Video Anomaly Detection Using Cycle-consistent Adversarial Networks

Zheyi Fan[*1], Mengjie Cui[*2], Di Wu[*3], Yu Song[*4], Zhiwen Liu[*5]

[*1] Beijing Institute of Technology, School of Information and Electronics, Institute of Signal and Image Processing, 5 South Zhongguancun Street, Haidian District, Beijing, China
E-mail: funye@bit.edu.cn

Abstract. Anomaly detection is a challenging work in the area of intelligent video surveillance. It aims to identify abnormal events from monitoring videos. The main challenge of this task is the ambiguity of anomaly definition. In recent years, many researchers exploit hand-crafted features to detect abnormal events, and all these methods follow a two-stage learning strategy, including feature extraction and model establishment. In this paper, we propose an end-to-end anomaly detection framework using cycle-consistent adversarial networks. In the training phase, the representation of regularity is learned from normal video frames and corresponding optical flow images. Our networks are trained with only normal frames, so our model is sensitive to abnormal behavior in abnormal frames. At testing time, the abnormal areas will have larger reconstruction errors than the normal. We can detect abnormal behaviors according to the error between the reconstructed frame and the original frame through a reasonable threshold. Experimental results in challenging datasets show that our method surpasses state-of-the-art methods.

Keywords: Anomaly detection, Cycle-consistent adversarial networks, Cycle-consistency loss, Optical flow, Reconstruction error

1. INTRODUCTION

Anomaly detection plays an important role in public monitoring systems. However, it faces many challenges because the definition of anomaly is ambiguous and environment-dependent. For example, "running" is abnormal on the high way while it is normal on the sports field. The ambiguity and scene dependence will largely affect the accuracy of anomaly detection. So far, many researchers have achieved gorgeous performance in solving this problem [1-4]. Feature reconstruction is usually performed by these works on training sets containing only normal behaviors. Some researchers [2, 5] use hand-crafted features to represent appearance or motion information, and then learn a dictionary to make the reconstruction errors of normal events as small as possible. Thus, the feature of the anomaly event will have a larger reconstruction error than that of a normal one. However, [2, 5] require manual selection of features and the computational complexity of the building process of the dictionary is high. With deep learning showing great advantages in computer vision tasks, Auto-Encoder is exploited to perform feature reconstruction. Hasan et al. [1] proposed a 3D convolutional autoencoder (Conv-AE) to learn the internal representations of normal frames. Nevertheless, the autoencoder requires a large amount of data, which is difficult to meet in actual situations. It usually has a large number of parameters, which increase the difficulty of fitting. Data augmentation is needed in autoencoder-based methods to increase the amount of input data, which means that redundant work for preprocessing the dataset is added for practical applications. Further, 3D convolution cannot characterize the spatial information very well, as shown in the activity recognition [6].

Recently, Generative Adversarial Networks (GANs) [7] have achieved success in the fields of image generation [8], image editing [9] and representation learning [10]. Zhu et al. [11] proposed cycle-consistent adversarial networks (CycleGAN) on the basis of GANs to realize image-to-image translation, such as transforming an image from one domain (e.g. images of zebras) to another (e.g. images of horses). Cycle-consistency loss is adopted in [11] to make the generated images authentic by constraining the reconstructed images to be close to the original ones.

A similar framework is used in our work. However, we do not aim at generating images which look realistic. Instead, we use generators to learn normal patterns in video frames. The cycle-consistency loss is minimized to train the generators and discriminators at the same time. In this way, the reconstructed normal frames is close to the original. Hence, in the testing phase, the reconstructed abnormal frame will be different. Blur will appear in the region corresponding to the abnormal event. Finally, reconstruction errors are calculated between the

Acknowledgements This study was funded by National Natural Science Foundation of China (grant number 61701029) and Beijing Natural Science Foundation-Haidian Original Innovation Joint Fund (grant number L192036), and Industry-University-Research Innovation Foundation of the Science and Technology Development Center of the Ministry of Education (grant number 201920548040).

reconstructed frames and the original ones to easily and robustly detect abnormal behaviors in the frames.

The contribution of our work can be listed as follows.

1. We use only positive samples to train the model, and filter out the abnormal behavior by a reasonable threshold of reconstruction error, solving the problem of fuzzy definition of anomaly and imbalance of positive and negative samples.

2. Optical flow and CycleGAN are simultaneously used for anomaly detection, making full use of appearance and motion information. Through the idea of CycleGAN, the reconstructed video frames is as similar to the original as possible, so as to effectively detect abnormal behavior.

The remainder of our paper is organized as follows. Section 2 gives a brief introduction to previous work on anomaly detection. In Section 3, we specify the proposed method in detail. Section 4 gives the experimental results and analysis. Finally, we draw our conclusions in Section 5.

2. RELATED WORK

2.1. Methods based on hand-crafted features

Methods based on hand-crafted features mainly contains two steps: feature extraction and model establishment. For the former, low-level trajectory features are typically used in early works [12, 13] to represent normal patterns. However, the trajectory features are extracted by tracking-based algorithms that will fail in crowded or occluded scenes. Hence, these methods are not robust in complex scenarios. Considering the limitations of trajectory features, low-level spatial-temporal features, such as Histograms of Oriented Gradients (HOG) and Histograms of Oriented Flows (HOF) are used. Based on spatial-temporal features, Zhang et al. [14] used Markov Random Fields (MRF) to model normal patterns. Adam et al. [15] exploited histograms to measure the probability of optical flow of local blocks. Kim et al. [16] modeled the local optical flow pattern with the Mixed Probability Principal Component Analysis (MPPCA) and enforced global consistency using MRF. Mehran et al. [18] described crowd behavior using a social force model and then used Latent Dirichlet Allocation (LDA) to detect abnormal events. Mahadevan et al. [3] fitted a Gaussian mixture model to Mixture Dynamic Texture (MDT). In addition to these statistical models, there are some works [2, 5, 17] encoding normal behaviors by sparse coding or dictionary learning. The basic assumption of these works is that any regular pattern can be expressed as a linear combination of the basis of a dictionary which encodes normal patterns on the training set. However, the computational cost of the optimization of sparse coefficients is expensive.

2.2. Deep learning based methods

Deep learning has been successfully applied to many computer vision tasks [18, 19], including anomaly detection. Xu et al. [20] designed a multi-layer autoencoder for feature extraction, demonstrating the effectiveness of deep learning features. However, their networks are very shallow. Moreover, additional one-class Support Vector Machines (SVMs) need to be trained on the top of the learned representation. Ravanbakhsh et al. [21] exploited a deep CNN for anomaly segmentation task, and proposed a binary quantization layer plugged as the final layer of the CNN to capture the temporal motion patterns in video frames. However, these works first trained CNNs for other tasks (e.g., object recognition) and then exploited the trained models for anomaly detection. Hasan et al. [1] proposed a 3D Conv-AE to model the video frames which only contain normal events, directly training the deep network for anomaly detection. However, their networks need redundant steps of preprocessing the datasets and they cannot extract the spatial information well.

In our work, we propose an end-to-end anomaly detection framework based on cycle-consistent adversarial networks, training a deep generation network directly for the task of anomaly detection. The proposed method avoids losing spatial information of the reconstructed normal frame by introducing cycle-consistency loss. The proposed method improves the performance of anomaly detection via two closed loops, i.e. video frame → optical flow → video frame and optical flow → video frame → optical flow.

3. METHOD

In this section, we describe our anomaly detection method in detail. First, in Section 3.1, the method of extracting motion features is determined by comparison. Section 3.2 gives an introduction to the cycle-consistent adversarial networks. Then, Section 3.3 introduces the training phase and presents how to identify abnormal events via reconstruction errors.

3.1. Motion feature extraction

Under normal circumstances, people move a short distance in adjacent video frames, and there is a large amount of redundant information in the video, which interferes with model training and leads to high complexity of calculation. Therefore, before extracting motion information, we first use the method in [22] to extract the key frames of the video.

Optical flow refers to the change of light pattern on the plane. In the field of computer vision, it refers to the movement of pixels at various points in a video image over time, reflecting the temporal and spatial changes in the brightness of pixels in the video sequence and the relationship between object motion and structure. It is widely used in motion estimation and behavior recognition for its rich motion information. In our work, optical flow images extracted from the video sequence are input into network as motion feature.

The dense optical flow algorithm is an image registration method proposed by Gunnar Farneback [23]

for point-by-point matching of images. FlowNet is the first attempt to use CNN to directly predict optical flow. It models the optical flow prediction problem as a supervised deep learning problem. FlowNet2.0 [24] makes improvements on FlowNet, with a small concession in speed in exchange for a substantial increase in performance. It stacks multiple FlowNet networks to achieve the effect of "coarse-to-fine" and solves the problem of inaccurate estimation of small displacement of FlowNet.

We compare the performance of the above three optical flow extraction methods on the CUHK Avenue dataset [2]. Some optical flow images are shown in Fig. 1. It can be seen that the optical flow images extracted by the dense optical flow algorithm is the most rough, because traditional methods need to make trade-offs between accuracy and speed. Compared with traditional methods, FlowNet has greatly improved the accuracy, while FlowNet2 algorithm has higher accuracy than FlowNet.

Therefore, in this paper, FlowNet2.0 is selected as the motion extraction method, and the obtained optical flow images and video frames are simultaneously input into the subsequent network for training.

3.2. Cycle-consistent Adversarial Networks

Cycle-consistent Adversarial Networks is a dual image-based image style conversion technology, proposed by Zhu et al. [11]. It uses a circular pair of generators and discriminators to achieve image style conversion, as shown in Fig. 2.

Through G , the mapping relationship between domain X and domain Y is realized, that is, after G , image x in domain X is mapped to image $G(x)$ in domain Y . $D(Y)$ is used to judge whether it is real image or generated one. This constitutes a single-generation adversarial process. However, generating confrontation only in this direction is not enough. It may appear that all the pictures in domain X are mapped to

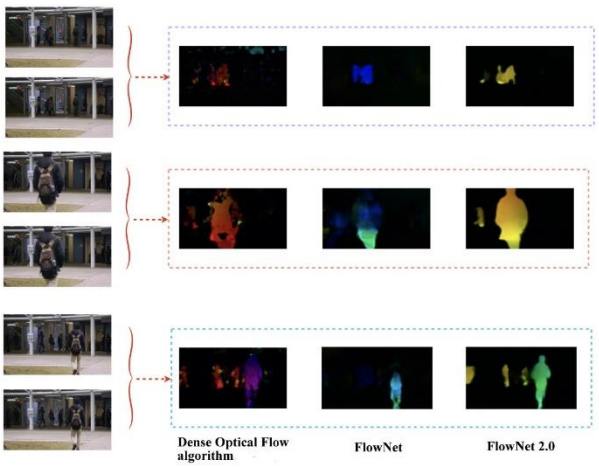

Fig.1 Optical flow images extracted by dense optical flow algorithm, FlowNet and FlowNet 2.0

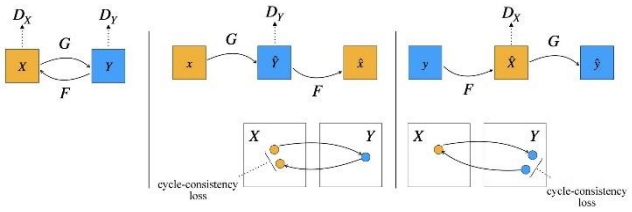

Fig.2 Cycle-consistent Adversarial Networks

the same picture in domain Y . In order to avoid this problem, the generation confrontation process in another direction is added, that is, the image y in domain Y is mapped to domain X to generate the image $F(y)$ through the mapping F . Similarly, $D(X)$ is used to determine whether it is real image or generated one.

The model needs to learn mapping G and mapping F , and at the same time, it must meet the requirement of cycle consistency, that is, $F(G(x)) \approx x$ and $G(F(y)) \approx y$.

3.3. Anomaly detection algorithm based on CycleGAN

In anomaly detection task, in order to learn the pattern of normal behavior, the generative model can be used to reconstruct the normal behavior with the smallest error, then the normal frame and the abnormal one can be distinguished according to the characteristics of the abnormal behavior with a larger reconstruction error.

The dual generation adversarial model of the CycleGAN can be regarded as a reconstruction process in two directions, $x \to G(x) \to \hat{x}$ and $y \to F(y) \to \hat{y}$. This paper reconstructs the video frames (considered as domain X) and its motion features (considered as domain Y) that only contain normal behavior. By reducing cycle-consistency loss, the reconstruction error of normal behavior is as small as possible.

The structure of the overall network to learn normal behaviors is shown in Fig.3. Given original video frames $\{x\}$ and optical flow images $\{y\}$, our goal is to learn mapping functions between domain X and Y . Specifically, we denote the distribution of original normal frames as $x \sim p_{data}(x)$, where $x \in X$, and the distribution of optical flow images obtained by FlowNet2.0 [24] as $y \sim p_{data}(y)$, where $y \in Y$. As illustrated in Fig. 3, our model contains two mappings, $G: X \to Y$ and $F: Y \to X$. Besides, we introduce two adversarial discriminators $D(X)$ and $D(Y)$. $D(X)$ aims to distinguish between the video frame $F(y)$ generated from the original optical flow and the original video frame x . In the same way, $D(Y)$ aims to distinguish between the optical flow $G(x)$ generated from the original video frame and the original optical flow y .

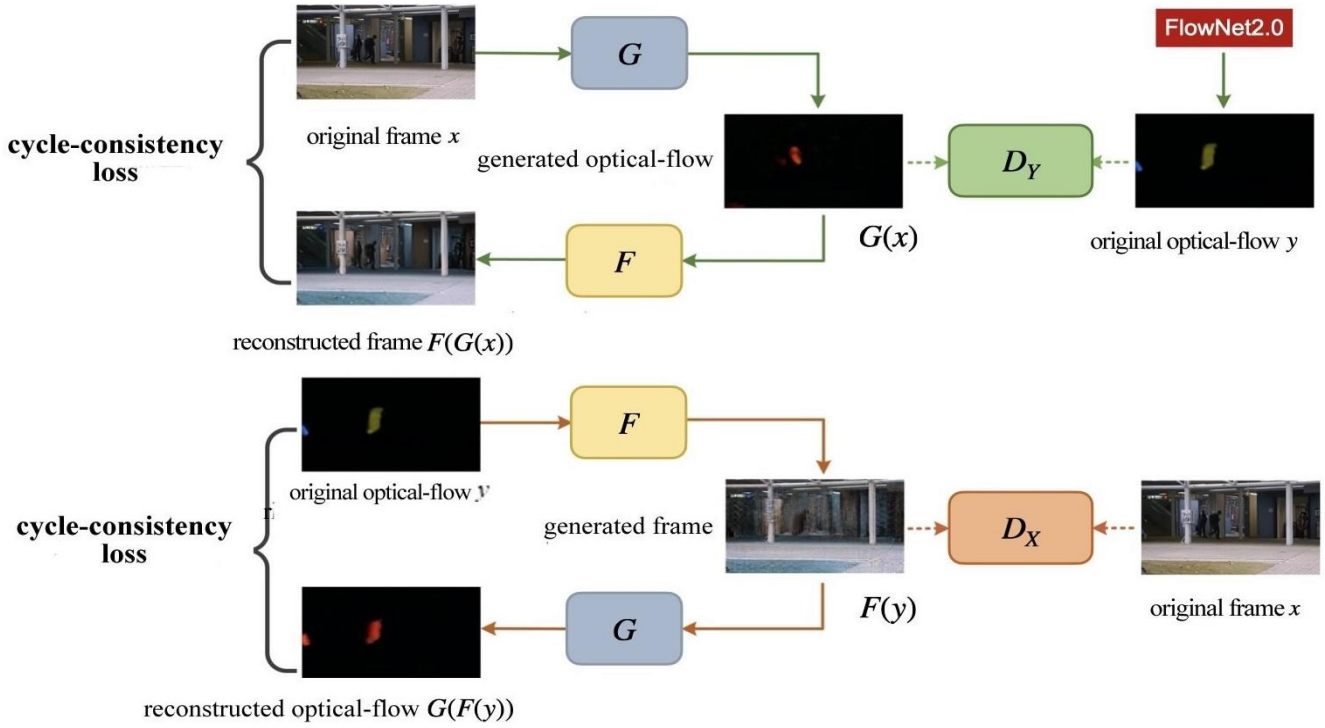

Fig. 3 Overall structure of the proposed method based on cycle-consistent adversarial networks

$G, F, D(X)$ and $D(Y)$ are trained by adversarial loss and cycle-consistency loss. The adversarial loss is used to match the distribution of the generated images with that of the target domain. The cycle-consistency loss is to keep the contour information of the input image and prevent the learned mappings G and F from contradicting each other.

We apply the adversarial loss to both mapping functions. For the mapping function $G: X \to Y$ and its discriminator $D(Y)$, the objective can be expressed as:

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim p_{data}(y)}\left[\log D_Y(Y)\right] + E_{x \sim p_{data}(x)}\left[\log(1 - D_Y(G(X)))\right] \quad (1)$$

where G attempts to generate the optical flow $G(x)$ from the original video frame x, while $D(Y)$ tries to differentiate the generated optical flow $G(x)$ from the original optical flow y. Consequently, G tries to reduce $L_{GAN}(G, D_Y, X, Y)$, while $D(Y)$ increase it. Similarly, for the mapping $F: Y \to X$ and its discriminator $D(X)$, the objective is:

$$L_{GAN}(F, D_X, Y, X) = E_{x \sim p_{data}(x)}\left[\log D_X(x)\right] + E_{y \sim p_{data}(y)}\left[\log(1 - D_X(F(y))\right] \quad (2)$$

Theoretically, adversarial training can learn the mapping G and F which produce outputs identically distributed as target domains X and Y, respectively. However, the adversarial loss is not sufficient to produce desired images, as it leaves the model under-constrained. Thus, cycle-consistency loss is introduced to reduce the possible space of the mapping function. Each video frame x passes through the generation loop, i.e. $x \to G(x) \to F(G(x)) \approx x$, ultimately making the reconstructed video frame $F(G(x))$ almost identical to the original video frame x. In this way, the important information in the video frame can be retained. The loop $x \to G(x) \to F(G(x)) \approx x$ is called forward cycle consistency. For the same reason, the optical flow y should be consistent with the backward cycle consistency, i.e. $y \to F(y) \to G(F(y)) \approx y$. The cycle-consistency loss is denoted as:

$$L_{cyc}(G, F) = E_{x \sim p_{data}(x)}\left[\left\|F(G(x)) - x\right\|_1\right] + E_{y \sim p_{data}(y)}\left[\left\|G(F(y)) - y\right\|_1\right] \quad (3)$$

In summary, our complete objective function is:

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F) \quad (4)$$

where λ controls the relative importance of adversarial loss and cycle-consistency loss.

Theoretical analysis shows that cycle-consistency loss coefficient plays an important role in image generation, and the quality of the generated image can be controlled by adjusting the coefficient λ.

Then, our goal is to solve:

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} L(G, F, D_X, D_Y) \tag{5}$$

In Section 4.3, we analyze the influence of different λ on the accuracy of anomaly detection and compare the effects of our method with ablations of the complete objective function for anomaly detection in the videos.

In the testing phase, we can use the difference between the reconstructed video frame and the original video frame for anomaly detection. Specifically, the testing video frames are input into the network and then reconstructed through the generators G and F sequentially. Since the network only learns normal behaviors in the training phase, the region representing anomaly will have a larger reconstruction error than that of the normal region. As shown in Fig. 4, the portion of the reconstructed frame corresponding to the bag thrown by a person is blurred. Following [25], we use Peak Signal-to-Noise Ratio (PSNR) which assesses image quality to measure the reconstruction error, which can be calculated by:

$$PSNR(x, \hat{x}) = 10\log_{10} \frac{\left[\max_{\hat{x}}\right]^2}{\frac{1}{N}\sum_{i=0}^{N}(x_i - \hat{x}_i)^2} \tag{6}$$

where x denotes the original frame, $\hat{x} = F(G(x))$ represents the reconstructed frame and N is the total number of pixels. We normalize the PSNR of all frames in the testing set to the range [0,1] and calculate the regular score via equation (7). M is the total number of the testing video frames and k means the k-th frame. Higher $S(k)$ implies that the reconstructed frame is more similar to the original frame, further implying that the frame is more likely to be normal. A threshold can be set to identify whether the frame is normal or not. The best threshold is obtained by training and cross-validation of a small amount of labeled data in the scene.

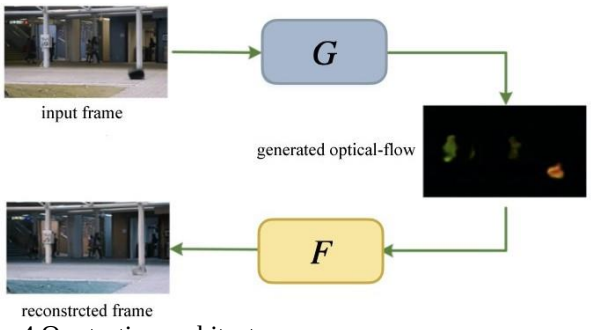

Fig. 4 Our testing architecture

$$S_k = \frac{PSNR(x_k, \hat{x}_k) - \min_{0\le l\le M} PSNR(x_l, \hat{x}_l)}{\max_{0\le l\le M} PSNR(x_l, \hat{x}_l) - \min_{0\le l\le M} PSNR(x_l, \hat{x}_l)} \tag{7}$$

4. EXPERIMENTS

4.1. Datasets

Avenue The CUHK Avenue dataset [2] contains 16 training and 21 testing videos. In total, there are 15328 frames in the training set and 15324 frames in the testing set. The resolution of each frame is 640×360. There are 47 abnormal events, including throwing, loitering and running. In the dataset, the size of the person changes with the position and angle of the camera.

UMN The UMN dataset [26] consists of three different crowded scenes, each with 1453, 4144, 2144 frames, respectively. The resolution is 320 × 240. In these three scenarios, the normal event is that people walking around, and the abnormal behavior is defined as people running in all directions. Consistent with [5], we use the first 400 normal frames of each scene for training.

4.2. Evaluation metric

In [2, 3], the Receiver Operating Characteristic (ROC) is calculated by gradually changing the threshold of the regular score, and then the Area Under Curve (AUC) is evaluated for performance. The higher the value of the AUC is, the better the performance of anomaly detection achieves. Following [1, 25, 27], we exploit frame-level AUC for performance evaluation.

4.3. Analysis of the loss function

First, we analyze the effects of different λ in equation (4) in our method. If the value of λ is large, the training process is dominated by the cycle-consistency loss, and the generated image can better retain the contour, but cannot be close to the distribution of the target domain. On the contrary, the generated image can better approximate the target domain distribution, but cannot well retain the contour of the input image. Therefore, the gap between the two needs to be balanced. In the original CycleGAN, the value of λ defaults to 10. In order to ensure that λ changes within a reasonable range, and to observe the trend of AUC with λ, we select six values of λ in our verification experiment, namely 0.5, 1, 5, 10, 15 and 20. On the one hand, we can verify the effectiveness of the cycle-consistency loss. On the other hand, we can intuitively observe the influence of the value of λ on the accuracy of anomaly detection within a reasonable range, so as to select an appropriate coefficient. Experimental results on the Avenue dataset and UMN dataset are shown in Fig. 5, where on (a) the Avenue dataset, (b) the first scene of the UMN dataset, (c) the second scene of the UMN dataset, (d) the third scene of the UMN dataset. It can be observed from equation (4) and Fig. 5 that the value of λ will affect the loss function and thus the final AUC, which is the highest when $\lambda = 10$ for all datasets.

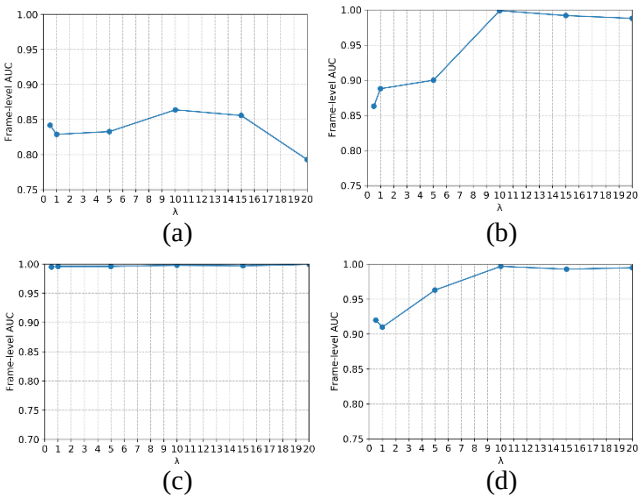

Fig. 5 Frame-level AUC with different λ

Table 1 AUC of different loss function on Avenue dataset

Loss	AUC
GAN alone	82.3 %
Cycle alone	75.6 %
GAN + forward cycle	85.4 %
GAN + backward cycle	83.9 %
CycleGAN(ours)	**86.4%**

Further, we compare the different parts of the loss function in Table 1. Taking the Avenue dataset as an example, it can be observed that removing the GAN loss or the cycle-consistency loss degrades the AUC score. Therefore, we conclude that the GAN loss and the cycle-consistency loss are essential to our anomaly detection method. We also evaluate the effect of the cycle loss in one direction, i.e. GAN+forward cycle loss $E_{x\sim p_{data}(x)}\left[\left\|F(G(x))-x\right\|_1\right]$, or GAN+backward cycle loss $E_{y\sim p_{data}(y)}\left[\left\|G(F(y))-y\right\|_1\right]$ in equation (3), and find that the AUC decreases somewhat compared to the full objective function. In summary, the cycle losses in both directions improve the performance of anomaly detection.

4.4. Results on the Avenue dataset

We compare the proposed method based on the cycle-consistent adversarial networks with the existing methods [1, 4, 25, 27-30] on the Avenue dataset. The AUC scores for the different methods are listed in Table 2. As can be seen from the table, our approach can surpass the performance of the latest methods.

In Fig. 6, we show the frame-level anomaly scores (between 0 and 1) of the 5th and 6th testing videos in the Avenue dataset produced by our framework. The anomaly score of each frame is given by:

$$S'(k)=1-S(k) \tag{8}$$

Table 2 AUC of different methods on Avenue Datasets

Method	AUC
Conv-AE [1]	80.0 %
DeepAppearance [4]	84.6 %
Liu et al. [25]	85.1 %
Stacked RNN [27]	81.7 %
Del et al. [28]	78.3 %
ConvLSTM-AE [29]	77.0 %
Unmasking [30]	80.6 %
Proposed method	**86.4 %**

where $S(k)$ is obtained by equation (7). Higher anomaly score means that the video frame is more likely to contain abnormal events. It can be noticed that our scores correlate well to the ground truth. We mark four video frames including normal and abnormal behaviors in Fig. 6. In the last place where the "throwing" behavior occurs, it is labeled as abnormal in the ground truth, while the anomaly score has a large decrease. In fact, by checking the testing set, we find that the person throwing a bag does not appear in several frames, so the anomaly score dropping sharply here is in line with the actual situation. The effectiveness of our method is demonstrated.

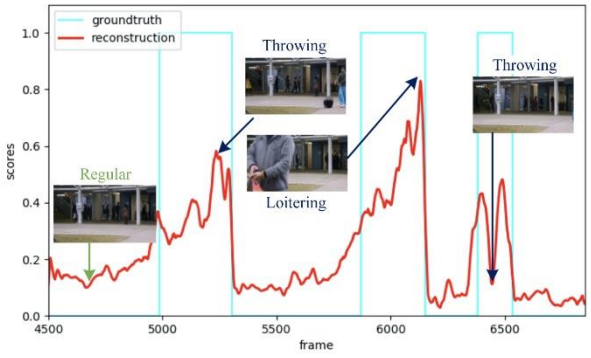

Fig. 6 Frame-level anomaly detection scores

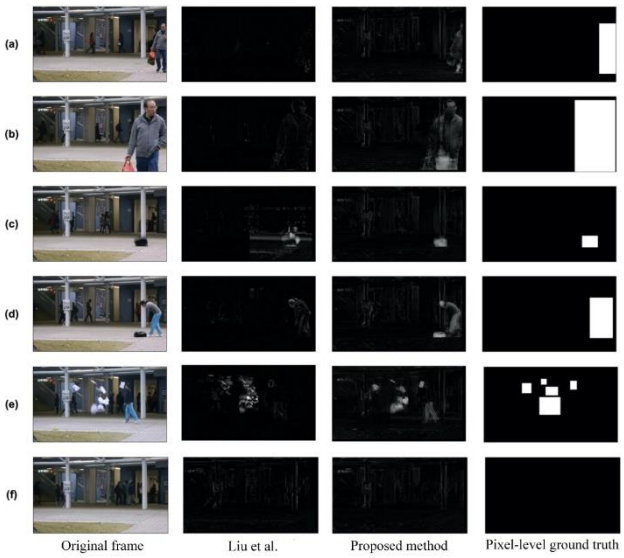

Fig. 7 Visual results in pixel-level

Fig. 7 shows the visual results of the algorithm of this paper and the method in [25] in pixel-level. They both draw on the idea of GAN, and use optical flow as a constraint on the network. The difference with the algorithm proposed in [25] is that our method uses the idea of reconstruction to detect abnormal behavior, and retains the details of the image through the reconstruction of frames and optical flow, which can better learn the normal behavior pattern.

In Fig. 7, the first column is the original frames. The second is the visualization results of [25], illustrating the reconstruction errors between the original video frames and the reconstructed video frames. And we present our visualization results in the third column. Finally, we show the pixel-level ground truth in the last column. Rows (a)-(f) represent (a) 364th frame in video 6. (b) 541st frame in video 6. (c) 823rd frame in video 6. (d) 961st frame in video 6. (e) 100th frame in video 20. (f) 1261st frame in video 12, respectively. As shown in Fig. 7(a) and (b), [25] cannot detect the beginning of the abnormal behavior. And it can be observed in Fig. 7(c), (d) and (e) that [25] fails in detecting bag and paper thrown by a person, while our method can clearly screen out the objects that have not been learned. Fig. 7(f) displays the results of the normal video frame, indicating that our method is comparable to [26] in reconstructing normal events, which has small reconstruction errors.

4.5. Results on the UMN dataset

In Table 3, we compare the proposed method with [4-5, 26, 31-34] on the UMN dataset. The frame-level AUC for each scene and the average value for the dataset are reported. From the table, our method outperforms the state-of-the-art methods on both the first scene and the second scene and achieves comparable results for the third scene.

In Fig. 8, we show the frame-level anomaly scores (between 0 and 1) calculated by equation (8) for the second scene. We can find that when the threshold is set to around 0.5, the abnormal behaviors can be accurately detected without any false positive detections. It can be observed that the frame has a local maximum score when people run in all directions. In addition, the scores increase before some abnormal events occur, for example, before the 1367th and 2086th frame of the second scene. By carefully studying the dataset, it is known that people begin to run in all directions before the occurrence of these two anomaly events, as shown in Fig. 9, where the left and right subgraphs is the 1348th and 2077th frame of the second scene. This proves the good performance of our method in real world. Because the proposed algorithm allows us to be aware of abnormal events before they occur, so that we can deal with them in time, meeting the real-time requirements of abnormal detection.

4.6. Running time

Our proposed method is implemented with TensorFlow. We benchmark the performance of our method on the

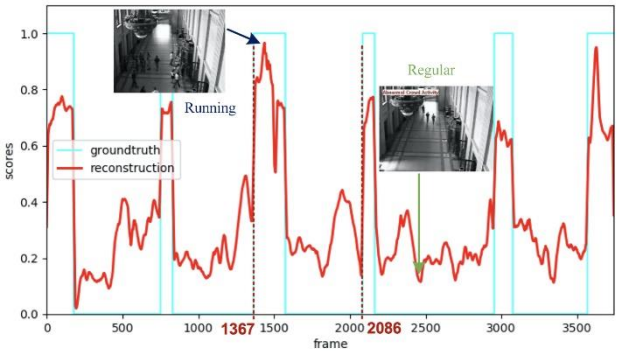

Fig. 8 Frame-level anomaly detection scores provided by our framework for the second scene in the UMN dataset

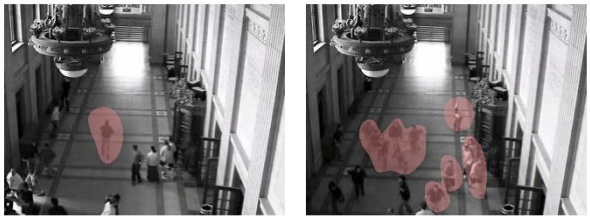

Fig. 9 Frames in the UMN datasets that people began to run in all directions before the two events are labeled as "abnormal"

Avenue and UMN dataset. All testing phases are performed on NVIDIA TITAN Xp GPUs with Intel Xeon(R) E5-2650 2.20GHz CPUs. The average running time is 13 fps, including the whole reconstruction process. We also report the running time of other methods such as 0.5 fps in [17], 20 fps in [30] and 25 fps in [25].

5. CONCLUSION

In this paper, we present an anomaly detection structure based on cycle-consistent adversarial networks. The proposed network learns the internal representation of the scene with normal video frames and corresponding optical flow images. Cycle-consistency loss is used to minimize the reconstruction errors of the normal video frames and the optical flow images. In the testing phase, the regions corresponding to the abnormal events will have a larger reconstruction error than the normal events. Our method makes full use of appearance and motion information to reconstruct video frames. After the confrontation process of CycleGAN, our model has the ability to accurately reconstruct normal behaviors, thereby accurately judging abnormal ones. A large number of experimental results on public datasets show that our proposed method outperforms existing approaches, which demonstrates the effectiveness of our anomaly detection method. In the future, we will reduce the computational complexity on the basis of ensuring the accuracy of the algorithm, and focus on improving the real-time performance of the algorithm to better apply it in actual scenarios.

Table 3 AUC of different methods on the UMN dataset

Method	Scene 1	Scene 2	Scene 3	All scenes
DeepAppearance [4]	98.8 %	93.6 %	98.9 %	97.1 %
Cong et al. [5]	99.5 %	97.5 %	96.4 %	97.8 %
Mehran et al. [25]	-	-	-	96.0 %
Saligrama et al. [30]	-	-	-	98.5 %
Sun et al. [31]	99.8 %	99.3 %	**99.9 %**	99.7 %
Wang et al. [32]	99.9 %	94.5 %	97.6 %	97.3 %
Zhang et al. [33]	99.2 %	98.3 %	98.7 %	98.7 %
Proposed method	**99.9 %**	**99.8 %**	99.7 %	**99.8 %**

REFERENCES:

[1] Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 733-742.

[2] Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: IEEE International Conference on Computer Vision, pp 2720-2727.

[3] Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1975-1981.

[4] Smeureanu S, Ionescu RT, Popescu M, Alexe B (2017) Deep appearance features for abnormal behavior detection in video. In: International Conference on Image Analysis and Processing, pp 779-789.

[5] Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3449-3456.

[6] Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1):221-231.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp 2672-2680.

[8] Denton E, Chintala S, Szlam A, Fergus R (2015) Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems, pp 1486-1494.

[9] Zhu JY, Kr¨ahenb¨uhl P, Shechtman E, Efros AA (2016) Generative visual manipulation on the natural image manifold. In: European Conference on Computer Vision, pp 597-613.

[10] Mathieu MF, Zhao JJ, Zhao J, Ramesh A, Sprechmann P, LeCun Y (2016) Disentangling factors of variation in deep representation using adversarial training. In: Advances in Neural Information Processing Systems, pp 5040-5048.

[11] Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, pp 2223-2232.

[12] Tung F, Zelek JS, Clausi DA (2011) Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. Image Vision Computing 29(4):230-240.

[13] Wu S, Moore BE, Shah M (2010) Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2054-2060.

[14] Zhang D, Gatica-Perez D, Bengio S, McCowan I (2005) Semi-supervised adapted HMMs for unusual event detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol 1, pp 611-618.

[15] Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3):555−560.

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

[17] Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3313-3320.

[18] Girshick R (2015) Fast R-CNN. In: IEEE International Conference on Computer Vision, pp 1440-1448.

[19] Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097-1105.

[20] Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding 156:117-127.

[21] Ravanbakhsh M, Nabi M, Mousavi H, Sangineto E, Sebe N (2018) Plugand-play CNN for crowd motion analysis: An application in abnormal event detection. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 1689-1698.

[22] Liu H, Meng W, Liu Z, et al. Key frame extraction of online video based on optimized frame difference[C]. fuzzy systems and knowledge discovery, 2012: 1238-1242.

[23] Farnebäck G. Two-frame motion estimation based on polynomial expansion[C]. Scandinavian Conference on Image Analysis. 2003: 363–370.

[24] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2462-2470.

[25] Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection{a new baseline. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 6536-6545.

[26] Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 935-942.

[27] Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked RNN framework. In: IEEE International Conference on Computer Vision, pp 341-349.

[28] Del Giorno A, Bagnell JA, Hebert M (2016) A discriminative framework for anomaly detection in large videos. In: European Conference on Computer Vision, pp 334-349.

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

[30] Tudor Ionescu R, Smeureanu S, Alexe B, Popescu M (2017) Unmasking the abnormal events in video. In: IEEE International Conference on Computer Vision, pp 2895-2903.

[31] Saligrama V, Chen Z (2012) Video anomaly detection based on local statistical aggregates. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2112-2119.

[32] Sun Q, Liu H, Harada T (2017) Online growing neural gas for anomaly detection in changing surveillance scenes. Pattern Recognition 64:187-201.

[33] Wang T, Qiao M, Zhu A, Niu Y, Li C, Snoussi H (2018) Abnormal event detection via covariance matrix for optical flow based feature. Multimedia Tools and Applications 77(13):17375-17395.

[34] Zhang Y, Lu H, Zhang L, Ruan X, Sakai S (2016) Video anomaly detection based on locality sensitive hashing filters. Pattern Recognition 59:302-311.