller T leads to that more pixels are recognized as the foreground, and the foreground can contain more details and noise.

In Fig. 8, remaining other settings the same, we report the AUCs of the proposed method when the threshold T ranges from 0.0 to 0.25. It is easily observed that when the threshold is over 0.05, the performance of FSM-GAN slightly dropdown with the increase of threshold but can always achieve AUCs over 70.5% and outperform the latest art MNAD. When the threshold is equal to 0.05, our model achieves the best performance. From the example frame shown in Fig. 7, we can see that, in this case, the foreground contains more details. It makes the small-scale events in the distance can be modeled precisely. When the threshold is equal to 0.0, all pixels in the frame are regarded as the foreground. In this way, the separated foreground degenerates into the entire scene, and the model cannot explicitly focus on the objects of interest. It makes the modeling of events suffer from background noise, thus producing poor performance. In practice, 0.05 is a suitable choice for most situations.

6. Conclusion and future work

In this paper, we propose the Foreground–Background Separation Mutual Generative Adversarial Network for abnormal event detection. To capture the nuances of abnormal events and suppress various background noise, we are the first to propose a GAN-based framework that permits the decomposition of the foreground and background to guide the modeling of events. For the first step, the foreground extractor permits to separate the foreground from the given scenes. Then, mutual generative adversarial networks are proposed to model raw-pixel images in optical-flow representations of the foreground under the background condition. Finally, the high-level presentation of the foreground is utilized to take the place of low-level visual primitives for identifying anomalies.

By applying the designed framework on the most commonly used dataset (Avenue dataset) and a low frame resolution benchmarking dataset (UCSD dataset), the experimental results show that the designed model can identify the spatio-temporal features of the foreground under the condition of background, and perform satisfactorily even in the situation of a large-scale dataset (ShanghaiTech dataset). In the future, we will seek to utilize attention mechanisms to separate the foreground evolved with events in video frames automatically.

CRediT authorship contribution statement

Zhi Zhang: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Sheng-hua Zhong:** Conceptualization, Validation, Resources, Data curation, Writing – review & editing, Project administration, Funding acquisition. **Ahmed Fares:** Writing – review & editing. **Yan Liu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province, China [No. 2019A1515011181], the Science and Technology Innovation Commission of Shenzhen, China under Grant [No. JCYJ20190808162613130], the Shenzhen high-level talents program, China, and the Open Research Fund, China from Guangdong Laboratory of Artificial Intelligence & Digital Economy (SZ) under Grant [No. GML-KF-22-28].

References

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C., 2019. Mixmatch: A holistic approach to semi-supervised learning. CoRR.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: fast and flexible image augmentations. Information 11, 125.

Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., Tishby, N., 2011. Detecting anomalies in people's trajectories using spectral graph analysis. Comput. Vis. Image Underst. 115 (8), 1099–1111.

Cohen, W.W., Schapire, R.E., Singer, Y., 1998. Learning to order things. In: NIPS. pp. 451–457.

Dan, X., Yan, Y., Elisa, R., Nicu, S., 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput. Vis. Image Underst. 156, 117–127.

Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F., 2020. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. Comput. Vis. Image Underst. 102920.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d., 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV. pp. 1705–1714.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein GANs. In: NIPS. pp. 5767–5777.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S., 2016. Learning temporal regularity in video sequences. In: CVPR. pp. 733–742.

Hinami, R., Mei, T., Satoh, S., 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge. In: ICCV. pp. 3619–3627.

Hong, J., Li, Y., Chen, H., 2019. Variant Grassmann manifolds: A representation augmentation method for action recognition. TKDD 13 (2), 1–23.

Idrees, H., Zamir, A.R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M., 2017. The THUMOS challenge on action recognition for videos "in the wild". Comput. Vis. Image Underst. 155, 1–23.

Ionescu, R.T., Khan, F.S., Georgescu, M.-I., Shao, L., 2019. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: CVPR. pp. 7842–7851.

Jeong, H., Chang, H.J., Choi, J.Y., 2011. Modeling of moving object trajectory by spatio-temporal learning for abnormal behavior detection. In: AVSS. pp. 119–123.

Jiang, F., Yuan, J., Tsaftaris, S.A., Katsaggelos, A.K., 2011. Anomalous video event detection using spatiotemporal context. Comput. Vis. Image Underst. 115 (3), 323–333.

Lee, S., Kim, H.G., Ro, Y.M., 2018. STAN: Spatio-temporal adversarial networks for abnormal event detection. In: ICASSP. pp. 1323–1327.

Liu, W., Luo, W., Lian, D., Gao, S., 2018. Future frame prediction for anomaly detection–a new baseline. In: CVPR. pp. 6536–6545.

Lu, C., Shi, J., Jia, J., 2013. Abnormal event detection at 150 fps in matlab. In: ICCV. pp. 2720–2727.

Luo, W., Liu, W., Gao, S., 2017a. Remembering history with convolutional lstm for anomaly detection. In: ICME. pp. 439–444.

Luo, W., Liu, W., Gao, S., 2017b. A revisit of sparse coding based anomaly detection in stacked rnn framework. In: ICCV. pp. 341–349.

Luo, W., Stenger, B., Zhao, X., Kim, T.-K., 2018. Trajectories as topics: Multi-object tracking by topic discovery. Trans. Image Process. 28 (1), 240–252.

Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N., 2010. Anomaly detection in crowded scenes. In: CVPR. pp. 1975–1981.

Park, H., Noh, J., Ham, B., 2020. Learning memory-guided normality for anomaly detection. In: CVPR. pp. 14372–14381.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N., 2017. Abnormal event detection in videos using generative adversarial nets. In: ICIP. pp. 1577–1581.

Ravanbakhsh, M., Sangineto, E., Nabi, M., Sebe, N., 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In: WACV. pp. 1896–1904.

Roshtkhari, M.J., Levine, M.D., 2013. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Comput. Vis. Image Underst. 117 (10), 1436–1452.

Rota, P., Conci, N., Sebe, N., 2012. Real time detection of social interactions in surveillance video. In: ECCV. pp. 111–120.

Saligrama, V., Chen, Z., 2012. Video anomaly detection based on local statistical aggregates. In: CVPR. pp. 2112–2119.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. Adv. Neural Inf. Process. Syst. 29, 2234–2242.

Sengupta, A., Budvytis, I., Cipolla, R., 2020. Synthetic training for accurate 3D human pose and shape estimation in the wild. In: BMVC.

Sultani, W., Chen, C., Shah, M., 2018. Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488.

Wan, J., Luo, W., Wu, B., Chan, A.B., Liu, W., 2019. Residual regression with semantic prior for crowd counting. In: CVPR. pp. 4036–4045.

Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C., 2013. DeepFlow: Large displacement optical flow with deep matching. In: ICCV. pp. 1385–1392.

Wiliem, A., Madasu, V., Boles, W., Yarlagadda, P., 2012. A suspicious behaviour detection using a context space model for smart surveillance systems. Comput. Vis. Image Underst. 116 (2), 194–209.

Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N., 2015. Learning deep representations of appearance and motion for anomalous event detection. CoRR.

Zaharescu, A., Wildes, R., 2010. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In: ECCV. pp. 563–576.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595.

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.-S., 2017. Spatio-temporal autoencoder for video anomaly detection. In: MM. pp. 1933–1941.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3D human pose regression. In: CVPR. pp. 3425–3435.

Zhu, X., Liu, J., Wang, J., Li, C., Lu, H., 2014. Sparse representation for robust abnormality detection in crowded scenes. Pattern Recognit. 47 (5), 1791–1799.