Generative Adversarial Networks for Abnormal Event Detection in Videos Based on Self-Attention Mechanism

WEICHAO ZHANG\textsuperscript{1,2}, GUANJUN WANG\textsuperscript{1,2,3}, MENGXING HUANG\textsuperscript{1,2}, (Member, IEEE), HONGYU WANG\textsuperscript{1,4}, (Member, IEEE), AND SHAOPING WEN\textsuperscript{1,2}

1State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University, Haikou 570228, China
2Department of Information and Communication Engineering, Hainan University, Haikou 570228, China
3Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China
4Department of Computer and Cyberspace Security, Hainan University, Haikou 570228, China

ABSTRACT

Unsupervised anomaly detection defines an abnormal event as an event that does not conform to expected behavior. In the field of unsupervised anomaly detection, it is a pioneering work that leverages the difference between a future frame predicted by a generative adversarial network and its ground truth to detect an abnormal event. Based on the work, we improve the ability of video prediction framework to detect abnormal events by enhancing the difference between prediction results for normal and abnormal events. We incorporate super-resolution and self-attention mechanism to design a generative adversarial network. We propose an auto-encoder as a generator, which incorporates dense residual networks and self-attention. Moreover, we propose a new discriminator, which introduces self-attention on the basis of a relativistic discriminator. To predict a future frame with higher quality for normal events, we impose a constraint on the motion in video prediction by fusing optical flow and gradient difference between frames. We also introduce a perception constraint in video prediction to enrich the texture details of a frame. The AUC of our method on CUHK Avenue and Shanghai Tech datasets reaches 89.2% and 75.7% respectively, which is better than most existing methods. In addition, we propose a processing flow that can realize real-time anomaly detection in videos. The average running time of our video prediction framework is 37 frames per second. Among all real-time methods for abnormal event detection in videos, our method is competitive with the state-of-the-art methods.

INDEX TERMS
Abnormal event detection, generative adversarial networks (GANs), self-attention, video understanding.

I. INTRODUCTION

With the development of intelligent security, more and more security cameras are used in various occasions to ensure public safety. However, it will consume a lot of labor and material resources to realize abnormal event detection in videos by only hand. There is an urgent need for a computer vision algorithm that can realize real-time anomaly detection in videos.

Abnormal event detection in videos is a long-standing and extremely challenging vision problem. There are two main challenges facing abnormal event detection: (1) The small probability of anomaly occurrence. We usually define events that conform to the expectation in real life as normal events, and events that do not conform to the expectation as abnormal events. Therefore, we do not have enough
abnormal data to train a classifier for binary video classification. Sultani et al. [2] proposed an anomaly dataset composed of 1900 real-world surveillance videos, which is currently the largest anomaly dataset. However, videos in this dataset are collected from video websites. They have different resolutions, some of which even contain watermark. All of these will affect the training of a model for anomaly detection. Therefore, the lack of abnormal data is still a troublesome problem in the field of abnormal event detection in videos.

(2) The definition of abnormality is subjective. For example, a car driving on the expressway is a normal event, while a car on a sidewalk is an abnormal event. Except for criminal behavior, which are regarded as abnormal events in any environment, whether other events are normal or abnormal depends on different environment. Therefore, some researchers develop anomaly detection algorithms for a specific abnormal event, such as violence detector [3], traffic accident detector [4], [5]. However, it is obvious that the solutions cannot be generalized to detect other abnormal events. Their practicality is limited.

In the early work, researchers modeled normal appearance and motion patterns using hand-crafted features. If an event deviates greatly from the normal patterns, it can be regarded as an abnormal event. The common low-level features include histogram of gradient [6], 3D gradient [7], histogram of optical flow [8], etc. However, it is difficult to define robust features suitable for a variety of complex scenes by hand-crafted features, which means that some prior information must be enhanced. Unfortunately, the prior information is hard to be defined in complex video scenes.

Deep learning has become the mainstream method in the field of abnormal event detection due to its excellent feature extraction capabilities. According to the type and the labeling method of training data, anomaly detection methods can be divided into supervised learning based methods, weakly supervised learning based methods and unsupervised learning based methods [9]. However, the difficulty of collecting abnormal data and the cumbersome manual labeling restrict the development of the first two methods. Unsupervised learning based methods does not require any label information. Based on the assumption that abnormal events are irregular, and normal events are predictable, unsupervised learning based methods train a network to learn normal behavior patterns in a certain environment. The network generates video frames by reconstructing or predicting. If the error between a generated frame and its ground truth is large, we can consider that an abnormal event has occurred. The advantage of the methods based on learned normal patterns is that only normal samples are needed for training, which solves the difficulty in obtaining abnormal data. On the other hand, the subjectivity problem of definition for abnormal events can be solved by learning normal patterns of certain environment. The method based on sparse coding [15], [16] is a very representative method among the methods based on unsupervised learning. They uses the initial part of a video to construct a normal event dictionary. Then, the main idea for anomaly detection is that abnormal events cannot be accurately reconstructed from the normal event dictionary. The method based on auto-encoder [17]–[19] uses similar ideas to reconstruct normal events. However, the most fatal flaw of reconstruction-based methods is over fitting. When the capacity of the deep neural network is high, large reconstruction errors for abnormal events do not necessarily happen. The emergence of Generative Adversarial Network(GAN) provides us with another possibility to detect abnormal events in videos. Ravanbakhsh et al. [20] used U-net as the generator for GAN for cross-modality reconstruction. The network is not to generate ground truth, but to learn how to realize the mutual transformation between a RGB image and its corresponding optical-flow image. But the essence of this method is still based on reconstruction. Liu et al. [1] introduce a future video frame prediction based anomaly detection method, which achieved good results. This is the first work that leverages video prediction for anomaly detection. Specifically, given a video clip, Liu et al. predict the future frame based on its historical observation. So a good predictor is a key to the task.

Motivated by Liu et al. [1], we conceive a video prediction framework that can predict a future frame with higher quality for normal events. Since abnormal events are irregular and unpredictable, our framework will not improve prediction results for abnormal events. So we increase the difference between prediction results for normal and abnormal events, thereby improving the accuracy of our framework for abnormal event detection. Due to the outstanding ability of attention mechanism to ignore irrelevant information and focus on key information, we propose an auto-encoder combined with self-attention mechanism. In addition, drawing on the idea of super-resolution, we introduce dense residual networks in the auto-encoder to improve the feature extraction ability of a video prediction framework. At the same time, we enforce the texture details of the predicted frame to be close to its ground truth by imposing a perceptual constraint on the auto-encoder. We use the auto-encoder as a generator. In order to avoid gradient vanishing, we propose a discriminator based on a Relativistic GAN, which combines with self-attention mechanism. The discriminator also ensures that the predicted frame is closer to its ground truth. Historical information is an important part of video prediction. In order to make full use of the historical information of a video, we propose a motion feature that incorporates optical flow and gradient difference between frames. The motion feature maximizes the use of historical information to represent temporal features of a video clip. By inputting the first four frames of a video into GAN, our generator generates the fifth frame that is enough to confuse the discriminator. The error between the fifth frame and its ground truth is used as the criterion for judging whether an abnormal event occurs in the fifth frame.

We summarize our contributions as follows:

(1) We propose a novel GAN. We propose an auto-encoder incorporating dense residual networks and self-attention. We use the auto-encoder that acts as a generator to increase the difference between prediction results for normal and
abnormal events. What’s more, we propose a new discriminator, which incorporates self-attention module on the basis of a Relativistic GAN. Our discriminator not only avoids gradient vanishing in training, but also guarantees that the generator generates a frame closer to its ground truth.

(2) In order to make full use of historical information of video clips, we propose a new motion feature, which combines optical flow and gradient difference between frames. In addition, we creatively leverage a perceptual constraint for anomaly detection, which highlights the abnormal existence of abnormal objects in videos.

(3) We propose a processing flow for real-time abnormal event detection, which ensures that the average running time of our model is 37 frames per second(FPS).

II. RELATED WORK

A. DEEP LEARNING BASED METHODS

Supervised learning based methods usually treat anomaly detection as a classification problem. The methods train a model by using detailed labeled data. Since videos contain spatial and temporal information, abnormal event detection in videos based on supervised learning requires a neural network that can obtain spatial-temporal features, such as C3D [10], two-stream network [11], T-CNN [12], etc. However, supervised learning based methods require sufficient prior information. The methods are only suitable for anomaly detection in situations where all types of abnormal events are known. The training data used by weakly supervised learning based methods is only labeled with normal or abnormal labels at the video level of training samples. For a weakly labeled training video, we only know whether there are an abnormal event in the video, but not the specific type and time location of an abnormal event. Abnormal event detection using weak label is a typical multi-instance learning problem. Sultani et al. [2] extracted C3D features for video clips and combined with a fully connected network to predict anomaly scores. But motion information is critical to abnormal event detection. In view of multi-instance learning ranking loss in [2], which ignores potential temporal structure, Zhu and Newsam [13] leveraged an attention module to strengthen the learning for motion features. Zhong et al. [14] devise a graph convolutional network to correct noisy labels. The corrected label is used to train an action classifier to detect abnormality, which greatly improves the frame-level AUC on UCF-Crime dataset. However, like supervised learning based methods, the difficulty of collecting abnormal data and the cumbersome manual labeling restrict the development of these two methods.

B. SUPER-SOLUTION

Super-resolution(SR) refers to obtaining a high-resolution image from single low-resolution image. SR can be divided into three categories: interpolation-based, reconstruction-based, and learning-based. Methods based on interpolation are simple to implement, such as linear, bicubic and Lanczos [23], which have been widely used. But these linear models are difficult to recover high-frequency details of an image. Methods based on sparse representation [24] assume that any image can be sparsely represented by the elements of a dictionary. From the dictionary, we can learn the mapping from low-resolution images to high-resolution images. However, the methods are computationally complex. SR based on learning can be roughly divided into two categories. One is an algorithm that pursues detail restoration, such as SRCNN [25]. SRCNN is the pioneering work of SR based on deep learning. This method introduced convolutional neural networks into SR for the first time. It only uses a three-layer network to achieve advanced results. Subsequently, various models based on deep learning appeared. The other is an algorithm that aims to reduce perceptual loss and emphasizes visual perception, the most representative of which is SRGAN [26]. This method can obtain images with high-frequency details based on perceptual loss. On the basis of SRGAN, Wang et al. [22] proposed an enhanced SRGAN, which can obtain better visual quality and more realistic natural textures.

C. SELF-ATTENTION MECHANISM

Attention mechanism mimics the internal process of biological observation behavior, that is, a mechanism that aligns internal experience and external sensation to increase the fineness of observation in some areas. Attention mechanism can quickly extract important features from small amounts of data, so it is widely used in natural language processing tasks, especially machine translation. Self-attention is improvement of attention mechanism, which reduces the dependence of models on external information. It is good at capturing the internal correlation of data or features.

The basic idea of self-attention in computer vision is to let the model learn to ignore irrelevant information and focus on the key points of information. For example, if an eagle is flying in a cloudless sky, human will focus on the eagle. The sky will naturally become background information in visual system. In neural networks, a convolutional layer obtains the output features through the linear combination of a convolution kernel and original features. Since convolution kernels are usually local, in order to increase the receptive field, a method of stacking convolution layers is often adopted. However, this method is not efficient. Self-attention [27], [28] shows a good balance between the ability to model long-range dependencies and the computational and statistical efficiency. Zhang et al. proposed SAGAN [29], which combines self-attention with GAN. With the help of self-attention, the generator can draw images in which fine details at every location are carefully coordinated with fine details in distant portions of the images.

III. METHOD

Unpredictability of abnormal events is a necessary premise for abnormal event detection in videos based on unsupervised learning. Based on this assumption, we use GAN to learn the development process of normal events. By imposing
FIGURE 1. The pipeline of our network. The black arrow indicates the training phase. In the training phase, the generator generates a future frame based on the historical clips of a video. Then we input the predicted future frame and its ground truth into the discriminator. If the requirement of the discriminator is not met, the network will continue to be trained until the generator generates a frame that is sufficient to confuse the discriminator. The green arrow indicates the testing phase. In the testing phase, we compare the error between the future frame generated by the generator and its ground truth. If the error is greater than the threshold, the generator fails to predict the development process of the event, which is determined that an abnormal event has occurred in the video.

A. NETWORK ARCHITECTURE
Generative adversarial networks [30] is a machine learning architecture proposed by Good Fellow in 2014. The main idea of GAN is that there are two competing neural networks. One is a generator that takes noise as input and generates samples. The other is a discriminator, which receives samples from the generator and training data. It must be able to distinguish the two data source. In the training phase, the generator learns to produce a sample that are close to its ground truth. The discriminator learns how to distinguish the generated data from its ground truth. These two networks are trained at the same time. Our goal is to make the generated samples indistinguishable from the real data through competition. Constraints on the generator and discriminator, the generator generates the future frame $I_t$ of the video according to the historical clips $I_1, I_2, I_3, \ldots, I_{t-1}$ of a video. According to many experiments, we found that when $t = 5$, our network has the best prediction results for future frames, as shown in Section IV. Our main aim is to improve the accuracy of abnormal event detection in videos by enhancing the difference between prediction results for normal events and abnormal events. We propose a novel GAN to improve the quality of prediction results for normal events. Since abnormal events cannot be predicted, our framework will not improve prediction results for abnormal events. Therefore, it is easier for our framework to distinguish between normal events and abnormal events in a video by comparing the error between the predicted frame and its ground truth. At the same time, we propose a processing flow that can realize real-time anomaly detection in videos. The details are shown in Section IV. In this section, we will introduce our method from network architecture and loss function. Fig.1 is the pipeline of our network.

FIGURE 2. Encoder-decoder architecture. The architecture contains an encoder and a decoder. The green block represents self-attention module. The blue arrow represents feature extraction by dense residual block. The black arrow represents feature fusion.

For the generator architecture, we discard skip-connection of U-net in [1]. Because we find that skip-connection is not compatible with our framework. See Section IV for detailed analysis. At the same time, in order to ensure detection efficiency of our framework, we do not use deconvolution in the generator. As shown in Fig.2, we propose an auto-encoder as a generator. The auto-encoder contains two modules. One is an encoder, which starts with the four convolution blocks. Each convolution blocks include two convolution layers. The number of output channels of the four convolution blocks are 64, 128, 256, 512, respectively. Except for the first convolution block, all other convolution blocks contain a max pooling layer. We uses LeakyReLU($\alpha = 0.2$) as an activation function. Based on super-resolution, we introduce dense residual networks in [22] to the encoder. Different from the residual module in [31], we increase the network capacity by connecting dense residual blocks on the main path. In addition, we introduce self-attention module to enhance the focus of our framework on key information in an image. The other is a decoder, which consists of 4 up-sampling layers. In order to ensure the detection efficiency of our method, each layer contains nearest interpolation operation and a convolution layer with 64 feature maps. Finally, we generate a three-channel...
image through a convolution layer with 3 feature maps. The kernel sizes of all convolution are set to 3 × 3.

B. DENSE RESIDUAL BLOCK

The dense residual block in Fig.3 is the core of our generator. Residual networks were first proposed by He et al. [32]. It has a strong ability to eliminate redundant information in a network. In order to ensure real-time abnormal event detection in videos, our dense residual block cannot directly stack multiple residual networks like super-resolution. Moreover, in the training phase, multi-layer back propagation of error signals is easy to cause gradient vanishing or gradient explosion. Therefore, the deeper the network, the more difficult it is to train a model. After many experiments, our dense residual network consists of 2 convolution layers with 32 feature maps and 1 convolution layer with 64 feature maps. The kernel sizes of all convolution are set to 3 × 3. We use LeakyReLU(α = 0.2) as an activation function. Section IV describes in detail the experiments on the influence of the number of convolutions in a dense residual block and the number of dense residual blocks in the network on our model performance. We improve network performance by cascading two dense residual blocks, which also ensures detection efficiency. In addition, we do not use Batch Normalization (BN) layers in the network to normalize features. Reference [22] pointed out that when deep network training is carried out under the framework of GAN, BN layers may bring artifacts and limit the generalization ability. Therefore, we remove BN layers to achieve stable training, which improve performance and reduce computational complexity.

![FIGURE 3. Dense residual block. The network consists of 3 convolutions with 3 × 3 convolution kernels, one of which contains 64 feature maps, and the remaining three contain 32 feature maps. We do not use BN layers to normalize features. We use LeakyReLU(α = 0.2) as an activation function. The output result of each layer is connected to the following layer by shortcut. The purple block represents the output result of the previous layer.](image)

C. DISCRIMINATOR

For the traditional discriminator based on Jensen-Shannon (JS) divergence, if there is no overlap between the probability distribution of predicted frames and that of ground truth, then the JS divergence is a constant at this time. This is fatal in learning algorithms. Because it will cause gradient vanishing. This situation is very likely to occur when the network is initialized. In order to avoid the situation where the gradient is 0, we propose a new discriminator, which introduces self-attention on the basis of a relativistic discriminator [33]. Instead of estimating the probability that input images are real, we predict the probability that real images are more realistic than fake images. The discriminator architecture consists of convolution layers, self-attention modules and fully connected layers. The low-level features of input images are extracted by the convolution layers. Two self-attention modules are added after the last two convolution layers to extract features that contain more global information. Finally, we can obtain the probability that a real image is more realistic than a fake image by a three-layer fully connected network. The architecture is shown in Fig. 4.

![FIGURE 4. Discriminator. We combine self-attention modules with convolution layers to extract features that contain more global information. Finally, we can obtain the probability that real images is more realistic than fake images by a three-layer fully connected network. The green block represents a self-attention module.](image)

D. SELF-ATTENTION MECHANISM

In order to extract features that contain more global information, we use self-attention mechanism on both the generator and the discriminator to reduce the dependence of our network on external information. This method effectively models long-range dependencies in images, so as to obtain a larger receptive field and more sufficient contextual information. We adopt the self-attention architecture proposed in [34]. We inherit the attention function in [28], which can be described as mapping a query and a set of key-value pairs to an output. The input features \( x \) are linearly mapped by \( 1 \times 1 \) convolution to compress the number of channels, thus we can obtain key \( f_1(x) \), query \( f_2(x) \), value \( f_3(x) \). Here, \( x \in \mathbb{R}^{c \times N} \), where \( c \) is the number of channels and \( N \) is the number of feature locations of features from the previous hidden layer. Then the corresponding attention weight is obtained by multiplying the key and the query, which are multiplied by the value to obtain the output. In addition, we combine the linear mapping of \( 1 \times 1 \) convolution to the output of the attention layer with the input feature map. Therefore, the final output is as follows:

\[
y = W_{f_4}(W_{f_3} \cdot \text{softmax}(s)) + x.
\]
where \( s = (W_{f_1})^T W_{f_2} \) indicates the degree to which the model pays attention to other areas when generating a certain area. In the above formula, \( W_{f_1} \in \mathbb{R}^{T \times C} \), \( W_{f_2} \in \mathbb{R}^{T \times C} \), \( W_{f_3} \in \mathbb{R}^{T \times C} \), \( W_{f_4} \in \mathbb{R}^{C \times P} \), \( \bar{C} \) is the number of compressed channels. We apply self-attention module to the generator and the discriminator (as shown in Fig. 3 and Fig. 4) to improve the accuracy of abnormal event detection in videos without greatly increasing the size of our model. The self-attention architecture is shown in Fig. 5.

**FIGURE 5.** Self-attention architecture. The architecture implements channel compression of feature maps through \( 1 \times 1 \) convolution, and then uses matrix multiplication between features to model long-range dependencies in feature maps. The softmax operation only operates on each row of the feature map.

### E. LOSS FUNCTION

Loss function is used to evaluate the degree of difference between the predicted value and its ground truth, which is an important part of neural networks. We set loss functions on the appearance and the motion, we can obtain frames we expect, that is, the error between the frame and its ground truth is small.

1) INTENSITY LOSS

In order to make predicted frames as close as possible to ground truth, we impose a constraint on the appearance by using intensity loss. Intensity loss guarantees the similarity of all pixels in RGB space between a predicted frame and its ground truth. We minimize \( L_2 \) distance between the three-channel pixel values of the predicted frame \( \hat{I}_t \) and that of ground truth \( I_t \), as follows:

\[
L_{int}(\hat{I}_t, I_t) = \frac{1}{2WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (\hat{I}_{t,i,j} - I_{t,i,j})^2.
\]

(2)

where \( i, j \) denotes the pixel indexes in a frame.

2) MOTION LOSS

Temporal features are an extremely important part of video understanding. Our video prediction framework can learn the development process of events by temporal features of video clips. Here we use optical flow and gradient differences between images to describe the temporal features of videos. Optical flow represents instantaneous speed of pixels. If the time interval is short enough, such as the front and back frames of a video, then we can regard optical flow as displacement of pixels. Optical flow estimation based on deep learning [35]–[38] not only reduces the computational complexity of traditional optical flow estimation, but also obtains more representative motion characteristics. In this study, we use LiteFlowNet [36] to calculate optical flow of videos. We enforce the optical flow between predicted frames to be close to their optical flow ground truth, which improve the ability of our framework to predict the development of normal events. The loss function for optical flow is shown in the following formula:

\[
L_{op}(\hat{I}_t, I_{t-1}) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (f_{i,j}(\hat{I}_t, I_{t-1}) - f_{i,j}(I_t, I_{t-1})).
\]

(3)

where \( f_{i,j}(\cdot) \) is LiteFlowNet pre-trained on a dataset [36]. In addition, historical information plays an important role in video prediction. Therefore, we need to make full use of historical information to predict future frames, so as to improve the accuracy of abnormal event detection. As mentioned above, we use four historical frames to predict the fifth frame. So we decide to use the relation between the third frame and the fifth frame to supplement the description of the impact of historical information on video prediction. If we continue to use optical flow to represent the relation between the fifth frame and the third frame, the training process will be too complicated and slow. Here, we use gradient difference between frames as part of the motion feature, which can maximize the use of historical information at the cost of minimal training time. We can obtain prediction results by minimizing the error between the gradient difference of the predicted frame and that of its ground truth. \( G_{\hat{I}} \) represents the gradient of a predicted frame \( \hat{I} \) at time \( t \). \( G_{I} \) represents the gradient of ground truth \( I \) at time \( t \). \( G_{I_{t-2}} \) represents the gradient of ground truth \( I \) at time \( t-2 \). \( G_{\hat{I}}, G_{I}, G_{I_{t-2}} \) are shown as follows:

\[
G_{\hat{I}_{t,i,j}} = |\hat{I}_{t,i,j+1} - \hat{I}_{t,i,j}| + |\hat{I}_{t,i,j+1} - \hat{I}_{t+1,i,j}|.
\]

(4)

\[
G_{I_{t,i,j}} = |I_{t,i,j+1} - I_{t,i,j}| + |I_{t,i,j+1} - I_{t+1,i,j}|.
\]

(5)

\[
G_{I_{t-2,i,j}} = |I_{t-2,i,j+1} - I_{t-2,i,j}| + |I_{t-2,i,j+1} - I_{t-3,i,j}|.
\]

(6)

So the loss function of gradient difference between frames is shown in the following formula:

\[
L_{g}(\hat{I}_t, I_{t-2}) = ||G_{\hat{I}} - G_{I_{t-2}}|| - ||G_{I} - G_{I_{t-2}}||_1.
\]

(7)

Therefore, the motion loss is as follows:

\[
L_m(\hat{I}_t, I_t, I_{t-1}) = L_{op}(\hat{I}_t, I_{t-1}) + L_g(\hat{I}_t, I_t, I_{t-2}).
\]

(8)

3) PERCEPTUAL LOSS

Perceptual loss was first proposed by Johnson \textit{et al}. [39]. It is widely used in image style transfer and super-resolution. We enforce the convolution features from predicted frames to be close to their convolution features ground truth, which enriches the texture details of predicted frames. The premise of abnormal event detection based on unsupervised learning is that normal events can be predicted and abnormal events
cannot be predicted. Therefore, we propose to replace the gradient loss used in [1] with perceptual loss. In this way, the prediction results on normal events obtained by our framework can obtain richer texture details and high-level features. However, due to the unpredictability of abnormal events, imposing a perceptual constraint on video prediction will not improve prediction results for abnormal events, which further highlight the abnormal existence of abnormal objects in videos. We use a pre-trained 19-layer VGG network [40] to extract high-level features of frames. Reference [22] pointed out that the activated features are very sparse, which can foretell that the activated features are very sparse, which can further highlight the abnormal existence of abnormal objects in videos. After the i-th maxpooling layer within the VGG19 network.

4) DISCRIMINATOR LOSS
As mentioned above, when there is no overlap between the probability distribution of a predicted frame and that of its ground truth, traditional JS divergence-based GAN will experience gradient vanishing, which makes neural networks difficult to train. Therefore, we propose a new discriminator, which incorporates self-attention module on the basis of a relativistic discriminator. We are not to estimate the probability that one input image is true, but to predict the probability that a real image is more realistic than a fake one. The discriminator loss is as follows:

\[
L^D_{adv}(\hat{I}, I) = \frac{1}{2} \sum_{i,j} (D(I)_{i,j} - D(\hat{I})_{i,j} - 1)^2 + \frac{1}{2} \sum_{i,j} (D(\hat{I})_{i,j} - D(I)_{i,j})^2. \tag{10}
\]

where \(D(\cdot)\) represents the output of our discriminator. \(i\) and \(j\) denote the pixel indexes in \(D(\cdot)\).

5) GENERATOR LOSS
The training procedure for our generator is to allow the generator to generate a frame sufficient to confuse the discriminator. Therefore, we hope that a generated image is close to its ground truth, so as to maximize the probability of the discriminator making a mistake. The adversarial loss is shown in the following formula:

\[
L^G_{adv}(\hat{I}, I) = \frac{1}{2} \sum_{i,j} (D(\hat{I}) - D(I) - 1)^2. \tag{11}
\]

So we combine all these constraints into our objective function. We can get the following objective function:

\[
L_G(\hat{I}, I, I_{t-1}, I_{t-2}) = \lambda_{int} L_{int}(\hat{I}, I) + \lambda_{m} L_{m}(\hat{I}, I, I_{t-1}, I_{t-2}) + \lambda_{p} L_{p}(\hat{I}, I) + \lambda_{G} L^G_{adv}(\hat{I}, I). \tag{12}
\]

When we train the discriminator, we use the following loss function:

\[
L_D(\hat{I}, I) = \lambda_{D} L^D_{adv}(\hat{I}, I). \tag{13}
\]

Our training process can be summarized in Algorithm 1.

**Algorithm 1 End-to-End Training for Our Network**

1: Initialize the network parameters with 0.
2: for numbers of training iterations do
3: Sample four consecutive frames of a video \(I^{(1)}, I^{(2)}, I^{(3)}, I^{(4)}\).
4: Update the generator \(G(I; \phi)\) by ascending its stochastic gradient.
5: Sample one example \(\hat{I}^{(5)}\) from data generating distribution.
6: Sample one example \(I^{(5)}\) from training set.
7: Update the discriminator \(D(I; \theta)\) by ascending its stochastic gradient.
8: end for

Output: the generator \(G(I; \phi)\)

### F. ANOMALY DETECTION

Abnormal event detection based on unsupervised learning assumes that normal events can be predicted, but abnormal events cannot be predicted. Therefore, we judge whether an abnormal event occurs based on the similarity between a predicted frame and its ground truth. We use peak signal-to-noise ratio (PSNR) [41] to evaluate the similarity between a predicted frame and its ground truth:

\[
PSNR(\hat{I}, I) = 10\log \frac{[\max(I)]^2}{\frac{1}{N} \sum_{i=0}^{N} (\hat{I}(i) - I(i))^2}. \tag{14}
\]

The higher PSNR, the more similar the predicted frame and its ground truth. We normalize PSNR of all testing video frames, which limits the range to \([0, 1]\), and use the following formula to calculate the regular score of each frame:

\[
S(t) = \frac{PSNR(\hat{I}, I) - \min_{I} PSNR(\hat{I}, I)}{\max_{I} PSNR(\hat{I}, I) - \min_{I} PSNR(\hat{I}, I)}. \tag{15}
\]

Therefore, we can set the threshold according to \(S(t)\) as a criterion for judging whether an abnormality occurs.

### IV. EXPERIMENTS

**A. DATASET**

We evaluate our proposed method on two publicly available anomaly detection datasets, including CUHK Avenue dataset [43] and ShanghaiTech Campus dataset [47].
CUHK Avenue dataset contains 16 training videos and 21 testing videos. The training videos are all normal events without labeling. The testing videos include normal events and abnormal events. Abnormal events which include throwing objects, running and crossing boundaries are labeled by frame-level labels in the testing videos.

Shanghai Tech Campus dataset contains 330 training videos and 107 testing videos. The training videos consist of 13 scenes, but the testing videos consist of 12 scenes. The dataset with various anomaly types is a challenging abnormal event detection dataset.

B. EVALUATION METRIC
In the field of anomaly detection, we usually calculate a Receiver Operating Characteristic (ROC) curve as an evaluation metric. The curve uses the probability of predicting a normal event to be wrong as the abscissa, and the probability of predicting an abnormal event to be correct as the ordinate. The closer the curve is to the upper left of the square area enclosed by (0, 0), (0, 1), (1, 0) and (1, 1), the better the detection performance of a model. Therefore, the Area Under Curve (AUC) which is a scalar can also be used for performance evaluation. If AUC is closer to 1, then the performance of the model is better.

C. TRAINING DETAILS
Unlike CUHK avenue dataset, Shanghai Tech Campus dataset contains 13 scenes with various camera angles. Our video prediction framework leverages global image features predict future frames, so environmental information has a great influence on our prediction framework. If we use all training videos containing 13 scenes to train our network, it will be difficult for our network to converge. Therefore, for the 12 scenes on the test set, we use the videos corresponding to the 12 scenes on the training set to train our network. Finally, we get 12 AUC scores. We take the average of these 12 AUC scores as the AUC of our network on Shanghai Tech Campus dataset. Since there is no corresponding testing data for the 13th scene in the training set, we only used 323 videos in the training phase. Our training method meets the requirements of applying our method to practice.

We normalize the intensity of pixels in all frames to $[-1, 1]$ and resize the size of each frame to $512 \times 512$. Due to the limited performance of our computer, we set the mini-batch size as 1. We initialize the parameters of our networks to 0. For optimization, we choose the Adam algorithm, where $\beta_1 = 0.9, \beta_2 = 0.999$. We alternately update the generator and the discriminator until our network converges.

The deep learning framework we used is Pytorch. All training and testing are completed on NVIDIA GeForce RTX 2080Ti GPU and Intel(R) Core(TM) i9-9900KF 3.6GHz CPUs.

D. RESULTS
1) AUC
We compare our method with multiple existing methods [1], [18], [42], [43], [45]–[47], [49]–[55]. Since our video prediction framework uses global image features to predict future frames, if we directly use the entire training set in Shanghai Tech dataset to train our network, it will be difficult for our network to converge. At this time, the AUC of our network on Shanghai Tech dataset is 71.6%. Therefore, we adopt the training method mentioned in Section IV-C, our network converges faster, and the AUC on Shanghai Tech dataset reaches 75.7%. Table 1 lists AUC of the above methods on CUHK avenue dataset and Shanghai Tech Campus dataset. From this table, we can see that our proposed method outperforms many existing methods. Fig. 6 is the visualization results of our method detecting abnormal events on the public datasets. We judge whether an abnormal event occurs based on the difference between prediction results and their ground truth, so we use the differential images obtained by subtracting prediction results from their ground truth to represent abnormal detection results. From the figure, we can see that the differential images related to normal events are almost completely black, while the differential images related to abnormal events have abnormal targets with artifacts. In other words, our framework can predict normal events, but cannot predict abnormal events. In summary, our network performs well on the public datasets.

| Method                        | CUHK Avenue | Shanghai Tech Campus |
|-------------------------------|-------------|----------------------|
| A.D. Giorno et al. [42]       | 78.3%       | N/A                  |
| M. Hasan et al. [18]          | 80.0%       | 60.9%                |
| W. Luo et al. [43]            | 77.0%       | N/A                  |
| S. Smeureanu et al. [46]      | 84.6%       | N/A                  |
| R. T. Ionescu et al. [45]     | 80.6%       | N/A                  |
| W. Luo et al. [47]            | 81.7%       | 68.0%                |
| W. Liu et al. [1]             | 85.1%       | 72.8%                |
| H. Park et al. [49]           | 88.5%       | 70.5%                |
| T. Ganokratana et al. [50]    | 87.9%       | N/A                  |
| D. Gong et al. [51]           | 84.9%       | 72.8%                |
| Y. Qiang et al. [52]          | 85.8%       | 73.7%                |
| H. Prawiro et al. [53]        | 86.4%       | N/A                  |
| F. Dong et al. [54]           | 83.7%       | 71.5%                |
| R. Cai et al. [55]            | 86.6%       | 73.7%                |
| Ours                          | **89.2%**   | **75.7%**            |

2) RUNNING TIME
We analyze the running time of our network. When we perform anomaly detection on each frame, our network needs to input the preprocessing results of four video frames. If we read four frames from the computer memory every time and perform preprocessing such as normalization, our network will not be able to detect abnormal events in real time. Because the process takes 20 milliseconds. Therefore, we propose a processing flow for real-time abnormal event detection, which shown in Fig. 7. In testing, we first preprocess the first three frames of a video read from the computer...
memory, and save the preprocessing results in the computer memory. Then we read the fourth frame of the video from the computer memory and save its preprocessing result in the computer memory. At this time, the four preprocessing results in the computer memory is the input of our network. When we read the fifth frame of the video, we will delete the preprocessing result of the first frame in the memory, and save the preprocessing result of the fifth frame, which ensure that there are always four preprocessing results in the memory. Finally, we compare the previous prediction result with the fifth frame to detect abnormal events. In this way, the time required for each abnormal event detection is only the time of preprocessing, the time of comparison between a predicted frame and its ground truth, and the time of prediction. We calculate the average running time of our network on the Avenue dataset. We start timing before preprocessing the fifth frame of each video. The average running time of our network is about 37 FPS. The average running time of the baseline [1] is 25 FPS, and our method shortens the running time by about 1.5 times. Most of the existing methods only consider the accuracy of detecting abnormal events, and ignore the importance of running time in real life. We also report the running time of other methods such as 20 FPS in [45], 3 FPS in [51]. What’s more, although AUC of method [48]
FIGURE 7. The processing flow for real-time abnormal event detection. We first preprocess the first three frames of a video read from the computer memory, and save the preprocessing results \( I_1, I_2, I_3 \) in the computer memory. Then we read the fourth frame of the video from the computer memory and save its preprocessing result \( I_4 \). At this time, the four preprocessing results \( I_1, I_2, I_3, I_4 \) in the computer memory is the input of our network. When we read the fifth frame of the video, we will delete the preprocessing result \( I_1 \) of the first frame in the memory, and save the preprocessing result \( I_4 \) of the fifth frame, which ensure that there are always four preprocessing results in the memory. In this way, the time required for each abnormal event detection is only the time of preprocessing, the time of comparison between a predicted frame and its ground truth, and the time of prediction. The gray block in the figure indicates the preprocessing process. The purple block is our network. The green block indicates the comparison between a predicted frame and its ground truth.

is 90.4% which is higher than ours, the running time of our method is better than 11 FPS of [48]. Among all real-time algorithms for abnormal event detection in videos, our method is competitive with the state-of-the-art methods.

3) ERRORS ANALYSIS
We analyze the reasons why our method misjudges abnormal events. Our method is to use the relationship between the PSNR of a predicted frame and a threshold as the basis for judging whether an abnormal event has occurred. Therefore, when an abnormal event is too far away from a camera, even if a few pixels of a predicted frame are very different from the corresponding pixels of its ground truth, the PSNR of the predicted frame is difficult to be greatly affected by the difference. This is why it is difficult for our method to detect abnormal events far away from a camera. In addition, we also analyze the reasons why our method misjudges normal events. Since we detect abnormal events through future frame prediction, when a moving object is too close to the camera, although we can predict the approximate state of the object, we cannot guarantee that all pixels of the object are exactly the same as its ground truth. This will reduce the PSNR of predicted frames and cause our algorithm to misjudge normal events. Nevertheless, this defect does not affect the application of our algorithm in real life. Because surveillance cameras are installed on a wall with a certain height in real life, there will be no moving objects too close to the camera.

E. ABLATION STUDY
We gradually ablate different improved parts to analyze the impact of the parts on anomaly detection. We conduct ablation experiment on CUHK Avenue dataset. First, based on the framework in [1], we use LiteFlowNet to replace FlowNet2 in [1]. In addition, we use gradient difference between frames to supplement the description of temporal features. The experimental results show that the AUC on the Avenue dataset is 87.2%, which is 2.1% higher than the method in [1]. Then we abandon the skip-connection in [1], and introduce dense residual networks. At the same time, we propose a new discriminator on the basis of a relativistic discriminator and modify discriminator loss and adversarial loss. The AUC of the improved framework on the Avenue dataset is 87.8%. What’s more, the size of our network is 21.1MB, which is 9MB smaller than the framework in [1]. Then we do not use gradient loss, and add self-attention modules in the discriminator and the encoding phase of the generator. The AUC of the improved network on the Avenue dataset is 88.8%. Here, the size of our network is 23.5MB. Finally, we impose a perceptual constraint on prediction results, which enhances the texture details of the generated images. The AUC of the improved model on the Avenue dataset is 89.2%. The results in Table 2 show that all of our contributions are crucial to obtain superior results.

F. SKIP-CONNECTION
Reference [1] uses U-net as a generator. U-net introduces the feature information of the encoding phase to the decoding phase through skip-connection. This is extremely important in image segmentation. Because deconvolution or upsampling process in the decoding phase needs to fill in a lot of blank content and generate feature information from scratch. This process lacks enough auxiliary information. The skip-connection provides multi-scale and multi-level information for image segmentation, so that a more refined segmentation effect can be obtained.
TABLE 2. Ablation study of our method on CUHK Avenue dataset.

| Method                                           | AUC   | Size   |
|--------------------------------------------------|-------|--------|
| Network architecture in [1] + \( L_{int}^*, L_{gd}^*, L_{adv}^*, L_{adv}^* \) | 85.1% | 30.1MB |
| Network architecture in [1] + \( L_{int}^*, L_{gd}^*, L_{adv}^*, L_{adv}^* \) \( L_m \) | 87.1% | 30.1MB |
| Our proposed method (Dense residual networks) + \( L_{int}^*, L_{gd}^*, L_{adv}^*, L_{adv}^* \) \( L_m \) | 87.8% | 21.1MB |
| Our proposed method (Dense residual networks + self-attention) + \( L_{int}^*, L_{adv}^*, L_{adv}^*, L_{adv}^* \) \( L_m \) | 88.8% | 23.5MB |
| Our proposed method (Dense residual networks + self-attention) + \( L_{int}^*, L_{adv}^*, L_{adv}^*, L_{adv}^* \) \( L_m \) \( L_p \) | 89.2% | 23.5MB |

\( L_{gd}^* \) indicates gradient loss. \( L_{adv}^* \), \( L_{adv}^* \) are the adversarial loss and the discriminator loss mentioned in [1].

![FIGURE 8. Skip-connection. The network is based on the structure of Fig.2, which use skip-connection to combine feature maps of different scales in the encoding phase with corresponding scale feature maps in the decoding phase.](image)

However, when we add skip-connection to our network (the architecture is shown in Fig.8), skip-connection does not work. We found through experiments that although the network with skip-connection can converge quickly, the AUC of the network on CUHK Avenue dataset is reduced by 10.1%. The experimental results are shown in Fig.9. Obviously, the network without skip-connection performs better than that with skip-connection on the Avenue dataset. In addition, we analyze the cause through the prediction results of the above two networks on a test video in the Avenue dataset. Fig.10a and Fig.10b are the PSNR curves of the prediction results of the network with skip-connection and that without skip-connection on an abnormal video, respectively. According to the PSNR curves, we can find that normal events part and abnormal events part of the PSNR curves in Fig.10b is more obvious. On the premise that the two networks have the same accuracy rate for abnormal event detection, the network without skip-connection has a lower rate of misjudgment for normal events. Although skip-connection improves the quality of prediction results, the network improves the prediction results for normal and abnormal events at the same time, which does not highlight the difference between normal and abnormal events. It seems that the network with skip-connection can predict abnormal events, but it is not. Skip-connection excessively provides the information of input frames for the up-sampling phase, so that a generated frame is not a prediction result of the future frame, but is closer to the previous frame. Due to the small changes from frame to frame of a video, this creates the illusion that the network with skip-connection can predict abnormal events. Through the analysis of experimental results, we believe that the co-occurrence of the self-attention mechanism and skip-connection will adversely affect anomaly detection. The function of skip-connection in U-net is to make the target image have the details of the source image. But we do not need to get too many details of the input frames for anomaly detection. Our purpose is to predict the next frame without abnormal events. The self-attention mechanism in our network has provided us with enough detailed information, and then using skip-connection is no more than superfluous. It will also produce the above-mentioned over-fitting.

FIGURE 9. Comparison of ROC curves of the network with skip-connection and that without skip-connection on the Avenue dataset. It can be seen that the ROC curve of the model without skip-connection is closer to the upper left of the square enclosed by (0, 0), (0, 1), (1, 0) and (1, 1). That is, the model without skip-connection performs better on the Avenue dataset.

![FIGURE 9. Comparison of ROC curves of the network with skip-connection and that without skip-connection on the Avenue dataset.](image)

G. DENSE RESIDENTIAL NETWORKS

A dense residual network is an extremely important part of our proposed method. We introduce dense residual networks in the auto-encoder to improve the feature extraction ability of our networks. Therefore, the number of dense residual networks and the number of convolutions in dense residual networks are critical. In theory, the deeper the convolutional network, the stronger the ability to extract features of videos. However, deep convolutional networks are not only difficult to train, but also brings unbearable delays in abnormal event detection. In this part, we use multiple sets of experiments to test the effect of the number of dense residual networks and the number of convolutions in the dense residual networks on the performance of our network. The experimental results are shown in Table 3. From the results shown in the table, we can find that the AUC of our method is lowest when \( N_d = 1 \). When \( N_d = 3 \), in theory, the AUC of our method should be higher than that of \( N_d = 2 \), but the network is too deep to be trained, which make it difficult to find the optimal solution.
FIGURE 10. Comparison of the PSNR curves of the prediction results of the network with skip-connection and that without skip-connection on an abnormal video. (a) and (b) are the prediction results for the third video in the Avenue test set by the network with and without skip-connection, respectively. The result is expressed by the normalized value of PSNR calculated by (15). The red area in the figure is the abnormal events range labeled in the dataset. The dashed line is the threshold we set. When the PSNR normalized value of a predicted frame is below the threshold, an abnormal event is considered to have occurred. We set the thresholds in (a) and (b) to 0.755 and 0.78 respectively to ensure that the two networks have the same accuracy for anomaly detection in the video. We can see from the yellow area in the figure that the network without skip-connection has a lower rate of misjudgment for normal events.

TABLE 3. The influence of the number of convolutions in dense residual networks and the number of dense residual networks on the performance of our method on CUHK Avenue dataset.

| $N_d$ | $N_c$ | AUC  | Size  | Runtime |
|-------|-------|------|-------|---------|
| 1     | 2     | 87.9%| 22.8 MB | 37 FPS  |
| 1     | 3     | 88.0%| 23.0 MB | 37 FPS  |
| 1     | 4     | 88.1%| 23.2 MB | 37 FPS  |
| 2     | 2     | 88.4%| 23.1 MB | 37 FPS  |
| 2     | 3     | 89.2%| 23.5 MB | 37 FPS  |
| 2     | 4     | 88.4%| 23.9 MB | 36 FPS  |
| 3     | 2     | 88.6%| 23.4 MB | 36 FPS  |
| 3     | 3     | 88.8%| 23.9 MB | 36 FPS  |
| 3     | 4     | 88.7%| 24.6 MB | 36 FPS  |

$N_d$ indicates the number of dense residual networks. $N_c$ indicates the number of convolutions in a dense residual network.

of the model. In addition, we find an interesting situation that the size of the network with $N_d = 2$ and $N_c = 2$ is smaller than that of the network with $N_d = 1$ and $N_c = 4$, but it has a higher AUC. The same situation exists between the model with $N_d = 3$ and $N_c = 2$ and the model with $N_d = 2$ and $N_c = 4$. Therefore, we think that the features extracted by multiple dense residual networks are more representative than that extracted by stacked convolution in a single dense residual network. Finally, we choose $N_d = 2$ and $N_c = 3$ as the optimal solution of our network.

H. HISTORICAL INFORMATION

A video is not only a stack of frames, but the interrelationship between frames is also an important part of a video. Therefore, we can use historical information to learn the development process of events. Our method makes full use of historical information to predict the future frames. When we predict the video frame $I_t$, the existing video frame $I_{t-n}$ ($t > n > 0$) will provide a lot of useful information that affects the prediction for $I_t$. Of course, this useful information will gradually decrease as $n$ becomes larger. Therefore, in this part, we use multiple sets of historical frames as the input of the generator, and finally get the best detection effect when $n = 5$. That is, the historical information contained in the video frame $I_{t-n}$ ($n > 5$) is not enough to affect the prediction for the video frame $I_t$. The experimental results are shown in Table 4. From the table, we can find that AUC decreases slightly when $N_h > 4$. So we think that too many historical frames will provide too much useless information, which interferes with the prediction for the next frame.

V. CONCLUSION

We leverage the difference between a future frame predicted by a video prediction framework and its ground truth to achieve unsupervised anomaly detection. We present a GAN that achieves better performance for abnormal event detection than previous video prediction frameworks. We formulate a novel auto-encoder containing dense residual networks and self-attention. We use the auto-encoder that acts as a generator to improve future frame prediction for normal events. Since abnormal events are irregular and unpredictable, our generator will not improve prediction results for abnormal events. So we increase the difference between prediction results for normal and abnormal events, thereby improving the accuracy of our network for abnormal event detection.

TABLE 4. The influence of historical information on the performance of our method on CUHK Avenue dataset.

| $N_h$ | AUC  | Runtime |
|-------|------|---------|
| 2     | 87.2%| 41fps   |
| 3     | 89.0%| 39fps   |
| 4     | 89.2%| 37fps   |
| 5     | 88.9%| 34fps   |
| 6     | 87.8%| 32fps   |

$N_h$ indicates the number of frames that provide historical information.
We also propose a new discriminator, which incorporates self-attention module on the basis of a relativistic discriminator. The discriminator not only avoids gradient vanishing in training, but also guarantees that the generator generates a frame closer to ground truth. In addition, we impose intensity, motion, and perception constraints on prediction results, which improves the quality of prediction results. We propose a real-time processing flow to detect abnormal events, which greatly improves the detection efficiency of anomaly detection algorithms based on a video prediction framework. Extensive experiments on two datasets show our method outperforms most existing methods. In future work, we aim to predict future behavior of an object in video based on object detectors. We will use a prediction framework to predict the future behavior of an object based on local features of the object obtained by object detectors, instead of predicting future frames based on global features of frames. By ignoring redundant environmental information, we will improve the performance of the prediction framework to detect abnormal events.

REFERENCES

[1] 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., Salt Lake City, UT, USA, Jun. 2018, pp. 6536–6545.

[2] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 6479–6488.

[3] S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio, “Angry crowds: Detecting violent events in videos,” in Proc. ECCV, Amsterdam, The Netherlands, 2016, pp. 3–18.

[4] L. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., Venice, Italy, Oct. 2015, pp. 5822–5831.

[5] Y. Zhu and S. Newsam, “Motion-aware feature for improved video anomaly detection,” in Proc. BMVC, Cardiff, Britain, U.K., 2019, pp. 1–12.

[6] J. Zhong, N. Li, W. Kong, S. Liu, T. Li, and G. Li, “Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection,” in Proc. CVPR, Long Beach, CA, USA, Jun. 2019, pp. 1237–1246.

[7] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 FPS in MATLAB,” in Proc. ICCV, Sydney, NSW, Australia, Dec. 2013, pp. 2720–2727.

[8] B. Zhao, L. Fei-Fei, and E. P. Xing, “Online detection of unusual events in videos via dynamic sparse coding,” in Proc. CVPR, Colorado Springs, CO, USA, Jun. 2011, pp. 3313–3320.

[9] Y. Chong and T. Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” in Proc. ISNN, Muroram, Hokkaido, Japan, 2017, pp. 189–196.

[10] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in Proc. CVPR, Las Vegas, NV, USA, Jun. 2016, pp. 733–742.

[11] J. R. Medel and A. Savakis, “Anomaly detection in video using predictive convolutional long short-term memory networks,” 2016, arXiv:1612.00390. [Online]. Available: https://arxiv.org/abs/1612.00390

[12] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, “Abnormal event detection in videos using generative adversarial nets,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, Sep. 2017, pp. 1577–1581.

[13] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention, Munich, Germany: Springer, 2015, pp. 234–241.

[14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Loy, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in Proc. ECCV, Munich, Germany, Sep. 2018, pp. 1–23.

[15] C. E. Duchon, “Lanczos filtering in one and two dimensions,” J. Appl. Meteorol., vol. 18, no. 8, pp. 1016–1022, 1979.

[16] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super-resolution as sparse representation of raw image,” in Proc. CVPR, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-solution using deep convolution networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016.

[18] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. CVPR, Honolulu, HI, USA, Jul. 2017, pp. 4681–4690.

[19] J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” in Proc. EMNLP, Austin, TX, USA, 2016, pp. 551–561.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017, arXiv:1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762

[21] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 7354–7363.

[22] 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.

[23] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, “Residual networks of residual networks: Multilevel residual networks,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 6, pp. 1304–1314, Jun. 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, Las Vegas, NV, USA, Jun. 2016, pp. 770–777.

[25] A. Jolicoeur-Martineau, “The relativistic discriminator: A key element missing from standard GAN,” 2018, arXiv:1807.00774. [Online]. Available: https://arxiv.org/abs/1807.00774

[26] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

[27] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. De, Smagt, D. Cremers, and T. Brox, “FlowNet: Learning optical flow with convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 2758–2766.

[28] T.-W. Hui, X. Tang, and C. C. Loy, “LifeflowNet: A lightweight convolutional neural network for optical flow estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 8981–8989.

[29] T. W. Hui and C. C. Loy, “LifeflowNet3: Resolving correspondence ambiguity for more accurate optical flow estimation,” in Proc. ECCV, 2020, pp. 169–184.
GUANJUN WANG received the B.S. degree in electronic science and technology and the M.S. degree in physical electronics from Zhengzhou University, Zhengzhou, China, in 2005 and 2008, respectively, and the Ph.D. degree in optical engineering from Beihang University, Beijing, China, in 2012. From 2013 to 2015, he was with the Institute of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, Shenzhen University, Shenzhen, China, as a Post-doctoral Researcher. He then was with the North University of China, Taiyuan, China, as an Associate Professor. In 2018, he joined Hainan University. He is currently an Associate Professor. His research interests include optical fiber sensor and demodulation systems, the Internet of Things, digital signal processing, digital image processing, and artificial intelligence.

MENGXING HUANG (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, in 2007. He then joined the Staff with the Research Institute of Information Technology, Tsinghua University, as a Post-doctoral Researcher. In 2009, he joined Hainan University. He is currently a Professor and a Ph.D. Supervisor of information and communication engineering, and the Dean of the School of Information Science and Technology. He is also the Executive Vice-President of the State Key Laboratory of Marine Resource Utilization in the South China Sea, and the Director of Hainan Key Laboratory of Big Data and Smart Services. He has published more than 60 academic articles as the first or corresponding author. He has reported 12 patents of invention, owns three software copyright, and published two monographs and two translations. His current research interests include big data and intelligent information processing, ocean information perception and fusion AI, and smart service. He is a Senior Member of Chinese Computer Federation (CCF), and a member of the Information System Committee in CCF. He has awarded two Second Class and one Third Class Prizes of The Hainan Provincial Scientific and Technological Progress.

HONGYU WANG (Member, IEEE) received the B.S. degree in computer science and technology from Zhengzhou University. He is currently pursuing the M.S. degree with the College of Computer and Cyber Security, Hainan University, Haikou, China. He went to study at the University of California at Berkeley, CA, USA, from December 2018 to January 2019. His research interests include computer vision, pattern recognition, deep learning, machine learning, SLAM, and formation control of multi-robot.

SHAOPING WEN is currently pursuing the M.S. degree with the College of Information and Communication Engineering, Hainan University, Haikou, China. His research interests include computer vision, pattern recognition, 3D object detection, and semantic image segmentation.

WEICHAO ZHANG received the B.S. degree from Yangzhou University, Yangzhou, China, in 2019. He is currently pursuing the M.S. degree with Hainan University, Haikou, China. His current research interests include computer vision, video understanding, and image processing.