Dual Discriminator Generative Adversarial Network for Video Anomaly Detection

FEI DONG1, YU ZHANG2, AND XIUSHAN NIE3, (Member, IEEE)

1School of Journalism and Communication, Shandong Normal University, Jinan, 250014, China
2School of Software, Shandong University, Jinan, 25101, China
3School of Computer Science and Technology, Shandong Jianzhu University, Jinan, 250014, China

Corresponding author: Xiushan Nie (niexsh@hotmail.com)

ABSTRACT Video anomaly detection is an essential task because of its numerous applications in various areas. Because of the rarity of abnormal events and the complicated characteristic of videos, video anomaly detection is challenging and has been studied for a long time. In this paper, we propose a semi-supervised approach with a dual discriminator-based generative adversarial network structure. Our method considers more motion information in video clips compared with previous approaches. Specifically, in the training phase, we predict future frames for normal events via a generator and attempt to force the predicted frames to be similar to their ground truths. In addition, we utilize both a frame discriminator and motion discriminator to adverse the generator to generate more realistic and consecutive frames. The frame discriminator attempts to determine whether the input frames are generated or original frames sampled from the normal video. The motion discriminator attempts to determine whether the given optical flows are real or fake. Fake optical flows are estimated from generated frames and adjacent frames, and real optical flows are estimated from the real frames sampled from original videos. Then, in the testing phase, we evaluate the quality of predicted frames to obtain the regular score, and we consider those frames with lower prediction qualities as abnormal frames. The results of experiments on three publicly available datasets demonstrate the effectiveness of our proposed method.

INDEX TERMS Anomaly detection, generative adversarial network, dual discriminator.

I. INTRODUCTION

Video anomaly detection has long been studied because of its essential applications in various areas, such as traffic monitors, violence alerts, and smoke alarms. However, finding the abnormal events in videos is a challenging task, for two reasons. First, abnormal events rarely occur, which results in extreme imbalance between normal and abnormal data. Second, anomalies in videos are more complicated compared with those in other data forms, such as image and text. For video anomaly detection, we cannot ignore the importance of time series. Learning the relationships between consecutive frames may greatly improve the anomaly detection performance in videos.

Many efforts have been made regarding video anomaly detection. Based on the different settings of training sets, the methods can be primarily divided into supervised, semi-supervised, and unsupervised methods.

For supervised anomaly detection, both normal and abnormal data are provided and labeled in the training set. A classic solution is to apply weakly supervised methods, on which only video-level labels are provided in the training set. Multiple-instance learning (MIL) has proven to be feasible as a weakly supervised anomaly detection method [1]. A video is usually considered as a bag, and snippets constituting the video are viewed as instances. Then, they learn instance-level anomaly labels via bag-level annotations. However, most of these MIL models need to extract offline features in advance, and they may classify anomalies as normal data, as the observed data only contain a few types of anomalies, instead of all kinds of abnormal events.

For unsupervised anomaly detection, the labels of training data are not given, and unsupervised clustering is applied to detect anomalies. However, the high-dimensional complexity of video data limits its performance.

Based on the definition of an anomaly, as an event that does not conform to expected behavior [2], semi-supervised methods have demonstrated their advantages. In these methods,
only normal data are provided in the training set, and those events deviating from the normal pattern are regarded as anomalies. Many experiments regarding semi-supervised anomaly detection have been performed. The goal of semi-supervised anomaly detection is to learn a model or a representation that captures normal motion and spatial appearance patterns [3]. For example, trajectory features have been used to represent the normal patterns of the object of interest [4], [5], and those outliers are considered to be anomalies. However, the trajectory-based method is not robust in complicated environments such as crowded scenes. Dictionary learning or sparse coding is another popular approach for video anomaly detection [6], [7]. In these methods, normal events are encoded and a dictionary capturing the normal patterns is learned. When an abnormal event is sent into the model, it results in a large reconstruction error. The shortcoming of this kind of method is the huge time consumption involved in optimizing sparse coefficients. Furthermore, these methods are mostly based on hand-crafted features, and the limited capability of video representations may result in a performance bottleneck.

Recently, with the success of deep learning methods in many fields, such as image classification, object detection, and video comprehension, some deep-learning-based anomaly detection approaches have been proposed. Hasan et al. [8] utilized a fully convolutional auto-encoder to learn regular patterns; their method takes short video snippets in a temporal sliding window as the input. In the test phase, they compute the reconstruction error of the frame intensity to obtain the regular score for each frame. To learn better spatial-temporal information, recurrent neural network (RNN) and its long short-term memory (LSTM) variant have also been applied to model normal spatial and motion patterns. Almost all of these reconstruction-based methods assume that normal events would cause large reconstruction errors; however, because of the huge capability of deep neural networks, abnormal events may correspond to small reconstruction errors and be classified as normal events.

In addition to the reconstruction method, the frame prediction-based method has been proven effective for video anomaly detection [9]. The basic assumption of the frame prediction-based method is that normal events can be easily predicted, while abnormal events have an extreme possibility of being unexpected. Recently, with the emergence of the generative adversarial network (GAN) [10], the performance of future frame prediction has been greatly improved, and the frame prediction-based method has achieved the state-of-the-art performance for semi-supervised video anomaly detection. Liu et al. [9] leveraged a U-Net as the generator to predict future frames, and a patch discriminator was used to augment the quality of generated frames. However, they did not take into account a sufficient number of motion characteristics.

In this paper, we propose a semi-supervised video anomaly detection approach with a dual discriminator-based GAN structure that aims to generate more realistic and consecutive future frames for normal events. Specifically, our framework is composed of a generator and two discriminators, which are optimized alternately. Given a video clip of normal events, we predict a future frame with the generator and determine whether it is real or fake with the frame discriminator. Additionally, we estimate the optical flow between consecutive frames as motion features and use the motion discriminator to determine their authenticity. In the testing phase, we compare the quality of predicted frames and their ground truths and then obtain the regular score for each frame.

II. RELATED WORK
A. ANOMALY DETECTION
Many methods have been designed for video anomaly detection, including hand-crafted feature-based and deep-learning-based approaches. For hand-crafted features, low-level trajectory features and spatial-temporal features, such as the histogram of oriented gradients (HOG), are used to capture the normal patterns. Kim and Grauman [11] modeled the optical flow with a mixture of probabilistic principal component analysis (MPPCA), which aims to better capture local motion patterns. According to the learned optical flow distribution of the normal behavior, a Markov random field (MRF) based on time and space information can be further constructed. Then, they mapped the nodes in the MRF graph one-to-one with the spatial-temporal grid in the frame to detect if there is any abnormal behavior in the video. Mehran et al. [12] used the social force (SF) model to detect anomalies in crowded scenes. Mahadevan et al. [13] modeled video via a mixture of dynamic textures (MDT).

In addition to these hand-crafted feature-based approaches, many deep-learning-based methods have been proposed. Hasan et al. [8] detected abnormal events according to the reconstruction errors of a convolutional auto-encoder. Chong and Tay [14] utilized a convolutional long short term memory (ConvLSTM) to capture temporal information regarding normal events. Sultani et al. [1] considered normal and anomalous videos as bags and video segments as instances in multiple instance learning (MIL) and used a 3D convolutional network to extract spatial-temporal information. Morais et al. [15] learned the skeleton trajectories with a message-passing encoder-decoder recurrent network.

B. FRAME GENERATION WITH GAN
Generative adversarial networks (GANs) have attracted the attention of a large number of research scholars in recent years. GANs usually consist of two models: a generator and a discriminator, which can be implemented with neural networks. The generator attempts to capture the distribution of data and generate new examples that are as realistic as possible, and the discriminator attempts to distinguish generated examples from the true examples [16]. Both the generator and discriminator improve through adversarial learning. The optimization goal of GANs is to reach Nash equilibrium [17]. Because of the excellent generation capability of GAN, it has been applied to many fields, including image generation...
and video prediction. Reed et al. [18] showed that the conditional GAN was able to synthesize plausible images from text descriptions. Isola et al. [19] applied the GAN to image-to-image translation tasks and achieved good performance. Villegas et al. [20] used an appearance discriminator to augment the prediction of future frames.

III. PROPOSED METHOD

An overview of our model is shown in Figure 1. We apply the idea of the GAN for video anomaly detection, and our framework is composed of a generator and two discriminators, which discriminate the true and false from the appearance and motion, respectively. The generator and discriminators are optimized alternately, which leads to better prediction performances.

A. NOTATION

The input dataset is represented by \( D \) and is split into a training set \( D_{\text{tra}} \) and a testing set \( D_{\text{tst}} \). Then, we train our model \( f \) on \( D_{\text{tra}} \) and evaluate the performance on \( D_{\text{tst}} \). The input video clips are sampled from the dataset \( D \) and are represented by \( \{I_{t-T}, \ldots, I_t\} \). In the training set \( D_{\text{tra}} \), each frame belongs to the normal class, and the testing set \( D_{\text{tst}} \) contains both normal and abnormal examples. We use \((I_m, y_m)\) to denote a frame example, where \( y_m = 0 \) denotes the normal class and \( y_m = 1 \) denotes the abnormal class. The generated frame at time \( t \) is symbolized as \( \hat{I}_t \), and its ground truth is denoted \( I_t \).

The training objective \( J \) of our model \( f \) is to learn the regularity of normal events, and those events deviating from it are considered to be anomalies. We utilize \( G \) to denote the generator, \( D_F \) to denote the frame discriminator, and \( D_M \) to denote the motion discriminator. In addition, we use \( \hat{F}_i \) to denote the optical flow of generated frame \( \hat{I}_i \) and \( F_i \) to denote its ground truth. Additionally, \( S(t) \) is used to denote the regular score for each frame obtained in the testing phase. A higher score means it is more likely to be a normal frame.

B. THE MODEL

Our model consists of a generator to predict future frames and dual discriminators, one of which is used to discriminate frames (appearance), and the other of which is used to distinguish the motion between true and false.

We implement the generator with U-net [21], a convolutional neural network with skip connections. The network does not have any fully connected layers and possesses a u-shaped architecture. Every step of the left side consists of the repeated two \( 3 \times 3 \) convolutions and a \( 2 \times 2 \) max pooling operation with stride 2 for downsampling. Each step of the right side consists of an upsampling of the feature map, a concatenation with the feature map from the corresponding left side, and two \( 3 \times 3 \) convolutions, each followed by a rectified linear unit (ReLU). Following a previous work [9], the architecture of our generator is slightly different from that of the original network, and the maximum number of feature channels is 512. The architecture of the generator is shown in Figure 2.

We implement the discriminators with PatchGAN following the framework of a previous study [19]. This discriminator attempts to classify if each \( N \times N \) patch in a frame or feature map is real or fake, and we average all responses to provide the ultimate output of the discriminator. Our frame discriminator \( D_F \) and motion discriminator \( D_M \) share the same architecture, although they have different inputs.

C. OBJECTIVE FUNCTIONS

Based on the assumption that abnormal events cannot be easily predicted, our optimization objective is to train a model that can easily predict future frames for normal data. Because the training data only contain normal events, when abnormal examples are given, it is difficult for the trained model to accurately predict. For video anomaly detection, spatial and temporal information are both important. To predict future
frames better, we consider both appearance and motion constraints. The appearance loss includes intensity loss, gradient loss, and adversarial training loss, and the motion loss includes optical flow loss and adversarial training loss with respect to motion information.

1) OBJECTIVE FUNCTION FOR GENERATOR
The main goal of our generator is to generate frames similar to the ground truths. Given a video snippet \( I_0, \ldots, I_{T-1} \) with length \( T \), our generator outputs a predicted frame \( \hat{I} \). To make \( \hat{I} \) consistent with its ground truth \( I_t \), we add penalties for appearance and motion.

To reduce the distance between \( \hat{I} \) and \( I_t \) in RGB space, we add an intensity loss, which can guarantee the basic appearance of objects to be similar. The intensity loss is estimated as

\[
L_{\text{int}}(\hat{I}, I_t) = \| \hat{I} - I_t \|_2^2, \tag{1}
\]

where \( \hat{I} \) denotes the generated frame, \( I_t \) denotes the corresponding ground truth, and \( \| \cdot \|_2 \) denotes the Euclidean distance.

In addition, we add the gradient penalty to preserve the sharpness of original frames following a previous study [9]. The gradient difference of two images along two spatial dimensions is calculated as

\[
L_{\text{gd}}(\hat{I}, I) = \sum_{i,j} \| \hat{I}_{i,j} - \hat{I}_{i-1,j} - (I_{i,j} - I_{i-1,j}) \|_1 + \| \hat{I}_{i,j} - \hat{I}_{i,j-1} - (I_{i,j} - I_{i,j-1}) \|_1, \tag{2}
\]

where \( i,j \) denote the spatial indices of a frame, \( \hat{I} \) denotes a generated frame, and \( I \) denotes the corresponding ground truth.

Intensity and gradient losses guarantee the appearance of predicted frames to be consistent with real frames, but the frames may be discontinuous from the previous frames. Therefore, we add a motion penalty to maintain the temporal continuity. Optical flow is a good solution to capture the temporal motion information from video clips. Here, we use a CNN-based approach (FlowNet) to estimate the optical flow [22], which can help us build a differentiable system. FlowNet is pre-trained, and all of its parameters are fixed. Following a previous study [9], we apply the \( l_1 \) distance loss to calculate the motion penalty:

\[
L_{\text{mot}}(\hat{F}, F_t) = \| \hat{F}_t - F_t \|_1, \tag{3}
\]

where \( \hat{F}_t \) is the estimated optical flow of the previous frame \( I_{t-1} \) and the generated frame \( \hat{I}_t \), and \( F_t \) is the ground truth flow of real frames \( I_{t-1} \) and \( I_t \).

In addition to these appearance and motion constraints, we also use an adversarial penalty for the generator. Here, we use the least squares GANs (LSGANs) [23]. In the adversarial training process, our generator attempts to generate frames that frame discriminator \( D_F \) cannot distinguish from the real frames. The generator also attempts to generate frames with optical flows that motion discriminator \( D_M \) cannot distinguish from the optical flows of real consecutive frames. The adversarial loss for generator \( G \) is defined as

\[
L_{\text{adv}}(\hat{I}, \hat{F}_t) = \lambda_{\text{img}}^G \sum_{i,j} \frac{1}{2} (D(\hat{I}_{i,j}) - 1)^2 + \lambda_{\text{flow}}^G \sum_{i,j} \frac{1}{2} (D(\hat{F}_{i,j}) - 1)^2, \tag{4}
\]

where \( i,j \) denotes the spatial patch indices. \( \lambda_{\text{img}}^G \) and \( \lambda_{\text{flow}}^G \) are the corresponding coefficients controlling weights.

All of the above constraints constitute our total objective function of \( G \), which is estimated as follows:

\[
L_G = \lambda_{\text{int}} L_{\text{int}}(\hat{I}, I_t) + \lambda_{\text{gd}} L_{\text{gd}}(\hat{I}, I_t) + \lambda_{\text{mot}} L_{\text{mot}}(\hat{F}_t, F_t) + \lambda_{\text{adv}} L_{\text{adv}}(\hat{I}, \hat{F}_t). \tag{5}
\]

where \( \lambda_{\text{int}}, \lambda_{\text{gd}}, \lambda_{\text{mot}}, \) and \( \lambda_{\text{adv}} \) are coefficient factors.
2) OBJECTIVE FUNCTION FOR DISCRIMINATORS

We design two discriminators, for frames and optical flows, respectively. Frame discriminator $D_F$ attempts to determine whether the given frames are real video frames or predicted frames generated from generator $G$. When optimizing $D_F$, we fix the weights of $G$. The adversarial penalty for $D_F$ is expressed as follows:

$$L_{adv}^{D_F}(\hat{I}_t, I_t) = \sum_{i,j} \frac{1}{2}(D_F(I_t)_{i,j} - 1)^2$$

$$+ \sum_{i,j} \frac{1}{2}(D_F(\hat{I}_t)_{i,j} - 0)^2,$$

(6)

where $i, j$ denotes the spatial patch indices.

For motion discriminator $D_M$, we use the optical flows estimated from two consecutive frames as motion information that needs to be learned. When optimizing $D_M$, we also fix the weights of $G$. In our implement, $D_M$ and $D_F$ have the same PatchGAN architecture [19]. The adversarial penalty for $D_M$ is expressed as follows:

$$L_{adv}^{D_M}(\hat{F}_t, F_t) = \sum_{i,j} \frac{1}{2}(D_M(F_t)_{i,j} - 1)^2$$

$$+ \sum_{i,j} \frac{1}{2}(D_M(\hat{F}_t)_{i,j} - 0)^2,$$

(7)

where $F_t$ is the optical flow calculated from frames $I_{t-1}$ and $I_t$, and $\hat{F}_t$ is the optical flow estimated from frame $I_{t-1}$ and predicted frame $\hat{I}_t$.

D. ANOMALY DETECTION

In the testing phase, we only utilize our trained generator to predict future frames. To detect anomalies, we provide a regular score $S(t)$ for every frame. If $S(t)$ is smaller, the corresponding frame has a greater possibility of being abnormal. Following a previous study [9], we use the similarity between the predicted frames and their ground truths to measure the regularity and obtain $S(t)$. The peak signal to noise ratio (PSNR) is an effective way to assess image quality and measure similarity. Therefore, we use the PSNR to measure the quality of the predicted frames:

$$PSNR(I, \hat{I}) = 10\log_{10}\frac{[max]_2}{\sum_{i=0}^{N}(I_i - \hat{I}_i)^2},$$

(8)

where $\hat{I}$ indicates the predicted frame and $I$ denotes the corresponding ground truth.

Higher PSNR values indicate better image quality, which means that the predicted frame is more similar to its ground truth and thus more likely to be normal. After calculating the PSNR of each frame, we normalize the PSNR of all frames in each testing video to the range [0,1] and obtain the regular score $S(t)$ of each frame; the formula for normalization is expressed as

$$S(t) = \frac{PSNR_t - min(PSNR)}{max(PSNR) - min(PSNR)},$$

(9)

where $max(PSNR)$ and $min(PSNR)$ are the maximum and minimum PSNR values of frames in the corresponding testing video.

IV. EXPERIMENTS

In this section, we evaluate our proposed method on three benchmark datasets of video anomaly detection, including the UCSD Ped2 [13], CUHK Avenue [24], and ShanghaiTech [25] datasets. These datasets are provided with frame-level ground truth. The training datasets contain normal events, and the testing datasets contain both abnormal and normal events.

A. DATASETS

Here, we provide a brief introduction to the datasets we used. Some examples are illustrated in Figure 3.

1) UCSD Ped2

UCSD Ped2 is a frequently used dataset that contains 16 training videos and 12 testing videos. The scenes in this dataset are captured by a static camera, which is parallel to the walking direction of pedestrians. In the normal cases, there are only pedestrians walking on the sidewalk. However, in the abnormal videos, bicycles or vehicles may appear on the sidewalk; these are considered abnormal events.

2) CUHK AVENUE

CUHK Avenue is a larger dataset than UCSD Ped2. It consists of 30652 frames that are split into 16 training videos and 21 testing videos. The scene in this dataset is a campus avenue, and people go in and out of a building. It contains 47 abnormal events, including loitering, running, and throwing objects into the air. Pedestrians walk in the direction parallel to the surveillance camera, and pedestrians also walk toward the camera.

3) ShanghaiTech

The ShanghaiTech dataset is a newer and more challenging video anomaly detection dataset compared with the previous two datasets. It contains 330 training videos and 107 testing videos with 130 abnormal events. ShanghaiTech contains more than 270,000 training frames and more than 42,000
testing frames. The abnormal events include appearance anomalies (e.g., vehicles) and motion anomalies (e.g., fighting).

B. IMPLEMENTATION AND EVALUATION METRICS

To train our network, we resize all of the frames to 256 × 256 and normalize the pixel values to the range [-1,1]. We sampled clips of 5 sequential frames randomly in the training set. We use the first 4 frames as input, and the last frame as the ground truth of the future frame that we want to predict. We update the generator and discriminators alternately, and use the Adam [26]-based stochastic gradient descent method to optimize parameters.

In the previous literature regarding video anomaly detection, the frame-level criterion is commonly used; this criterion counts a frame with any detected anomalous pixels as a positive frame and all other frames as negative [27]. Here, we first calculate the receiver operation characteristic (ROC) by gradually changing the threshold of the regular scores. Then, we utilize the frame-level area under the curve (AUC) to evaluate the performance.

C. RESULTS

In this section, we compare our approach with different methods, including hand-crafted-based methods [11], [13] and deep-learning-based methods [25], [28]–[31]. MPPCA, MPPCA+SFA, and MDT belong to hand-crafted-based methods. Unmasking [28] is an unsupervised deep learning method that requires no training sequences. Conv-AE [8] and ConvLSTM-AE [29] leverage auto-encoders for video anomaly detection. StackRNN [25] is an approach that combines a deep neural network and sparse coding. MemAE [30] memorizes normality to detect anomalies via a memory-augmented deep auto-encoder. LSA [31] adds latent space autoregression to an auto-encoder for frame reconstruction. FFP [9] determines whether the frames are normal or abnormal by predicting future frames.

The AUC values of different methods are listed in Table 1. Our proposed approach achieves effective performance. The performance is slightly improved with respect to the UCSD Ped2 dataset, and the performance on the ShanghaiTech dataset is more obvious. We think this is because ShanghaiTech is a larger-scale dataset and has more complicated movement, on which the motion discriminator can achieve more obvious performance improvements.

We compared our proposed method with FFP [9] on the ShanghaiTech dataset, which predicts future frames using a single discriminator, and the results are listed in Table 2. In addition to the AUC metric, we report the frame-level equal error rate (EER) of our method and the compared method. A smaller EER corresponds to better performance. We also report the PSNR gap and regular score gap between normal and abnormal events. From the table, we can see that our method achieves larger gaps, which correspond to a small false alarm rate and higher detection rate. Thus, our method performs better than the compared method with respect to all of these evaluation metrics on ShanghaiTech.

Figure 4 illustrates our ROC curve compared with that of the baseline [9] on the ShanghaiTech dataset. The red curve denotes the ROC of our method, and the blue one denotes that of the baseline. It is clear that our method outperforms the baseline at most thresholds.

V. CONCLUSION

In this paper, we propose a dual discriminator-based GAN structure for video anomaly detection and perform experiments on datasets with different scales, which proves the effectiveness of our approach. Specifically, we first predict future frames for normal events, and we use a frame discriminator and motion discriminator to augment the quality of predicted frames. This provides a comprehensive consideration of appearance and motion and can better capture normal patterns. In the testing phase, the events that cannot be predicted well are considered anomalies deviating from normality. In the future, we will attempt to evaluate the quality of predicted frames in a better way, as high PSNR values may not correspond to better prediction quality. Taking into account both appearance and motion information for evaluation can lead to better detection results.

| Method      | Avenue | UCSD Ped2 | ShanghaiTech |
|-------------|--------|-----------|--------------|
| MPPCA       | N/A    | 0.693     | N/A          |
| MPPCA+SFA   | N/A    | 0.613     | N/A          |
| MDT         | N/A    | 0.829     | N/A          |
| Unmasking   | 0.806  | 0.822     | N/A          |
| Conv-AE     | 0.800  | 0.850     | 0.609        |
| ConvLSTM-AE | 0.770  | 0.881     | N/A          |
| StackRNN    | 0.817  | 0.922     | 0.680        |
| MemAE       | 0.833  | 0.941     | 0.712        |
| LSA         | N/A    | 0.954     | 0.725        |
| FFP         | 0.849  | 0.954     | 0.728        |

Our Method: 0.849, 0.956, 0.737
REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jan. 2018, pp. 6479–6488.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, Jul. 2009.

[3] B. Kiran, D. Thomas, and R. Parakkal, “An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos,” J. Imag., vol. 4, no. 2, p. 36, 2018.

[4] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, “Anomalous video event detection using spatiotemporal context,” Comput. Vis. Image Understand., vol. 115, no. 3, pp. 323–333, Mar. 2011.

[5] S. Wu, B. E. Moore, and M. Shah, “Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes,” in Proc. IEEE. Comput. Soc. Conf. Comput. Vis. Pattern Recognit. , Jun. 2010, pp. 2054–2060.

[6] B. Zhao, L. Fei-Fei, and E. P. Xing, “Online detection of unusual events in videos via dynamic sparse coding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR, Jun. 2011, pp. 3313–3320.

[7] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR, Jun. 2011, pp. 3449–3456.

[8] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 733–742.

[9] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for abnormal event—A new baseline,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. , Jun. 2018, pp. 6536–6545.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[11] 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., Jan. 2009, pp. 2921–2928.

[12] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection and video event detection in crowded scenes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1975–1981.

[13] Y. S. Chong and Y. H. Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” in Proc. Int. Symp. Neural Netw. Cham, Switzerland: Springer, 2017, pp. 189–196.

[14] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, “Learning regularity in skeleton trajectories for anomaly detection in videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, p. 11.

[15] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A review on generative adversarial networks: Algorithms, theory, and applications,” 2020, arXiv:2001.06937. [Online]. Available: http://arxiv.org/abs/2001.06937

[16] L. J. Ratliff, S. A. Burden, and S. S. Sastry, “Characterization and computation of local Nash equilibria in continuous games,” in Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton), Oct. 2013, pp. 917–924.

[17] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” 2016, arXiv:1605.05796. [Online]. Available: http://arxiv.org/abs/1605.05796

[18] P. Isoola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[19] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, “Learning to generate long-time future via hierarchical prediction,” in Proc. 34th Int. Conf. Mach. Learn., Aug. 2017, pp. 3560–3569.

[20] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention (Lecture Notes in Computer Science), Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[21] 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.

[22] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2794–2802.

[23] 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.

[24] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked RNN framework,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 341–349.

[25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[26] B. Ramachandra and M. Jones, “Street Scene: A new dataset and evaluation protocol for video anomaly detection,” in Proc. IEEE WACV, Mar. 2020, pp. 2569–2578.

[27] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, “Unmasking the abnormal events in video,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2895–2903.

[28] 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.

[29] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1705–1714.

[30] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, “Latent space autoregression for novelty detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 481–490.

FEI DONG is a Lecturer with Shandong Normal University, Jinan, China. His research interests include computer vision and computer aided design.

YU ZHANG received the B.S. degree from the Dalian University of Technology, Dalian, China, in 2018. She is currently pursuing the M.Sc. degree with Shandong University, Jinan, China. Her research interests include multimedia security and computer vision.

XUISHAN NIE (Member, IEEE) received the Ph.D. degree from Shandong University, in 2011. He was a Research Fellow with the University of Missouri-Columbia, USA, from 2013 to 2014, under the supervision of Prof. Wenjun (Kevin) Zeng. He is currently a Full Professor with the School of Computer Science and Technology, Shandong University of Finance and Economics, China. His research interests include multimedia retrieval and indexing, multimedia security, and computer vision.