IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 30, 2021, accepted July 20, 2021, date of publication July 27, 2021, date of current version August 9, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3100678*

# Bidirectional Retrospective Generation Adversarial Network for Anomaly Detection in Videos

**ZHIWEI YANG, JING LIU, (Senior Member, IEEE), AND PENG WU**
Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China

Corresponding author: Jing Liu (neouma@163.com)

This work was supported in part by the Key Project of Science and Technology Innovation 2030 through the Ministry of Science and Technology of China under Grant 2018AAA0101302, and in part by the General Program of National Natural Science Foundation of China (NSFC) under Grant 61773300.

**ABSTRACT** Anomaly detection in videos is the task of identifying frames from a video sequence that depict events that do not conform to expected behavior, which is an extremely challenging task due to the ambiguous and unbounded properties of anomalies. With the development of deep learning, video anomaly detection methods based on deep neural networks have made great progress. The existing methods mainly follow two routes, namely, frame reconstruction and frame prediction. Due to the powerful generalization ability of neural networks, the application of reconstruction-based methods is limited. Recently, anomaly detection methods based on prediction have achieved advanced performance. However, their performance suffers when they cannot guarantee lower prediction errors for normal events. In this paper, we propose a novel future frame prediction model based on a bidirectional retrospective generation adversarial network (BR-GAN) for anomaly detection. To predict a future frame with higher quality for normal events, first, we propose a bidirectional prediction combined with a retrospective prediction method to fully mine the bidirectional temporal information between the predicted frame and the input frame sequence. Then, the intensity and gradient loss between the predicted frame and the actual frame together with an adversarial loss are used for appearance (spatial) constraints. In addition, we propose a sequence discriminator composed of a 3-dimensional (3D) convolutional neural network to capture the long-term temporal relationships between frame sequences composed of predicted frames and input frames; this network plays a crucial role in maintaining the motion (temporal) consistency of the predicted frames for normal events. Such appearance and motion constraints further facilitate future frame prediction for normal events, and thus, the prediction network can be highly capable of distinguishing normal and abnormal patterns. Extensive experiments on benchmark datasets demonstrate that our method outperforms most existing state-of-the-art methods, validating the effectiveness of our method for anomaly detection.

**INDEX TERMS** Anomaly detection, future frame prediction, generation adversarial network, video surveillance.

## I. INTRODUCTION

Detecting abnormal behavior in video surveillance plays a vital role in maintaining public safety. To date, video surveillance systems have been gradually deployed across our whole society, and the amount of surveillance video data has increased dramatically. It is quite time-consuming and inefficient to rely on manual detection of abnormal events from massive surveillance video data. Therefore, the study of automatic detection algorithms for abnormal behavior in video surveillance has become a research hotspot. However, it is an extremely challenging task because abnormal events are ambiguous and unbounded, which makes it difficult to clearly define abnormal events. In addition, abnormal events rarely occur, and thus, it is extremely difficult to collect abnormal samples from a large amount of surveillance video data for the learning of algorithms. In contrast, the acquisition of ordinary moments in video surveillance is much easier. Thus, semi-supervised learning is one of the most common methods for anomaly detection, where only normal data

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


This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

are available in the training set under the semi-supervision. To determine whether an abnormal event occurs, a common method is to exploit normal patterns according to their appearance and motion in the training set. Any pattern that is inconsistent with the normal patterns is classified as an abnormal pattern.

Semi-supervised anomaly detection methods can be roughly divided into two categories: reconstruction-based methods and prediction-based methods. The previous reconstruction methods [2], [3], [5], [6], [10] rely on hand-crafted appearance and motion features. Then, a dictionary is learned to sparsely encode all normal events with small reconstruction errors by learning those features. During testing, the features that correspond to abnormal events could lead to a large reconstruction error. However, the optimization of sparse coefficients is extremely time-consuming, and the performance of anomaly detection is governed by hand-crafted features, and these two reasons make the method of dictionary learning suffer from limitations. Later, with the development of deep learning, some studies [7]–[9], [11]–[13], [24], [25] have used deep neural networks to automatically extract deep features instead of hand-crafted features. For example, Hasan *et al*. in [9] proposed to use an auto-encoder to reconstruct normal events with small reconstruction errors. However, the deep neural networks have strong generalization capabilities, with the result that some abnormal frames can also be reconstructed with small errors. With regard to recently proposed prediction-based methods, their core idea is to build a model that can accurately predict the future frame of normal events, whereas the unpredictability of abnormal events will cause large prediction errors. Inspired by the good performance of the video prediction model in [14], Liu *et al*. in [15] proposed a future frame prediction framework (FFP) for anomaly detection. The FFP uses U-Net to predict the future frame based on the observed past frames. The model is optimized according to the intensity and gradient difference between the predicted frame and actual frame as well as the optical flow together with an adversarial loss. A larger difference between the predicted future frame and the actual future frame indicates a possible abnormal event in the frame. This method achieves good performance in public datasets, which manifests the potential of prediction-based methods for anomaly detection.

Although prediction-based methods have achieved impressive results on public datasets, their performance suffers when they cannot guarantee lower prediction errors for normal events. Their performance has mainly suffered from the following limitations. 1) The inverse temporal information between the predicted frame and the input frames is ignored. As a kind of time-series image data, the video frame sequence contains both forward and inverse mapping relationships. However, the previous work only performed forward prediction in the training phase and ignored the inverse mapping relationships among the video frame sequences, which limited the quality of the predicted frame. 2) The motion constraint on the prediction frame is not considered from the perspective of long-term temporal consistency. The previous methods mainly pretrained an optical flow extraction network to estimate the optical flow and used the optical flow loss as a motion constraint. However, the optical flow can only represent the short-term temporal relationship between two adjacent frames, and computing the optical flow is time-consuming. The long-term temporal relationship of events that occurs in a video sequence is usually ignored.

In this paper, we propose a novel future frame prediction framework called a bidirectional retrospective generative adversarial network (BR-GAN) for anomaly detection, which moves beyond these limitations. The BR-GAN consists of a generator, a frame discriminator, and a sequence discriminator composed of 3-dimensional (3D) convolutional neural networks, which can be trained end-to-end. The overall architecture of BR-GAN is shown in Fig. 1. This framework has the following traits: 1) it can fully explore the bidirectional mapping relationship among video frame sequences to more accurately establish the mapping model from a few frames in the past to a future frame for normal events; 2) the motion constraint can be performed from the perspective of long-term temporal consistency to ensure that the predicted frame and the real frame are consistent with normal events, in terms of the motion.

Specifically, to fully mine the bidirectional mapping relationship among video frame sequences, we propose a bidirectional prediction combined with a retrospective prediction method. The bidirectional prediction means that the generator performs both forward and backward prediction. The retrospective prediction uses the previously predicted frames as input for review verification, and the main idea of retrospective prediction is that if the predicted frame is realistic, the frame that is predicted again using the previously predicted frame as input should also be realistic. More specifically, in the bidirectional prediction process, forward prediction is to predict a future frame by observing a few frames in the past, and backward prediction is to predict a past frame by observing future frames. Then, the forward predicted frame and the backward predicted frame are respectively added to the corresponding input sequence to perform retrospective prediction for review verification. Next, the appearance constraint is imposed by minimizing the intensity loss and gradient loss between the predicted frame and the actual frame. To further improve the robustness of the generator and the quality of the predicted frames, we also combine a frame discriminator to determine whether the image produced by the generator is real or fake, which discourages the generator from producing obviously fake images.

Furthermore, to make the predicted frame consistent with the real object in motion, we propose a sequence discriminator composed of 3D convolutional neural networks to capture the long-term temporal information among video frame sequences. More specifically, we input a video cube

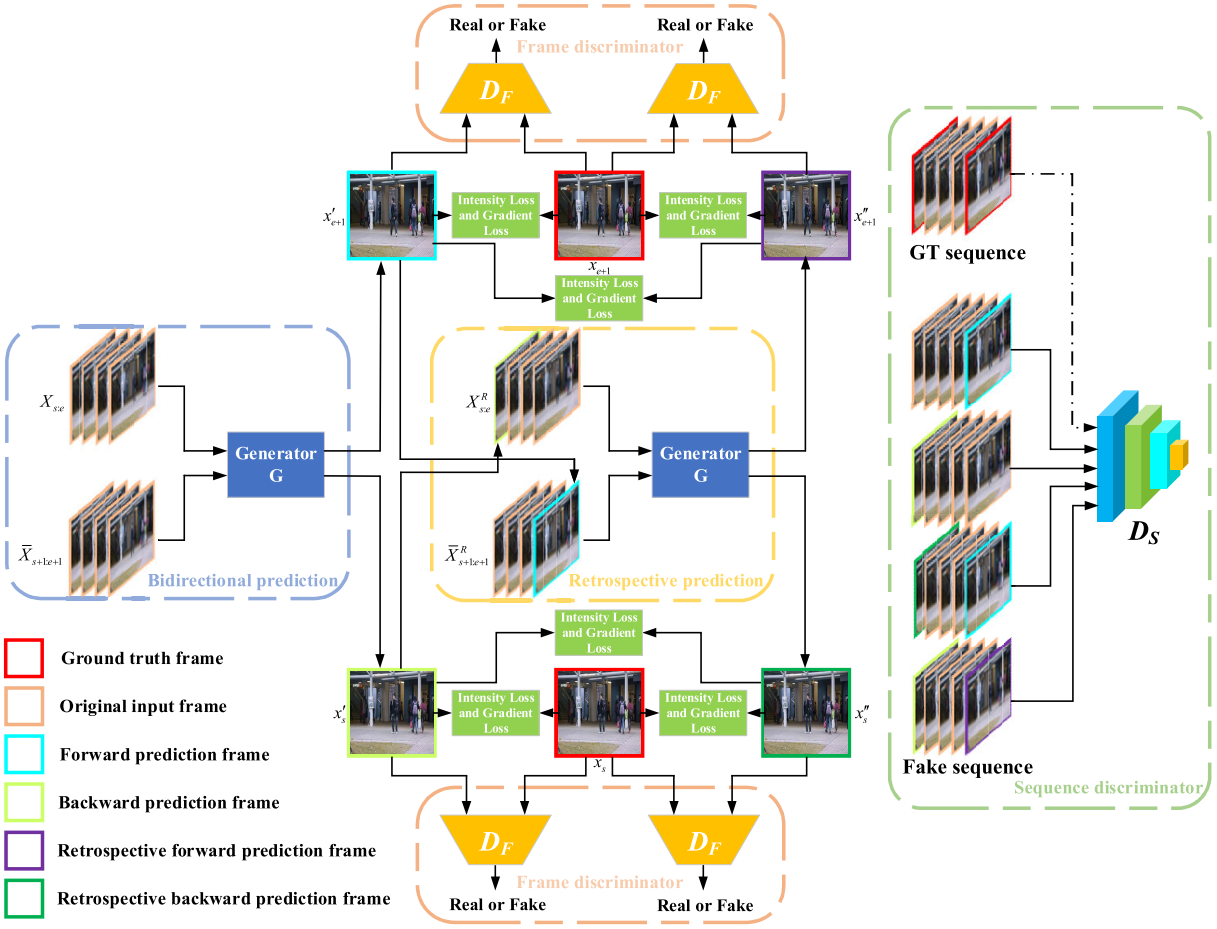



**FIGURE 1.** An overview of our proposed BR-GAN model. The BR-GAN consists of a generator $G$, a frame discriminator $D_F$, and a sequence discriminator $D_S$ composed of a 3D convolutional neural network. First, the generator performs bidirectional prediction, including forward and backward prediction. Then, the predicted frames are added to the corresponding input sequence to perform retrospective prediction for review verification. The intensity loss and gradient loss are used to minimize the prediction errors between the predicted frames and the actual frames. Furthermore, the frame discriminator is used to determine whether the image produced by the generator is real or fake, and it discourages the generator from producing obviously fake images. The sequence discriminator is used to promote the temporal consistency of the predicted frame by identifying whether the sequence contains fake frames or not.

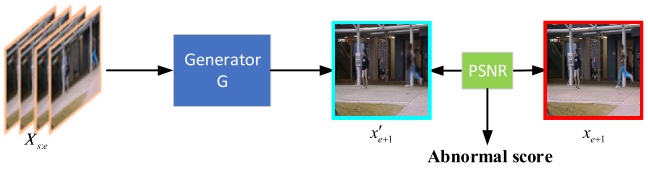



**FIGURE 2.** Pipeline of our proposed anomaly detection method in the testing stage.

composed of a predicted frame and the corresponding input frames into the sequence discriminator, and then, the authenticity evaluation loss is used to constrain the motion. In the testing phase, the generator U-Net predicts the future frame based on its history observation, as shown in Fig. 2. Because of the unpredictability of abnormal events, a large error between the predicted future frame and the actual future frame indicates the presence of possible abnormal events in the frame.

To summarize, the main contributions of our work are listed as follows.

1) We propose a novel future frame prediction framework based on a bidirectional retrospective generative adversarial network that can be trained end-to-end for anomaly detection.
2) We propose a bidirectional prediction combined with a retrospective prediction method to fully mine the bidirectional mapping relationship between video frame sequences, and our method achieves more accurate future frame prediction of normal events.
3) We propose a sequence discriminator composed of 3D convolutional neural networks to capture the long-term temporal relationships between frame sequences composed of predicted frames and input frames, which play a crucial role in maintaining the motion consistency of predicted frames for normal events.

4) Our method outperforms most existing state-of-the-art methods, and we also provide an extensive experimental analysis with ablation studies.

The remainder of this paper is organized as follows. Section II reviews the related work on the detection of anomalous events. Our approach for anomaly detection is introduced in detail in Section III. The experiments on four benchmark datasets are given in Section IV. Finally, our conclusions are given in Section V.

## II. RELATED WORK

Over the past few decades, a large number of research studies [1]–[13], [15], [18]–[23], [26], [32]–[36], [40], [48], [49], [53]–[55] have been devoted to solving the video anomaly detection problem. All of these methods can be roughly categorized into traditional methods, deep learning-based methods, and other classic methods.

### A. TRADITIONAL ANOMALY DETECTION

Traditional anomaly detection methods mainly use hand-crafted features or features learned from the training set by traditional machine learning techniques to construct feature space, and then, they characterize the distribution of normal or anomaly scenarios and, finally, identify isolated clusters or outliers as abnormal. Statistical models and sparse coding are common modeling methods that are used in traditional anomaly detection.

#### 1) STATISTICAL MODEL-BASED ANOMALY DETECTION METHODS

Early work based on statistical models usually utilizes dynamic trajectory features [56], [57] of an object to represent the normal patterns. For example, Hu *et al*. in [57] proposed a method to automatically learn motion patterns for anomaly detection from trajectory features that are clustered hierarchically using spatial and temporal information. Piciarelli *et al*. in [58] presented an online trajectory cluster method for the identification of abnormal events, and the proposed method builds the clusters as the data are acquired by the tracking system. However, the trajectory features based on object tracking can fail in crowded scenes, and thus, these methods are not robust in complex scenes. To alleviate this shortcoming, low-level spatial-temporal features, such as the histogram of oriented gradients (HOG) [28] and histogram of oriented flows (HOF) [29], have gradually replaced trajectory features for anomaly detection. Based on low-level spatial-temporal features, Adam *et al*. [59] characterized the normal local histograms of optical flows based on statistical monitoring of low-level observations at multiple spatial locations. Kim and Grauman [22] modeled the local optical flow pattern with a mixture of probabilistic principal component analyzers (MPPCA) and trained a space-time Markov random field (MRF) to infer abnormalities. Mahadevan *et al*. [33] fitted a Gaussian mixture model to a mixture of dynamic textures (MDT).

#### 2) SPARSE CODING-BASED ANOMALY DETECTION METHODS

The sparse coding-based methods proposed early [2], [3], [6], [10], [48], [49] usually learn a dictionary from hand-crafted features to reconstruct normal events with small reconstruction errors, and the events with high reconstruction errors are marked as abnormal. For example, Cong *et al*. in [2] introduced the sparse reconstruction cost (SRC) over the normal dictionary to measure the normality of testing samples. Zhao *et al*. in [3] proposed a fully unsupervised dynamic sparse coding approach for detecting unusual events in videos based on the online sparse reconstruction of query signals from an atomically learned event dictionary. To solve the time-consuming problem of optimizing sparse coefficients to accelerate training and testing, Lu *et al*. in [10] proposed an efficient sparse combination learning (SCL) framework.

These traditional methods have achieved good results in specific scenarios, but due to their weak feature expression capabilities, the performance of these methods is greatly reduced in some complex scenarios. In addition, the high computational complexity of some traditional methods also makes it difficult to meet the requirements of anomaly detection from videos in real-time.

### B. DEEP LEARNING-BASED ANOMALY DETECTION

In recent years, deep learning methods have achieved great success in various fields, such as image classification [27], [60], object detection [68], [69], and video retrieval [61], [62]. Many methods based on deep learning have been proposed for anomaly detection from videos. In general, these methods can be roughly divided into two categories: reconstruction-based methods and prediction-based methods.

#### 1) RECONSTRUCTION-BASED ANOMALY DETECTION METHODS

The reconstruction-based method is similar to the earlier sparse coding-based method. Owing to the superiority of deep features [9], [16], [25], [27], [53] over previous hand-crafted features [28], [29], [50], recent years have witnessed advancements in anomaly detection by adopting deep features in reconstruction. For example, Wu *et al*. in [21] proposed a two-stream neural network to extract spatial-temporal fusion features (STFF) in hidden layers and used a fast sparse coding network (FSCN) to build a normal dictionary based on the STFF. Hasan *et al*. in [9] used the extracted feature as input to a fully connected neural network-based autoencoder to learn the temporal regularity in the videos. A regularity score was calculated according to the reconstruction error and used to determine whether an abnormality occurs. However, this method does not mine the motion information due to the 2D convolution. Then, Xu *et al*. in [1] proposed a stacked denoising autoencoder to separately learn both the appearance and the motion features. Luo *et al*. presented a convolution autoencoder combined with long short-term

memory (LSTM) in [11]. Although deep neural networks have strong feature extraction capabilities, they also have good capacity and generalization capabilities, which make the reconstruction errors of normal and abnormal events have little difference. To address this problem, Park *et al*. [12] used a memory module to record a prototypical pattern of normal data, while lessening the representation capacity of CNNs.

#### 2) PREDICTION-BASED ANOMALY DETECTION METHODS

Video prediction methods have developed rapidly in recent years because they can use massive unlabeled data to learn the internal representation of the video, and they have a wide range of applications in the fields of robot decision-making, autonomous driving, and video understanding. The goal of the video prediction task is to predict future frames based on the past frames observed in the video. In [14], Mathieu *et al*. proposed a multi-scale network with adversarial training to generate future frames given an input sequence. D'Avino *et al*. [30] proposed a recurrent autoencoder using an LSTM that models temporal dependences between patches from a sequence of input frames to detect video forgery. The promising performance of these video prediction methods shows that regular events can be predicted. In contrast, abnormal events are rare events that are difficult to predict. Under this assumption, the idea of video prediction began to be applied to the field of video anomaly detection. For example, a convolution feature representation was fed into an LSTM model to predict the latent space representation, and its prediction error was used to evaluate anomalies in a robotics application [31]. An end-to-end trainable composite convolutional long short-term memory (Conv-LSTM) network was proposed in [7] to predict the evolution of a video sequence. Recently, Liu *et al*. in [15] presented a video anomaly detection method that predicts the future frame with the U-Net architecture, and the network is optimized according to the difference between the predicted frame and the actual frame as well as the optical flow together with an adversarial loss. Ye *et al*. in [32] proposed a predictive coding network for anomaly detection, which consists of a predictive coding module and an error refinement module. Our proposed anomaly detection method is also based on the prediction framework. However, the difference is the following: 1) we have performed both forward prediction and backward prediction combined with retrospective prediction, which can more fully mine the bidirectional mapping relationship between video frame sequences. 2) We use a sequence discriminator composed of 3D convolutional neural networks to capture the long-term temporal relationship between frame sequences composed of predicted frames and input frames, which promotes the temporal consistency of the predicted frame.

### C. OTHER CLASSIC METHODS

Sun *et al*. [34] introduced a novel unsupervised model called online growing neural gas (online GNG) to learn surveillance scenes by continuously updating the learning parameters. Chaker *et al*. [35] modeled the local and global behaviors pattern with a social network to detect abnormal behaviors. Sabokrou *et al*. [51] introduced a DNN-based cascade classifier for fast anomaly detection by hierarchically modeling normal patches using deep features and Gauss distributions. In another work [20], Wu *et al*. proposed a novel deep one-class (DeepOC) neural network, which can simultaneously learn a one-class classifier and compact features using convolutional neural networks. Recently, Pang *et al*. [4] introduced a novel end-to-end approach to unsupervised video anomaly detection based on self-trained ordinal regression. Cai *et al.* [63] model the consistency between normal video appearances and motion to tackle the video anomaly detection by an appearance-motion memory consistent network. Georgescu *et al.* [64] proposed a novel multi-task learning method that integrates a set of self-supervised and knowledge distillation proxies in a single object-centric architecture for video anomaly detection. Lu *et al.* in [65] proposed a few-shot scene-adaptive approach based on meta-learning to detect anomalies in a previously unseen scene with only a few frames.

## III. PROPOSED METHOD

The anomaly detection method based on future frame prediction usually trains a future frame prediction network by learning the video data of normal behavior, which can predict the future frame of normal behavior well according to the previous frames. In contrast, because abnormal events are unpredictable, the future frame generated through the prediction network can generate larger prediction errors, and thus, abnormal events can be detected. At present, some existing studies [15], [32], [38] have adopted a method based on the prediction of future frames to detect abnormal events. The promising performance of these methods manifests the potential of prediction-based methods for anomaly detection. However, their performance suffers when they cannot guarantee lower prediction errors for normal events. If the future frames of normal events with higher quality can be predicted, the prediction network can be more capable of distinguishing normal and abnormal patterns. Thus, a novel future frame prediction framework called BR-GAN for anomaly detection is proposed. In this section, the overall architecture of BR-GAN and the specific implementation process are described in detail.

### A. NETWORK ARCHITECTURES

The BR-GAN consists of a generator, a frame discriminator, and a sequence discriminator composed of a 3D convolutional neural network. The overall architecture of BR-GAN is shown in Fig. 1.

#### 1) GENERATOR

We use the U-Net [39] network architecture as the generator of our prediction framework. U-Net is transformed from a fully convolutional network and consists of a symmetrical down-sampling and up-sampling network. The details of the

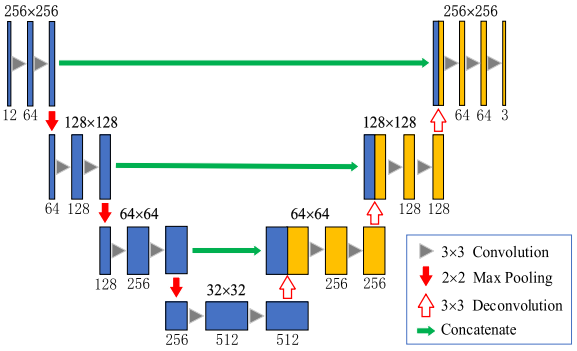



**FIGURE 3.** The architecture of our generator network (U-Net). The resolutions of the images of the same layer are equal.

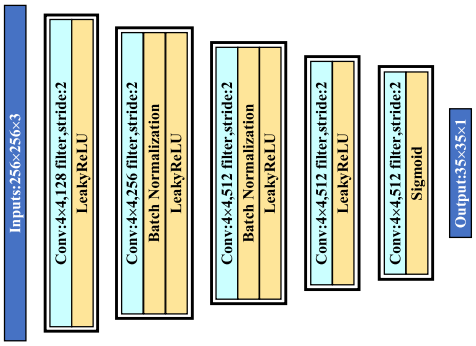



**FIGURE 4.** The architecture of the frame discriminator network.

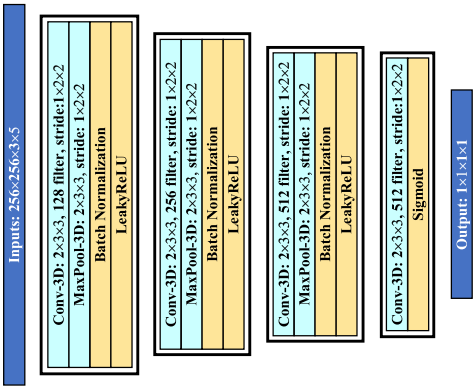



**FIGURE 5.** The architecture of the sequence discriminator network.

U-Net network are illustrated in Fig. 3. First, the left half of the network is a down-sampling component composed of multiple convolutional layers and pooling layers, which are used to extract features from the input video sequence. The right half of the network is symmetrical with the left and is an up-sampling component, which is used to generate a predicted image from the features by gradually increasing the spatial resolution. In addition, there is a skip connection operation in the U-Net network; in other words, the feature map of each downsampling on the left side and the corresponding feature map on the right side are concatenated. This design enables the output to share the low-level information of the input images and avoids the problem of gradient vanishing and information imbalance in each layer, which contributes to the prediction since the adjacent frames usually have the same background and similar foreground.

#### 2) FRAME DISCRIMINATOR

The role of frame discriminator is to compare the predicted future frame with the actual future frame to determine whether the input comes from a real distribution or is generated by the generator, which can effectively improve the robustness of the generator and the quality of the predicted frame through a dynamic game with the generator. Following the work in [37], we use the patchGAN as the frame discriminator in our work. The difference from the conventional GAN discriminator is that patchGAN produces a matrix in which each element determines whether the patch is real or fake instead of mapping an input image to a single scalar output in the range of [0, 1]. We set the size of the output matrix to $35 \times 35$ in our implementation. The details of the frame discriminator network are illustrated in Fig. 4.

#### 3) SEQUENCE DISCRIMINATOR

A sequence discriminator is used to promote the temporal consistency of the predicted frame by identifying whether the frame sequence contains fake frames or not, which can encourage the generator to predict video frames that are consistent with the real sequence in the temporal relationship. Inspired by the work in [52], we design a sequence discriminator that consists of a four-layer 3D convolutional network, which can capture the long-term temporal relationship among frame sequences composed of predicted frames and input frames. The input of the sequence discriminator is a video cube composed of predicted frames and input frames, and the output is a scalar to indicate the possibility that the input is a realistic sequence. To reduce the number of parameters, we performed a max pooling operation on each layer. The kernel sizes are set to be $2 \times 3 \times 3$ for both the convolution layers and max pooling layers. The details of the sequence discriminator network are illustrated in Fig. 5.

### B. BIDIRECTIONAL RETROSPECTIVE PREDICTION OF THE GENERATOR

In this subsection, we describe the implementation process of the bidirectional retrospective prediction of the generator in detail. For a clearer explanation, the generator, frame discriminator, and sequence discriminator are represented by $G$, $D_F$, and $D_S$, respectively. First, $G$ performs bidirectional prediction, including forward and backward prediction. Mathematically, the input sequences of the forward prediction and backward prediction are defined as follows:

$$X_{s:e} = \{x_s, x_{s+1}, \ldots, x_{e-1}, x_e\} \quad \text{s.t. } s < e \tag{1}$$

$$\overline{X}_{s+1:e+1} = \{x_{e+1}, x_e, \ldots, x_{s+2}, x_{s+1}\} \quad \text{s.t. } s < e \tag{2}$$

where $x_i \in \mathbb{R}^2$ represents a frame image and $s$ and $e$ are indices of the start frame and end frame. To be consistent

with [15], we set the input sequence length to 4, *i.e.*, $e - s + 1 = 4$. $X_{s:e}$ is organized in chronological order, and $\overline{X}_{s+1:e+1}$ is organized in reverse chronological order. In the forward prediction, $X_{s:e}$ is fed into $G$ to predict a future frame $x_{e+1}$. Here, the predicted future frame is denoted as $x'_{e+1}$. In the backward prediction, $\overline{X}_{s+1:e+1}$ is fed into $G$ to predict a past frame $x_s$. Similarly, the predicted past frame is denoted as $x'_s$. Then, $x'_{e+1}$ and $x'_s$ are added to the corresponding input sequence to perform retrospective prediction. Specifically, the input sequence that contains $x'_{e+1}$ and $x'_s$ is respectively defined as follows:

$$\overline{X}^R_{s+1:e+1} = \left\{x'_{e+1} \cup \overline{x}_{s+1:e}\right\} \tag{3}$$

$$X^R_{s:e} = \left\{x'_s \cup x_{s+1:e}\right\} \tag{4}$$

$\overline{X}^R_{s+1:e+1}$ is organized in reverse chronological order, which is fed into $G$ to predict the past frame $x_s$ again. The predicted past frame is denoted as $x''_s$. $X^R_{s:e}$ is organized in chronological order, which is fed into $G$ to predict the future frame $x_{e+1}$ again. Similarly, the predicted future frame is denoted as $x''_{e+1}$.

### C. CONSTRAINTS ON THE APPEARANCE AND MOTION

During the bidirectional and retrospective prediction process, multiple prediction frames are generated. To make these predicted frames close to actual frames, following the work in [14], intensity loss and gradient loss are used to constrain the appearance of the predicted frames. The intensity loss penalizes pixel differences between the predicted frame and the actual frame to guarantee the similarity of all pixels in RGB space, and the gradient loss penalizes gradient differences between them to sharpen the predicted frame. Specifically, the intensity loss function is defined as follows:

$$L_{int} = \sum_{(p,g)\in I^{pair}_{s,e}} f_1(p, g) \tag{5}$$

where $f_1(\cdot, \cdot)$ represents the L2 distance between two images, which is defined in (6). $I^{pair}_{s,e}$ is a set of image pairs, which is defined in (7).

$$f_1(p, g) = \|p - g\|_2^2 \tag{6}$$

$$\begin{aligned} I^{pair}_{s,e} = \{&(x_s, x'_s), (x_s, x''_s), (x'_s, x''_s), \\ &(x_{e+1}, x'_{e+1}), (x_{e+1}, x''_{e+1}), (x'_{e+1}, x''_{e+1})\} \end{aligned} \tag{7}$$

Here, the intensity loss function minimizes the L2 distance of six pairs of images in $I^{pair}_{s,e}$. The pair $(x_{e+1}, x'_{e+1})$, which consists of the forward prediction frame and the actual frame, is used to minimize the forward prediction errors. Similarly, $(x_s, x'_s)$ is used to minimize the backward prediction errors. Because $x''_s$ is predicted by $x'_{e+1}$ and $x''_{e+1}$ is predicted by $x'_s$, $(x_{e+1}, x''_{e+1})$ and $(x_s, x''_s)$ are used to minimize the retrospective prediction errors. In other words, if the predicted frame $x'_{e+1}$ is realistic, then the past frame $x''_s$ predicted by $x'_{e+1}$ in the reverse-look-back form should also be realistic. Both $x'_{e+1}$ and $x''_{e+1}$ are future frames generated by forward prediction and retrospective prediction, respectively, and thus, they should be close. here, $(x'_{e+1}, x''_{e+1})$ is used to minimize the error between them, and similarly, for $(x'_s, x''_s)$. These six pairs of constraints enable the generator to fully mine the bidirectional mapping relationship between the frame sequences composed of the predicted frames and input frames.

Further, the gradient loss function is defined as follows:

$$L_{gd} = \sum_{(p,g)\in I^{pair}_{s,e}} f_2(p, g) \tag{8}$$

$$\begin{aligned} f_2 = &\sum_{i,j} \left\| \left|p_{i,j} - p_{i-1,j}\right| - \left|g_{i,j} - g_{i-1,j}\right| \right\|_1 \\ &+ \left\| \left|p_{i,j} - p_{i,j-1}\right| - \left|g_{i,j} - g_{i,j-1}\right| \right\|_1 \end{aligned} \tag{9}$$

where $i, j$ denote the spatial index of a video frame, and $f_2(\cdot, \cdot)$ represents the gradient differences between two images. Similarly, the gradient loss function minimizes the gradient differences of six pairs of images in $I^{pair}_{s,e}$.

Furthermore, in the task of video anomaly detection based on the future frame prediction, the motion consistency of objects in the predicted frame is an important factor for modeling the motion pattern of normal events. In the previous work [15], the motion constraint is usually performed by enforcing the optical flow between predicted frames and ground truth frames to be consistent. However, the optical flow can only represent the short-term temporal relationship between two adjacent frames, and it cannot reflect the long-term temporal consistency. In addition, pre-training the optical flow extraction network to calculate the optical flow is time-consuming. Therefore, we design a sequence discriminator that is composed of 3D convolutional neural networks to directly capture the long-term temporal relationship between the frame sequences, which are composed of predicted frames and input frames. The sequence discriminator encourages the generator to predict the video frames that are consistent with the real sequence by identifying whether the frame sequence containing the predicted frame is real or fake, which plays a crucial role in maintaining the motion consistency of normal future frames.

### D. ADVERSARIAL TRAINING

The GAN has a wide range of applications in the field of image and video generation [41]–[46]. Usually, a GAN is composed of a generator and a discriminator. The discriminator identifies the frames produced by the generator, and the generator produces frames that can deceive the discriminator. The model is trained by alternately updating the discriminator and generator. In our proposed method, BR-GAN consists of a generator $G$ and two discriminators $D_F$ and $D_S$. Similar to training the usual GAN, we use alternate training to optimize our model BR-GAN, and the specific training process is described as follows:

**Training *G*.** The goal of training $G$ is to make the frame predicted by it be consistent with the actual frame in appearance and motion by observing the past few frames. Ideally, both $D_F$ and $D_S$ would classify the predicted frame as 1, where 1 represents real labels. When we train $G$, the parameters of $D_F$ and $D_S$ are fixed, and two mean square error (MSE)

loss functions from $D_F$ and $D_S$ are imposed as follows:

$$L^G_{adv_D_F} = \sum_{\hat{x}\in\varsigma}\sum_{i,j}\frac{1}{2}L_{MSE}(D_F(\hat{x})_{i,j}, 1) \tag{10}$$

$$L^G_{adv_D_S} = \sum_{\hat{X}\in\zeta}\frac{1}{2}L_{MSE}(D_S(\hat{X}), 1) \tag{11}$$

in (10), where $\varsigma = \{x'_s, x''_s, x'_{e+1}, x''_{e+1}\}$, and $i, j$ denote the spatial patch indexes. $L_{MSE}$ is an MSE function, which is defined as follows:

$$L_{MSE}(\hat{y}, y) = (\hat{y} - y)^2 \tag{12}$$

where $y$ takes values in $\{0, 1\}$ and $\hat{y} \in [0, \quad 1]$. In (11), $\zeta$ is defined as follows:

$$\begin{aligned}\zeta = \{&X_{s:e} \cup x'_{e+1}, x'_s \cup X_{s+1:e+1},\\ &x''_s \cup X_{s+1:e} \cup x'_{e+1}, x'_s \cup X_{s+1:e} \cup x''_{e+1}\}\end{aligned} \tag{13}$$

where $\zeta$ includes four sets of sequences composed of the prediction frames and the input frames, and $\cup$ represents the concatenation operation.

**Training $D_F$** The goal of training $D_F$ is to classify actual frames $x_s$ and $x_{e+1}$ into class 1 and predicted frames $x'_s, x''_s, x'_{e+1}$ and $x''_{e+1}$ into class 0, where 0 and 1 represent fake and real labels, respectively. When we train $D_F$, the parameters of $G$ and $D_S$ are fixed, and a frame adversarial loss function $L^{D_F}_{adv}$ composed of MSE is imposed, as follows:

$$L^{D_F}_{adv}(\hat{x}, x) = \sum_{(\hat{x},x)\in\chi} l_{D_F}(\hat{x}, x) \tag{14}$$

in (14), $\chi$ is defined as follows:

$$\chi = \{(x'_{e+1}, x_{e+1}), (x'_s, x_s), (x''_{e+1}, x_{e+1}), (x''_s, x_s)\} \tag{15}$$

Additionally, $l_{D_F}$ is defined as follows:

$$l_{D_F}(\hat{x}, x) = \sum_{i,j}\frac{1}{2}L_{MSE}(D(x)_{i,j}, 1) + \sum_{i,j}\frac{1}{2}L_{MSE}(D(\hat{x})_{i,j}, 0) \tag{16}$$

where $i, j$ denote the spatial patch indexes.

**Training $D_S$** The goal of training $D_S$ is to classify a real sequence $X_{s:e+1}$ into class 1 and a fake sequence in (13) into class 0, where 0 and 1 are similar to the definition above. Similarly, when training $D_S$, the parameters of $G$ and $D_F$ are fixed, and a sequence adversarial loss function $L^{D_S}_{adv}$ composed of the MSE is imposed as follows:

$$\begin{aligned}L^{D_S}_{adv}(\hat{X}, X) = \sum_{X=X_{s:e+1},\hat{X}\in\zeta} &(\frac{1}{2}L_{MSE}(D_S(X), 1)\\ &+ \frac{1}{2}L_{MSE}(D_S(\hat{X}), 0))\end{aligned} \tag{17}$$

### E. OBJECTIVE FUNCTION

The overall objective functions include three parts: generation loss, frame adversarial loss, and sequence adversarial loss. When we train $G$, combining the appearance constraint, motion constraints, and adversarial loss, the generation loss function is defined as follows:

$$\begin{aligned}L_G = \lambda_{int}L_{int} + \lambda_{gd}L_{gd} + \lambda_{adv_D_F}L^G_{adv_D_F}\\ + \lambda_{adv_D_S}L^G_{adv_D_S}\end{aligned} \tag{18}$$

where $\lambda_{int}$, $\lambda_{gd}$, $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ are non-zero hyperparameters for balancing the four loss terms. When we train $D_F$, the frame adversarial loss function is defined as follows:

$$L_{D_F} = L^{D_F}_{adv}(\hat{x}, x) \tag{19}$$

when we train $D_S$, the sequence adversarial loss function is defined as follows:

$$L_{D_S} = L^{D_S}_{adv}(\hat{X}, X) \tag{20}$$

### F. ANOMALY DETECTION

#### 1) ANOMALY DETECTION ON TESTING DATA

In the testing phase, only the generator $G$ is used to predict future frames. By learning from a large amount of normal video data, generator $G$ has learned to predict future frames of normal behavior based on the past few frames. For abnormal behavior, for example, a car suddenly appears on the sidewalk; in this case, the corresponding prediction frame will have a large error. The error between the predicted frame and the real frame is used for anomaly detection. Following the work in [14], we use the peak signal-to-noise ratio (PSNR) to evaluate the quality of the predicted images. The calculation of PSNR is shown as follows:

$$PSNR(x, \hat{x}) = 10\log_{10}\frac{[\max_{\hat{x}}]^2}{\frac{1}{N}\sum_{i=0}^{N}(x_i - \hat{x}_i)^2} \tag{21}$$

where $\max_{\hat{x}}$ is the maximum value of the image pixels, and $N$ is the total number of image pixels. A higher PSNR indicates that the predicted frame is normal behavior, and vice versa. To further quantify the possibility of abnormalities, following the work in [14], we normalize the PSNR to obtain an anomaly score $S(t)$ in the range of [0, 1] for each frame in the test video by using the following equation:

$$S(t) = 1 - \frac{PSNR(x_t, \hat{x}_t) - \min_t PSNR(x_t, \hat{x}_t)}{\max_t PSNR(x_t, \hat{x}_t) - \min_t PSNR(x_t, \hat{x}_t)} \tag{22}$$

#### 2) ANOMALY DETECTION ON REAL DATA

The ultimate goal of studying anomaly detection algorithms is to apply them in actual scenarios. Since real-time detection is required in practical applications, we cannot normalize the PSNR to obtain an anomaly score in the range of [0,1] according to (22). We directly detect whether an abnormal event occurs based on the PSNR. To set an appropriate threshold, we pre-calculate the average PSNR on the normal video clips of the training set by using the following equation:

$$PSNR_{ave} = \frac{1}{T}\sum_{t=0}^{T} PSNR(x_t, \hat{x}_t) \tag{23}$$

where $T$ is the total number of frames of normal video clips. Then, we set an appropriate threshold $k$, which is approximately 85% of the $PSNR_{ave}$ based on experimental observations. When the PSNR of a certain video frame is lower than the preset threshold, we consider that the frame is abnormal. As mentioned in [48] for the event-level evaluation criterion, an abnormal event will last for several consecutive frames; if more than one frame is detected as abnormal and the position of the frame locates within the ground truth range, then it is considered to be an abnormal event. In contrast, if the abnormal frame appears in the non-ground truth range, it is regarded as a false alarm. In addition, to reduce the noisy and meaningless alarm value in the anomaly score, following [8], [9], we assume that local alarm values within 50 frames belong to the same abnormal event.

## IV. EXPERIMENTS

In this section, we first describe our experimental setup in Section $A$, including the datasets, evaluation metric, and training details. Then, we show the experimental results of the proposed method on four popular datasets and compare them with the performance of the existing methods in Section $B$. Next, we perform extensive ablation studies in Section $C$ to analyze the functionalities of different components of our method. Then, we verify the real application performance of our method in Section $D$ and discuss the advantages and limitations of our proposed method in Section $E$. The running time is analyzed in Section $F$.

### A. EXPERIMENTAL SETUP

#### 1) DATASETS

We evaluate our method on four popular datasets, which are widely used in the field of video anomaly detection. (1) CUHK Avenue (Avenue) dataset [10]: this dataset consists of 16 training videos and 21 testing videos with a total of 30652 frames. There are 47 abnormal events, including throwing objects, loitering, and running in the CUHK Avenue dataset. (2) UCSD Pedestrian 1 (Ped1) dataset [33]: this dataset includes 34 training videos and 36 testing videos with 40 abnormal events. All of these abnormal events concern vehicles such as bicycles and cars. (3) UCSD Pedestrian 2 (Ped2) dataset [33]: this dataset includes 16 training videos and 12 testing videos with 12 abnormal events, which contain the same types of abnormal events as Ped1. The main difference between Ped1 and Ped2 is that Ped1 has perspective distortion because pedestrians walk toward and away from the camera. (4) ShanghaiTech [26]: this dataset includes 330 training videos and 107 testing videos, which has a total of 130 abnormal events, distributed in 13 different scenes, and thus, this dataset is a very challenging anomaly detection dataset.

#### 2) EVALUATION METRIC

We use the Receiver Operation Characteristic (ROC) curve and the Area Under the Curve (AUC) to evaluate the effectiveness of our proposed method; these are widely used as performance metrics for classification tasks. A higher AUC value means better performance in anomaly detection. In this paper, to ensure comparability between different methods, we use frame-level AUC for performance evaluation. In addition, the equal error rate (EER) of the ROC curve is also used as the performance evaluation metric.

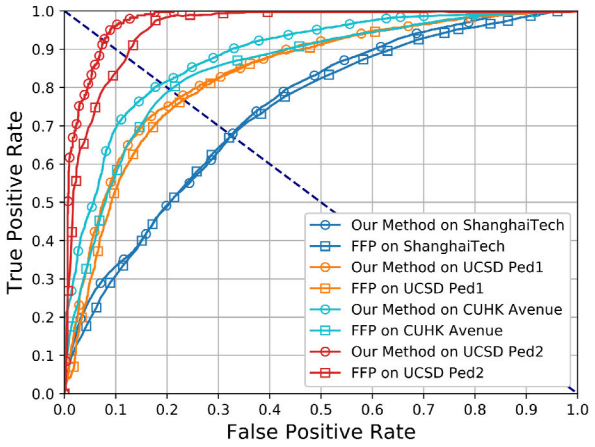



**FIGURE 6.** Comparison in terms of frame-level ROC curves between FFP [15] and our method on the four datasets.

#### 3) TRAINING DETAILS

To train the BR-GAN, following the work in [15], we first set the length of the input frame sequence $L$ to 4, and the size of each frame is resized to $256 \times 256$, while the values of the pixels in all frames are normalized to $[-1, 1]$. Then, we use the Adam optimizer based on the stochastic gradient descent method to optimize the parameters of the network. The batch size is set to 4. The learning rate of the generator is set to 0.0002, and the learning rates of both the frame discriminator and sequence discriminator are set to 0.00002. To choose a set of optimal hyperparameters for balancing the four datasets' losses, the grid search strategy is used for the hyperparameter search. We first determine the rough hyperparameter search space based on empirical values and then search for parameters from coarse to fine. Finally, the optimal hyperparameter settings are given as follows. For the Ped1 and Ped2 datasets, $\lambda_{int}$, $\lambda_{gd}$, $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ are set to 1.0, 1.0, 0.003, and 0.003, respectively. For the Avenue and ShanghaiTech datasets, $\lambda_{int}$, $\lambda_{gd}$, $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ are set to 1.0, 1.0, 0.025, and 0.025, respectively.

### B. EXPERIMENTAL RESULTS

#### 1) COMPARISON WITH EXISTING METHODS

We compare our method with a dozen state-of-the-art methods in Table 1. It is obvious that our method outperforms almost all existing methods on the four datasets. Although the methods proposed in [38], [32] integrate prediction and reconstruction branches, their AUC still lags behind that of our method. Especially, compared with the baseline

**TABLE 1.** AUC and EER obtained by different methods on the avenue, Ped1, Ped2 and shanghaitech datasets.

| Methods | Avenue | | Ped1 | | Ped2 | | ShanghaiTech | |
|---|---|---|---|---|---|---|---|---|
| | AUC | EER | AUC | EER | AUC | EER | AUC | EER |
| MPPCA [22] | N/A | N/A | 59.0% | 40.0% | 69.3% | 30.0% | N/A | N/A |
| MPPC+SFA [33] | N/A | N/A | 66.8% | 32.0% | 61.3% | 36.0% | N/A | N/A |
| MDT [33] | N/A | N/A | 81.8% | 25.0% | 82.9% | 25.0% | N/A | N/A |
| SCL [10] | 80.3% | 27.5% | N/A | N/A | N/A | N/A | N/A | N/A |
| Conv-AE [9] | 70.2% | 25.1% | 81.0% | 27.9% | 90.0% | 21.7% | 60.9% | N/A |
| ConvLSTM-AE [11] | 77.0% | N/A | 75.5% | N/A | 88.1% | N/A | N/A | N/A |
| Unmasking [18] | 80.6% | N/A | 68.4% | 31.0% | 82.2% | N/A | N/A | N/A |
| Stacked RNN [26] | 81.7% | N/A | N/A | N/A | 92.2% | N/A | 68.0% | N/A |
| AnomalyNet [13] | 86.1% | 22.0% | 83.5% | 25.2% | 94.9% | 10.3% | N/A | N/A |
| AnoPCN [32] | 86.2% | N/A | N/A | N/A | 96.8% | N/A | 73.6% | N/A |
| DeepOC [20] | 86.6% | **18.5%** | 83.5% | 23.4% | 96.9% | 8.8% | N/A | N/A |
| FSCN [21] | 85.5% | 20.7% | 82.4% | 25.2% | 92.8% | 12.5% | N/A | N/A |
| Tang *et al.* [38] | 85.1% | N/A | **84.7%** | N/A | 96.3% | 10.0% | 73.0% | N/A |
| CDDA [23] | 86.0% | N/A | N/A | N/A | 96.5% | N/A | 73.3% | N/A |
| MNAD [12] | 88.5% | N/A | N/A | N/A | 97.0% | N/A | 70.5% | N/A |
| FFP [15] | 84.9% | 21.2% | 83.1% | 23.5% | 95.4% | 12.0% | 72.8% | 31.8% |
| **BR-GAN** | **88.6%** | 19.0% | **84.7%** | **22.5%** | **97.6%** | **7.6%** | **74.5%** | **31.6%** |

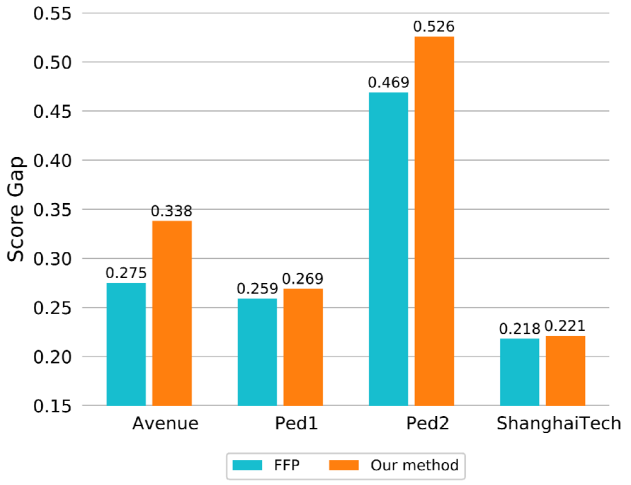



**FIGURE 7.** Comparison in terms of score gap between our method and FFP.

prediction method, FFP [15], our method provides an absolute gain of 3.7%, 1.6%, 2.2%, and 1.7% in terms of the AUC on Avenue, Ped1, Ped2, and ShanghaiTech, respectively. The ROC curves of FFP and our method on these datasets are shown in Fig. 6. These results fully demonstrate that our proposed prediction framework for anomaly detection is better than FFP. Furthermore, following the work in [15], we also compute the gap between the average score of normal frames and that of abnormal frames. A large score gap means that the network can be more capable of distinguishing normal and abnormal patterns. We compare the score gap between our method and FFP on the four datasets in Fig. 7. It can be seen that our method always achieves higher gaps than FFP, which further validates the effectiveness of our proposed prediction framework for anomaly detection.

**TABLE 2.** AUC obtained by different prediction methods on the avenue and Ped2 dataset.

| | Avenue | Ped2 |
|---|---|---|
| U-Prediction | 85.4% | 95.8% |
| B-Prediction | 87.1% | 96.7% |
| BR-Prediction | 88.6% | 97.6% |

#### 2) QUALITATIVE RESULTS

To further demonstrate the effectiveness of our method, we present several qualitative results in this subsection. Fig. 8 shows the future frame prediction results and the prediction error map of our prediction framework for normal events and abnormal events. For normal events, it can be seen that the predicted future frames tend to be close to the actual future frame, which results in a lower error in the residual map. However, for abnormal events, the predicted future frames are blurred and distorted compared with the actual future frame, which results in a higher error in the residual map. Moreover, we also show the anomaly score curves of several test video clips from the four datasets in Fig. 9. It is easy to observe that the anomaly scores rise dramatically when anomalies occur and decrease when anomalies disappear. Such an observation indicates that our method can accurately detect the occurrence of anomalies.

### C. ABLATION STUDY

In this section, we perform ablation experiments on the components and loss functions of the proposed method to validate the effectiveness of our method.

#### 1) BIDIRECTIONAL RETROSPECTIVE PREDICTION VS UNIDIRECTIONAL PREDICTION

To analyze the effectiveness of bidirectional retrospective prediction, we implement our approach with three different

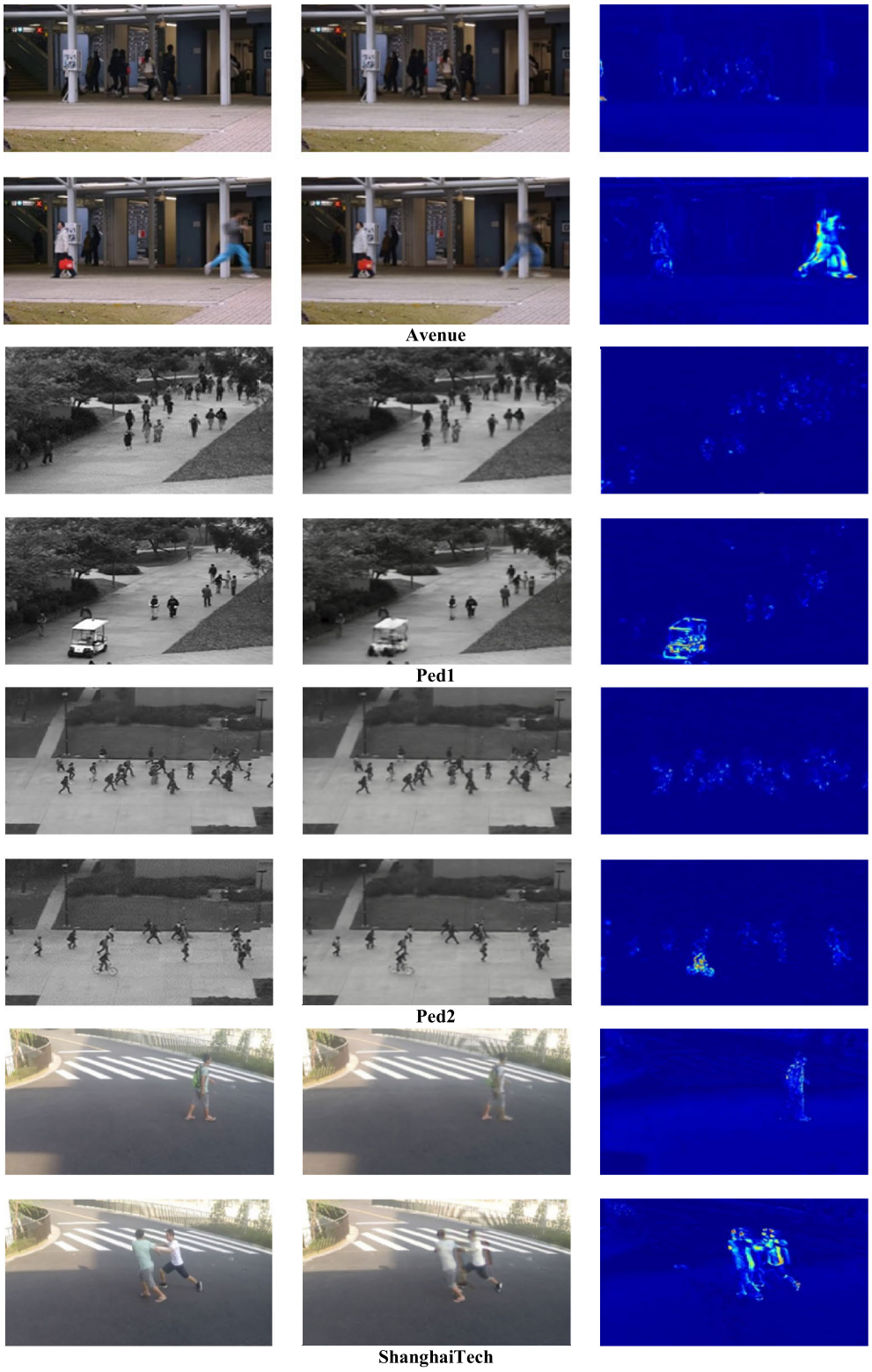



**FIGURE 8.** Examples of frame predictions on the four datasets. For each dataset, the first row is a prediction example of a normal event, and the second row is a prediction example of an abnormal event. Left column: the real frame. Mid column: the prediction frame. Right column: the prediction error map.

prediction methods: bidirectional retrospective prediction (BR-Prediction), bidirectional prediction (B-Prediction), and unidirectional prediction (U-Prediction). Table 2 shows the AUC of these three methods on the Avenue and Ped2 datasets.

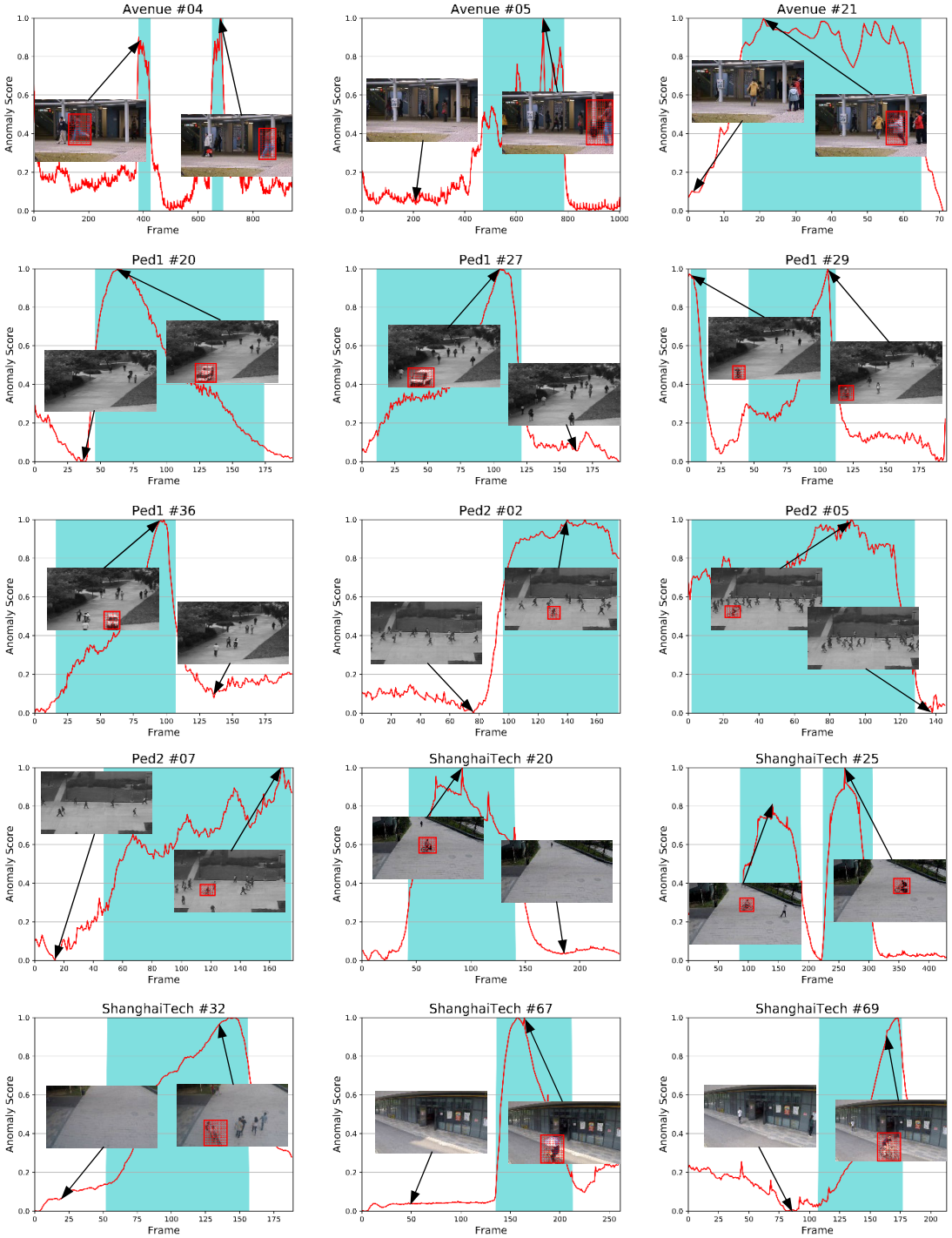



**FIGURE 9.** Anomaly score curves of fifteen test video clips of our method on four datasets. Cyan regions represent ground truth anomalous frames. It shows that the anomaly scores rise dramatically when anomalies occur and decrease when anomalies disappear. Best viewed in color.

It can be seen that the performance of bidirectional retrospective prediction is the best, followed by bidirectional prediction without retrospective, and the performance of unidirectional prediction comes last. This finding demonstrates that bidirectional retrospective prediction can more fully mine the mapping relationship between video frame sequences to better establish the pattern of normal events.

### 2) SEQUENCE DISCRIMINATOR VS OPTICAL FLOW

To show that the sequence discriminator in our proposed method can also play a vital role in motion constraints compared to the method of extracting optical flow, we replace the sequence discriminator in BR-GAN with the method of constraining the optical flow as that in FFP [15] to conduct comparative experiments. Specifically, we use the pre-trained

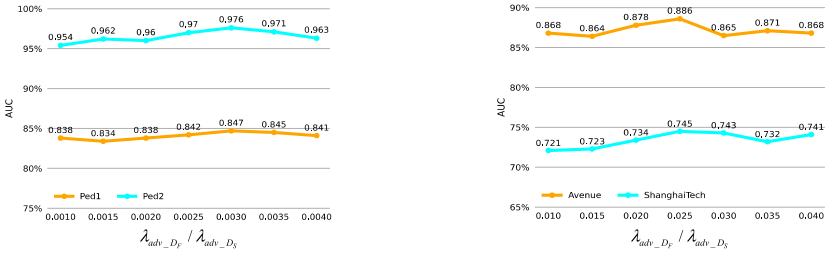



**FIGURE 10.** The change in the AUC with different $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ on (a) Ped1 and Ped2 datasets and (b) Avenue and ShanghaiTech datasets.

**TABLE 3.** AUC obtained by BR-GAN with the optical flow or sequence discriminator for motion constraints on the avenue and Ped2 dataset.

| | Avenue | Ped2 |
|---|---|---|
| BR-GAN (with optical flow) | 87.3% | 96.4% |
| BR-GAN (with sequence discriminator) | 88.6% | 97.6% |

FlowNet [17] to estimate the optical flow of predicted frames generated by bidirectional prediction and retrospective prediction, and perform motion constraints based on optical flow loss between these predicted frames and real frames. Table 3 shows that the model trained with the sequence discriminator for motion constraint achieves a higher AUC than the model trained with the estimated optical flow for motion constraint on Avenue and Ped2 datasets. This finding demonstrates that it is more effective to use the sequence discriminator instead of relying on low-level optical flow features.

### 3) IMPACT OF HYPERPARAMETERS

In our method, $\lambda_{int}$, $\lambda_{gd}$, $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ are the key hyperparameters used to balance the four loss terms of $L_{int}$, $L_{gd}$, $L^G_{adv_D_F}$, and $L^G_{adv_D_S}$, respectively. In this subsection, we explore the effects of these hyperparameters and conduct several experiments on the Avenue, Ped1, Ped2, and ShanghaiTech datasets. $L_{int}$ and $L_{gd}$ are the main loss items used to optimize the generator, and their loss values are on the same order of magnitude; thus, we set $\lambda_{int}$ and $\lambda_{gd}$ to 1 by default. $L^G_{adv_D_F}$ and $L^G_{adv_D_S}$ are two similar loss items, which are respectively combined with frame discriminator and sequence discriminator to assist in optimizing the generator. To determine the appropriate weights to balance them, we first unify all of the loss items to the same order of magnitude. For the Ped1 and Ped2 datasets, where the video frames are grayscale images, we set the initial $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ to 0.001. For the Avenue and ShanghaiTech datasets, where the video frames are color images, we set the initial $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ to 0.01. Then, we use grid search to search for an optimum value within certain limits. To make the frame discriminator and sequence discriminator have the same influence on the generator, we keep the values of $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ the same. As shown in Fig. 10, for the Ped1 and Ped2 datasets, larger $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ harm the performance, and the performance is relatively good when $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ are set to 0.003. Similarly, for the Avenue and ShanghaiTech datasets, setting $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ too large or too small will reduce the performance of the method. When $\lambda_{adv_D_F}$ and $\lambda_{adv_D_S}$ are set to 0.025, the performance is at a relatively optimal level.

**TABLE 4.** The evaluation of different loss components in our future frame prediction network on the avenue dataset.

| $L_{\text{int}}$ | $L_{gd}$ | $L^{frame}_{adv}$ | $L^{seq}_{adv}$ | AUC | Score Gap |
|---|---|---|---|---|---|
| √ | √ | √ | √ | 88.6% | 0.338 |
| √ | √ | √ | × | 87.2% | 0.320 |
| √ | √ | × | × | 86.1% | 0.309 |
| √ | × | × | × | 85.3% | 0.289 |

**TABLE 5.** The number of detected abnormal events and false alarms on the shanghaitech dataset.

| Dataset | Anomalous Event | True Positive | False Alarm |
|---|---|---|---|
| ShanghaiTech | 130 | 125 | 32 |

### 4) ANALYSIS OF LOSSES

The loss functions play a key role in the performance of the method. To analyze the impact of different loss functions in our method, we conduct ablation experiments on the Avenue dataset by gradually removing the four loss terms $L_{int}$, $L_{gd}$, $L^{frame}_{adv}$, and $L^{seq}_{adv}$. We utilize the AUC and score gap to evaluate how different loss terms influence the performance of the anomaly detection. A high AUC and a large score gap indicate better anomaly detection performance. The results are shown in Table 4. It can be seen that the absence of each loss item degrades the performance in anomaly detection. When all of the loss terms are used, our method achieves the best performance, which means that the combination of all of the loss terms is critical to the anomaly detection performance.

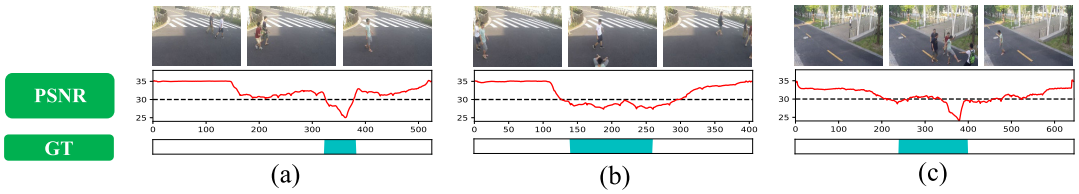



**FIGURE 11.** Qualitative results for the event-level of our method on the ShanghaiTech dataset. (a) robbery, (b) affray, (c) fighting. In addition, "GT" is the ground truth at the frame-level. Best viewed in color and high resolution.

**TABLE 6.** Comparisons in terms of the inference time between our method and other existing methods.

| Method | Platform | CPU | GPU | Inference time/frame (GPU) | Inference time/frame (CPU) |
|---|---|---|---|---|---|
| MDT [33] | C | 2.8GHz | N/A | 1.11-1.38s | N/A |
| AMDN [1] | Matlab/Caffe | 2.1GHz | NVIDIA Quadro K4000 | 9.4-12.8s | N/A |
| SRC [2] | Matlab | 2.6GHz | N/A | 3.8-4.6s | N/A |
| SCL [10] | Matlab | 3.4GHz | N/A | 0.0067s | N/A |
| Unmasking [18] | N/A | 2.3GHz | N/A | 0.050s | N/A |
| DAP [66] | N/A | 2.3GHz | N/A | 0.050s | N/A |
| NMC [67] | N/A | 2.3GHz | N/A | 0.042s | N/A |
| DeepOC [20] | Python/Tensorflow | 4.2GHz | NVIDIA GeForce 1080TI | 0.025s | N/A |
| FFP [15] | Python/Tensorflow | N/A | NVIDIA GeForce TITAN | 0.040s | N/A |
| **Our** | Python/Pytorch [47] | I7-10700k@3.8GHz | NVIDIA GeForce RTX 2070 Super | 0.022s | 0.2s |

### D. REAL APPLICATION RESULTS

In this subsection, we verify the real application performance of our method in terms of the evaluation criterion of the event-level mentioned in Section III.F.2. We conduct an experiment on the ShanghaiTech dataset, and we calculate the $PSNR_{ave}$ to be 35 on the normal video clips of the training Set; then, we set the threshold to 30, which is approximately 85% of the $PSNR_{ave}$ based on experimental observations. The ShanghaiTech dataset has a total of 130 abnormal events distributed in 13 different scenes, which is a very challenging anomaly detection dataset. From the experimental results in Table 5, it can be seen that the false alarm rate is slightly high due to the large changes in the scene, but almost all abnormal events can be detected. In Fig. 11, we show an event-level detection result for real abnormal events such as a robbery, affray, and fighting in the ShanghaiTech dataset. As can be seen from Fig. 11, our method can accurately detect abnormal events by setting an appropriate threshold.

### E. RUNNING TIME

In this section, we analyze the inference time of BR-GAN on a GPU and CPU and compare it with state-of-the-art methods. In the inference stage, BR-GAN contains both future frame prediction and anomaly detection. On the CPU, the average inference time of BR-GAN on the Avenue, Ped1, Ped2, and ShanghaiTech datasets is 0.2 s (5 FPS), which exceeds the inference speed of MDT [33], AMDN [1], and SRC [2] on the GPU, as shown in Table 6. On the GPU, the inference time of BR-GAN on the Avenue, Ped1, Ped2, and ShanghaiTech datasets is 37 FPS, 64 FPS, 56 FPS, and 28 FPS, respectively, and the average inference time is 0.022s (46 FPS). It is easy to see from Table 6 that the average inference time of our method on the GPU ranks second among the ten methods, which shows that our method can detect abnormal events in real time.

### F. DISCUSSION

As can be seen from the above experimental results, our framework, which adopts the bidirectional and retrospective prediction and combines the frame discriminator and sequence discriminator to fully mine the spatio-temporal features of video sequences, has achieved better performance than existing methods. The BR-GAN easily detects anomalous events, such as robbery, affray, and fighting. In addition, our method also has advantages in its detection speed, which can meet the requirements of real-time detection.

However, in the inference stage, the first $L$ frames of the video cannot be detected ($L = 4$ in our method) because these frames must be used as the original input of the prediction model, which is also the shortcoming of all prediction-based methods. In addition, if the video has perspective distortion, such as on the Ped1 and ShanghaiTech datasets, it is difficult for BR-GAN to process small objects that appear at the far end of the camera and objects that have seemingly normal movement. In future work, we will further explore this issue.

## V. CONCLUSION

In this paper, we propose a novel future frame prediction framework based on a bidirectional retrospective generative adversarial network (BR-GAN) that can be trained

end-to-end for anomaly detection in video. By adopting the bidirectional prediction combined with the retrospective prediction, the prediction network can more fully mine the bidirectional mapping relationship between video frame sequences to achieve more accurate future frame prediction of normal events. In addition, we propose a sequence discriminator composed of 3D convolutional neural networks that can capture the long-term temporal relationship between frame sequences composed of predicted and input frames, which plays a crucial role in maintaining the motion consistency of predicted frames. Extensive experiments on four benchmark datasets demonstrate that our method outperforms most existing state-of-the-art methods, which validates the effectiveness of our method for anomaly detection.

## REFERENCES

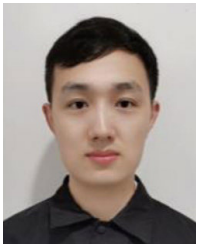
[1] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2016.
[2] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3449–3456.
[3] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, Jun. 2011, pp. 3313–3320.
[4] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," 2020, *arXiv:2003.06780*. [Online]. Available: http://arxiv.org/abs/2003.06780
[5] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.
[6] J. K. Dutta and B. Banerjee, "Online detection of abnormal events using incremental coding length," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 3755–3761.
[7] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*. [Online]. Available: http://arxiv.org/abs/1612.00390
[8] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.*, vol. 10262. Long Beach, CA, USA: Springer, Dec. 2017, pp. 189–196.
[9] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. CVPR*, Jun. 2016, pp. 733–742.
[10] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.
[11] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.
[12] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14360–14369.
[13] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.
[14] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.
[15] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.
[16] C. Xia, F. Qi, and G. Shi, "Bottom–up visual saliency estimation with deep autoencoder-based sparse reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1227–1240, Jun. 2016.
[17] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[18] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. ICCV*, Oct. 2017, pp. 2914–2922.
[19] P. Wu, J. Liu, F. Shen, 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 *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 322–339.
[20] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2609–2622, Jul. 2020.
[21] P. Wu, J. Liu, M. Li, Y. Sun, and F. Shen, "Fast sparse coding networks for anomaly detection in videos," *Pattern Recognit.*, vol. 107, pp. 1–30, Nov. 2020.
[22] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2921–2928.
[23] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 329–345.
[24] J. Li, X. Mei, D. Prokhorov, and D. Tao, "Deep neural network for structural prediction and lane detection in traffic scene," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 690–703, Mar. 2017.
[25] A. Stuhlsatz, J. Lippel, and T. Zielke, "Feature extraction with deep neural networks by a generalized discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 596–608, Apr. 2012.
[26] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. ICCV*, Oct. 2017, pp. 341–349.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.
[29] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. 9th Eur. Conf. Comput. Vis.* Vienna, Austria: Springer, May 2006, pp. 428–441.
[30] D. D'Avino, D. Cozzolino, G. Poggi, and L. Verdoliva, "Autoencoder with recurrent neural networks for video forgery detection," in *Proc. IS&T Int. Symp. Electron. Imag., Media Watermarking, Secur. Forensics*, Jan. 2017, pp. 92–99.
[31] A. Munawar, P. Vinayavekhin, and G. De Magistris, "Spatio-temporal anomaly detection for industrial robots through prediction in unsupervised feature space," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Santa Rosa, CA, USA, May 2017, pp. 1017–1025.
[32] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1805–1813.
[33] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.
[34] Q. Sun, H. Liu, and T. Harada, "Online growing neural gas for anomaly detection in changing surveillance scenes," *Pattern Recognit.*, vol. 64, pp. 187–201, Apr. 2017.
[35] R. Chaker, Z. Al Aghbari, and I. N. Junejo, "Social network model for crowd anomaly detection and localization," *Pattern Recognit.*, vol. 61, pp. 266–281, Jan. 2017.
[36] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.
[37] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Jul. 2017, pp. 5967–5976.
[38] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognit. Lett.*, vol. 129, pp. 123–130, Jan. 2020.
[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[40] W. Liu, W. Luo, Z. Li, P. Zhao, and S. Gao, "Margin learning embedded prediction for video anomaly detection with a few anomalies," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3023–3030.
[41] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future-flow embedded video prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1762–1770.

[42] V. C. Tim Salimans, I. Goodfellow, and W. Zarema, "Improved techniques for training GANs," *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 19, no. 1, pp. 1–9.
[43] S. Mahdizadehaghdam, A. Panahi, and H. Krim, "Sparse generative adversarial network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3063–3071.
[44] J. Li, J. Jia, and D. Xu, "Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 1–16.
[45] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2016, pp. 613–621.
[46] F. Liu, L. Jiao, and X. Tang, "Task-oriented GAN for PolSAR image classification and clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2707–2719, Sep. 2019.
[47] *PyTorch*. Accessed: Jun. 2021. [Online]. Available: https://github.com/pytorch/
[48] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, no. 7, pp. 1851–1864, 2013.
[49] X. Zhu, J. Liu, J. Wang, C. Li, and H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognit.*, vol. 47, no. 5, pp. 1791–1799, 2014.
[50] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2054–2060.
[51] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.
[52] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 82–90.
[53] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 8.1–8.12,
[54] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," 2020, *arXiv:2007.02500*. [Online]. Available: http://arxiv.org/abs/2007.02500
[55] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3513–3527, 2021.
[56] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 397–408, Jun. 2005.
[57] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, Sep. 2006.
[58] C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognit. Lett.*, vol. 27, no. 15, pp. 1835–1842, Nov. 2006.
[59] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[61] X. Nie, W. Jing, C. Cui, C. J. Zhang, L. Zhu, and Y. Yin, "Joint multi-view hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 10, pp. 1951–1965, Oct. 2020.
[62] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 401–410, Jan. 2021.
[63] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proc. AAAI Artif. Intell.*, May 2021, vol. 35, no. 2, pp. 938–946.
[64] 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," 2020, *arXiv:2011.07491*. [Online]. Available: http://arxiv.org/abs/2011.07491
[65] Y. Lu, F. Yu, M. Reddy, and Y. Wang, "Few-shot scene-adaptive anomaly detection," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12350, Oct. 2020, pp. 125–141.
[66] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe, "Deep appearance features for abnormal behavior detection in video," in *Proc. Int. Conf. Image Anal. Process.*, vol. 10485, Oct. 2017, pp. 779–789.
[67] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, "Detecting abnormal events in video using narrowed normality clusters," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1–17.
[68] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[69] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

**ZHIWEI YANG** received the B.S. degree in communication engineering from the Zhongyuan University of Technology, Zhengzhou, China, in 2019. He is currently pursuing the master's degree with Guangzhou Institute of Technology, Xidian University, China. His current research interests include computer vision, video anomaly detection, and deep learning.

**JING LIU** (Senior Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in circuits and systems from Xidian University, in 2000 and 2004, respectively.

She joined Xidian University, as a Lecturer, in 2005, and was promoted to a Full Professor, in 2009. From April 2007 to April 2008, she worked as a Postdoctoral Research Fellow with The University of Queensland, Australia. From July 2009 to July 2011, she worked as a Research Associate with The University of New South Wales, Australian Defence Force Academy. She is currently a Full Professor with Guangzhou Institute of Technology, Xidian University. Her research interests include evolutionary computation, complex networks, fuzzy cognitive maps, and computer vision. From 2017 to 2018, she was the Chair of the Emerging Technologies Technical Committee (ETTC) of the IEEE Computational Intelligence Society. From 2015 to 2020, she was an Associate Editor of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. Please see her homepage for more information (https://faculty.xidian.edu.cn/LJ22/zh_CN/index.htm).



**PENG WU** received the B.Eng. degree from Xidian University, Xi'an, China, in 2017, where he is currently pursuing the Ph.D. degree with Guangzhou Institute of Technology. His current research interests include video anomaly detection, video retrieval, temporal action detection, weakly supervised learning, and deep learning. Please see his homepage for more information (https://roc-ng.github.io/).