IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 13, 2021, accepted April 27, 2021, date of publication May 4, 2021, date of current version May 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077577

Anomaly Detection Based on Latent Feature Training in Surveillance Scenarios

YONG QIANG, SHUMIN FEI, AND YIPING JIAO
Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
Corresponding author: Shumin Fei (smfei@seu.edu.cn)

ABSTRACT Anomaly detection in videos is challenging due to the scarcity and variance in positive samples. Current anomaly detection methods can be categorized into reconstruction models and future frame prediction-based models. However, reconstruction models might be exceptionally adapted to abnormal events due to the learning capacity and generalization ability of deep neural networks, whereas prediction-based methods can be sensitive to noise. In this study, we propose an anomaly detection model based on the latent feature space, which combines advantages from both sides. We argue that the constraints in the latent feature space can promote reconstruction; moreover, the optical flow is also considered to introduce temporal information. We use SPyNet for accurate and efficient optical flow estimation. We extensively validate our method on the UCSD Ped1, UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets. The results demonstrated the feasibility of the proposed method and the benefit of utilizing information in the latent feature space.

INDEX TERMS Anomaly detection, GAN, latent feature vector, SPyNet.

I. INTRODUCTION

Video anomaly detection is an important research field in computer vision. Typically, samples with normal behavior represent the majority of the dataset, whereas only limited abnormal samples are available. The imbalance of samples, as well as the variance in abnormal behaviors and the complexity of monitoring scenarios, lead to difficulties in anomaly detection. Normally, deep learning models are trained in a supervised manner, which requires considerable annotated data and computing resources. However, this is not always feasible in the anomaly detection field.

Therefore, some studies considered anomaly detection based on reconstruction. Specifically, the model is trained with only normal samples for the reconstruction task. Various methods have been proposed to enhance the reconstruction. Reference [1] trained a fully convolutional autoencoder with manually annotated temporal and spatial data, and anomaly detection was based on reconstruction loss. Reference [2] used time-coherent sparse coding to encode two adjacent frames with similar reconstruction coefficients, thereby reducing the number of calculations, optimizing, and accelerating anomaly detection. In this type of method, the anomaly is recognized by monitoring the reconstruction error. Since the model is trained with normal samples only, it is expected to have high loss when anomalous samples are presented. However, due to the high learning capacity and the generalization ability of a convolutional neural network, it might fit abnormal events unexceptionally, leading to failure in anomaly detection.

Different from the reconstruction-based methods, the future frame prediction model compared the predicted frame with the real frame for anomaly detection. To realize future frame prediction in normal samples, GANs are usually used to enhance the predictive ability [3]–[10]. Moreover, constraints in motion and gradient are also proven effective. Reference [11] proposed a framework based on future frame prediction to detect anomalies. However, the future frame prediction method can be sensitive to noise and perturbation, especially in scenes with illumination changes, leading to inferior robustness in anomaly detection.

The double U-net [12] structure used in [13] has achieved good experimental results. In the reconstruction experiment, we found that although the U-net structure can recognize the

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

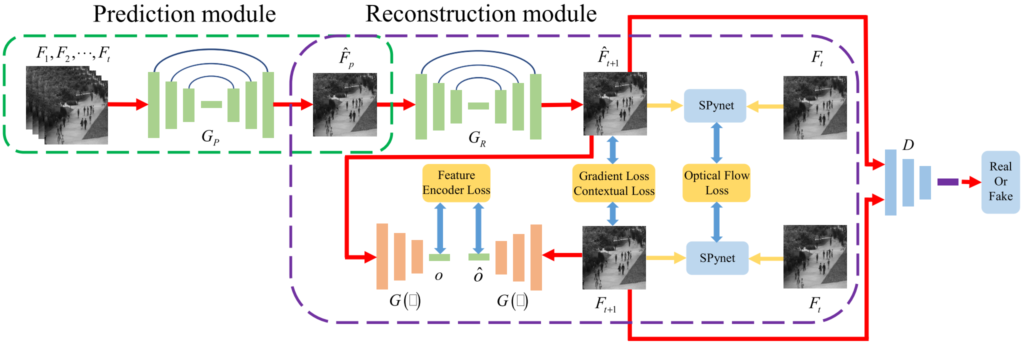


FIGURE 1. **The pipeline of our method.**

multiscale distribution in the image space, the low-latitude information of the image cannot be fully recognized. Inspired by the literature [11], [13], this paper proposes an anomaly detection model based on latent feature constraints. The model combines the advantages of the reconstruction method and the future frame prediction method. At the same time, it minimizes the reconstruction error of the potential feature vector in the image so that the global information can be more fully identified and reconstructed. In addition, the motion constraint in the previous literature uses the FlowNet [14] optical flow method. In our method, we use SPyNet [15] to constrain the predicted frame to be consistent with the ground truth. Since SPyNet is a pyramid structure, the model structure is smaller and more straightforward, only 4% of FlowNet, and its accuracy and calculation speed are better than FlowNet.

The frame of our method is shown in Figure 1. The main contributions of this paper are as follows: (1) An anomaly detection model based on latent feature constraints is designed, which combines the advantages of prediction-based methods and reconstruction-based methods. (2) We use SPyNet for accurate and efficient optical flow estimation, which enhances the reconstruction and future frame prediction modules. (3) The proposed method is validated and compared with several competing methods on multiple public datasets, demonstrating the effectiveness of the utilization of constraints in the latent feature space and optical flow in anomaly detection.

II. RELATED WORK

A. TRADITIONAL ANOMALY DETECTION

In the past, the anomaly detection field usually adopted three types of methods, including feature extraction-based methods, classification-based methods and clustering-based methods, and sparse reconstruction-based methods.

The feature extraction-based method manually designs a description of a video. For example, in a crowd scene, the moving speed of a target can be considered. Other descriptions were also used in the literature. For example, [16] used optical flow as an effective target motion description feature. Reference [17] used context information to use space-time blocks as basic events. Many pieces of literature used different features for anomaly detection. Some documents used features such as target position, target contour, and target trajectory [18]–[21].

The literature [22] used various features to express the essential events jointly and then used the multicore learning method to train the classifier for abnormal event detection. The literature [23] defined anomalies as having local temporal and spatial features. The distance between the normal video and the test video K-nearest neighbors (K-NN) is used to calculate the anomaly score. The literature [24] proposed an algorithm for trajectory clustering, which clustered the detected objects' motion trajectories and then distinguished between normal behavior and abnormal behavior according to the different degrees of trajectory clustering.

Sparse Representations. The usual practice for abnormal event detection is to build a normal event dictionary. The criterion for judging whether the event is abnormal is to use the weighted sum of the reconstruction error and the sparse constraint of the solution coefficient as the objective function and calculate whether the function's minimum value exceeds a predefined threshold. The literature [25] treated anomalous event detection as a low-rank matrix reconstruction problem and decomposed each column of the matrix with low rank and then reconstructed it. A threshold was defined based on the reconstruction error for anomaly detection.

B. RECONSTRUCTION-BASED METHODS

Deep neural networks can autonomously learn to express features from images and videos [26]. Researchers no longer need to spend much time and energy manually labeling data, bringing a high degree of convenience and speed to the field of anomaly detection research. The reconstruction method trains the model on a given normal behavior sample and lets the model generate images that are as consistent as possible with the normal sample. When testing the model, the abnormal event samples that have not been trained are input, and the reconstruction error is relatively large. Conversely, when inputting normal behavior samples, the reconstruction error value is relatively small. This distinguishes abnormal frames from normal frames [1], [3]. Reference [27] improved MemAE, a self-encoder, which was reconstructed by querying related memory items. During the training phase, the memory content was constantly updated, and elements representing normal sample data were encouraged. In the test phase, the memory content was fixed, making the model more distinguishable between normal and abnormal frames. Reference [28] used CNN for appearance coding, and ConvLSTM memorized motion information. Combining the two modules with the ConvLSTM-AE autoencoder can well reconstruct normal frames. Reference [29] designed a reconstruction structure of the appearance decoder and a motion decoder that shares an encoder by learning the correspondence between the appearance of the target and its related motion.

C. PREDICTION-BASED METHODS

Due to the rapid development of reconstruction methods, their learning ability is very strong. Sometimes, they can achieve a good reconstruction of abnormal frames and obtain smaller reconstruction errors to make the system misjudge. Therefore, anomaly detection methods based on predicting future frames are attracting attention. Because predicting future frames trains and learns from consecutive frames in the video, the motion characteristics and appearance characteristics of the predicted frames and the ground truth are minimized through some constraints. We Judge things beyond the forecast as abnormal. Therefore, based on the prediction method, it is expected that the reconstruction error of abnormal frames increases and the accuracy of abnormal detection is proven. Reference [30] trained a convolutional network to predict future frames from video sequences and proposed three different feature learning strategies. This research promoted future frame prediction research. Reference [31] proposed a spatiotemporal autoencoder, which learns spatiotemporal features through three-dimensional convolution and introduces a prediction loss to generate future frames. Reference [11] proposed for the first time a method for predicting future frames based on U-net's generation of confrontation network structure, which is a benchmark for predicting future frame anomaly detection. The method in [13] is based on the improvement of [11], which connects the prediction module and the reconstruction module in series.

III. ANOMALY DETECTION BASED ON LATENT FEATURE CONSTRAINTS

According to the previous introduction, on the one hand, anomaly detection based on the reconstruction method is a network structure with strong modeling ability and strong generalization ability. Nevertheless, with such a powerful ability, it is easy to reconstruct abnormal frames. The reconstruction error of the abnormal frame is not large. On the other hand, anomaly detection based on predicting future frames attempts to improve the defects of the reconstruction method. It defines the unexpected event as an abnormal event, inputs consecutive frames, and, through some constraints, forces the future frames to be consistent with the ground truth. In the experiment, we found that if we add Gaussian noise to the training samples and test samples, the AUC value of [11] drops faster. Its anti-noise ability is poor. It is impossible to obtain an accurate reconstruction error value for the test video with considerable noise, which easily causes misjudgment of normal frames. Thus, it is not competent for anomaly detection in more complex monitoring scenarios.

The U-net network structure in the reconstruction module does not consider the factors of the potential feature vector of the reconstructed frame [13], which makes the reconstruction result not as expected.

In summary, this article combines the prediction module and the reconstruction module in the generative adversarial network training framework and imposes latent feature constraints and SPyNet constraints in the reconstruction module, minimizes the reconstruction error value of the image and latent feature vectors on the reconstructed frame, which helps the model learn according to the normal distribution and completes better reconstruction work. The model structure of this article is introduced as follows.

A. STRUCTURE DESCRIPTION

Figure 1 shows the entire anomaly detection model. The whole model contains three parts: the prediction module, the reconstruction module, and the generative adversarial network module.

1) FUTURE FRAME PREDICTION MODULE

The prediction module is a U-net network with input frames $F = (F_1, F_2, \cdots F_t)$ to generate an intermediate frame $\hat{F}_p$. $\hat{F}_p$ contains some vital information in the prediction module. Figure 2 shows the U-net network structure [12], [13]. After adding a layer of layers, the convolution and deconvolution kernel size is 3×3, and the size of the maximum pooling layer is set to 2×2. We adjust the shape of all frames in the datasets to 256×256.

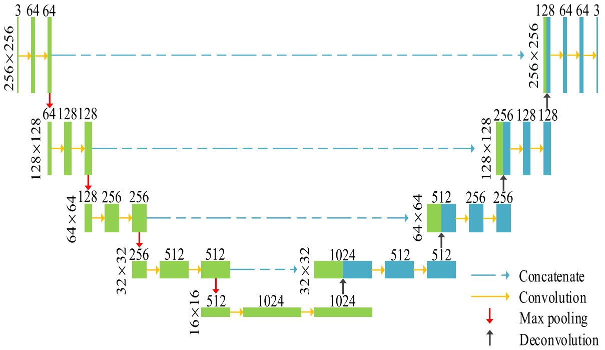


FIGURE 2. U-net network structure.

2) RECONSTRUCTION MODULE

The reconstruction module encodes and reconstructs the frame input sequence, which can be used for future frame prediction $\hat{F}_{t+1}$.

We use the U-net structure to reconstruct and retain the multiscale distribution information, minimize the distance between $\hat{F}_{t+1}$ and the ground truth under the constraints of latent features, and $\hat{F}_{t+1}$ conforms to the normal distribution learning of the ground truth F_{t+1}. The motion information can also be better reconstructed under the constraints of SPyNet. In addition, we also used other constraints. The structure of the latent feature encoder is shown in Figure 3. The encoder reads the input frame through the convolutional layer, batch norm, and leaky ReLU activation. $\hat{F}_{t+1}$ and F_{t+1} are compressed into latent feature vectors o and $\hat{o}$, respectively.

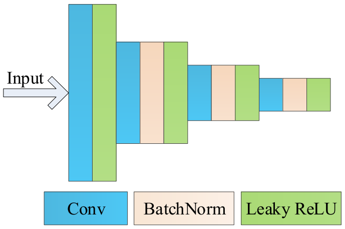


FIGURE 3. The structure of the latent feature encoder.

3) GAN MODULE

The prediction module and the reconstruction module form the generator. D in Figure 1 is the discriminator, and the role of D is to compare the reconstructed frame $\hat{F}_{t+1}$ with the ground truth F_{t+1} to determine whether the input is real or fake. The discriminator structure refers to the patch discriminator [32].

B. LOSS FUNCTIONS

Therefore, the entire structure needs to be carefully designed with the loss function. The following loss functions are combined into an objective function to make the image training more accurate and effective by means of adversarial training and reducing distances regarding latent vectors.

1) FEATURE ENCODER LOSS

To make the reconstructed frame $\hat{F}_{t+1}$ coincide with the context of the ground truth F_{t+1}, the latent feature vector o in the anomaly detection frame is approximated to the feature vector $\hat{o}$, where $o = G(\hat{F}_{t+1})$ and $\hat{o} = G(F_{t+1})$. We use the feature encoder loss L_{FE} to minimize the distance between o and $\hat{o}$. The L2 loss function can make the training converge faster when the discrepancy is large and is formulated as:

$$L_{FE} = \left\| G(\hat{F}_{t+1}) - G(F_{t+1}) \right\|_2 . \tag{1}$$

2) GRADIENT LOSS AND CONTEXTUAL LOSS

The gradient constraint aims to make the reconstructed frame retain the gradient with the ground truth. The gradient loss is defined as the difference between the absolute gradient along with the reconstructed frame and the ground truth in the space dimension. The contextual loss will make the predicted frame approach the ground truth from each pixel value. Reference [32] shows that the L1 loss can improve the sharpness of the generated image compared to the L2 loss [33].

According to the following formula:

$$L_{GT} = \sum_{i,j} \left\| |\hat{F}_{i,j} - \hat{F}_{i-1,j}| - |F_{i,j} - F_{i-1,j}| \right\|_1 + \left\| |\hat{F}_{i,j} - \hat{F}_{i,j-1}| - |F_{i,j} - F_{i,j-1}| \right\|_1 , \tag{2}$$

where i, j denote the spatial index.

$$L_{CL} = \left\| \hat{F}_{t+1} - F_{t+1} \right\|_1 . \tag{3}$$

3) OPTICAL FLOW LOSS

Although the feature encoder loss and gradient loss described above can make the reconstruction and the encoded latent vector similar to the ground truth, it is not guaranteed that the motion estimation in the video prediction frame is reasonable. The SPyNet optical flow calculation method [15] is a method that combines a spatial pyramid structure and deep learning to calculate optical flow. The advantages of this method are listed as follows. First, the model has few parameters, which makes the operation more efficient. Second, the convolution filter is improved compared to FlowNet making the optical flow estimation more accurate. In this paper, SPyNet is used to estimate the optical flow, and then the temporal loss is used to constrain the optical flow of the predicted frame and the ground truth. Defining $f\,()$ as the optical flow estimation performed by SPyNet, the optical flow loss can be formulated as follows:

$$L_{OF} = \left\| f(\hat{F}_{t+1}, F_t) - f(F_{t+1}, F_t) \right\|_1 . \tag{4}$$

4) ADVERSARIAL LOSS

GAN is composed of a generator G and a discriminator D. D is trained to determine whether the input is from a real distribution or is generated by G. In contrast, G is trained to deceive D by generating samples that are indistinguishable from real samples.

D is expected to classify the ground truth video frame F_{t+1} as class 1 (1 represents the 'real' label) and classifies the video prediction frame $\hat{F}_{t+1} = G(F)$ generated by G as class 0 (0 represents the 'fake' label). The mean absolute error (MAE) loss function is shown below:

$$L^D_{adv}(\hat{F}, F) = \sum_{i,j} \frac{1}{2} L_{MAE}(D(F)_{i,j}, 1) + \sum_{i,j} \frac{1}{2} L_{MAE}(D(\hat{F})_{i,j}, 0), \tag{5}$$

where i, j is the spatial index and the L_{MAE} function is:

$$L_{MAE}(\hat{Z}, Z) = \left| \hat{Z} - Z \right| , \tag{6}$$

where $\hat{Z} \in [0, 1]$, and $Z = 1$ or $Z = 0$.

G is trained to generate video prediction frames $\hat{F}_{t+1}$ and make the discriminator D output 'real' when a generated frame is shown. The loss for adversarial training is

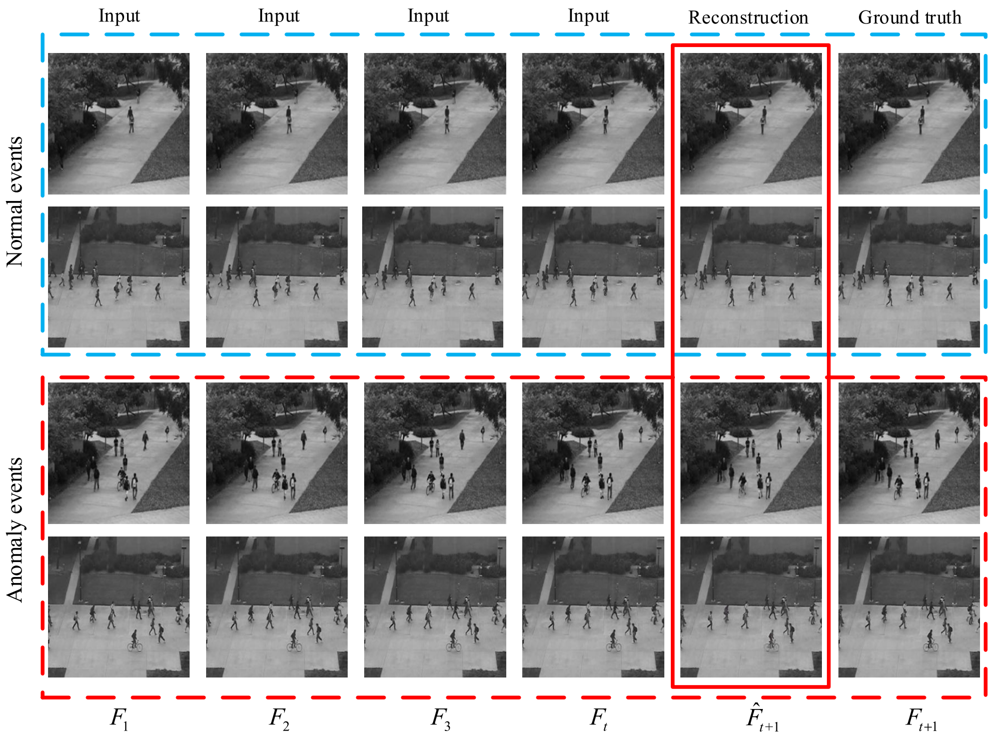


FIGURE 4. Examples of predicted future frames.

formulated as:

$$L_{adv}^{G}(\hat{F}) = \sum_{i,j} \frac{1}{2} L_{MAE}(D(\hat{F})_{i,j}, 1). \tag{7}$$

C. OBJECTIVE FUNCTION

The above constraints are combined: feature encoder loss, gradient loss, optical flow loss, and adversarial loss to form the objective function:

$$\begin{aligned} L_G &= \lambda_{FE} L_{FE}(\hat{F}_{t+1}, F_{t+1}) + \lambda_{GT} L_{GT}(\hat{F}_{t+1}, F_{t+1}) \\ &\quad + \lambda_{CL} L_{CL}(\hat{F}_{t+1}, F_{t+1}) + \lambda_{OF} L_{OF}(\hat{F}_{t+1}, F_{t+1}, F_t) \\ &\quad + \lambda_{adv} L_{adv}^{G}(\hat{F}_{t+1}), && (8) \\ L_D &= L_{adv}^{D}(\hat{F}_{t+1}, F_{t+1}), && (9) \end{aligned}$$

where λ_{FE}, λ_{GT}, λ_{CL}, λ_{OF}, λ_{adv} are weight coefficients, which balance the influence of each loss in the objective function.

D. ANOMALY CRITERION

After training, the entire model can distinguish abnormal behavior from normal behavior. The test video is input to check the detection effect of the model. We use the Euclidean distance to calculate the reconstruction error of all pixel values between the ground truth and the video prediction frame, as shown in formulas (10) and (11).

$$e(t) = \frac{1}{N} \sum_{t=1}^{N} (F_t - \hat{F}_t)^2, \tag{10}$$

$e(t)$ is further normalized to the interval [0,1] using:

$$s(t) = 1 - \frac{e(t) - e(t)_{\min}}{e(t)_{\max}}, \tag{11}$$

where $s(t)$ is the regularized value.

The test experiment $s(t)$ can determine when an abnormal event occurs in the video sequence. When a normal event occurs, the value $s(t)$ corresponds to a larger value, and when an abnormal event occurs, the value $s(t)$ corresponds to a lower value.

Figure 4 is a graphical illustration where some normal events, anomalous events, predicted frames, and ground truth in the dataset are shown. The video scene is a pedestrian zone that can be well predicted when pedestrians walk normally. When some abnormal behavior occurs, such as someone riding a bicycle, the prediction is blurred.

IV. EXPERIMENTS

In the experiments, multiple public video datasets for anomaly detection were used to validate our method. All the experiments were run on a server equipped with an Intel Xeon E5-2683 v3 processor and NVIDIA GeForce GTX 1080Ti GPUs. The experimental framework used TensorFlow [34]. Before training, the size of the video frame in the training sample was set to 256 × 256. Then, the pixels of all video frames were normalized to [−1,1]. As shown in Figure 1, we took five consecutive video frames, and T = 4, according to the algorithm proposed by Adam [35]. In the public dataset, UCSD Ped1 and UCSD Ped2 are grayscale images.

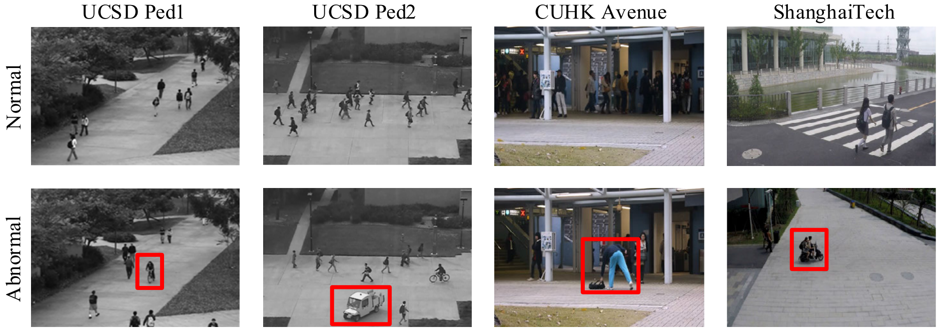


FIGURE 5. Normal and abnormal events of public datasets.

The learning rate of G and D ware 0.0001 and 0.00001. The video frames in CUHK Avenue and ShanghaiTech are color images. The learning rate of G and D ware 0.0002 and 0.00002. The values λ_{FE}, λ_{GT}, λ_{CL}, λ_{OF}, λ_{adv} in each public dataset were different. The values were set according to the experiment.

The frame-level discriminator followed the patch discriminator [13], [32] scheme and consisted of four convolutional layers and a fully connected layer. The core size of each layer was 5 × 5, and the activation function was leaky ReLU. The output value of the discriminator was used as the basis for judging the abnormal score of the frame. When the score was low, it indicated that the frame contained anomalous events.

Figure 5 shows several normal events and abnormal events as examples in the UCSD Ped1, UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets, where the red boxes indicate the abnormal behavior that appeared in the video.

A. DATASETS

This article uses three public datasets to train the model: the UCSD Ped1 and Ped2 datasets [36], the CUHK Avenue dataset [37], and the ShanghaiTech campus dataset [11]. These public datasets include normal event videos and abnormal event videos. The model in this paper uses the normal event videos in these data for training and then uses the abnormal event videos as detection samples for anomaly detection. The feasibility and accuracy of the model were tested according to the above work.

- UCSD Ped1. The video scenes collected by this dataset are pedestrians walking on a road in the lawn, and the pedestrian walking direction is perpendicular to the surveillance video frame. Ped1 consists of 34 training video samples and 36 test video samples. The abnormal events are cart, wheelchair, skater, and biker shuttled between pedestrians.
- UCSD Ped2. The surveillance scene in UCSD Ped2 is a horizontal sidewalk. The dataset has 12 test samples and 16 training samples. Abnormal events include skaters, wheelchairs, bicycles, and trolleys.
- CUHK Avenue. This dataset was taken on the campus avenue of the Chinese University of Hong Kong. It has 16 training video samples and 21 test video samples. The training video records normal events, while the test video includes normal events and abnormal events. Abnormal events are divided into three categories: strange action, wrong direction, and abnormal object.
- ShanghaiTech Campus dataset. The ShanghaiTech campus dataset contains 13 different scenarios and has more than 270,000 training frames and 130 abnormal behaviors. Abnormal behaviors include chasing, running, cycling, and wheelbarrow.

B. ANOMALY DETECTION

Referring to [1], [37], the general evaluation scheme was to change the regularization score threshold to calculate the receiver operation characteristic (ROC). Then, the area under the curve (AUC) was used as the evaluation standard for abnormal event detection performance. The higher the value was, the better the detection performance. This paper used the calculated AUC to test the performance of the experiment.

Reference [11] proposed a prediction-based method for anomaly detection. It has excellent performance, and we set it as the baseline of this study. In addition, we also compare the method of [13]. We conducted ablation experiments in the UCSD Ped1, UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets. The experimental results are shown in Table 1. We found that the AUC value of our method gained 2.1%, 1.7%, 0.7%, and 0.9% compared to [11]. Compared with [13], our method gained 0.5%, 0.8%, 0.7%, 0.7%. Therefore, the effectiveness of our method is verified, and it is superior to the existing advanced methods.

The whole experiment normalized the detected video frame's reconstruction error and judged whether the video behavior was abnormal according to the rule score of formula (11). Thresholds were set according to different surveillance video scenes so that the accuracy and effectiveness of anomaly detection were guaranteed.

As shown in Figure 6, the scene number represents the sequence number of the tested video frame, and the abnormal event is distinguished from the normal event according to the regularity score. The yellow area represents the ground truth

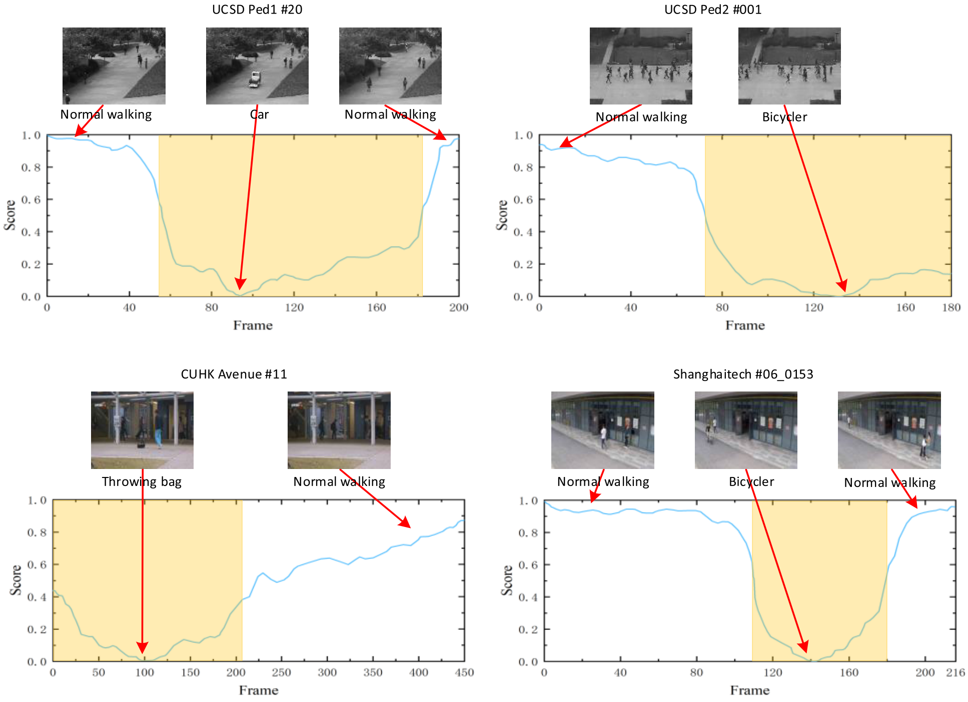


FIGURE 6. Schematic diagram of the score curves of the four public datasets.

TABLE 1. AUC values of different methods in each dataset.

Method	UCSD Ped1	UCSD Ped2	CUHK Avenue	Shanghai Tech
MPPCA [38]	59.0	69.3	None	None
SF [39]	67.5	55.6	None	None
MPPCA+ SF [36]	66.8	61.3	None	None
MDT [36]	81.8	82.9	None	None
ConvL-AE [28]	75.5	88.1	77.0	None
Unmasking [40]	68.4	82.2	80.6	None
sRNN [2]	None	92.2	81.7	68.0
Conv-AE [1]	81	90	70.2	None
Baseline [11]	83.1	95.4	85.1	72.8
IPR+OF [13]	84.7	96.3	85.1	73.0
Our method	85.2	97.1	85.8	73.7

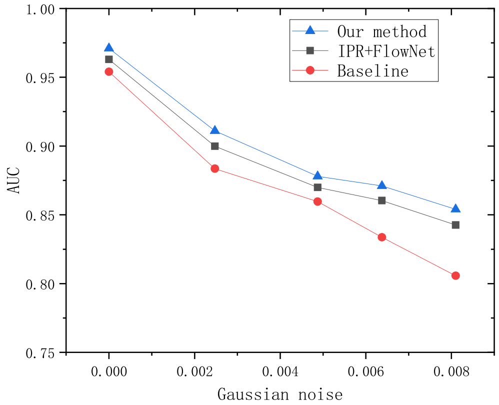


FIGURE 7. AUC value of different models in different Gaussian noise.

anomaly frame. The score is high when normal events occur, and the score is low when abnormal events occur.

C. ABLATION STUDIES

In this study, the prediction frame module and the reconstruction module are combined to make the reconstructed frame closer to the ground truth and improve the noise resistance and accuracy of the anomaly detection model. It uses latent feature loss, gradient loss, contextual loss, optical flow loss, and adversarial loss. We conducted a series of ablation experiments. The feasibility of our proposed method was verified.

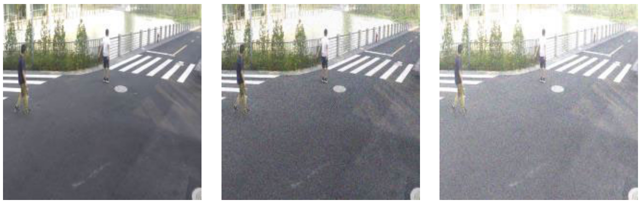
1) ANTI-NOISE PERFORMANCE

The previous part of this article introduced that to solve the characteristic of poor noise immunity of the model in the baseline, the advantages of the strong generalization ability of the reconstruction module were combined with the prediction module. The reconstruction module strengthens the anti-noise ability of the predicted frame and improves the quality of the predicted frame generation.

We added Gaussian noise to the datasets, and on this basis, we conducted multiple models of anti-noise performance experiments.

As shown in Figure 7, the experimental results for UCSD Ped2 show that with increasing Gaussian noise, the AUC values of all methods gradually decrease. Nevertheless, our method has better noise immunity than the baseline and IPR with optical flow.

We added Gaussian noise to the video frame, as shown in Figure 8. The visual effects here are Gaussian noise of 0.02 and 0.2.



(a) Gaussian noise 0 (b) Gaussian noise 0.02 (c) Gaussian noise 0.2

FIGURE 8. (a) Is the original frame, and frames (b) and (c) are added with different Gaussian noise.

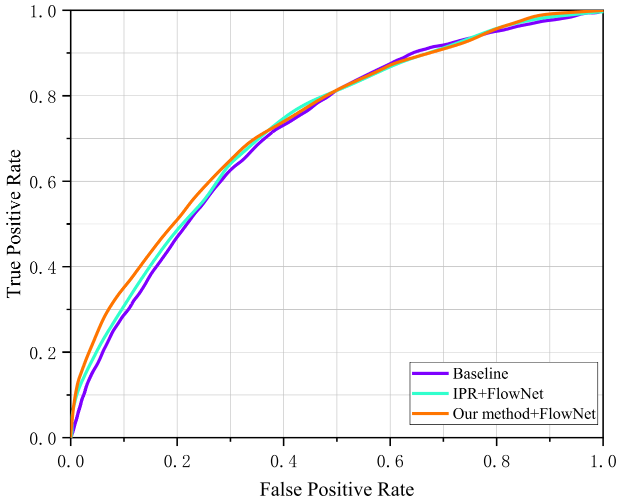


FIGURE 9. ROC curves of the three models on the ShanghaiTech dataset.

TABLE 2. AUC value for different optical flow estimators.

Method	UCSD Ped1	UCSD Ped2	CUHK Avenue	Shanghai Tech
Our method + FlowNet	85.1	96.9	85.4	73.5
Our method + SPyNet	85.2	97.1	85.8	73.7

2) CONSTRAINT ON LATENT FEATURE

To test the improvement of the anomaly detection effect of the potential feature constraints we proposed, we conducted an ablation experiment. Our model is divided into including latent feature constraints and not including latent feature constraints. In addition, to verify the fairness of the method, we also used the same FlowNet as IPR as our optical flow constraint in the ablation experiment.

Figure 9 shows three model tests on the ShanghaiTech dataset. The ROC curve shows that the application of latent feature constraints to the model improves the performance of anomaly detection. A better true positive rate is obtained, especially in the low range of false positive rates (0.0-0.4).

As shown in Figure 10, we conducted ablation experiments on different public datasets. Our proposed latent feature constraint method was verified in ablation experiments to improve the performance of anomaly detection. LFL is an abbreviation for latent feature loss.

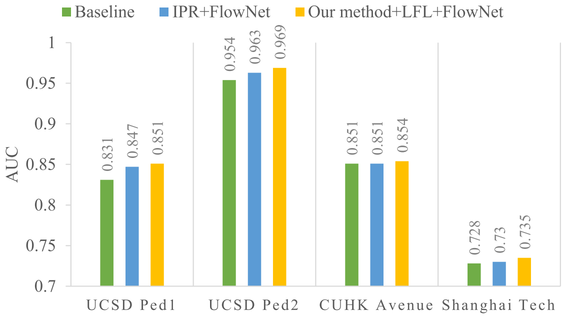
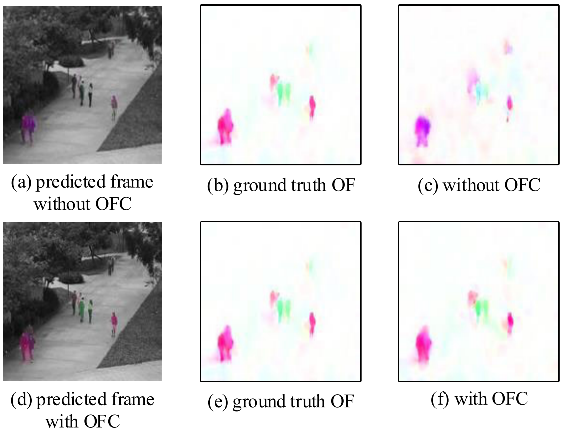


FIGURE 10. Latent feature constraint ablation experiment on public datasets.



(a) predicted frame without OFC (b) ground truth OF (c) without OFC
(d) predicted frame with OFC (e) ground truth OF (f) with OFC

FIGURE 11. Optical flow comparison of the predicted frame with or without the optical flow constraints.

3) CONSTRAINT ON OPTICAL FLOW

The optical flow constraint is an effective method for motion estimation of predicted frames. As shown in Figure 11, to verify the applicability of the optical flow constraint, we performed a visual comparison of optical flow between the predicted frame without the optical flow constraint and the predicted frame with the optical flow constraint. (f) compared to (c), the optical flow is closer to the ground truth. The optical flow constraint is more conducive to the motion estimation of the predicted frame. OF is an abbreviation for optical flow, and OFC is an abbreviation for optical flow constraint.

To verify the superior performance of SPyNet compared to FlowNet, we conducted an ablation experiment to compare the experimental results of the two optical flow constraints. It is proven in Table 2 that the model in this paper uses SPyNet optical flow constraints to improve the performance of anomaly detection.

V. CONCLUSION

In the video anomaly detection model based on frame prediction, the quality of future frame prediction is of vital importance. Due to its poor noise immunity and ignoring the constraints of latent features, this paper combines the prediction module and the reconstruction template to improve the model's noise immunity. Latent feature constraints and SPyNet optical flow constraints are used so that the predicted frame can obtain better reconstruction. The importance of each component was verified by ablation experiments in several public datasets. Experimental results show that our method is effective and robust and is superior to the existing advanced methods. The anomaly detection tasks of more complex monitoring scenarios and more challenging public data sets are future research work.

REFERENCES

[1] 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, doi: 10.1109/CVPR.2016.86.

[2] 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.

[3] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 622–637.

[4] Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1811–1820.

[5] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019, doi: 10.1016/j.media.2019.01.010.

[6] G.-J. Qi, "Loss-sensitive generative adversarial networks on Lipschitz densities," 2017, *arXiv:1701.06264*. [Online]. Available: http://arxiv.org/abs/1701.06264

[7] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1896–1904, doi: 10.1109/WACV.2019.00206.

[8] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101938, doi: 10.1016/j.artmed.2020.101938.

[9] B. R. Kiran, D. M. 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, pp. 1–15, 2018, doi: 10.3390/jimaging4020036.

[10] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[11] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545, doi: 10.1109/CVPR.2018.00684.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[13] 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, doi: 10.1016/j.patrec.2019.11.024.

[14] 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, doi: 10.1109/ICCV.2015.316.

[15] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2720–2729, doi: 10.1109/CVPR.2017.291.

[16] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 988–998, Jun. 2014, doi: 10.1109/TIFS.2014.2315971.

[17] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2458–2465, doi: 10.1109/CVPRW.2009.5206686.

[18] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, Apr. 2007, doi: 10.1007/s11263-006-0009-9.

[19] L. Wang and M. Dong, "Real-time detection of abnormal crowd behavior using a matrix approximation-based approach," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2701–2704, doi: 10.1109/ICIP.2012.6467456.

[20] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 55–61, doi: 10.1109/CVPRW.2011.5981799.

[21] F. Tung, J. S. Zelek, and D. A. Clausi, "Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance," *Image Vis. Comput.*, vol. 29, no. 4, pp. 230–240, Mar. 2011, doi: 10.1016/j.imavis.2010.11.003.

[22] X. Zhu, J. Liu, J. Wang, Y. Fang, and H. Lu, "Anomaly detection in crowded scene via appearance and dynamics joint modeling," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2705–2708, doi: 10.1109/ICIP.2012.6467457.

[23] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2112–2119, doi: 10.1109/CVPR.2012.6247917.

[24] 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, doi: 10.1016/j.patrec.2006.02.004.

[25] 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, Jul. 2013, doi: 10.1016/j.patcog.2012.11.021.

[26] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*. [Online]. Available: http://arxiv.org/abs/1901.03407

[27] 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," 2019, *arXiv:1904.02639*. [Online]. Available: http://arxiv.org/abs/1904.02639

[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, doi: 10.1109/ICME.2017.8019325.

[29] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," 2019, *arXiv:1908.06351*. [Online]. Available: http://arxiv.org/abs/1908.06351

[30] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2015, *arXiv:1511.05440*. [Online]. Available: http://arxiv.org/abs/1511.05440

[31] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal AutoEncoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1933–1941, doi: 10.1145/3123266.3123451.

[32] P. Isola, 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. 5967–5976, doi: 10.1109/CVPR.2017.632.

[33] D. Pathak, J. Donahue, and A. A. Efros. (2018). *Context Encoders: Feature Learning by Inpainting*. [Online]. Available: https://www.cvfoundation.org/openaccess/content_cvpr_2016/papers/Pathak_Context_Encoders_Feature_CVPR_2016_paper.pdf%0Apapers3://publication/uuid/9E05080B-9457-4DFE-B5DA-C42DC2CFEE40

[34] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[35] K. Diederik and J. L. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[36] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981, doi: 10.1109/CVPR.2010.5539872.

[37] 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, doi: 10.1109/ICCV.2013.338.

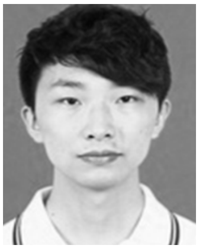
[38] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2921–2928, doi: 10.1109/CVPRW.2009.5206569.

[39] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942, doi: 10.1109/CVPRW.2009.5206641.

[40] 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. 2914–2922, doi: 10.1109/ICCV.2017.315.

YONG QIANG received the M.S. degree in electrical engineering from Anhui Polytechnic University, China, in 2013. He is currently pursuing the Ph.D. degree in automatic control with Southeast University, Nanjing. His research interests include deep learning and computer vision.

SHUMIN FEI received the Ph.D. degree from the Beijing University of Aeronautics and Astronautics, in 1995. From 1995 to 1997, he was a Postdoctoral Research Fellow with Southeast University, where he is currently a Professor and a Doctoral Advisor with the School of Automation. He has published more than 100 journal articles. His research interests include nonlinear systems, stability theory of delayed systems, complex systems, and computer vision.



YIPING JIAO received the B.S. degree from Jiangnan University, China, in 2013, and the M.S. degree from Southeast University, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Automation, Southeast University. He is mainly working on deep learning-based digital pathology image analysis. He is also interested in other machine learning applications, including near-infrared spectrum calibration.