Multi-Branch GAN-based Abnormal Events Detection via Context Learning in Surveillance Videos

Daoheng Li, Xiushan Nie, *Senior Member, IEEE*, Rui Gong, Ximing Lin and Hui Yu *Senior Member, IEEE*

***Abstract*—Video anomaly detection is an important task in the field of intelligent security. However, existing methods mainly detect and analyze videos from a single time direction, ignoring the semantic information of the video context, which adversely affects the detection accuracy. To address this issue, we design a multi-branch generative adversarial network with context learning (MGAN-CL) to detect abnormal events. In particular, we combine video context information to generate predicted frames, and determine whether an anomaly occurs by comparing the predicted frame with the actual frame. Different from the existing GAN-based methods, in the anomaly event detection stage, we use the discriminator to judge the video frames generated by the generator, which improves the accuracy of anomaly detection. In order to improve the ability of the discriminator, a pseudo-anomaly module is added to the discriminator for data augmentation to improve the robustness of the model. An extensive set of experiments performed on public datasets demonstrate the method's superior performance.**

***Index Terms*—video anomaly detection, video context information, bidirectional prediction, generative adversarial network, pseudo-anomaly module**

I. Introduction

VIDEO abnormal events detection is an essential task in the field of intelligent security. However, traditional manual detection is inefficient and cumbersome in the detection process due to the complexity. Therefore, it is necessary to conduct video anomaly detection automatically. Due to the rarity and diversity of abnormal events, video anomaly event detection is still a challenging task. Early studies mainly model normal and abnormal events by handcrafted extracted features, to classify abnormal events and complete the detection task [1]–[8]. Considering weak generalizability, the traditional feature extraction methods cannot be adjusted according to different application scenarios, which don't meet the current application requirements.

Daoheng Li, Xiushan Nie, Rui Gong and Ximing Lin are with School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China;

Hui Yu is with School of Creative Technology, University of Portsmouth, Portsmouth, UK.

Corresponding author: Xiushan Nie.

This work was supported in part by the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), National Natural Science Foundation of China (62176141), Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2021ZD15), Major science and technology innovation project of Shandong Province (2021CXGC011204), and Taishan Scholar Project of Shandong Province (tsqn202103088).

With the development of deep learning-based technology, the research on VAD has made great progress [9]–[15], [26]–[28]. Video anomaly detection methods based on deep learning can be divided into three categories: supervised learning, unsupervised learning and semi-supervised learning. Supervised learning is to label normal and abnormal data; unsupervised learning is to use unlabelled information in the training process; semi-supervised method is to adopt only normal samples in the training process.

Owing to the particularity of VAD as a binary classification task, the classification of normal and abnormal samples is extremely unbalanced. Toward this end, semi-supervised anomaly detection has gradually attracted more attention. Semi-supervised methods can be divided into two categories: reconstruction model-based methods [29] [34] and prediction model-based methods [31] [36]. Most reconstruction model-based methods have already mature. For example, if the video contains abnormal events during testing, the reconstruction error will become larger. With sparse reconstruction techniques, Mo et al. [9] proposed a linear sparse model with superior class separability for anomaly detection. Based on deep learning, the reconstruction model-based methods can obtain strong learning ability. Specifically, it can even well reconstruct abnormal events that have never occurred in some cases, which is a fatal disadvantage of those methods.

In order to solve the shortcomings of the reconstruction model-based method, the prediction model-based methods are proposed. Since anomalous events are not easy to predict, it is possible to analyze whether an anomalous event occurs by the error between the predicted frame and the ground truth. In particular, first, a continuous video frame containing only normal events is used as input, and a predicted frame is obtained by calculation [36]. Then, the error between the ground truth and the predicted frame is evaluated, based on which the abnormal event is judged. Prediction methods based on Generative Adversarial Networks (GAN) [48] have achieved good performance.

Although reconstruction and prediction methods have obtained good performance, most of them cannot adequately consider the information of the video context. It is noted that adopting the contextual semantic information of the video can detect abnormal events more accurately. As shown in Fig. 1, without the context information of the video, it is difficult to judge the abnormal behavior of object from the first 4 frames. However, using the 8 frames before and after the frame, that is, the video's context information, I_t can be easily recognized

as an anomaly event.

Fig. 1. With contextual information of video frames, we can easily identify this object is an anomaly event in current frame.

To use video contextual information effectively, we propose a multi-branch GAN with context learning (MGAN-CL) for video abnormal event detection. In the training phase, we only use normal continuous video frames as input, to generate the frames of normal events. When an abnormal event occurs, the ground truth will have a large error with the generated frame, so it will be considered as an abnormal event. The proposed MGAN-CL uses multi-branch GAN as the base network, and adopts video context information through two generators. The multi-branch GAN and the video context information are utilized to generate prediction frames and retrospective frames. To improve the performance of the discriminator and the robustness of our network, we use both the generators and the discriminator in the detection phase to detect abnormal events. Besides, a pseudo-anomaly module is added to create abnormal events that do not exist.

In the detection phase, we first use two generators to generate video frames, and then calculate the errors with the ground truth. In addition, we input the two generated frames into the discriminator, and compare the features of the generated frame with the ground truth. Finally, the normal event score of the video frame can be calculated with weights. If the normal event score of the frame is greater than that of the threshold, the video frame will be considered as a normal event, otherwise, an abnormal event.

The contributions of this research are summarized as follows:

- We design a multi-branch GAN to predict and retrospect video frames, and then learn the video contextual information to determine whether an abnormal event occurs. Different from existing methods that detect abnormal events from a single time direction, the proposed network can more adequately capture the event clues in the task of anomaly detection.
- We adopt a pseudo-anomaly module in the proposed network, and employ both the generator and the discriminator in the detection phase. In this way, the performance of the discriminator is improved in the anomaly detection phase, and the robustness of the proposed method is enhanced.

II. Related Work

With the increasing demand for intelligent security, video anomaly event detection has become a hot research topic recently. Handcrafted feature-based methods need to model abnormal events in the video. However, due to the ambiguous definition of abnormal events and the rarity of abnormal events, they cannot perform well enough. There are a growing number of deep learning-based VAD methods [16]–[24]. As technology advances, deep learning-based methods have gradually become the mainstream for abnormal event detection, which can be divided into three categories: unsupervised learning, supervised learning, and semi-supervised learning.

A. Unsupervised Anomaly Detection

In unsupervised methods, the proportion of normal events is assumed to be much larger than that of abnormal ones, and there are no labels in the training set. Ren et al. [25] proposed a method to detect anomalies in the clustering situation in the feature space through support vector data description, which uses non-negative matrix factorization and feature learning. GAN-based video anomaly detection mehtods always learn the appearance and optical flow features of normal events through the generator, so as to obtain the feature representation of normal behavior. Besides, they compare the features of objects in the current video frame with normal events in the testing phase. Both appearance features and optical flow features of abnormal events are different from those of normal events, thereby detecting abnormal events in videos [53]. Clustering is performed according to the correlation among samples. When an event cannot be clustered into any category, it is judged as an abnormal event. For example, Ionescu et al. [39] learned the latent features of video frames, during the training phase, they extracted the latent features and clustered the feature space to detect abnormal ones.

However, due to the lack of prior knowledge in unsupervised learning methods, overfitting may occur during the training process, resulting in the inability to effectively distinguish normal events from abnormal events [30].

B. Supervised Anomaly Detection

During training, supervised learning video anomaly detection methods require labels for normal and abnormal events. Zhou et al. [56] proposed an adaptive learning to detect anomalies, which removes noisy backgrounds, capture motion and alleviates the advantages of insufficient data. GAN is also utilized in supervised learning. Ravanbakhsh et al. [33] proposed a special approach using GAN, which generates only the normal distribution, the discriminator acts as a supervisor for the generator. Fan et al. [59] proposed a variational autoencoder based on gaussian mixtures that can learn the feature representation of normal samples as a gaussian mixture model (GMM) trained using deep learning. Xu et al. [17] proposed an adaptive intra-frame classification network (AICN) to transform this task to a multi-class classification problem. AICN is adaptive to frames with different resolutions and is easier to be applied for other scenes.

However, supervised learning methods need to cost a lot for labeling all video frames, especially for pixel-level anomaly detection. Therefore, Sultani et al. [35] proposed a weakly-supervised learning approach. In this method, abnormal events are learned through a deep multi-instance ranking framework

by exploiting weakly labeled training videos, i.e. training labels (abnormal or normal) at video-level rather than clip-level. Besides, normal and anomalous videos are treated as a bag, video clips are treated as instances in multiple instance learning (MIL), and a deep anomaly ranking model that predicts high anomaly scores is automatically learned. Doshi et al. [21] proposed an online anomaly detection method for surveillance videos using transfer learning and continuous learning, which continuously learns from recent data without suffering from catastrophic forgetting. Tian et al. [30] proposed a method to train a feature magnitude learning function for efficiently identifying positive instances, which significantly improves the robustness of MIL methods to negative instances from anomalous videos.

C. Semi-supervised Anomaly Detection

In semi-supervised anomaly detection, only normal samples are available as training data, which has attracted a lot of research interest. Since the training phase only learns the characteristics of normal events, if there are abnormal events in the test set, there will be a huge error. When the error is large, it is very likely that an abnormal event has occurred, otherwise, it is more likely to be a normal event. Hasan et al. [29] proposed a generative model that uses multiple sources with very limited supervision to learn regular motion patterns (referred to as rules) for addressing ambiguous definitions of meaning and confusion in scenes. Novelty detection proposed by Sabokrou et al. [32] implemented the task of data classification for test data, that is, to identify a new or uninterpretable set of data to determine whether these data conform to data features during training. Nguyen el al. [34] proposed a reconstruction network that shares the same encoder in combination with an image translation model. The former sub-network gets the most important cues that appear in the video frame, while the latter tries to associate motion templates with these cues. Although the reconstruction based methods have achieved excellent performance, these methods are still limited by feature space design [54]. Ye et al. [16] decomposed the reconstruction into prediction and refinement, introducing ERM to reconstruct the current prediction error and refine the coarse prediction. Liu et al. [18] proposed a hybrid framework named HF2-VAD, reconstructing normal patterns by memorizing optical flow. This approach captures the high correlation between video frames and optical flow to predict the next frame given several previous frames.

Liu et al. [36] proposed a prediction model-based method by learning features of video frames segment to generate future frames, which combines both object appearance and motion features. For this method, a sequence of consecutive video frames are input, and a prediction model is used to generate the predicted frame. Yu et al. [37] proposed an adversarial event prediction method that can generate a discriminative model to detect abnormal events without complementary information. Multi-space feature learning was proposed by Zhang et al. [40], which obtains the image features and latent features of generated frames. The multi-space features can add constraints to improve the performance of anomalous event detection.

III. Proposed Approach

In this study, we design a new network called MGAN-CL, which consists of two generators and a discriminator. The two generators first generate the video frame through context learning, and then input the generated frame to the discriminator to determine whether it is a real frame. Furthermore, due to the lack of abnormal samples, motivated by the study in Zaheer et al. [41], the pseudo-abnormal data are used to train the model in MGAN-CL, where we add a pseudo module and old-epoch generator module to the discriminator to improve the performance of the discriminator. The proposed network structure of MGAN-CL is shown in Fig. 2.

During the training phase, two U-Nets are used as the generators for the prediction and retrospective branches, respectively. Specifically, U-Nets permit not only low-level features to be retained but also high-level features to be utilized, resulting in more realistic predicted frames. Only one discriminator is included in the model network in MGAN-CL, which is used to distinguish whether the generated frames of the two branches are actual. In addition, the latent features of the generated frames and the ground truth are extracted, and the multi-space learning method is used to analyze whether the video frames are actual ones. It is worth mentioning that we use the pseudo-anomaly module to process two discontinuous video frames into one pseudo-anomaly frame. The pseudo-abnormal frame is not a real abnormal event, which can be used to increase the number of abnormal event samples for improving discriminant's ability.

A. Symbols and Notions

In our method, G and D denote the generator and discriminator in the network, respectively. L_G represents the loss of the generator G of our method, while the loss of the discriminator D of the network is denoted by L_D. Samples from the data set aims to consecutive $2T+1$ frame video samples $\{I_{t-T},\cdots,I_t,\cdots,I_{t+T}\}$, which is used as the input of the network structure, where I_t is to be detected. $\{I_{t-T},\cdots,I_{t-1}\}$ will be used as the input of the prediction branch generator G_p, and the output prediction frame is I_t^p. Similarly, $\{I_{t+1},\cdots,I_{t+T}\}$ will be inverted to $\{I_{t+T},\cdots,I_{t+1}\}$, and then used as the input of the retrospective branch generator G_b, and I_t^b represents the output retrospective frame. The detailed definition of symbols are shown in Table I.

B. Formulation

In the proposed method, we use a multi-branch network structure with dual generators, and add constraints to the predicted frame and the retrospective frame. Given the output frames I_t^p and I_t^b of the prediction and retrospective branch generators, they should come from the same frame I_t, and the image information and multispace features of I_t^p and I_t^b should be similar, including image attributes such as intensity and gradient. Towards this end, we design five objective function terms for the generators, including gradient loss, intensity loss, optical flow loss, adversarial loss, and latent loss. As for the discriminator, a block discriminator [42] is adopted,

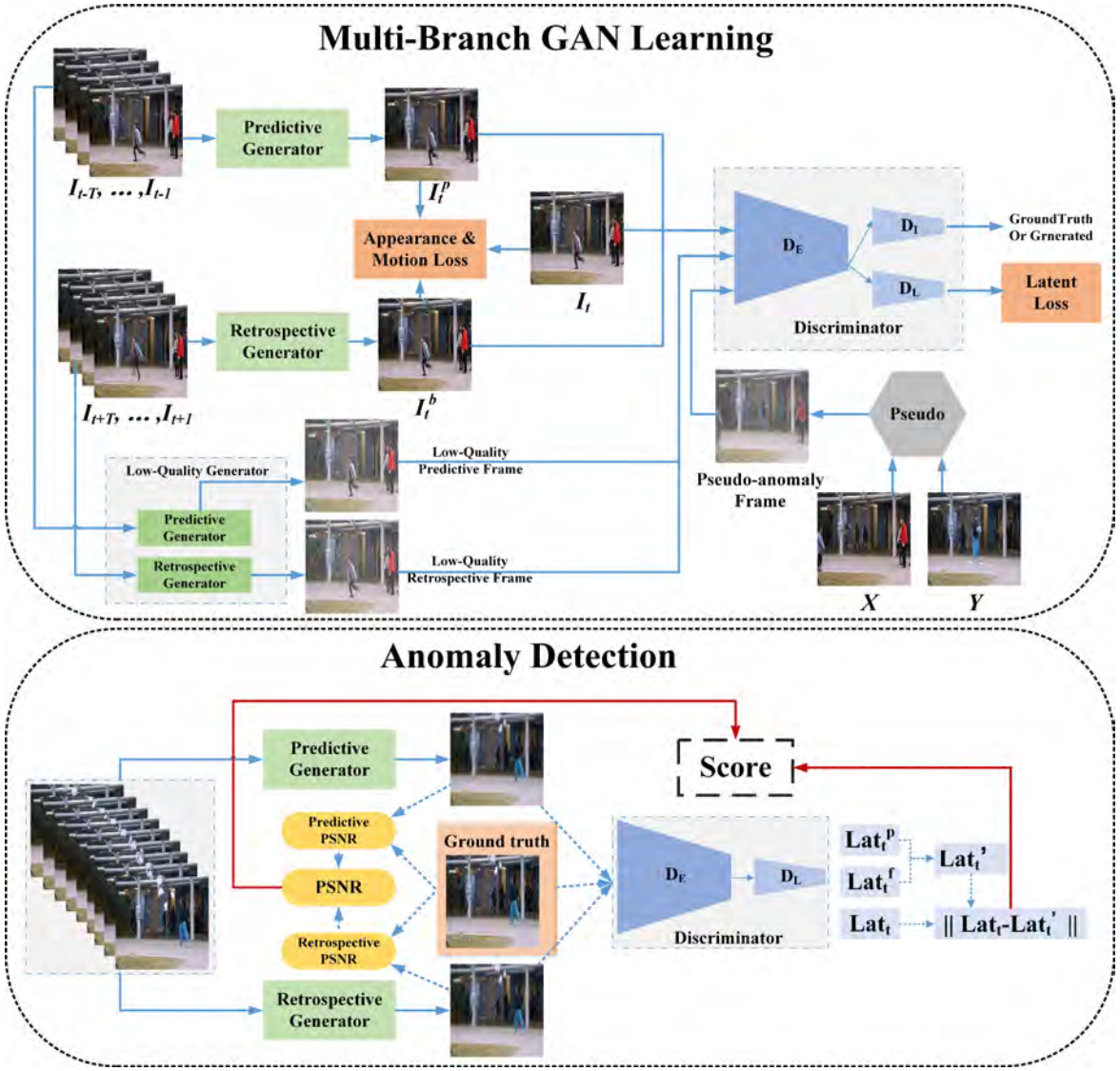


Fig. 2. The overall network structure of MGAN-CL.

TABLE I
SYMBOLS AND NOTIONS.

Symbol	Notion
G	Generators in the network
D	Discriminator in the network
G_p	Forward predict generator
G_b	Back retrospective generator
D_E	The main branch of the discriminator
D_L	The latent feature extraction branch of the discriminator
D_I	The image discrimination branch of the discriminator
I_t	Video frame at time t
I_t^p	The frame generated by the forward branch at time t
I_t^b	The frame generated by the retrospective branch at time t
I_t^{Pse}	Artificially generated pseudo anomalous video frames
I_{old}^p	The prediction frame of the low-quality generator
I_{old}^b	The retrospective frame of the low-quality generator

it first divides the video frame into $N \times M$ blocks, and then discriminates each block, and finally calculates the average score of the whole image as the final output. In our method, the discriminator consists of two branches: the image discrimination branch(D_E - D_I) and the latent feature branch(D_E - D_L). These two branches respectively design objective functions to optimize the discriminator, including adversarial loss and latent loss. Furthermore, we also add pseudo-anomaly loss to improve the robustness of the discriminator.

1) Training G: **Gradient loss**. The gradient of the image is the aspect ratio of the 2D image, adding a gradient loss could keep the dimensions of all frames. By comparing forward generated frames with retrospective generated frames, the generators can be encouraged to obtain more realistic video frames. So, we also add the error between the two generators in the loss function. The loss function of the two two-dimensional images is shown in formula (1):

$$\begin{aligned} L_{grad}(I_t, I_t^p, I_t^b) = \sum_{d\in\{x,y\}} (& \left\| |g_d(I_t^p)| - |g_d(I_t)| \right\|_1 \\ & + \left\| |g_d(I_t^b)| - |g_d(I_t)| \right\|_1 \\ & + \left\| |g_d(I_t^p)| - |g_d(I_t^b)| \right\|_1), \end{aligned} \tag{1}$$

where I_t^p and I_t^b respectively represent the prediction frame of the prediction branch generator and the retrospective frame of the retrospective branch generator at time t. I_t is the actual frame at time t, and d is an axis of the 2D image ($d \in x, y$).

Intensity loss. In grayscale images, the image intensity is the grayscale of the image. In the RGB color space, it can be considered as the pixel gray value of the R channel, the pixel gray value of the G channel, or the pixel gray value of the B channel. The image of the generated frame should be close to the image color of the ground truth, that is, constrain the distribution of pixels in the same RGB space. If the image intensity error between the generated frames and the ground truth is large, it means that the quality of the video frames generated by the generators is poor. Like gradient loss, we increase the error between forward generated frames and retrospective generated frames in intensity loss. The intensity loss function is shown in formula (2):

$$\begin{aligned} L_{int}(I_t, I_t^p, I_t^b) = &\left\|I_t^p - I_t\right\|_2^2 + \left\|I_t^b - I_t\right\|_2^2 \\ &+ \left\|I_t^p - I_t^b\right\|_2^2, \end{aligned} \tag{2}$$

where $\|\bullet\|_2^2$ represents the Euclidean distance.

Since the two losses are directly judged based on the image content of the video frame, the above two losses can be combined into an appearance loss of objects contained in the video frame. Combining the gradient loss and intensity loss, formula can be expressed as:

$$L_{appear}(I_t, I_t^p, I_t^b) = L_{grad}(I_t, I_t^p, I_t^b) + L_{int}(I_t, I_t^p, I_t^b), \tag{3}$$

constraints on the gradient and intensity loss terms ensure that the generated frames by the generator are closer to the real frames at the pixel level.

Optical Flow Loss. When an abnormal event occurs in an object in the video, the change of its optical flow will be particularly obvious compared with normal events. In our approach, optical flow represents the movement of a displaced object between two video frames in a video. Therefore, the optical flow can be considered as motion features of objects in video frames. If the quality of the video frames generated by the generators is low, resulting in an increase in the error of the optical flow, the optical flow loss can constrain the generators to generate more realistic video frames. Besides, we estimate optical flow based on an optical flow network (FlowNet2) [36], which is pre-trained and tuned. The loss function for calculating optical flow using the l_1 paradigm is shown in formula (4):

$$L_{flow}(F_t, F_t^p, F_t^b) = \left\|F_t^p - F_t\right\|_1 + \left\|F_t^b - F_t\right\|_1, \tag{4}$$

where F_t^p represents the optical flow between frames I_t^p and I_{t-1}, F_t^b denotes the optical flow between frames I_t^b and I_{t-1}, and F_t means the optical flow between frames I_t and I_{t-1}.

Adversarial loss. To enable the generator to obtain more realistic video frames, an adversarial loss is introduced in our method. The least squares GAN (LSGAN) is used for generative adversarial training, which can generate higher quality images than general GANs. Moreover, it is stable during learning [43]. Our proposed discriminator consists of two branches, the image discrimination branches and the latent feature branches. The former one includes D_E and D_I, while the latter one comprises D_E and D_L. As shown in Fig. 2, the discriminator in the training contains two branches. A 256×256 video frame is input into the discriminator. The outputs of the image discrimination branches are 35×35 image feature pixels, which are used to judge whether the video frame is ground truth. The output of the latent feature branch is a 128-dimensional latent feature vector, which completes the mapping of video frames in the latent space. The loss function of the adversarial loss of the generator G is shown in formula (5):

$$\begin{aligned} L_{adv}^G(I_t^p, I_t^b) = \sum_{x,y} \frac{1}{2} &((D_I(D_E(I_t^p))_{x,y} - 1)^2 \\ &+ (D_I(D_E(I_t^b))_{x,y} - 1)^2), \end{aligned} \tag{5}$$

where x, y represents the index number of the video frame block, and $D_I(D_E(\bullet))$ is the function of the discriminator backbone to extract features and judge whether it is a real frame by the image space branch.

Latent loss. The features of the latent space are also very important. If noise such as lens shake and light changes between two video frames that disturbs the feature extraction, the judgment in the image space will become very blurred. We can understand the pattern or structural similarity between video frames and ground truth by analyzing the features in the latent space, and even avoid noisethe influence. The latent space refers to the space of latent variables that are used to generate video frames in the GAN. The goal of latent loss is to encourage the generator network to produce video frames that are similar to the real frames in the latent space. The latent loss can help to capture the underlying patterns and structures in video data. Furthermore, latent loss can contribute to reducing the domain gap between the synthetic and the real video frames. The latent space contains the latent features of the image extracted by the self-encoder, which are advanced features. Under the constraints of multiple spaces, the latent features of the generator increase the discriminative ability of the discriminator, thus improving the authenticity of the generated video frames [44]. Using the l_2 paradigm to minimize the latent feature distance between the ground truth and the generated frame, the latent feature of the video frame is extracted by D_E and D_L, and the loss function of the latent feature of G is shown in formula (6):

$$\begin{aligned} L_{lat}^G(I_t, I_t^p, I_t^b) = &\left\|D_L(D_E(I_t^p)) - D_L(D_E(I_t))\right\|_2^2 \\ &+ \left\|D_L(D_E(I_t^b)) - D_L(D_E(I_t))\right\|_2^2 \\ &+ \left\|D_L(D_E(I_t^p)) - D_L(D_E(I_t^b))\right\|_2^2, \end{aligned} \tag{6}$$

where $D_L(D_E(\bullet))$ is the function of extracting features from the latent space after the discriminator backbone extracts features.

2) Training D: To improve the robustness of video anomaly detection, we adopt both the generator and the discriminator in the anomaly detection. Specifically, we add constraints of multi-space feature learning to the discriminator to improve the ability of the discriminator, and use the pseudo-anomaly module to obtain more anomalous data.

Multi-space features. The discriminator in a GAN has two tasks. The first task is to distinguish between the real and the fake images in the "image space". This means that it focuses on the actual image and decides whether it is real or fake. The second taks is to compare real and fake images in the "latent space". The latent space is a lower-dimensional space that contains the important features of the image, and it is learned by the generator. The discriminator compares the real and the fake images in this space to learn how similar they are. The combination of these two tasks constitutes multi-space learning. This means that the discriminator is learning to distinguish between the real and the fake images in different spaces (the image space and the latent space), which can help it to become more accurate to differentiate between real and fake images overall.

In the image space of the video frame, we input the video frame into the image discrimination branches of the discriminator to distinguish whether the video frame is an actual frame. If the discriminator judges the generated frame as ground truth, the corresponding loss will be huge, so the parameters of the discriminator will be strongly updated until the discriminator can distinguish the authenticity of the video frame well. The constraints on the image space are part of the multi-space features learning of the discriminator, and we adopt an adversarial loss, the adversarial loss function of discriminator D is shown in formula (7):

$$\begin{aligned} L_{adv}^{D}(I_t, I_t^p, I_t^b) = \sum_{x,y} \frac{1}{2}((D_I(D_E(I_t))_{x,y} - 1)^2 \\ +(D_I(D_E(I_t^p))_{x,y} - 0)^2 \\ +(D_I(D_E(I_t^b))_{x,y} - 0)^2), \end{aligned} \tag{7}$$

where x, y represents the index number of the video frame block.

The latent feature branches D_E and D_L of the discriminator extract the latent features of the two generated frames, and we compare them with the ground truth to obtain the error. Adding a latent loss is to make the latent feature encoding of the generated frame same with the ground truth, thereby constraining the connection between the generated frame and the ground truth. In multi-space features learning, the latent space of the image is constrained by the latent loss, the latent feature loss function of the discriminator D is shown in formula (8):

$$\begin{aligned} L_{lat}^{D}(I_t, I_t^p, I_t^b) = \left\| D_L(D_E(I_t^p)) - D_L(D_E(I_t)) \right\|_2^2 \\ + \left\| D_L(D_E(I_t^b)) - D_L(D_E(I_t)) \right\|_2^2 \\ + \left\| D_L(D_E(I_t^p)) - D_L(D_E(I_t^b)) \right\|_2^2. \end{aligned} \tag{8}$$

Pseudo-anomaly. For the discriminator, the data source mainly depends on the output of the generator. If the input of the discriminator increases through pseudo abnormal images, the discriminator will get better performance after training. Therefore, to improve the ability of the discriminator, we generate abnormal data based on the pseudo-anomaly module. Among them, the pseudo-anomaly module is divided into two parts. One are the outlier data we artificially create, and the other is to use the video frames predicted by the old generator, which is essentially a generator with poor performance early in the training, as fake anomalies.

Although these two parts are different in the way of generation, the essence of both is to improve their performance by training the discriminator. It is worth noting that artificially added anomalies and low-quality video frames are easily discriminated in image space. In the early stage of training, the discriminator has been trained to discriminate the image space of the video frame. If the discriminator continues to train the discriminator to discriminate the image space in the pseudo anomaly module, the performance of the discriminator will not be significantly improved, especially in the late training period. The discriminator can already tell whether a video frame is a generated frame or not. Therefore, in the pseudo-anomaly module, the discriminator only needs to train the ability of the discriminator to discriminate the latent space. Besides, the parameters of the old generator,which are saved, will not continue to be updated, and the discriminator has been updating the weight parameters, so it is of little significance to retrain the discriminator to the image space.

In the pseudo-anomaly module, the discriminator does not need to train the authenticity of video frames in the image space, because it focuses on distinguishing between the normal and the abnormal video frames in the latent space.

Given two random non-consecutive video frames I_x and I_y, the pseudo-abnormal data generation method is shown in formula (9):

$$I_t^{Pse} = \frac{I_x + I_y}{2}, x \neq y, \tag{9}$$

where x, y represents different video frame index numbers. The loss function of the pseudo-anomaly training discriminator is shown in formula (10):

$$L_{pseudo}(I_t, I_t^{Pse}) = \left\| D_L(D_E(I_t^{Pse})) - D_L(D_E(I_t)) \right\|_2^2, \tag{10}$$

where I^{Pse} is a randomly generated and constantly changing pseudo-anomaly frame, I_x and I_y must be non-consecutive different video frames to ensure that the generated spurious anomalies do not resemble normal events.

In addition to the pseudo-anomaly generated by the data of video frames, Zaheer et al. [41] believed that the low-quality video frames generated by low-quality generator can also be used as pseudo-anomaly data to train the discriminator. The parameters of the earlier low-quality generators are retained in our method, including the prediction branch G_{old}^{p} and the retrospective branch G_{old}^{b}. The loss function of training the discriminator which uses the video frames by the low-quality generator as pseudo-anomalies is shown in Equation (11):

$$\begin{aligned} L_{oldNum}(I_t, I_{old}^p, I_{old}^b) = \left\| D_L(D_E(I_{old}^p)) - D_L(D_E(I_t)) \right\|_2^2 \\ + \left\| D_L(D_E(I_{old}^b)) - D_L(D_E(I_t)) \right\|_2^2, \end{aligned} \tag{11}$$

where I_{old}^p and I_{old}^b are the prediction frame and the retrospective frame of the low-quality generator, respectively, and their generation methods are shown in formulas (12) and (13):

$$I_{old}^p = G_{old}^p(I_{t-T}, \cdots, I_{t-1}), \tag{12}$$

$$I_{old}^b = G_{old}^b(I_{t+T}, \cdots, I_{t+1}), \tag{13}$$

where T is the same as the input range T of the generators, $G^p_{old}(\cdots)$ and $G^b_{old}(\cdots)$ are functions of the old generators' prediction and backtracking video frames, respectively.

3) Objective Function: Combining all the loss function terms, the loss function G is shown in formula (14):

$$\begin{aligned} L_G = &\lambda_{appear} L_{appear}(I_t, I_t^p, I_t^b) \\ &+ \lambda_{flow} L_{flow}(F_t, F_t^p, F_t^b) \\ &+ \lambda_{adv}^G L_{lat}^G(I_t, I_t^p, I_t^b) \\ &+ \lambda_{lat}^G L_{lat}^G(I_t, I_t^p, I_t^b), \end{aligned} \tag{14}$$

where λ_{appear}, λ_{flow}, λ_{adv}^G, λ_{lat}^G are the weights of appearance loss, optical flow loss, generator adversarial loss and generator multi-space loss in the loss function L_G, respectively. The loss function D is shown in formula (15):

$$\begin{aligned} L_D = &\lambda_{adv}^D L_{adv}^D(I_t, I_t^p, I_t^b) \\ &+ \lambda_{lat}^D L_{lat}^D(I_t, I_t^p, I_t^b) \\ &+ \lambda_{pseudo} L_{pseudo}(I_t, I_t^{Pse}) \\ &+ \lambda_{oldNum} L_{oldNum}(I_t, I_{old}^p, I_{old}^b), \end{aligned} \tag{15}$$

where λ_{adv}^D, λ_{lat}^D, λ_{pseudo}, λ_{oldNum} are the weights of the discriminator adversarial loss, the discriminator multispace loss, the pseudo-anomaly module loss and the low-quality generator module loss in the loss function L_D, respectively.

C. Anomaly Detection

In our method, whether the video frame is abnormal is judged by the normal score, which is jointly calculated by peak signal-to-noise ratio (PSNR) and the error of latent features. When the similarity between the generated frame and the ground truth is higher, the normal score of the video frame is higher. On the contrary, the smaller the similarity between the generated frame and the actual frame is, the lower the normal score. In context learning, we calculate score in both the image and latent spaces.

We refer to previous work [36] and use PSNR as the discrepancy score between ground truth and generated frames in image space. We also utilize the PSNR to assess the quality of the generated frames:

$$\begin{aligned} PSNR_t(I_t, I_t^p, I_t^b) = 10(&\frac{1}{2}\log_{10}\frac{[max_{I_t^p}]^2}{\frac{1}{H\times W}\sum_{i,j}(I_t(x,y) - I_t^p(x,y))^2} \\ &+ \frac{1}{2}\log_{10}\frac{[max_{I_t^b}]^2}{\frac{1}{H\times W}\sum_{i,j}(I_t(x,y) - I_t^b(x,y))^2}), \end{aligned} \tag{16}$$

where x and y represent the spatial index of the frames, and $[max_I]$ is the maximum of I. Besides, H and W are the height and width of the frame, respectively.

Furthermore, we add the latent feature error as part of the normal score to judge the score of the video frame. We apply the l_2 distance to calculate the distance of ground truth and generated frames:

$$Dis_t(Lat_t, Lat_t^p, Lat_t^b) = \frac{1}{2}\left\|Lat_t^p - Lat_t\right\|_2^2 + \frac{1}{2}\left\|Lat_t^b - Lat_t\right\|_2^2, \tag{17}$$

where Lat_t^p, Lat_t^b and Lat_t represent the latent vector of I_t^p, the latent vector of I_t^b and the vector of the corresponding real frame I_t, respectively.

In order to facilitate the identification of abnormal events, we normalize the $PSNR_t$ and Dis_t, and calculate the final normal score by weighted:

$$Score(I_t) = \lambda PSNR_t(I_t, I_t^p, I_t^b) + (1-\lambda) Dis_t(Lat_t, Lat_t^p, Lat_t^b), \tag{18}$$

where λ is the weighting parameter controlling the importance of the score functions. Considering that the image space and latent space of the video frame are equally important in the anomaly detection stage, we set λ as 0.5. The $PSNR_t$ and Dis_t are normalized elements. If the normal score is greater than a pre-defined threshold, the video frame is determined as a normal event; otherwise, it is determined as an abnormal event.

IV. Experiments

A. Experiments Setting

These are surveillance video datasets for anomaly detection: UCSD Dataset, CHUK Avenue Dataset, UMN Dataset, Subway entrance/exit dataset, ShanghaiTech Dataset. In our method, three publicly accessible datasets in video anomaly detection are used, namely CHUK Avenue [45], UCSD Ped2 [46] and ShanghaiTech Campus [47]. This is because these few datasets are commonly found in semi-supervised methods. Among them, the UCSD data set contains two data sets, Ped1 and Ped2. There are two reasons why our method only selects the Ped2 data set for experiments: 1) The people in the Ped1 data set are facing or away from the camera, and the optical flow calculation network FlowNet2 used in our method is not suitable for objects with long distances [34]; and 2) many researchers use Ped1 as a dataset for pixel-level VAD evaluation [47], while this work focuses on frame-level abnormal event detection. The Fig. 3 shows the normal and abnormal events in the datasets.

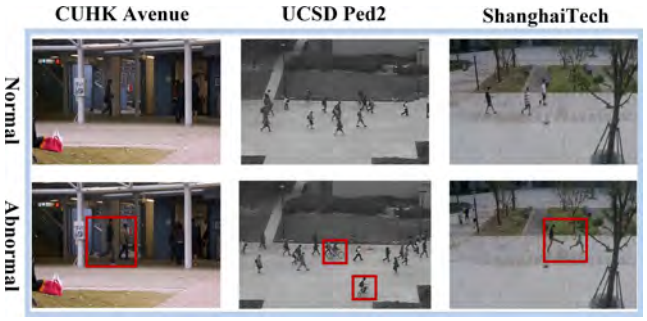
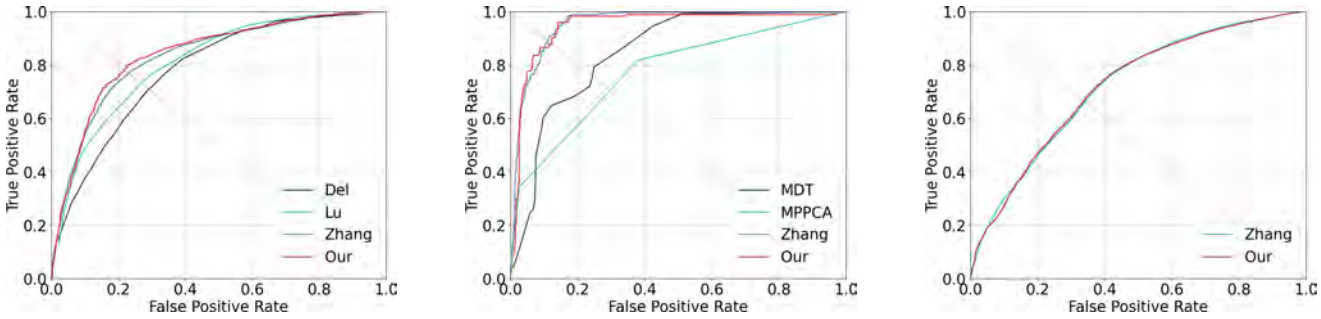


Fig. 3. Normal and abnormal events in the datasets. The abnormal event that occurred is indicated in the red box.

- CUHK Avenue dataset contains 16 training videos and 21 test videos with a total of 47 abnormal events, including throwing objects, running, and hanging out.
- UCSD Ped2 data set is a small-scale sample data set, including 16 training videos and 12 test videos with 12 abnormal events.



(a) CUHN Avenue (b) UCSD Ped2 (c) ShanghaiTech Campus

Fig. 4. ROC on the dataset CUHN Avenue, UCSD Ped2, and ShanghaiTech Campus.

- ShanghaiTech Campus data set is a large-scale data set. There are a total of 437 videos with 130 abnormal events, a total of 330 videos in the training set, and a total of 107 videos in the test set. In addition, there are a total of 13 different scenes in the video, in which abnormal events include slapstick, sprinting, skateboarding and even bicycles.

In this experiment, all the video frame size is 256×256, and the channel intensities of the images are normalized to $[-1, 1]$. The batchsize is set to 4, and the network is trained using the *Adam* algorithm. The length T of the video segment is set to 4. At this time, the total length of the video frame input is 9.

For the dataset, the learning rate of the generator is set to 2×10^{-4} for the color video dataset, while the learning rate of the discriminator is set to 2×10^{-5}. For datasets with only gray channels, the learning rate is set to 1×10^{-4} for the generator and 1×10^{-5} for the discriminator. During the experiments in MGAN-CL, λ_{lat}^{D} in L_D is set to 1.0, λ_{adv}^{D}, λ_{adv}^{G} are set to 0.05, and λ_{pseudo} is set to 5.0. The parameters λ_{appear} and λ_{lat}^{G} and λ_{lat}^{D} are all set to 1.0, while λ_{flow} is set to 2, λ_{oldNum} is set to 0.5, and the Epoch of the low-quality generator is set to 50.

TABLE II
COMPARISON OF AUC RESULTS IN CUHKAVENUE, UCSD PED2 AND SHANGHAITECH.

Methods	CUHK Avenue	UCSD Ped2	ShanghaiTech Campus
MPPCA [49]	N/A	0.683	N/A
MDT [46]	N/A	0.829	N/A
Lu et al. [45]	0.809	N/A	N/A
Del et al. [50]	0.783	N/A	N/A
Unmasking [51]	0.806	0.822	N/A
Conv-AE [29]	0.800	0.850	0.609
ConvLSTM-AE [52]	0.770	0.881	N/A
StackRNN [47]	0.817	0.922	0.680
AbnormalGAN [53]	N/A	0.935	N/A
MemAE [54]	0.833	0.941	0.712
Autoregressive [55]	N/A	0.954	0.725
STAE [34]	0.809	0.912	N/A
Liu et al. [36]	0.849	0.954	0.728
sRNN-AE [57]	0.835	0.922	0.696
Fan et al. [59]	0.834	0.922	N/A
Zhang et al. [40]	0.868	0.954	0.736
Chang et al. [58]	0.860	**0.965**	0.733
Georgescu et al. [60]	0.869	0.924	**0.835**
FastAno [61]	0.853	0.963	0.722
Li et al. [62]	**0.871**	0.963	0.736
MGAN-CL	**0.871**	**0.965**	0.736

B. Results and Analysis

The experiments in our method use Area Under the Curve (AUC) and Receiver Operating characteristic Curve (ROC) as evaluation metrics. In frame-level VAD, it is very intuitive to use AUC to evaluate the pros and cons of a method [41]. Higher AUC means better video anomaly detection ability. We evaluate the method using frame-level AUC. The results of AUC and ROC are respectively shown in Table II and Fig. 4, where it can be seen that our method achieves a better AUC performance on CHUK Avenue and UCSD Ped2. On the ShanghaiTech Campus dataset, there are too many scenes that are constantly switched, including scenes similar to the camera in the Ped1 dataset, and our FlowNet2 works poorly for objects with long distances and small targets [34]. The optical flow feature cannot exert its maximum advantage, making our results inferior to other methods on the ShanghaiTech Campus dataset.

In the proposed method MGAN-CL, score can clearly determine whether an abnormal event has occurred in the current video frame. As is shown in Fig. 5, when a person suddenly throws his item, or a sprinter appears in the square, the normal score will fluctuate strongly. Thus, something unusual has happened in the current frame. Besides, we use score gap to evaluate the performance of our method, the larger the score gap, the more obvious the effect of distinguishing normal events from abnormal events. High score gap indicates algorithm is better in distinguishing between normal and abnormal behavior. Score gap is useful for optimizing the threshold value for detecting anomalies. The results of score gap are shown in Table III, where the normal and abnormal scores are the average values in different datasets, and the scores are varying from 0 to 1. It can be seen that the score gaps are obvious in all datasets, indicating that the proposed method has good performance in distinguishing normal and abnormal events.

In addition, we compared the previous methods [62], after normalizing the score gap of the previous method, the score gap of the three datasets Avenue, UCSD Ped2, and

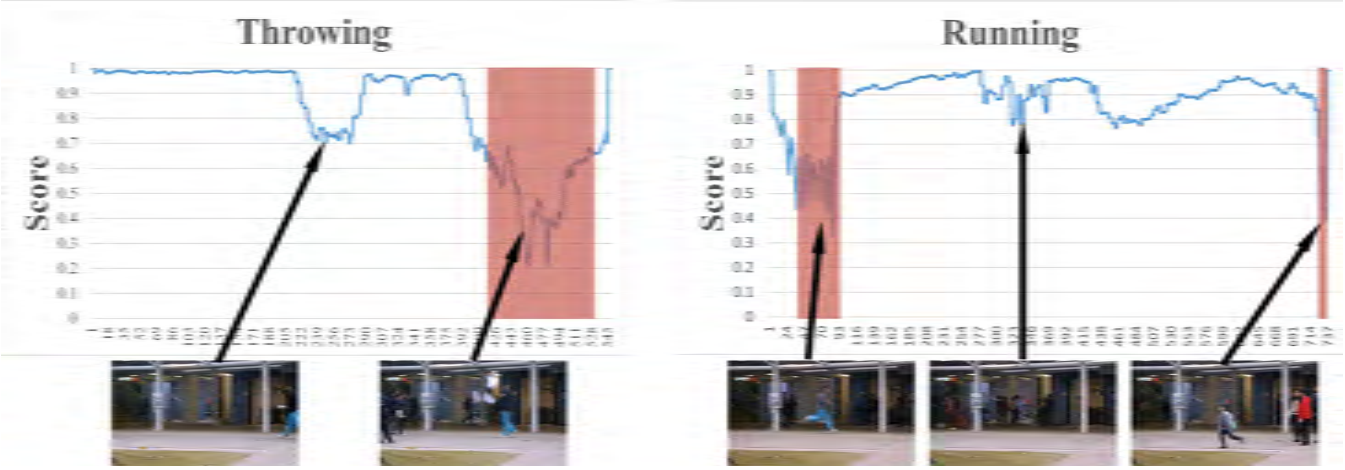


Fig. 5. (left column)When the person in the video walks normally, the score will remain at a high value, but when he suddenly throws into the air, the score changes abruptly. (right column)When people walk normally on the square, the score fluctuates but does not change drastically. However, when there are sprinters, the score begins to shake violently. Where the horizontal axis is the index of the video frame, and the vertical axis is the value of normal score. The higher the score, the more likely it is to be a normal event, and vice versa. Best viewed in color.

TABLE III
THE DIFFERENCE IN SCORES BETWEEN NORMAL AND ABNORMAL EVENTS IN EACH DATASET.

Datasets	Normal Score	Abnormal Score	Score Gap
Avenue	0.79	0.48	**0.31**
UCSD Ped2	0.87	0.36	**0.51**
ShanghaiTech	0.71	0.49	**0.22**

ShanghaiTech are 0.29, 0.54, and 0.21, respectively, without adding the pseudo-abnormal module. In contrast, after we added the fake exception module, the score gap reaches 0.31, 0.51 and 0.22 respectively. Among them, the score gap of Avenue and ShanghaiTech have been improved, which shows the effectiveness of the method.

Furthermore, we compared the speed of alternate optimization iterations during the training phase. Epoch refers to the unit for our model to be alternately optimized, that is, 1 Epoch represents that our model has been alternately optimized once. When the model reaches convergence, the smaller the Epoch, the faster the model converges. In the previous method [62], when the three data sets, Avenue, UCSD Ped2, and ShanghaiTech reach 44,000, 60,000, and 10,000 Epochs during training, the model is converged and achieves better accuracy. However, in our method, the three datasets Avenue, UCSD Ped2, and ShanghaiTech reach convergence when training to 39,000, 58,000, and 10,000 Epochs. It is obvious that our method improves the speed of model convergence after adding the pseudo-anomaly module.

C. Ablation Study

Ablation experiments are carried out in this study, and the results are shown in Table IV. The AUC_{single} is the result without multi-branch GAN, meaning that only the forward generator is used. The $AUC_{Bilateral}$ is the result with multi-branch GAN, which means that both the forward generator and the retrospective generator are combined, but there is no pseudo anomaly modules. The $AUC_{MGAN-CL}$ refers to the use of multi-branch GAN context learning and pseudo anomaly modules. In the ablation experiments, we used the same set of parameter settings for the three methods. Comparing AUC_{single} and $AUC_{MGAN-CL}$, it can be clearly seen that when we use same loss function, the performance of the multi-branch GAN is significantly improved, which can prove that our proposed method is meaningful to the performance improvement. Moreover, all constraints of MGAN-CL have a certain impact on the results. For each additional constraint, the performance of our method can be improved by 0.2% to 0.3%. It can be seen that each constraint is significant, and their roles are indispensable.

For the loss terms $\||g_d(I_t^p)|-|g_d(I_t^b)|\|_1$ and $\|I_t^p-I_t^b\|_2^2$ in Equations 1 and 2, we have added this constraint content and given the specific ablation experimental results in Table V. In order to avoid errors, we did a total of five experiments, allowing us to get a result range. And these results are obtained after the model is converged. As can be seen from this table, in the absence of both $\|I_t^p-I_t^b\|_2^2$ and $\||g_d(I_t^p)|-|g_d(I_t^b)|\|_1$, the result can only be between 0.954 and 0.958. When there is only $\||g_d(I_t^p)|-|g_d(I_t^b)|\|_1$, the result reaches between 0.956 and 0.960. Similarly, when only $\|I_t^p-I_t^b\|_2^2$ is used, the result reaches between 0.957 and 0.962.

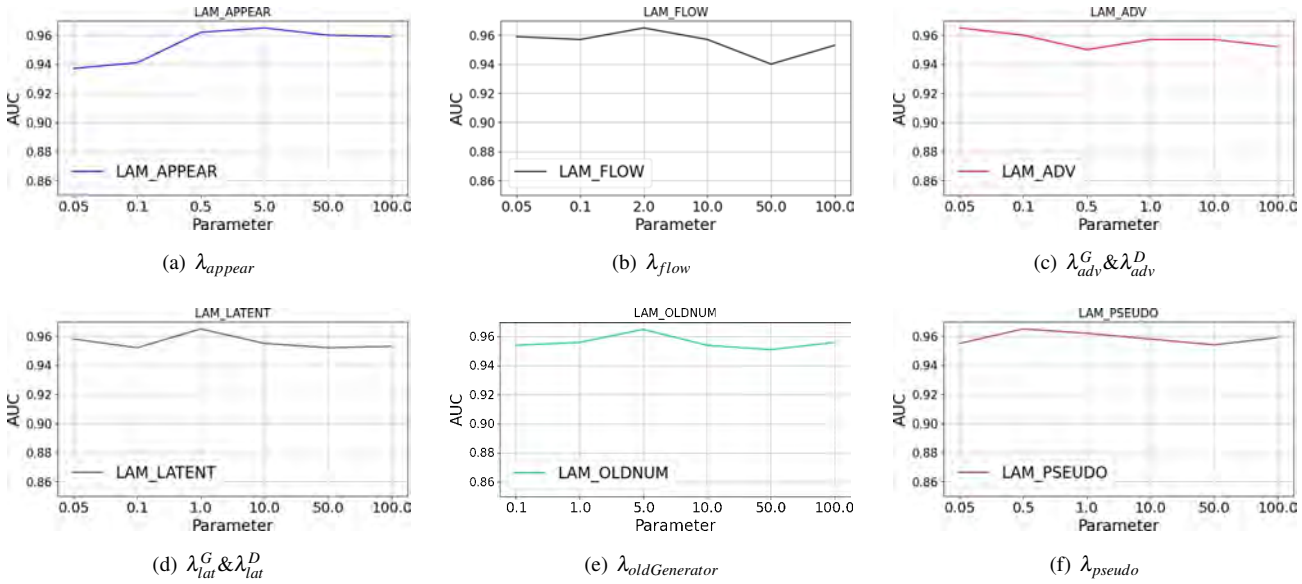
D. Parameter Sensitivity

In our method, the sensitivity experiment of model parameters is conducted. Generally speaking, with the change of parameters, the smaller the method performance changes, the stronger the stability of the method. The results of parameter sensitivity experiment are shown in Fig. 6. Among them, the changes of λ_{appear} and λ_{flow} have a great influence on the experimental results, indicating that appearance and action have a great effect on the detection of abnormal video events. In order to enhance the ability of the discriminator, we used the pseudo-anomaly module for training. In addition,

TABLE IV

COMPARISON OF AUC RESULTS IN CUHKAVENUE, UCSD PED2 AND SHANGHAITECH. THE AUC_{single} IS THE RESULT WITHOUT MULTI-BRANCH GAN, THE $AUC_{Bilateral}$ IS THE RESULT WITH MULTI-BRANCH GAN, AND $AUC_{MGAN-CL}$ IS THE RESULT WITH MULTI-BRANCH GAN AND PSEUDO ANOMALY MODULES. AMONG THEM, L_{pseudo} REPRESENTS THE PSEUDO-ABNORMAL MODULE, WHICH CONTAINS TWO PARTS OF L_{pseudo} AND L_{oldNum}.

L_{appear}	L_{flow}	L_{adv}	L_{lat}	L_{pseudo}	AUC_{Single}	$AUC_{Bilateral}$	$AUC_{MGAN-CL}$
✓	✗	✗	✗	✗	0.936	0.949	**0.954**
✓	✓	✗	✗	✗	0.938	0.953	**0.957**
✓	✓	✓	✗	✗	0.942	0.954	**0.960**
✓	✓	✓	✓	✗	0.951	0.959	**0.962**
✓	✓	✓	✓	✓	-	-	**0.965**



(a) λ_{appear} (b) λ_{flow} (c) λ^G_{adv}&λ^D_{adv}

(d) λ^G_{lat}&λ^D_{lat} (e) $\lambda_{oldGenerator}$ (f) λ_{pseudo}

Fig. 6. Parameter Sensitivity Experiment. In the Ped2 dataset, changes in the values of hyperparameters λ_{appear}, λ_{flow}, λ^G_{adv}, λ^D_{adv}, λ^G_{lat}, λ^D_{lat}, $\lambda_{oldGenerator}$, and λ_{pseudo} correspond to changes in AUC.

TABLE V

COMPARISONS BETWEEN AUC RESULTS IN UCSD PED2. $|||g_d(I_t^p)| - |g_d(I_t^b)|||_1$ AND $\|I_t^p - I_t^b\|_2^2$ ARE TERMS FROM EQUATIONS 1 AND 2, RESPECTIVELY.

$\|I_t^p - I_t^b\|_2^2$	$\|\|\|g_d(I_t^p)\| - \|g_d(I_t^b)\|\|\|_1$	$AUC_{MGAN-CL}$
✗	✗	$0.954 \sim 0.958$
✗	✓	$0.956 \sim 0.960$
✓	✗	$0.957 \sim 0.962$
✓	✓	**0.965**

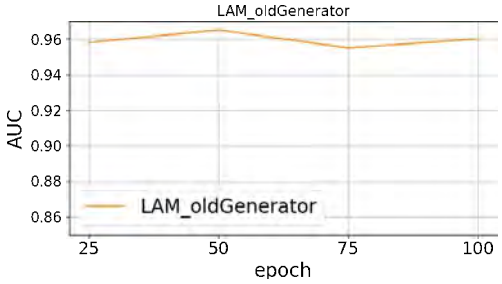


Fig. 7. Accuracy of old generators for different epochs.

the low-quality video frames generated by the early low-quality generators G^p_{old} and G^b_{old} can also be used as pseudo-anomaly data to train the discriminator in our method [41]. We conducted a series of experiments on how to select early low-quality generators G^p_{old} and G^b_{old}. As shown in Fig. 7, among the multiple generator choices of the early epoch, the experimental results vary less.

E. Detection Time

Using early low-quality generators and pseudo-anomaly modules to train the discriminator, the training time increases to a certain extent. This method can enhance the ability to discriminate abnormal events, and the stability of the method is improved. It is worth mentioning that the average detection rate of all data sets can reach 30 frames per second(FPS) with NVIDIA GeForce RTX 2080 Ti, and the detection efficiency can meet the application requirements.

V. CONCLUSION

In this study, we propose a multi-branch GAN with video context learning for abnormal event detection, which can capture the features of normal events based on the bidirectional contextual. In the proposed method, two generators are designed for predictive and retrospective tasks, two branches are

designed in the discriminator to extract original image features and latent semantics, and the pseudo-anomaly module is added to enhance the discriminative anomaly event capability. In addition, in VAD, we adopt the trained discriminator to improve the accuracy of the model for video anomaly event detection. Nevertheless, our method still has some limitations, for example, it is a frame-level anomaly event detection, which cannot accurately detect anomalous objects in the video frame. In the future work, we will focus on pixel-level anomaly event detection by combining video object detection, and further, use pre-trained optical flow technology to improve detection efficiency. Moreover, we will configure more data set parameters and adopt more public data sets to conduct more experimental comparisons.

REFERENCES

[1] C. Piciarelli, C. Micheloni and G. L. Foresti, "Trajectory-Based Anomalous Event Detection", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544-1554, Nov. 2008.

[2] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes", *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18-32, 2013.

[3] J. Wang and Z. Xu, "Crowd anomaly detection for automated video surveillance", in *6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15)*, pp. 1-6, 2015.

[4] A. S. Rao, J. Gubbi, S. Rajasegarar, S. Marusic and M. Palaniswami, "Detection of Anomalous Crowd Behaviour Using Hyperspherical Clustering", in *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-8, 2014.

[5] W. Huan, H. Guo and X. Wu, "Saliency attention based abnormal event detection in video", in *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*, pp. 1039-1043, 2014.

[6] T. Wang and H. Snoussi, "Detection of Abnormal Visual Events via Global Optical Flow Orientation Histogram", *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988-998, June 2014.

[7] S. Wang, E. Zhu, J. Yin, and F. Porikli, "Video anomaly detection and localization by local motion based joint video representation and OCELM", *Neurocomputing*, vol. 277, pp. 161-175, 2018.

[8] C. Li, Z. Han, Q. Ye, and J. Jiao, "Visual abnormal behavior detection based on trajectory sparse reconstruction analysis", *Neurocomputing*, vol. 119, pp. 94-100, 2013.

[9] X. Mo, V. Monga, R. Bala and Z. Fan, "Adaptive Sparse Representations for Video Anomaly Detection", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 631-645, April 2014.

[10] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade and W. R. Schwartz, "Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 673-682, March 2017.

[11] C. G. Blair and N. M. Robertson, "Video Anomaly Detection in Real Time on a Power-Aware Heterogeneous Platform", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2109-2122, Nov. 2016.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1933-1941, 2016.

[13] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition", in *European conference on computer vision (ECCV)*, pp. 20-36, Springer, Cham, October 2016.

[14] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, "Deep local video feature for action recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPR)*, pp. 1-7, 2017.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", in *Proceedings of the IEEE international conference on computer vision*, pp. 4489-4497, 2015.

[16] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "Anopcn: Video anomaly detection via deep predictive coding network", in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1805-1813, October 2019.

[17] K. Xu, T. Sun, and X Jiang, "Video anomaly detection and localization based on an adaptive intra-frame classification network", in *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 394-406, Feb 2020.

[18] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pp. 13588-13597, 2021.

[19] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? a new dataset for anomaly detection in driving videos", in *arXiv preprint*, arXiv:2004.03044, 2020.

[20] W. Wang, F. Chang, and H. Mi, "Intermediate fused network with multiple timescales for anomaly detection", in *Neurocomputing*, 433, 37-49, 2021.

[21] K. Doshi, and Y. Yilmaz, "Continual learning for anomaly detection in surveillance videos", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops(CVPR)*, pp. 254-255, 2020.

[22] B. Ramachandra, and M. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2569-2578, 2020.

[23] M. Pranav, and L. Zhenggang, "A day on campus-an anomaly detection dataset for events in a single camera", in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[24] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision", in *European conference on computer vision(ECCV)*, Springer, Cham, pp. 322-339, August, 2020.

[25] W. Ren, G. Li, B. Sun, and K. Huang, "Unsupervised kernel learning for abnormal events detection", *The Visual Computer*, vol. 31, no. 3, pp. 245-255, 2015.

[26] M. Sabokrou, M. Fayyaz, M. Fathy and R. Klette, "Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes", in *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992-2004, April 2017.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", *Advances in neural information processing systems*, vol. 28, 2015.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 779-788, 2016.

[29] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 733-742, 2016.

[30] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4975-4986, 2021.

[31] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. S. Hua, "Spatio-temporal autoencoder for video anomaly detection", in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1933-1941, October, 2017.

[32] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3379-3388, 2018.

[33] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds", in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1896-1904, January, 2019.

[34] T. N. Nguyen, and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence", in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 1273-1283, 2019.

[35] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6479-6488, 2018.

[36] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline", in *Proceedings of the IEEE*

conference on computer vision and pattern recognition (CVPR), pp. 6536-6545, 2018.

[37] J. Yu, Y. Lee, K. C. Yow, M. Jeon and W. Pedrycz, "Abnormal Event Detection and Localization via Adversarial Event Prediction", in *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241, Springer, Cham, October, 2015.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842-7851, 2019.

[40] Y. Zhang, X. Nie, R. He, M. Chen and Y. Yin, "Normality Learning in Multispace for Video Anomaly Detection", in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3694-3706, Sept. 2021.

[41] M. Z. Zaheer, J. H. Lee, M. Astrid, and S. I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14183-14193, 2020.

[42] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1125-1134, 2017.

[43] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks", in *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 2794-2802, 2017.

[44] 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 *International conference on information processing in medical imaging*, pp. 146-157, Springer, Cham, June, 2017.

[45] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab", in *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 2720-2727, 2013.

[46] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, "Anomaly detection in crowded scenes", in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1975-1981, 2010.

[47] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework", in *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 341-349, 2017.

[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, ..., and Y. Bengio, "Generative adversarial nets", *Advances in neural information processing systems*, vol. 27, 2014.

[49] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates", in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2928, 2009.

[50] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos", in *European conference on computer vision (ECCV)*, pp. 334-349, Springer, Cham, October, 2016.

[51] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video", in *Proceedings of the IEEE international conference on computer vision*, pp. 2895-2903, 2017.

[52] W. Luo, W. Liu and S. Gao, "Remembering history with convolutional LSTM for anomaly detection", in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 439-444, 2017.

[53] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni and N. Sebe, "Abnormal event detection in videos using generative adversarial nets", in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1577-1581, 2017.

[54] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. V. D. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1705-1714, 2019.

[55] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 481-490, 2019.

[56] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu and R. S. M. Goh, "AnomalyNet: An Anomaly Detection Network for Video Surveillance", in *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537-2550, Oct. 2019.

[57] W. Luo et al., "Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1070-1084, 1 March 2021.

[58] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection", in *European Conference on Computer Vision (ECCV)*, pp. 329-345, Springer, Cham, August 2020.

[59] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder", *Computer Vision and Image Understanding*, vol. 195, pp. 102920, 2020.

[60] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12742-12752, 2021.

[61] C. Park, M. Cho, M. Lee, and S. Lee, "FastAno: Fast anomaly detection via spatio-temporal patch transformation", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2249-2259, 2022.

[62] D. Li, X. Nie, X. Li, Y. Zhang, and Y Yin, "Context-related video anomaly detection via generative adversarial network", in *Pattern Recognition Letters (PRL)*, vol. 156, pp. 183-189, 2022.