Self-Supervised Attentive Generative Adversarial Networks for Video Anomaly Detection

Chao Huang, Jie Wen, Yong Xu, *Senior Member, IEEE*, Qiuping Jiang, *Member, IEEE*, Jian Yang, *Member, IEEE*, Yaowei Wang, *Member, IEEE*, and David Zhang, *Life Fellow, IEEE*

***Abstract*— Video anomaly detection (VAD) refers to the discrimination of unexpected events in videos. The deep generative model (DGM)-based method learns the regular patterns on normal videos and expects the learned model to yield larger generative errors for abnormal frames. However, DGM cannot always do so, since it usually captures the shared patterns between normal and abnormal events, which results in similar generative errors for them. In this article, we propose a novel self-supervised framework for unsupervised VAD to tackle the above-mentioned problem. To this end, we design a novel self-supervised attentive generative adversarial network (SSAGAN), which is composed of the self-attentive predictor, the vanilla discriminator, and the self-supervised discriminator. On the one hand, the self-attentive predictor can capture the long-term dependences for improving the prediction qualities of normal frames. On the other hand, the predicted frames are fed to the vanilla discriminator and self-supervised discriminator for performing true–false discrimination and self-supervised rotation detection, respectively. Essentially, the role of the self-supervised task is to enable the predictor to encode semantic information into the predicted normal frames via adversarial training, in order for the angles of rotated normal frames can be detected. As a result, our self-supervised framework lessens the generalization ability of the model to abnormal frames, resulting in larger detection errors for abnormal frames. Extensive experimental results indicate that SSAGAN outperforms other state-of-the-art methods, which demonstrates the validity and advancement of SSAGAN.**

***Index Terms*— Generative adversarial network (GAN), self-supervision, video anomaly detection (VAD).**

I. Introduction

WITH the increasing demand for addressing the challenge of a public security problem, surveillance cameras have been widely equipped in public, such as shopping malls, campuses, and train stations. These surveillance cameras can acquire huge amounts of surveillance videos for detecting abnormal events (such as fighting, crimes, and traffic accidents). However, it is extremely labor-intensive and time-consuming to manually detect abnormal events by watching all the surveillance videos, because the abnormal events rarely happen. Therefore, it is urgent to develop a video anomaly detection (VAD) method to automatically recognize abnormal events in surveillance videos timely and effective.

Theoretically, it is an intuitive solution to learn a binary classifier to directly discriminate a frame is or not abnormal. However, it is still a challenging task because of the unbounded and rare property of abnormal events. Specifically, it seems infeasible to accurately define all possible abnormal cases in various environments, which differs VAD from traditional classification tasks. In addition, it is an arduous task to collect sufficient abnormal samples for fully modeling abnormal patterns because of the sporadic occurrence of abnormal events. Thus, VAD is generally modeled as an outlier detection problem [1], [2]. Under this unsupervised setting, only normal data are available to train a model of normal patterns. At the test phase, frames not confirming the model of normal patterns are identified as abnormal frames.

Currently, the VAD approaches can be categorized into weakly supervised and unsupervised methods according to the supervisory signals at the training phase. The weakly supervised VAD [3]–[6] needs both normal and abnormal data to train the model. "Weak supervision" means that the model needs to be trained to achieve the fine (segment-level) anomaly detection with the supervision of coarse (video-level) annotations. In this article, we focus on achieving the unsupervised VAD, which does not require the abnormal video that is difficult to collect in real-life situations. In general, the existing unsupervised VAD methods can be categorized into three types: 1) hand-crafted features-based,

Manuscript received 10 February 2021; revised 28 October 2021; accepted 4 March 2022. Date of publication 5 April 2022; date of current version 30 October 2023. This work was supported in part by the Establishment of Key Laboratory of Shenzhen Science and Technology Innovation Committee under Grant ZDSYS20190902093015527, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20190806142416685, and in part by the Shenzhen Research Institute of Big Data and the Shenzhen Institute of Artificial Intelligence and Robotics for Society. *(Corresponding author: Yong Xu.)*

Chao Huang, Jie Wen, and Yong Xu are with the Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: huangchao_08@126.com; jiewen_pr@126.com; yongxu@ymail.com).

Qiuping Jiang is with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jiangqiuping@nbu.edu.cn).

Jian Yang is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@mail.njust.edu.cn).

Yaowei Wang is with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: wangyw@pcl.ac.cn).

David Zhang is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China, and the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China (e-mail: davidzhang@cuhk.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3159538.

Digital Object Identifier 10.1109/TNNLS.2022.3159538

2162-237X © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

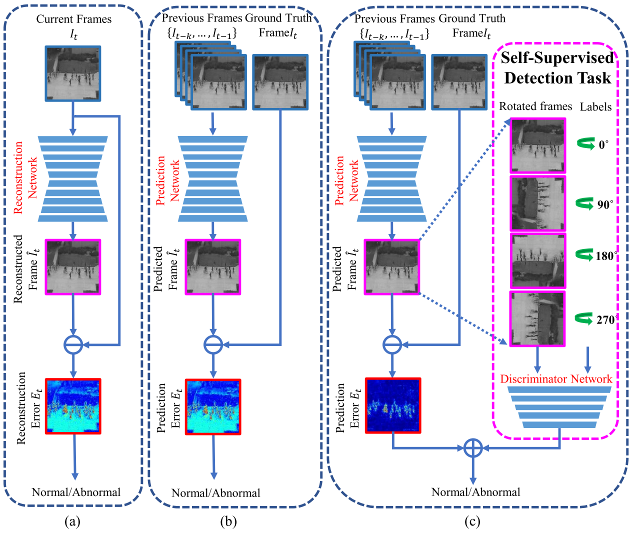

Current Frames
I_t
Reconstruction Network
Reconstructed Frame $\hat{I}_t$
Reconstruction Error E_t
Normal/Abnormal
(a)
Previous Frames Ground Truth
$\{I_{t-k}, \dots, I_{t-1}\}$ FrameI_t
Prediction Network
Predicted Frame $\hat{I}_t$
Prediction Error E_t
Normal/Abnormal
(b)
Previous Frames Ground Truth
$\{I_{t-k}, \dots, I_{t-1}\}$ FrameI_t
Self-Supervised Detection Task
Rotated frames Labels
0°
90°
180°
270°
Discriminator Network
Prediction Network
Predicted Frame $\hat{I}_t$
Prediction Error E_t
Normal/Abnormal
(c)

Fig. 1. Different frameworks of VAD. (a) Reconstruction-based framework. (b) Prediction-based framework. (c) Our framework introduces a self-supervised task into the prediction-based framework. The model is trained to produce rotation-detectable frames for normal data. In our framework, both prediction and self-supervised losses are used to detect anomalies, which can enlarge the gap of anomaly scores between normal and abnormal frames.

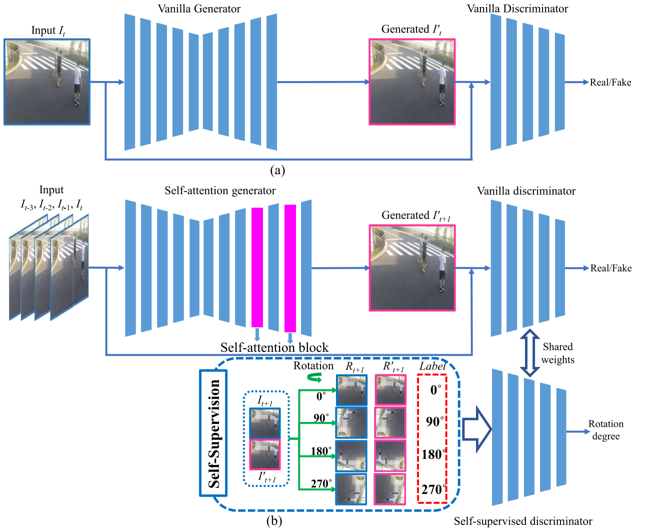


Fig. 2. Architectures of GAN. (a) Vanilla GAN. (b) Proposed SSAGAN. Self-attention block is embedded into generator to capture long-range contextual information. Notably, SSAGAN contains two discriminators, which are composed of Siamese networks with the shared weights to perform different tasks. The vanilla discriminator performs true versus fake discrimination task, while self-supervised discriminator detects rotation degree.

2) trajectory-based, and 3) deep generative model (DGM)-based methods. The hand-crafted features-based VAD approaches [7]–[11] usually represent spatiotemporal patterns of videos using some low-level features, such as histogram of oriented gradient (HOG) [8] and histograms of oriented optical flow (HOF) [9], [10]. The model is only trained on normal videos to reconstruct the normal frames with smaller reconstruction loss. Meanwhile, it is expected that this model would yield larger reconstruction errors for abnormal frames at the test phase. In this way, the reconstruction error is used to distinguish between normal and abnormal frames. However, this kind of method needs prior knowledge to design appropriate features. In addition, the generalization ability of such methods is poor, which makes them unable to detect anomalies in scenes not included in training data. The trajectory-based VAD [12]–[16] is another common approach, which first utilizes tracking algorithms to obtain motion information of objects in scenes. And then the statistical models are usually employed to obtain the normal patterns in videos according to the tracking results. Unfortunately, trajectory-based VAD cannot obtain satisfactory performance in complex scenes since tracking algorithms are difficult to accurately extract trajectories of objects. 3) With the recent advances in deep learning, DGM-based VAD [17]–[44] has become the most popular unsupervised paradigm of VAD. In general, the prior DGM-based methods include two types, reconstruction- and prediction-based methods. Specifically, they usually adopt generative adversarial networks (GAN) [45] or autoencoder (AE) [46] to reconstruct or predict the input frames and utilize the differences between input and generated frames to detect anomalies, as shown in Fig. 1(a) and (b). However, these methods usually capture the shared patterns between normal and abnormal samples due to the strong representation ability of the deep neural network, which leads to the model yielding similar generation errors for all types of samples. Thus, directly regarding generation error as the anomaly score of each frame like the existing DGM-based methods will yield a large overlap between the anomaly scores of normal and abnormal samples, resulting in less discrimination.

In this article, we propose a novel self-supervised framework for VAD to tackle the above-mentioned problem. As shown in Fig. 1(c), the proposed approach improves the existing DGM-based methods by enlarging the gap of abnormal scores between normal and abnormal frames from two perspectives. On the one hand, we develop a self-supervised mechanism for VAD, which enlarges the gap of anomaly scores between normal and abnormal frames by lessening the generalization of the model for abnormal frames. Specifically, our framework introduces a self-supervised task, i.e., rotation degree detection, which enables the model to detect the rotation degree of normal samples, while the abnormal samples will cause larger errors on this self-supervised task. Essentially, the role of self-supervision is to enable the generator to embed semantic information into the generated frames of normal frames via adversarial training, which enables the model to detect the rotation angle of normal frames. On the other hand, we insert the self-attention mechanism into our SSAGAN to capture long-range contextual information for improving the prediction qualities of normal frames. As a result, our model can produce a larger gap of abnormal scores between normal and abnormal frames, which improves the discrimination ability of the model.

In order to achieve the above-mentioned purposes, we present a novel self-supervised attentive GAN (SSAGAN). Different from the vanilla GAN, the proposed SSAGAN includes three modules: self-attentive generator, vanilla discriminator, and auxiliary self-supervised discriminator, as shown in Fig. 2(b). It is noted that the vanilla discriminator

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

performs true versus fake binary classification, while the self-supervised discriminator conducts the self-supervised task, i.e., rotation degree detection. Specifically, the generator receives consecutive frames to predict a future frame. Then, the generated future frame will be rotated four degrees, i.e., 0°, 90°, 180°, and 270°, and then they are fed to the self-supervised discriminator for conducting the rotation detection task. As a result, both the vanilla prediction errors and auxiliary rotation detection loss can be used to obtain a more discriminative anomaly score. To the best of our knowledge, this is the first work to solve the problem of VAD by using the self-supervised adversarial learning mechanism.

The main contributions of this article are listed as follows.

1) We propose a novel end-to-end self-supervised video anomaly detection framework. Our proposed framework enlarges the gap of anomaly scores between normal and abnormal frames by jointly using both prediction and self-supervised losses to detect anomalies, which outperforms the existing methods that only use prediction errors.
2) Different from the vanilla GAN, we embed the self-supervision mechanism into our proposed SSAGAN via adversarial training for enabling the auxiliary self-supervision discriminator to detect the rotation degrees of normal frames, which can improve the discriminative ability of the model by lessening the generalization ability for abnormal frames.
3) A self-attention mechanism is embedded into our generator, which can improve the discriminative ability of the model by facilitating the generator to better predict normal frames.

The remainder of this article is organized as follows. Section II introduces the related works. Section III details our proposed SSAGAN. Section IV presents the experimental results and discussions. Finally, Section V concludes this work.

II. Related Work

In this section, we briefly review several topics related to this work, including self-supervised learning (SSL), GANs, and video anomaly detection.

A. Self-Supervised Learning

As a popular unsupervised learning method, SSL has attracted wide attention since it obtains superior performance meanwhile avoiding the use of manually annotated data by introducing self-supervised signals. Theoretically, the earliest SSL can be traced back to AE [46], which learns feature representation by utilizing the input itself as supervision. By using noise as the supervision signal, denoising AE [47] can learn latent representations that are robust to the input noise. In general, SSL can be classified into three types according to the source of supervision signal: 1) several works leverage restoring ***part of the data itself*** as self-supervised task, such as image patch inpainting [48], [49], cross-channel prediction [50], and colorization of grayscale image [51], [52]. 2) ***Intrinsic structure information*** of image is used as supervision signal in some researches, such as pixel position [53], solving jigsaw puzzle [54], counting [55], predicting rotation degree [56], [57], and image instance identification [58]. 3) Some works explore the use of ***self-labels*** obtained from existing technologies [59], [60]. Instead of learning feature representations as to the earlier approaches, this work aims to enable the generator to generate rotation-detectable frames for normal frames via an adversarial learning scheme. To this end, we leverage supervision signals from image rotation to regularize the feature representation of self-supervision discriminator; thereby, the generator is enabled to generate rotation-detectable frames via adversarial training.

B. Generative Adversarial Networks

Recently, GAN has achieved great success in various computer vision tasks, such as inpainting [61], [62], image translation [63], video prediction [64], and image generation [57], [65]. In general, GAN contains two independent components, generator and discriminator, and utilizes discriminator against generator via a min–max two player game in order to both sides improve over training. In the architecture of GAN, generator generates a sample $x = \mathcal{G}(z)$ following the probability distribution $P_{\mathcal{G}}(x)$, when it is fed a noise z sampled from the latent space distribution. Then, the generator $\mathcal{G}$ is trained to enable the generative data distribution $P_{\mathcal{G}}(x)$ to be as close as possible to the distribution $P_{\text{data}}(x)$ of real samples $x_1, \ldots, x_n$. The optimization of generator $\mathcal{G}$ can be described as

$$\mathcal{G}^* = \arg\min_{\mathcal{G}} \Psi(P_{\mathcal{G}}(x), P_{\text{data}}(x)) \tag{1}$$

where $\Psi(\cdot)$ denotes the divergence between the distribution $P_{\mathcal{G}}(x)$ of generative data and distribution $P_{\text{data}}(x)$ of real data. This formulation means that the divergence between $P_{\mathcal{G}}(x)$ and $P_{\text{data}}(x)$ is expected to be as small as possible. Whereas, the role of discriminator $\mathcal{D}$ is to discriminate a sample is sampled from $P_{\text{data}}(x)$, rather than $P_{\mathcal{G}}(x)$. Therefore, the Jensen–Shannon divergence between $P_{\mathcal{G}}(x)$ and $P_{\text{data}}(x)$ is calculated as

$$\mathcal{V}(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log \mathcal{D}(x)] + \mathbb{E}_{x\sim P_{\mathcal{G}}(x)}[\log(1-\mathcal{D}(x))] \tag{2}$$

where the first term of right side indicates the expectation that a real data is discriminated as real. The second term of right side denotes the expectation that a generative data is discriminated as false. The optimization of discriminator is expressed as

$$\mathcal{D}^* = \arg\max_{\mathcal{D}} \mathcal{V}(\mathcal{G}, \mathcal{D}). \tag{3}$$

A larger value of $\mathcal{V}(\mathcal{G}, \mathcal{D})$ means a greater difference between $P_{\mathcal{G}}(x)$ and $P_{\text{data}}(x)$, and it is easier to discriminate whether a sample is generative data or real sample. Hence, the optimization of generator can be represented as

$$\begin{aligned}\mathcal{G}^* = \arg\min_{\mathcal{G}} \max_{\mathcal{D}} \big[& \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log \mathcal{D}(x)] \\ & + \mathbb{E}_{x\sim P_{\mathcal{G}}(x)}[\log(1-\mathcal{D}(x))]\big].\end{aligned} \tag{4}$$

When training the generator $\mathcal{G}$, the discriminator $\mathcal{D}$ is fixed, as well as $\mathcal{G}$ is fixed when training $\mathcal{D}$. In this way, the generator $\mathcal{G}$ is trained to generate a more realistic sample for fooling the

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

discriminator $\mathcal{D}$, while $\mathcal{D}$ is trained to distinguish the generated samples.

In different types of GAN, the conditional GAN [66] is usually applied in VAD methods. For instance, Liu *et al.* [17] proposed a GAN-based anomaly detection method, in which the generator is trained as future frames predictor of normal samples. Different from [17], our SSAGAN contains two discriminators (vanilla discriminator and auxiliary discriminator), in which the auxiliary discriminator is used to enable the generator to generate frames with certain information embedded. As a result, the gap in the generator's ability to generate frames embedded with specific information for different types of samples can also be used to detect anomalies.

C. Video Anomaly Detection

In general, the existing unsupervised VAD can be categorized into three classes: 1) hand-crafted features-based, 2) trajectory-based, and 3) DGM-based methods. For the first type of method, early researchers use low-level spatiotemporal features to learn normal patterns of videos. For example, Saligrama and Chen [7] propose a statistical probability model-based VAD method, in which the motion direction and magnitude are extracted to detect spatiotemporal anomalies in videos. Dalal and Triggs [8] take HOG as the feature to model the normal patterns. In [9] and [10], HOF is introduced into the VAD task. In addition to the statistic model [33], sparse coding is another popular method to model the normal patterns [67]. However, this type of method requires prior knowledge of scenes for designing appropriate features. Instead, our proposed SSAGAN is a data-driven method, and it does not require any prior knowledge of scenes to design features manually.

Trajectory-based methods usually utilize tracking algorithms to obtain motion information of objects and detect anomalies by analyzing the acquired trajectories. For instance, Santhosh *et al.* [12] propose a trajectory-based VAD method, which utilizes the Dirichlet process mixture model to cluster trajectories. Hu *et al.* [13] detect anomalies by clustering motion trajectories obtained from a multi-object tracking algorithm. Jiang *et al.* [14] utilize the hidden Markov model to represent object trajectories. Mo *et al.* [15] adopt a novel general trajectory-based adaptive sparse representations method. However, trajectory-based methods cannot obtain satisfying performance in complex scenes, because tracking algorithms are not robust to these scenes.

Deep neural networks have demonstrated remarkable performances in VAD, especially the DGMs. In some works [18]–[23], the appearance and motion features are extracted by the spatiotemporal AE to detect abnormal events. Luo *et al.* [24], [42] augments the sparse coding with temporally coherent to extract the similarity between adjacent frames, which obtain well performance for VAD. Sabokrou *et al.* [1], [25] put forward a deep one-class classifier-based VAD, in which denoising AE is used to distinguish normal and abnormal samples. Recently, GAN-based methods have achieved remarkable performance for VAD, including video prediction framework [17], adversarial one-class classifier [32], [35], and spatiotemporal GAN [36]. However, the above-mentioned DGM-based methods usually capture the shared patterns between normal and abnormal events due to the strong representation ability of DNNs, which leads to the trained model yielding similar reconstruction or prediction errors for abnormal and normal samples. Different from the previous DGM-based methods, our approach improves them by exploiting a self-supervised VAD framework, which can improve the discriminative ability of the model by enlarging the gap of abnormal scores between normal and abnormal frames.

III. Methodology

A. Problem Formulation

Prediction-based unsupervised VAD typically learns a frame prediction model on normal videos for capturing normal patterns. Then, anomalies are identified by poor prediction based on the assumption that the learned normal model can capture patterns of normal videos in order to predict the normal future frames well while poorly predict abnormal frames. Let $\mathbf{I}_{t-\Delta t+1:t} = \{I_i\}_{i=t-\Delta t+1}^{t}$ denote a video sequence containing Δt consecutive frames from $(t - \Delta t + 1)$th to tth frame, and $I_t \in \mathbb{R}^{H\times W\times C}$ represents the tth frame in a video. Given previous $\Delta t - 1$ frames $\mathbf{I}_{(t-\Delta t+1):(t-1)}$, the goal is to train a predictor $\mathcal{P}$ that can capture the discriminative semantics of normal patterns to minimize the differences between predicted normal frame $\widehat{I_t} = \mathcal{P}(\mathbf{I}_{(t-\Delta t+1):(t-1)})$ and actual normal frame I_t, which can be expressed as

$$\mathcal{P}^* = \underset{\mathcal{P}}{\arg\min}\ \|I_t - \mathcal{P}(\mathbf{I}_{(t-\Delta t+1):(t-1)})\|. \tag{5}$$

However, predicting the next frame I_t with tiny difference from the input sequence cannot enable the predictor $\mathcal{P}$ to learn how to capture the discriminative patterns of normalities. As a result, the abnormal frames can be also predicted well due to the powerful generalization ability of DNNs. Therefore, we propose a novel framework for VAD, which embeds the self-supervised mechanism into prediction-based architecture for obtaining more the discriminative anomaly score. To this end, we design the self-supervised attentive GANs (SSAGAN), which learns to capture the discriminative high-level semantics across normal training videos. Specifically, the proposed SSAGAN consists of three modules: the self-attentive frame predictor $\mathcal{P}$, the vanilla discriminator $\mathcal{D}^{\mathcal{V}}$, and the self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$, as shown in Fig. 3. A self-attention mechanism is embedded into $\mathcal{P}$ for encouraging $\mathcal{P}$ to capture the long range dependencies across all frame regions, which aims to reduce the prediction error between the predicted frame $\widehat{I_t}$ and actual frame I_t for normal frames. The vanilla discriminator $\mathcal{D}^{\mathcal{V}}$ performs true versus fake discrimination task, which can facilitate the predicted frame $\widehat{I_t}$ to approximate I_t by the constant adversarial process between $\mathcal{P}$ and $\mathcal{D}^{\mathcal{V}}$. Whereas, the self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$ predicts the rotation degrees of the rotated predicted frames. Given a set of rotational transformation $\mathcal{T} = \{\mathcal{R}(\widehat{I}; r)\}_{r=1}^{N}$ for each predicted frame $\widehat{I_t}$. The tth frame with the rth rotation is denoted by $\widehat{I_t^r}$, and $\widehat{I_t^r} = \mathcal{R}(\widehat{I_t}; r)$. The self-supervision

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

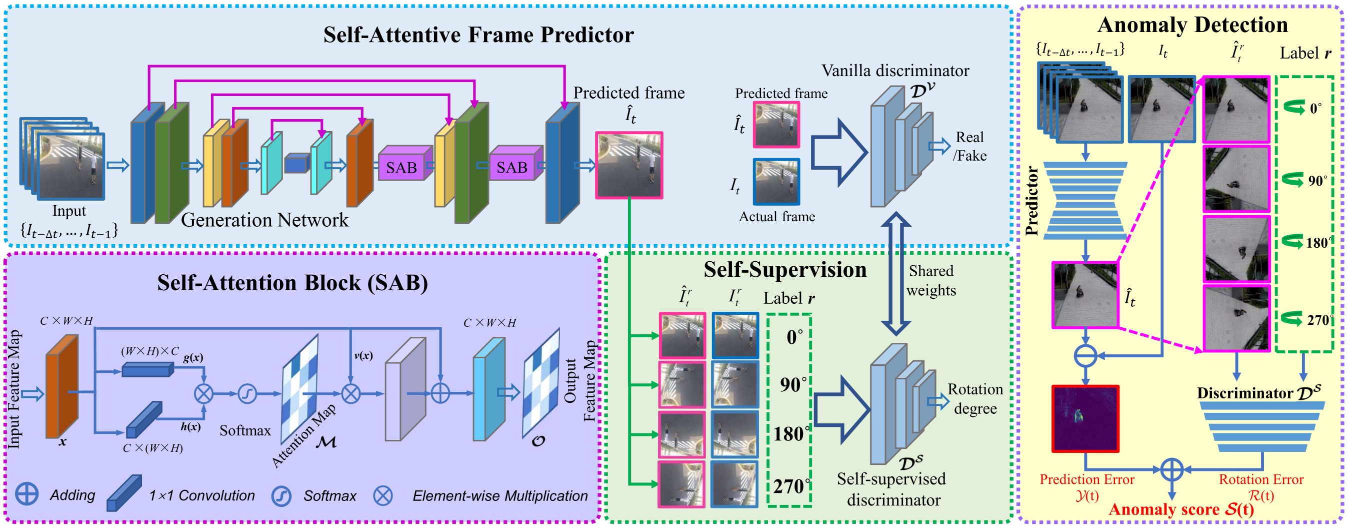


Fig. 3. SSAGAN contains three submodules: self-attentive frame predictor, vanilla discriminator, and self-supervision discriminator. Specifically, the frame predictor is trained with two discriminators in an adversarial manner for learning normal patterns. The vanilla discriminator performs true vs. fake binary classification, while self-supervision discriminator performs rotation degree detection. In addition, self-attention block is embedded into the frame predictor for obtaining predicted frames with high quality. At testing phase, only frame predictor and self-supervision discriminator are used. And the prediction error and self-supervised rotation detection error are used to calculate the anomaly score.

discriminator $\mathcal{D}^{\mathcal{S}}$ is trained to classify each rotated frame to one of the transformations, which can be expressed as

$$\mathcal{D}^{\mathcal{S}*} = \arg\min_{\mathcal{D}^{\mathcal{S}}} \frac{1}{N}\sum_{r=1}^{N} \ell\big(\mathcal{D}^{\mathcal{S}}\big(\widehat{I}_t^r\big), r\big) \tag{6}$$

where ℓ indicates the cross-entropy loss, and $N = 4$. $\widehat{I}_t^r = \mathcal{R}(\widehat{I}_t; r)$ indicates rotating frame $\widehat{I}_t$ counterclockwise by $(r-1)\cdot 90$ degree. Through the constant adversarial training process between $\mathcal{P}$ and $\mathcal{D}^{\mathcal{S}}$, the $\mathcal{P}$ is encouraged to generate rotation-detectable normal frames. During the test phase, $\mathcal{D}^{\mathcal{S}}$ can accurately detect the rotation degrees of rotated predicted frames for normal samples, while yields larger rotation detection errors for abnormal frames. As a result, our SSAGAN can utilize both the prediction error and rotation detection error to calculate the more discriminative anomaly score, which can be calculated as

$$\mathcal{S}(I_t) = \big\|I_t - \widehat{I}_t\big\|_2^2 + \frac{1}{N}\sum_{r=1}^{N} \ell\big(\mathcal{D}^{\mathcal{S}}\big(\widehat{I}_t^r\big), r\big). \tag{7}$$

B. Self-Attentive Frame Predictor

According to the existing works [13], [63], U-Net is an effective network architecture for image generation tasks since it can avoid the problems of gradient vanishing and information imbalance between feature layers in vanilla encoder–decoder network. Inspired by U-Net, our frame predictor is composed of an encoder-decoder structure with skip connections for achieving multiscale feature fusion, as shown in Fig. 3. Considering the discrepancy between previous frames and the future frame, a self-attention module is embedded into our frame predictor to capture the long-range contextual information, in order to model the relationships between local regions and widely spatial regions more efficiently. In general, nodes in a convolution layer are usually only connected with a small local neighborhood of nodes in the previous layer. Therefore, it is difficult to capture the long-range dependencies by only using the convolution layer. Whereas, self-attention block can be used to address this problem, which calculates the feature value at a position as the weighted sum of all feature values from different spatial positions. As a result, it enables $\mathcal{P}$ to predict the future frames with high quality.

1) Self-Attention in $\mathcal{P}$: Given a feature map $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ from the former layer, it is first fed to two 1×1 convolution layers to transform it into two corresponding feature space $\boldsymbol{g}$ and $\boldsymbol{h}$, yielding two new feature maps $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}^{C\times H\times W}$. Then, they are reshaped to $\mathbb{R}^{C\times K}$, where $K = H \times W$. Finally, the attention map $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{K\times K}$ is calculated by conducting matrix multiplication to $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})^{\top}$, where the operator $\top$ represents matrix transpose. The attention matrix $\boldsymbol{\mathcal{M}}$ is calculated as

$$\boldsymbol{\mathcal{M}}_{j,i} = \frac{\exp\big(\boldsymbol{g}(\boldsymbol{x}_j)\cdot \boldsymbol{h}(\boldsymbol{x}_j)^{\top}\big)}{\sum_{i=1}^{K}\exp\big(\boldsymbol{g}(\boldsymbol{x}_i)\cdot \boldsymbol{h}(\boldsymbol{x}_j)^{\top}\big)} \tag{8}$$

where $\boldsymbol{\mathcal{M}}_{j,i}$ represents the degree of attention that the model pays to the ith position when generating the jth region. In addition, $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ is also fed to another 1×1 convolution layer to yield feature map $\boldsymbol{v}(\boldsymbol{x}) \in \mathbb{R}^{C\times K}$. Then, $\boldsymbol{v}(\boldsymbol{x})$ is multiplied by the attention map $\boldsymbol{\mathcal{M}}$ and reshaped to $\mathbb{R}^{C\times H\times W}$. Finally, we multiply $\boldsymbol{\mathcal{M}}$ by a scalar factor, which is then added to $\boldsymbol{x}$. The final output of the self-attention module $\boldsymbol{\mathcal{O}} \in \mathbb{R}^{C\times H\times W}$ can be computed as

$$\boldsymbol{\mathcal{O}}_j = \boldsymbol{x}_j + \beta \cdot \sum_{i=1}^{K} \boldsymbol{\mathcal{M}}_{j,i}\cdot \boldsymbol{v}(\boldsymbol{x}_i) \tag{9}$$

where i and j are positions of the maps, and β is a scalar factor which is initialized as 0 and adjusted during training.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

2) Future Frame Prediction: In order to enable the frame predictor $\mathcal{P}$ to generate a desirable predicted frame $\widehat{I_t}$ which is close to the actual frame I_t, we define a conditional loss $\mathcal{L}_{\mathcal{C}}$ composed of three components: intensity loss ℓ_I, gradient loss ℓ_G, and motion loss ℓ_M. Specifically, we constrain the distance regarding gradient and intensity between $\widehat{I_t}$ and I_t. The intensity constraint ensures the similarity of pixels between predicted frames and target frames, and the gradient constraint can sharpen the predicted frames. Let $\mathcal{P}_\Phi$ and $\mathcal{D}_\Lambda^{\mathcal{V}}$ indicate the responses of frame predictor with parameter Φ and vanilla discriminator with parameter Λ. Then, the intensity loss ℓ_I is calculated by

$$\ell_I(\Phi; t) = \|\mathcal{P}(\mathbf{I}_{(t-\Delta t+1):(t-1)}) - I_t\|_2^2 \tag{10}$$

where $\mathbf{I}_{(t-\Delta t+1):(t-1)}$ denotes the input sequence containing previous $(t-1)$ frames of current frame I_t. Furthermore, the gradient loss ℓ_G is defined as

$$\ell_G(\Phi; t) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \left|\nabla \widehat{I_t}(i, j) - \nabla I_t(i, j)\right|. \tag{11}$$

The motion loss is defined by minimizing the difference of optical flows between generated frame and corresponding ground truth for constraining the temporal coherence between adjacent frames. In this article, the optical flow is estimated by LiteFlowNet [68] which is denoted as $\mathcal{F}$. The motion loss ℓ_M is calculated as

$$\ell_M(\Phi; t) = \left\|\mathcal{F}(\widehat{I_t}, I_{t-1}) - \mathcal{F}(I_t, I_{t-1})\right\|_1. \tag{12}$$

In summary, the total conditional loss $\mathcal{L}_{\mathcal{C}}$ for the predictor can be calculated as follows:

$$\mathcal{L}_{\mathcal{C}}(\Phi) = \sum_{t \in \mathcal{B}} \lambda_I \ell_I(\Phi; t) + \lambda_G \ell_G(\Phi; t) + \lambda_M \ell_M(\Phi; t) \tag{13}$$

where $\mathcal{B}$ indicates the mini batch size. λ_I, λ_G, and λ_M are hyperparameters balancing different components of objective function.

Through the constant adversarial training process between the frame predictor $\mathcal{P}$ and vanilla discriminator $\mathcal{D}^{\mathcal{V}}$, the frames predicted by $\mathcal{P}$ will gradually approximate the actual frames. Specifically, by minimizing the generative adversarial loss, $\mathcal{P}$ will generate more realistic frame to fool the vanilla discriminator for failing to discriminate it as fake. The generative adversarial loss of the frame predictor $\mathcal{P}$ from the vanilla discriminator $\mathcal{D}^{\mathcal{V}}$ is defined as

$$\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{V}}}(\Phi) = \sum_{t \in \mathcal{B}} -\log\left(\mathcal{D}_\Lambda^{\mathcal{V}}\left(\widehat{I_t}\right)\right) \tag{14}$$

where $\mathcal{D}_\Lambda^{\mathcal{V}}(\cdot)$ indicates the probability that the predicted frame $\widehat{I_t}$ is identified as real by the vanilla discriminator.

In order to enhance the discriminative ability of the vanilla discriminator, a discriminative adversarial loss is defined to train it. The vanilla discriminator is trained to discriminate the real frames from the generated frames by minimizing the discriminative adversarial loss. The loss function of the vanilla discriminator is defined as

$$\ell_{\mathcal{D}^{\mathcal{V}}}(\Lambda) = \sum_{t \in \mathcal{B}} \underbrace{-\log\left(1 - \mathcal{D}_\Lambda^{\mathcal{V}}\left(\widehat{I_t}\right)\right)}_{\text{(a)}} \underbrace{-\log\left(\mathcal{D}_\Lambda^{\mathcal{V}}(I_t)\right)}_{\text{(b)}} \tag{15}$$

in which (a) forces the vanilla discriminator to discriminate the generated frame as fake, while (b) encourages the vanilla discriminator to discriminate the real frame as real.

C. Augmenting With Self-Supervision

For tackling the problem of less discrimination in the existing DGM-based methods, our main idea is to facilitate the generator to obtain some semantically meaningful representations, which can be used by the additional self-supervised detection task for distinguishing normal and abnormal samples. In order to facilitate the generator to generate frames in accordance with the above purpose, the proposed SSAGAN investigates geometric transformations of generated frames, specifically rotations of generated frames by multiples of 90°, as supervisory signals and adds an additional discriminator to detect their transformations. In this way, the generator can generate transformation-detectable frames for normal samples through adversarial training. Whereas, the abnormal frames will obtain large detection errors in the self-supervised transformation detection task since the model did not learn how to detect their rotation angles. In addition to the prediction error between the predicted frame and its ground truth, the self-supervised transformation detection error between detected transformation and its label can be also used by our SSAGAN to differentiate normal and abnormal frames.

Similar to [56] and [57], we add the state-of-the-art self-supervision, frame rotation detection to the frame prediction-based VAD. In our SSAGAN, both the predicted frames and their corresponding original frames are rotated 0°, 90°, 180°, and 270°, and the rotation angles are used as the label of the self-supervised detection task. Furthermore, an additional self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$ is added to detect the rotation angle of each frame. Actually, the vanilla discriminator and the added self-supervision discriminator have the same network architecture, and they are shared with the same network weights. In other words, we use a single-discriminator network with two output heads, which perform different tasks, as shown in Fig. 3. In particular, the role of the vanilla discriminator head is to discriminate whether the nonrotation predicted frame is true or fake, while self-supervision discriminator head performs the rotation angle detection task. Moreover, the frame predictor aims to predict the future frame matching its ground truth, meanwhile, the predicted future frame allows the self-supervision discriminator head to detect its rotation angle after being rotated. To this end, we design a rotation detection loss for regularizing the frame predictor and the discriminator networks. The self-supervision adversarial loss of frame predictor $\mathcal{P}$ is defined as

$$\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{S}}}(\Phi) = \sum_{t \in \mathcal{B}} \sum_{r \in \mathcal{R}} -\log \mathcal{D}_\Lambda^{\mathcal{S}}\left(\widehat{I_t^r}\right) \tag{16}$$

where $r \in \mathcal{R} = \{0°, 90°, 180°, 270°\}$ is the rotation angle, and $\widehat{I_t^r}$ indicates rotating the predicted frame $\widehat{I_t}$ by r degrees. $\mathcal{D}_\Lambda^{\mathcal{S}}(\widehat{I_t^r})$ is the probability that the rotation degree r of rotated prediction frame $\widehat{I_t^r}$ is correctly detected by the self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$. The self-supervision adversarial loss of self-supervision

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

discriminator $\mathcal{D}^{\mathcal{S}}$ is defined as

$$\ell_{\mathcal{D}^{\mathcal{S}}}(\Lambda) = \sum_{t\in\mathcal{B}}\sum_{r\in\mathcal{R}} -\log \mathcal{D}^{\mathcal{S}}_{\Lambda}\big(I_t^r\big) \tag{17}$$

where I_t^r indicates rotating the actual frame I_t counterclockwise by r degrees. $\mathcal{D}^{\mathcal{S}}_{\Lambda}(\widehat{I_t^r})$ is the probability that the rotation degree r of rotated prediction frame $\widehat{I_t^r}$ is correctly detected by the self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$. Therefore, the total adversarial loss of frame predictor $\mathcal{P}$ can be described as

$$\begin{aligned}\mathcal{L}_A^{\mathcal{P}}(\Phi) &= \underbrace{\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{V}}}(\Phi)}_{\text{Vanilla loss}} + \underbrace{\alpha\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{S}}}(\Phi)}_{\text{Self-Supervised loss}}\\ &= \sum_{t\in\mathcal{B}}\left[-\log\big(\mathcal{D}^{\mathcal{V}}_{\Lambda}\big(\widehat{I_t}\big)\big) + \alpha\sum_{r\in\mathcal{R}} -\log \mathcal{D}^{\mathcal{S}}_{\Lambda}\big(\widehat{I_t^r}\big)\right] \end{aligned}\tag{18}$$

where α is the weight factor of self-supervision adversarial loss for frame predictor.

During the training phase, the overall loss of the frame predictor $\mathcal{P}$ is the combination of conditional loss $\mathcal{L}_{\mathcal{C}}$ and adversarial loss $\mathcal{L}_A^{\mathcal{P}}$, which can be expressed as

$$\begin{aligned}&\mathcal{L}_{\mathcal{P}}(\Phi)\\ &\quad= \underbrace{\mathcal{L}_{\mathcal{C}}(\Phi)}_{\text{Condition loss}} + \underbrace{\lambda_A\mathcal{L}_A^{\mathcal{P}}(\Phi)}_{\text{Adversarial loss}}\\ &\quad= \sum_{t\in\mathcal{B}}\lambda_I\ell_I(\Phi;t) + \lambda_G\ell_G(\Phi;t) + \lambda_M\ell_M(\Phi;t)\\ &\qquad+\lambda_A\sum_{t\in\mathcal{B}}\left[-\log\big(\mathcal{D}^{\mathcal{V}}_{\Lambda}\big(\widehat{I_t}\big)\big) + \alpha\sum_{r\in\mathcal{R}} -\log \mathcal{D}^{\mathcal{S}}_{\Lambda}\big(\widehat{I_t^r}\big)\right]\end{aligned}\tag{19}$$

where λ_A is the weight factor of generative adversarial loss of frame predictor. In this article, λ_I, λ_G, λ_M, and λ_A are set as to 1, 1, 2, and 0.05, respectively. Notably, our SSAGAN contains two discriminators, i.e., the vanilla discriminator $\mathcal{D}^{\mathcal{V}}$ and self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$, which are shared parameters. Therefore, the total loss of discriminator can be expressed as

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\Lambda) &= \ell_{\mathcal{D}^{\mathcal{V}}}(\Lambda) + \beta\ell_{\mathcal{D}^{\mathcal{S}}}(\Lambda)\\ &= \sum_{t\in\mathcal{B}} -\log\big(1-\mathcal{D}^{\mathcal{V}}_{\Lambda}\big(\widehat{I_t}\big)\big) - \log\big(\mathcal{D}^{\mathcal{V}}_{\Lambda}(I_t)\big)\\ &\quad+\beta\sum_{t\in\mathcal{B}}\sum_{r\in\mathcal{R}} -\log\mathcal{D}^{\mathcal{S}}_{\Lambda}\big(I_t^r\big)\end{aligned}\tag{20}$$

where β is the weight factor of self-supervision adversarial loss for discriminator.

In our SSAGAN, the frame predictor $\mathcal{P}$ and the vanilla discriminator head $\mathcal{D}^{\mathcal{V}}$ are adversarial with respect to the true versus fake discrimination loss $\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{V}}}(\Phi)$, while $\mathcal{P}$ and the self-supervision discriminator head $\mathcal{D}^{\mathcal{S}}$ are collaborative with respect to the self-supervised rotation detection loss $\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{S}}}(\Phi)$. On the one hand, $\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{V}}}(\Phi)$ biases the prediction toward realistic frames, while $\ell_{\mathcal{P}}^{\mathcal{D}^{\mathcal{S}}}(\Phi)$ enables $\mathcal{P}$ to generate future frames which allow $\mathcal{D}^{\mathcal{S}}$ to detect their rotation angles after being rotated. The frame predictor $\mathcal{P}$ is also augmented with the conditional loss $\mathcal{L}_{\mathcal{C}}$, which encourages $\mathcal{P}$ to predict the future frames as close as possible to their actual frames. On the other hand, the discriminator is trained to detect rotation angles only based on the actual frames. In other words, the self-supervised rotation detection loss on ground truth frames, i.e., $\ell_{\mathcal{D}^{\mathcal{S}}}(\Lambda)$, is used to update the parameters of the discriminator. This prevents the negative collaboration between frame predictor and discriminator whereby the frame predictor generates frames whose subsequent rotation angle is easily detected by the discriminator. The frame predictor is trained to generate rotation-detectable future frames since the predicted frames and the actual frames share features that can be utilized to detect rotation angles. During the test phase, the frame predictor cannot generate frames with the above-mentioned shared features for abnormal frames, and thus, their rotation angles cannot be accurately detected by the discriminator.

D. Anomaly Detection

This work aims to achieve frame-level VAD providing an anomaly score for each frame. In the existing works, such scores usually are reconstructed or predicted errors between reconstructed or predicted frames and their ground truth. In general, there are two common scores applied in DGM-based VAD: Peak signal-to-noise ratio (PSNR) and ℓ_p distance [26], [43], [44]. In essence, the existing DGM-based approaches usually utilize the different reconstruction or prediction abilities of the learned model for normal and abnormal samples to distinguish them. As discussed in Section I, directly using the reconstruction or prediction error as the anomaly score will yield a large overlap between the anomaly scores of abnormal and normal samples, resulting in less discrimination. Different from the existing DGM-based methods, our method jointly uses the prediction error and self-supervised rotation detection error as the anomaly score for enlarging the gap of anomaly scores between normal and abnormal samples. On the one hand, we apply PSNR to calculate the prediction error

$$\mathcal{Y}(t) = 10\log_{10}\frac{\max^2\big(\widehat{I_t}\big)}{\big\|I_t-\widehat{I_t}\big\|_2^2/P} \tag{21}$$

where P is the number of pixels of actual frame I_t, and $\max(\widehat{I_t})$ is the maximum value of the predicted frame $\widehat{I_t}$. It is noted that a low PSNR between I_t and $\widehat{I_t}$ indicates that I_t is more likely to be an abnormal frame.

On the other hand, the predicted frame $\widehat{I_t}$ will be rotated and fed to the discriminator for conducting self-supervised rotation detection task. And the rotation detection error can be calculated as

$$\mathcal{R}(t) = \frac{1}{N}\sum_{r=1}^{N}\ell\big(\mathcal{D}^{\mathcal{S}}\big(\widehat{I_t^r}\big), r\big) \tag{22}$$

where $r \in \mathcal{R} = \{0°, 90°, 180°, 270°\}$ is the rotation angle, and $N = 4$. $\widehat{I_t^r}$ represents that rotating the predicted frame $\widehat{I_t}$ by r degrees. ℓ indicates the cross-entropy loss. $\mathcal{D}^{\mathcal{S}}(\widehat{I_t^r})$ is the probability that the rotation degree r of rotated prediction frame $\widehat{I_t^r}$ is correctly detected by the self-supervision discriminator $\mathcal{D}^{\mathcal{S}}$. A high $\mathcal{R}(t)$ represents that I_t is more likely to be an abnormal frame.

Following the existing works [17], [19], both $\mathcal{Y}(t)$ and $(\mathcal{R}(t))$ are normalized in the range of [0,1] by a min–max normalization [17]. The final anomaly score for each

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DESCRIPTION OF DATASETS FOR VIDEO ANOMALY DETECTION

Dataset	Frame Number					Abnormal Events	Scenes Number
	Training	Testing	Total	Normal	Abnormal		
UCSD Ped1	6,800 frames	7,200 frames	14,000 frames	9,995 frames	4,005 frames	40	1
UCSD Ped2	2,550 frames	2,010 frames	4,560 frames	2,924 frames	1,636 frames	12	1
Avenue	15,324 frames	15,328 frames	30,652 frames	26,832 frames	3,820 frames	47	1
ShanghaiTech	274,515 frames	42,883 frames	317,398 frames	300,308 frames	17,090 frames	130	13

frame is defined as

$$\mathcal{S}(t) = \eta\mathcal{N}(\mathcal{R}(t)) + (1-\eta)(1-\mathcal{N}(\mathcal{Y}(t))) \tag{23}$$

where η is the weight factor for rotation detection error. $\mathcal{N}(\cdot)$ represents the min–max normalization over whole video frames, which is calculated as

$$\mathcal{N}(\mathcal{S}(t)) = \frac{\mathcal{S}(t) - \min_t \mathcal{S}(t)}{\max_t \mathcal{S}(t) - \min_t \mathcal{S}(t)} \tag{24}$$

where $\max_t \mathcal{S}(t)$ and $\min_t \mathcal{S}(t)$ denote the maximum and the minimum values of $\mathcal{S}(t)$ in a testing video, respectively. Thus, for a given threshold, the proposed model can automatically discriminate whether a frame is or not an abnormal frame based on the anomaly score $\mathcal{N}(\mathcal{S}(t))$.

IV. Experiments

In this section, we conduct extensive experiments on four datasets, UCSD Ped1, UCSD Ped2, CUHK Avenue, and ShanghaiTech to evaluate the performance of our proposed SSAGAN. Meanwhile, we compare our SSAGAN with other VAD approaches for validating the superiority of SSAGAN.

A. Datasets Description

The detailed descriptions of datasets are given in Table I. Notably, the training videos only contain normal videos, while the testing videos include normal and abnormal videos.

1) UCSD Dataset: This dataset contains two subdatasets: UCSD Ped1 and UCSD Ped2. The main difference between these two subdatasets is that videos in the former have object scale variation. The UCSD Ped1 totally contains 14 000 frames which are further divided into 34 and 36 videos for training and testing, respectively. In this subdataset, there are 40 types of abnormal behaviors, such as walking in the wrong direction, cars, skateboarding, and cycling among pedestrians. Whereas, the Ped2 contains 4 560 frames which are divided into 16 and 12 videos for training and testing, respectively. Each video contains less than 200 frames. In UCSD Ped2, there are 12 abnormal events, including: carts, cars, skateboarding, and cycling among pedestrians.

2) CUHK Avenue Dataset: There are 37 videos with the resolution of 640 × 360 pixels in CUHK Avenue. Among them, 16 and 21 videos are used for training and testing, respectively. In testing videos, there are 47 types of different abnormal events, such as throwing objects, dancing, and running.

3) ShanghaiTech Dataset: ShanghaiTech is the largest dataset for unsupervised VAD, which contains 437 video clips with a resolution of 856 × 480 pixels. In this dataset, the training set and testing set consist of 330 and 107 clips, respectively. This dataset is the most challenging dataset since it is captured from 13 scenes with various camera angles and illumination conditions. In addition, ShanghaiTech contains 130 real-world abnormal events, such as bicycles, skateboards, motorbikes, cars, fighting, chasing, jumping, and so on.

B. Implementation Details

The input frames are resized to frames with the resolution of 256 × 256 and the intensity of $[-1, 1]$ for all datasets. For the frame predictor, we adopt an encoder–decoder structure with skip connection which contains three blocks for the encoder and three blocks for the decoder. Each block includes a max pooling or deconvolution layer, after two convolution layers with the kernel size of 3 × 3. In this article, the sizes of max pooling and deconvolution layers are defined as 3 × 3 and 2 × 2, respectively. In addition, the self-attention block is embedded into the frame predictor to capture the long-range contextual information, in order to improve the qualities of predicted future frames. Let Conv(i, j, k) represent a convolution layer, in which i, j and k are kernel size, stride and the number of kernels, respectively. The architecture of discriminator is designed as: Conv(4, 2, 128)-Conv(4, 2, 256)-Conv(4, 2, 512)-Conv(4, 2, 512)-Output. Notably, the output layer contains two heads, $\mathcal{O}_1$ and $\mathcal{O}_2$. $\mathcal{O}_1$ is composed of a Conv(1, 1, 1) layer and a sigmoid layer to perform a true versus fake discrimination task, while $\mathcal{O}_2$ is composed of a Conv(1, 1, 4) layer to perform rotation degree classification. We adopt Adam optimizer with a batch size of 8 to optimize the parameters of our SSAGAN. The learning rates of frame prediction and discriminator are 2e-4 and 2e-5, respectively.

C. Comparison With State-of-the-Art Methods

Our proposed SSAGAN discriminates whether a frame is normal or abnormal based on the anomaly score, as shown in Fig. 4. The normal frames obtain low anomaly scores, while the abnormal frames obtain significantly higher anomaly scores. When a frame obtains an anomaly score that exceeds the set threshold, it will be identified as an abnormal frame. By setting different thresholds, we can get the receiver operating characteristic curve (ROC) curve. In previous works, the area under ROC (AUC) is used as an evaluation metric for measuring the performance of VAD approaches. Following previous works [17], [42], the frame-level AUC is applied

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

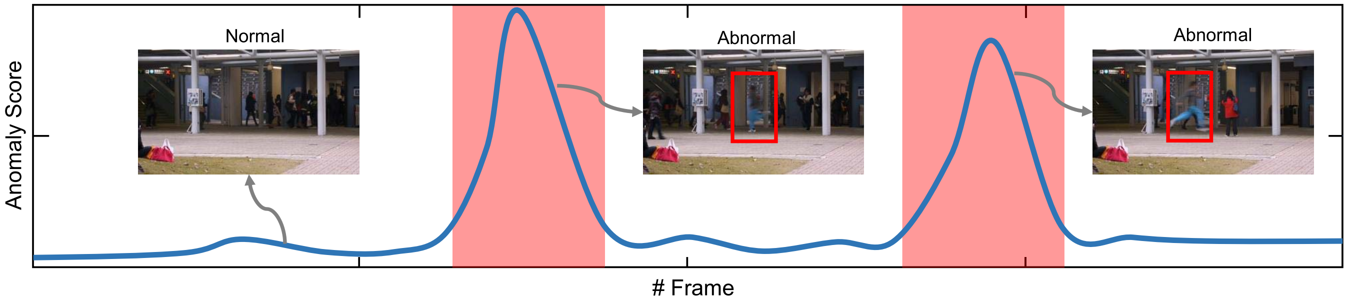
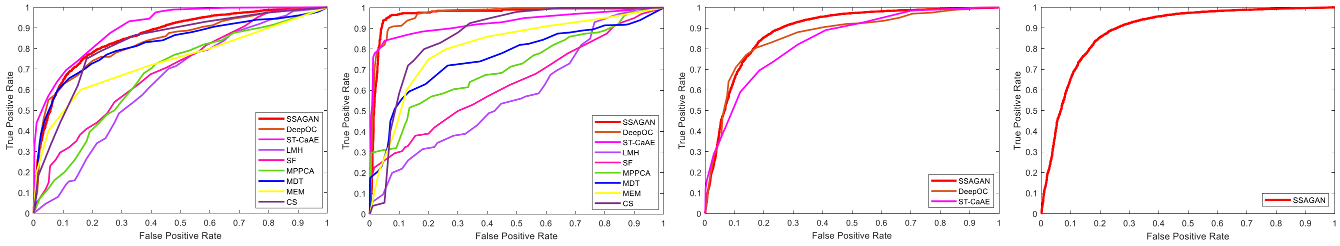


Fig. 4. Diagram of anomaly score curve. The red regions represent abnormal events.



(a) (b) (c) (d)

Fig. 5. ROC curves of different approaches on four datasets. (a) UCSD Ped1. (b) UCSD Ped2. (c) Avenue. (d) ShanghaiTech.

as the evaluation metric in this article. The performance comparison results on four datasets are shown in Tables II–IV. The ROC curves of different approaches on four datasets are shown in Fig. 5.

1) Results on the UCSD Ped1: Fig. 5(a) shows the ROC curves of SSAGAN and other compared approaches on Ped1, and the detailed results are listed in Table II. On Ped1, SSAGAN without preprocessing can obtain a competitive performance with the average frame-level AUC of 84.2%, which can be ranked the top three among all compared methods and only less than that of ST-CaAE [21] and AdaNet [39]. Notably, both ST-CaAE [21] and AdaNet [39] preprocess the video into multiple video cuboids, which enables them to obtain better performance on Ped1 than methods without the preprocessing. The reason is that videos in UCSD have perspective distortions, and dividing a video into multiple video cuboids can efficiently mitigate this problem. This point can be proved in [26]. When our SSAGAN adopts the preprocessing, it can obtain the best performance with the average frame-level AUC of 92.1%. In addition, the performance of SSAGAN without preprocessing is better than the frame prediction-based baseline future frame prediction (FFP) [17], which validates the effectiveness of our self-supervised architecture.

2) Results on the UCSD Ped2: On UCSD Ped2, SSAGAN is superior to all the baselines involved in the comparison with an average frame-level AUC of 97.6%, which is higher 0.6% of AUC than the existing state-of-the-art method dynamic prototype unit (DPU) [80]. Notably, SSAGAN without preprocessing obtains an average AUC of 96.9%, which is also superior to other frame prediction framework-based approaches, FFP [17], BMAN [19], ADLNet [75], and RUVAD [77]. The reason is that SSAGAN augments this framework with self-supervision, which enables our model to use both the prediction error and self-supervised detection error to calculate the anomaly score. Thus, SSAGAN can enlarge the gap between normal and abnormal frames and

TABLE II
COMPARISON WITH OTHER METHODS ON UCSD

Year	Method	Ped1 ↑	Ped2 ↑	Mean
2008	LMH [69]	63.4	58.1	60.8
2009	SF [70]	68.8	70.2	69.5
2009	MPPCA [71]	67.4	71.0	69.2
2014	MDT [72]	82.5	76.5	77.5
2016	Conv-AE [23]	75.0	90.0	82.5
2017	CS [73]	82.0	86.6	84.3
2017	CLSTM-AE [18]	75.5	88.1	81.8
2017	STAE [22]	-	91.2	-
2017	sRNN [24]	-	92.2	-
2017	Unmasking [40]	68.4	82.2	75.3
2018	MEM [74]	75.0	81.0	78.0
2018	FFP [17]	83.1	95.4	89.3
2019	AdaNet* [39]	90.4	90.3	90.4
2019	MemAE [34]	-	94.1	-
2019	AnomalyNet [28]	83.5	94.9	89.2
2019	AMAE [36]	-	96.2	-
2020	ISTL [20]	75.2	91.1	83.2
2020	ADLNet [75]	83.9	96.0	90.0
2020	BMAN [19]	-	96.6	-
2020	DeepOC* [26]	83.5	96.9	90.2
2020	MNAD [31]	-	97.0	-
2020	SIGnet [43]	86.0	96.2	91.1
2021	sRNN-AE [42]	-	92.2	-
2021	ST-CaAE* [21]	90.5	92.9	91.7
2021	MESDnet [76]	-	95.6	-
2021	RUVAD [77]	83.4	96.3	89.9
2021	AMMCN [78]	-	96.6	-
2021	ITAE [79]	-	96.8	-
2021	DPU [80]	85.1	96.9	91.0
2021	**SSAGAN***	**92.1**	**97.6**	**94.9**
2021	SSAGAN	84.2	96.9	90.6

[1] The methods masked with ‘*’ adopt the preprocessing.

TABLE III

COMPARISON WITH OTHER METHODS ON AVENUE

Year	Method	Frame-level AUC (%) ↑
2016	Conv-AE [23]	70.2
2017	CLSTM-AE [18]	77.0
2017	Unmasking [40]	80.6
2017	STAE [22]	80.9
2018	FFP [17]	85.1
2019	MemAE [34]	83.3
2019	AnomalyNet [28]	86.1
2019	AMAE [36]	86.9
2020	ISTL [20]	76.8
2020	AICNet [27]	77.2
2020	IPR [81]	85.1
2020	FSCN [82]	85.5
2020	ClusterAE [83]	86.0
2020	ADLNet [75]	86.0
2020	CL [30]	86.4
2020	DeepOC [26]	86.6
2020	SIGnet [43]	86.8
2020	Multispace [84]	86.8
2020	MNAD [31]	88.5
2021	sRNN-AE [42]	83.4
2021	ST-CaAE [21]	83.5
2021	DDGAN [85]	85.5
2021	MESDnet [76]	86.3
2021	AMMCN [78]	86.6
2021	ITAEGM [79]	86.0
2021	RUVAD [77]	88.3
2021	**SSAGAN**	**88.8**

TABLE IV

COMPARISON WITH OTHER METHODS ON SHANGHAITECH

Year	Method	Frame-level AUC (%) ↑
2016	Conv-AE [23]	60.9
2017	CLSTM-AE [18]	55.0
2017	TSC [24]	67.9
2018	RW [6]	71.5
2018	FFP [17]	72.8
2019	AdaNet [39]	70.0
2019	MemAE [34]	71.2
2019	PDE-AE [37]	72.5
2019	LSAG [37]	72.8
2019	PCM [86]	73.2
2019	MPED-RNN [16]	73.4
2019	AnoPCN [86]	73.6
2020	MNAD w/o Mem. [31]	66.8
2020	MNAD [31]	70.5
2020	CL [30]	71.6
2020	IPR [81]	73.0
2020	ClusterAE [83]	73.3
2020	Multispace [84]	73.6
2021	sRNN [24]	68.0
2021	sRNN-AE [42]	69.6
2021	Online [87]	70.9
2021	ITAE [79]	71.8
2021	ITAEGM [79]	73.0
2021	MESDnet [76]	73.2
2021	AMMCN [78]	73.7
2021	DPU [80]	73.8
2021	**SSAGAN**	**74.3**

improves the performance. Fig. 5(b) shows the ROC curves of our SSAGAN and other compared approaches on UCSD Ped2.

3) Results on the Avenue: Fig. 5(c) shows the ROC curves of SSAGAN and other compared approaches on Avenue, and the detailed experimental results are shown as Table III. Compared with the existing approaches, SSAGAN obtains the best performance with an average AUC of 88.8%, which is higher 0.3% than that of the previous best results of 88.5% AUC reported by memory-guided normality for anomaly detection (MNAD) [31]. It is noteworthy that the result of object-centric-CAE (88.9% AUC) reported in the article [29] is evaluated using a different metric, and the frame-level AUC evaluated with the common metric is 86.5% [77], which is 2.3% less than that of our SSAGAN. Compared with the frame prediction-based framework method FFP [17], SSAGAN obtains a performance gain of 3.1% AUC.

4) Results on the ShanghaiTech: On ShanghaiTech, SSAGAN is superior to all the approaches involved in the comparison with an average AUC of 74.3%, which is higher 0.5% than that of the existing state-of-the-art approach DPU [80] on this dataset. Compared with frame prediction framework-based methods FFP [17], MNAD [31], AnoPCN [86], and predictive coding module (PCM) [86], our self-supervision framework-based SSAGAN can obtain performance gain of 1.5%, 3.8%, 0.7%, and 1.1% in terms of frame-level AUC, respectively. In addition, Fig. 5(d) shows the ROC curve of SSAGAN on ShanghaiTech.

D. Analysis of SSAGAN

1) Gap of Anomaly Scores: The goal of our work is to tackle the problem that the model trained on normal data produces similar anomaly scores for normal and abnormal events. To this end, we introduce self-supervision into video anomaly detection for enlarging the gap $\Delta\mathcal{S}$ of anomaly scores between normal and abnormal frames. A higher value of the gap $\Delta\mathcal{S}$ indicates a more discriminative framework for distinguishing normal and abnormal samples. Therefore, we calculate the gap $\Delta\mathcal{S}$ of anomaly scores to verify the superiority of our framework, although the frame-level AUC has indicated the advanced performance of our approach. As shown in Table V, our self-supervision-based framework can obtain higher values of the gap $\Delta\mathcal{S}$ than the existing state-of-the-art frameworks (multiscale, AE, U-Net, future frame prediction, and so on), which indicates that our SSAGAN can achieve the expectation that enlarging the gap $\Delta\mathcal{S}$ of anomaly scores and also validates the effectiveness of the newly introduced self-supervision strategy.

TABLE V

GAP $\Delta\mathcal{S}$ OF ANOMALY SCORES

Method	Ped1	Ped2	Avenue	ShanghaiTech
CAE [23]	0.243	0.384	0.256	0.173
UNet [17]	0.243	0.435	0.270	0.174
FFP [17]	0.259	0.469	0.275	0.176
AnoPCN [86]	-	0.517	0.287	0.178
RUVAD [77]	0.260	0.512	0.344	**0.182**
SSAGAN	**0.262**	**0.519**	**0.347**	**0.180**

2) Ablation Study: In this section, we investigate the contribution of each component in SSAGAN to the overall performance, and the results on Avenue are listed in Table VI. The baseline model only contains the frame predictor, and

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

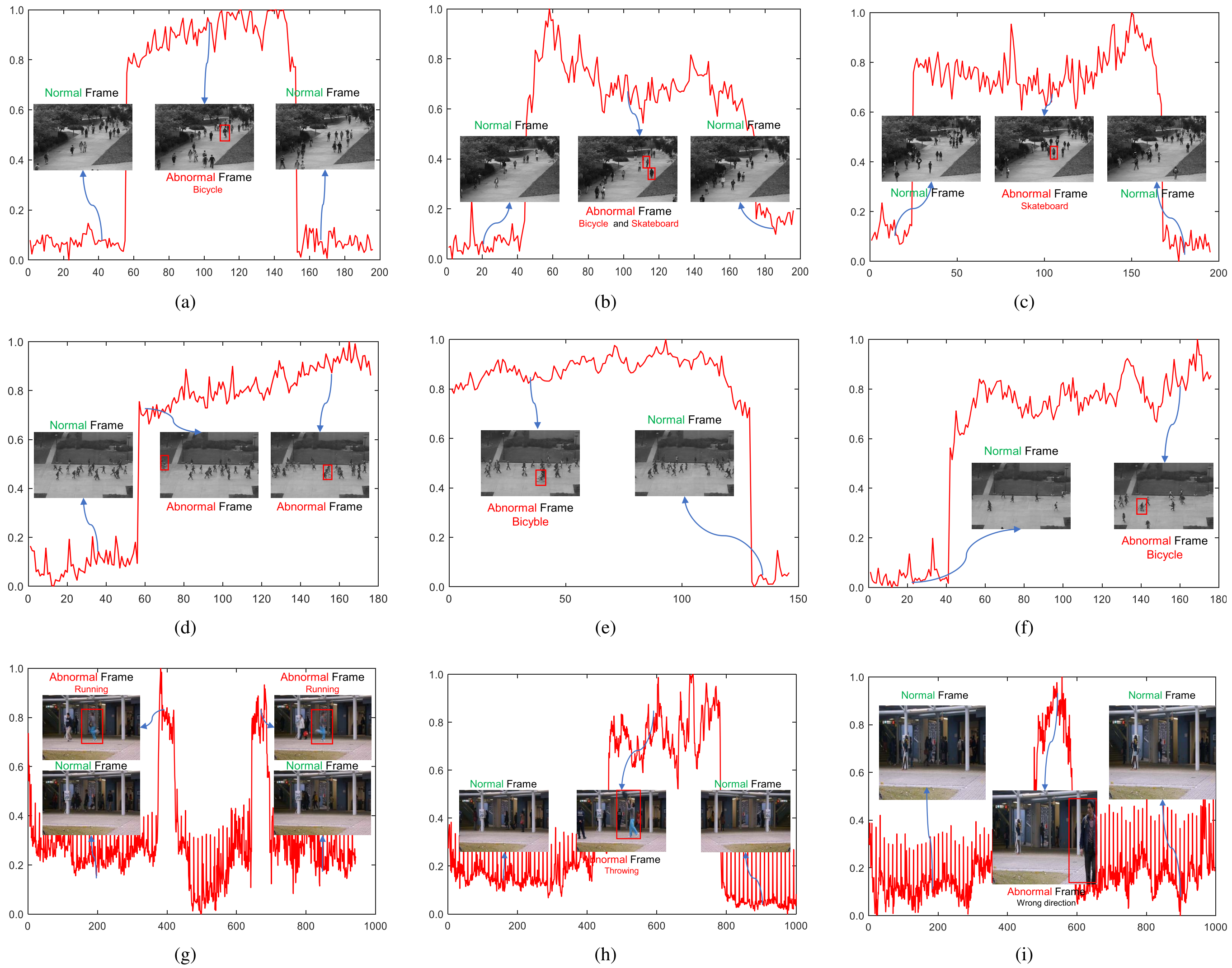


Fig. 6. Examples of anomaly score curves obtained by SSAGAN and representative video frames. (a) UCSD Ped1 testing video 1. (b) UCSD Ped1 testing video 2. (c) UCSD Ped1 testing video 4. (d) UCSD Ped2 testing video 1. (e) UCSD Ped2 testing video 5. (f) UCSD Ped2 testing video 7. (g) Avenue testing video 4. (h) Avenue testing video 5. (i) Avenue testing video 15.

TABLE VI
RESULTS OF THE ABLATION STUDIES ON AVENUE DATASET

Components	**Baseline**	**Network Design**		
Frame Predictor	✓	✓	✓	✓
Adversarial Loss	✗	✓	✓	✓
Self-Attention	✗	✗	✓	✓
Self-Supervision	✗	✗	✗	✓
Freme-level AUC	83.2	84.8	86.2	**88.8**

its training process is supervised by the losses of intensity, gradient, and motion. This basic model obtains a frame-level AUC of 83.2% on the Avenue. The adversarial loss, the self-attention module, and the self-supervision are further added step by step into our framework. First, we add a vanilla discriminator into the baseline to train the frame predictor. After adding the vanilla discriminator, the model can obtain an AUC of 84.8% with a performance gain of 1.6%. Then, we further insert the self-attention module into the frame predictor to capture the long-range contextual information for predicting normal frames better. As shown in Table VI, the self-attention module can help the model obtain a performance gain of 1.4% AUC. This competitive performance indicates the effectiveness of the self-attention module. Finally, we insert the self-supervision mechanism into our framework, i.e., the self-supervised discriminator is added to detect the rotation degrees of the predicted frames after being rotated. The complete SSAGAN obtains the average AUC of 88.8% on the Avenue with a performance improvement of 1.6%. The significant performance gain indicates that the self-supervision mechanism is the essential part, which improves the discriminative ability of SSAGAN.

3) Effect of Input Frame Number: In SSAGAN, the number of input frames T is a hyperparameter. On the one hand, it affects the qualities of predicted frames, thereby impacting the subsequent anomaly detection. On the other hand, it also influences the running speed of the frame predictor. In other words, it can be used to adjust the tradeoff between running speed and detection performance. Therefore, we conduct an

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
PERFORMANCE COMPARISON OF SSAGAN WITH DIFFERENT T

Input frames T	AUC Gain from $T = 4$		Average speed (fps)
	Ped2	Avenue	
2	-1.1%	-3.8%	45.2
4	**96.9%**	**88.8%**	**40.5**
6	+0.3%	+0.2%	36.3
8	+0.1%	+0.1%	33.5
10	+0.0%	+0.1%	32.3

TABLE VIII
TIME AND SPACE COMPLEXITY OF SSAGAN

Complexity	$\mathcal{P}$	$\mathcal{D}^{\mathcal{V}}$	$\mathcal{D}^{\mathcal{S}}$	Total
Time (FLOPs)	54.9 G	5.1 G	5.1 G	65.1 G
Space	44.4 M	10.1 M	10.6 M	65.1 M

[1] PLOPs indicate the number of floating point operations per second.
[2] M and G indicate million and gillion, respectively.

experiment on Ped2 and Avenue to show the effect of it on the performance, the result is listed in Table VII. It is clear that more input frames can provide more context and motion information, resulting in better performance for anomaly detection. Meanwhile, the speed of anomaly detection decreases with the increase in the number of input frames. When $T = 4$, SSAGAN can obtain a good tradeoff between running speed and performance of VAD with sufficiently fast running speed of 40.5 frames per second (fps), and more input frames can only obtain limited performance gain of no more than 0.3% of AUC.

4) Visualization of Anomalous Event Detection: Fig. 6 shows some instances of anomaly score curves from our SSAGAN and representative video frames of normal or abnormal events. It can be observed that our SSAGAN can make the correct responses to normal and abnormal events in time. Specifically, the anomaly score curve rises sharply when an abnormal object suddenly appears. And the anomaly score curve continuously remains at a quite high level when the abnormal event is in progress. When the objects causing the anomalous events to disappear or the abnormal behaviors are over, the curves quickly drop to a fairly low level.

5) Time and Space Consumption: Furthermore, we analyze the time and space complexities of our SSAGAN. Similar to the existing works [26], [76], we approximate the time and space complexity of the convolution and deconvolution layers in our network to that of our SSAGAN. On the one hand, the time complexity of each layer is computed as

$$TC \sim O(H_F W_F \cdot H_K W_K \cdot C_F C_K) \tag{25}$$

where H_F and W_F indicate the height and width of feature map; H_K and W_K indicate height and width of the kernel; and C_F and C_K indicate the number of channels for the kernel and output feature map, respectively. Consequently, the total time complexity of our SSAGAN is calculated as

$$TC \sim O\left(\sum_{i=1}^{n} H_F^i W_F^i \cdot H_K^i W_K^i \cdot C_F^i C_K^i\right). \tag{26}$$

On the other hand, there are two factors that cause spatial complexity, i.e., parameters and feature maps. And the spatial complexity of each layer can be defined as

$$SC \sim O(H_K W_K \cdot C_F C_K + H_F W_F \cdot C_F). \tag{27}$$

Therefore, the total space complexity of our model is calculated as

$$SC \sim O\left(\sum_{i=1}^{n} H_K^i W_K^i \cdot C_F^i C_K^i + H_F^i W_F^i \cdot C_F^i\right). \tag{28}$$

According to formula (26) and (28), the calculated time and space complexities of our SSAGAN are shown in Table VIII, where $\mathcal{P}$, $\mathcal{D}^{\mathcal{V}}$, and $\mathcal{D}^{\mathcal{S}}$ represent frame predictor, villa discriminator, and self-supervised discriminator, respectively. Notably, SSAGAN only uses $\mathcal{P}$ and $\mathcal{D}_{\mathcal{V}}$ in the testing phase; thus, its time and space complexities are 60 G and 55 M, respectively. Theoretically, GeForce RTX 2080 Ti can provide the operation speed of 112 640 GFLOPs; thus, our SSAGAN can achieve real-time abnormal events detection theoretically.

V. CONCLUSION

In this article, we propose a novel SSAGAN for video anomaly detection. Our SSAGAN is an end-to-end and data-driven unsupervised VAD method, which does not need prior knowledge of scenes to design appropriate features. To tackle the problem of less discrimination in the existing methods, SSAGAN adopts a self-supervised framework, which can enlarge the gap of anomaly scores between normal and abnormal frames by introducing a self-supervision mechanism. In addition, the self-attention mechanism is embedded into our generator, which enables the model to better predict the normal frames. As a result, SSAGAN can utilize both prediction and self-supervised detection errors to detect abnormal events. The experimental results on four datasets indicate SSAGAN outperforms the existing unsupervised VAD methods, which validates the effectiveness of SSAGAN. In the future work, we will explore to solve the problem of video anomaly detection using temporal transformation-based self-supervision.

REFERENCES

[1] M. Sabokrou, M. Fathy, G. Zhao, and E. Adeli, “Deep end-to-end one-class classifier,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 675–684, Feb. 2021.

[2] Q. Yu and K. Aizawa, “Unsupervised out-of-distribution detection by maximum classifier discrepancy,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9517–9525.

[3] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, “Localizing anomalies from weakly-labeled videos,” *IEEE Trans. Image Process.*, vol. 30, pp. 4505–4515, 2021.

[4] P. Wu and J. Liu, “Learning causal temporal relation and feature discrimination for anomaly detection,” *IEEE Trans. Image Process.*, vol. 30, pp. 3513–3527, 2021.

[5] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, “MIST: Multiple instance self-training framework for video anomaly detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14009–14018.

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

[7] 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.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[9] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1932–1939.

[10] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017.

[11] C. Huang *et al.*, "Online learning-based multi-stage complexity control for live video coding," *IEEE Trans. Image Process.*, vol. 30, pp. 641–656, 2021.

[12] K. K. Santhosh, D. P. Dogra, P. P. Roy, and B. B. Chaudhuri, "Trajectory-based scene understanding using Dirichlet process mixture model," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 4148–4161, Aug. 2021.

[13] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, Sep. 2006.

[14] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 907–913, Apr. 2009.

[15] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 631–645, Apr. 2014.

[16] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11988–11996.

[17] 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.

[18] 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.

[19] S. Lee, H. G. Kim, and Y. M. Ro, "BMAN: Bidirectional multi-scale aggregation networks for abnormal event detection," *IEEE Trans. Image Process.*, vol. 29, pp. 2395–2408, 2020.

[20] R. Nawaratne, D. Alahakoon, D. D. Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 393–402 Jan. 2020.

[21] N. Li, F. Chang, and C. Liu, "Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Trans. Multimedia*, vol. 23, pp. 203–215, 2021.

[22] 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.

[23] 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.

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

[25] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[26] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2609–2622, Jul. 2020.

[27] K. Xu, T. Sun, and X. Jiang, "Video anomaly detection and localization based on an adaptive intra-frame classification network," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 394–406, Feb. 2020.

[28] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[29] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7834–7843.

[30] K. Doshi and Y. Yilmaz, "Continual learning for anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 254–255.

[31] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.

[32] M. Zaigham Zaheer, J.-H. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14183–14193.

[33] E. Epaillard and N. Bouguila, "Variational Bayesian learning of generalized Dirichlet-based hidden Markov models applied to unusual events detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1034–1047, Apr. 2019.

[34] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[35] J. Wang and A. Cherian, "GODS: Generalized one-class discriminative subspaces for anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8200–8210.

[36] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.

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

[38] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12173–12182.

[39] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2138–2148, Aug. 2020.

[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.

[41] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[42] W. Luo *et al.*, "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1070–1084, Mar. 2021.

[43] Z. Fang, J. Liang, J. T. Zhou, Y. Xiao, and F. Yang, "Anomaly detection with bidirectional consistency in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1079–1092, Mar. 2022.

[44] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz, "Abnormal event detection and localization via adversarial event prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 3, 2021, doi: 10.1109/TNNLS.2021.3053563.

[45] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[46] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6, 1994, pp. 1–8.

[47] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[48] S. Jenni, H. Jin, and P. Favaro, "Steering self-supervised feature learning beyond local pixel statistics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6408–6417.

[49] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[50] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[51] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 840–849.

[52] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 645–654.

[53] T. N. Mundhenk, D. Ho, and B. Y. Chen, "Improvements to context based self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9339–9348.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

[54] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[55] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5899–5907.

[56] Z. Feng, C. Xu, and D. Tao, "Self-supervised representation learning by rotation feature decoupling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10356–10366.

[57] T. Chen, X. Zhai, M. Ritter, M. Lucic, and N. Houlsby, "Self-supervised GANs via auxiliary rotation loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12146–12155.

[58] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6203–6212.

[59] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 139–156.

[60] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9359–9367.

[61] Y. Shin, M. Sagong, Y. Yeo, S. Kim, and S. Ko, "PEPSI++: Fast and lightweight network for image inpainting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 252–265, Jan. 2021.

[62] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Recurrent temporal aggregation framework for deep video inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1038–1052, May 2020.

[63] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4066–4079, Aug. 2018.

[64] S. Li, J. Fang, H. Xu, and J. Xue, "Video frame prediction by deep multi-branch mask network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1283–1295, Apr. 2021.

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

[66] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–21.

[67] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017.

[68] T.-W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8981–8989.

[69] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[70] 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.

[71] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1446–1453.

[72] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[73] R. Leyva, V. Sanchez, and C.-T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3463–3478, Jul. 2017.

[74] T. Chen, C. Hou, Z. Wang, and H. Chen, "Anomaly detection in crowded scenes using motion energy model," *Multimedia Tools Appl.*, vol. 77, no. 11, pp. 14137–14152, Jun. 2018.

[75] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, "Attention-driven loss for anomaly detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4639–4647, Dec. 2020.

[76] Z. Fang, J. T. Zhou, Y. Xiao, Y. Li, and F. Yang, "Multi-encoder towards effective anomaly detection in videos," *IEEE Trans. Multimedia*, vol. 23, pp. 4106–4116, 2021.

[77] X. Wang *et al.*, "Robust unsupervised video anomaly detection by multipath frame prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 4, 2021, doi: 10.1109/TNNLS.2021.3083152.

[78] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proc. AAAI Artif. Intell.*, May 2021, vol. 35, no. 2, pp. 938–946.

[79] M. Cho, T. Kim, I.-J. Kim, and S. Lee, "Unsupervised video anomaly detection via normalizing flows with implicit latent features," pp. 1–11, 2021, *arXiv:2010.07524*.

[80] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15425–15434.

[81] 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.

[82] P. Wu, J. Liu, M. Li, Y. Sun, and F. Shen, "Fast sparse coding networks for anomaly detection in videos," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107515.

[83] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 329–345.

[84] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3694–3706, Sep. 2021.

[85] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, "Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5546–5554.

[86] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1805–1813.

[87] K. Doshi and Y. Yilmaz, "Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate," *Pattern Recognit.*, vol. 114, Jun. 2021, Art. no. 107865.

Chao Huang received the B.S. degree from Ningbo University, Ningbo, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China.

He has authored or coauthored over ten technical papers at prestigious international journals and conferences, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. His research interests include visual anomaly detection, video analysis, object detection, image/video coding, and deep learning.

Jie Wen received the Ph.D. degree in computer science and technology from Harbin Institute of Technology (HIT), Shenzhen, China, in 2019.

He is currently an Assistant Professor with HIT. He has authored or coauthored over 50 technical papers at prestigious international journals and conferences, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON MULTIMEDIA, European Conference on Computer Vision, the Association for the Advance of Artificial Intelligence, International Joint Conference on Artificial Intelligence, and ACM International Conference on Multimedia. His current research interests include biometrics, pattern recognition, and machine learning.

Yong Xu (Senior Member, IEEE) received the B.S. and M.S. degrees in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, 2015.

He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. He has authored or coauthored over 70 papers in top-tier academic journals and conferences. His research interests include pattern recognition, deep learning, biometrics, machine learning, and video analysis.

Dr. Xu has served as a Co-Editor-in-Chief for the *International Journal of Image and Graphics*, an Associate Editor for the *Chinese Association for Artificial Intelligence Transactions on Intelligence Technology*, and an Editor for the *Pattern Recognition and Artificial Intelligence*. His articles have been cited more than 5 800 times in the Web of Science and 13 000 times in Google Scholar.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.

Qiuping Jiang (Member, IEEE) received the Ph.D. degree in signal and information processing from Ningbo University, Ningbo, China, in 2018.

From 2017 to 2018, he was a Visiting Student with Nanyang Technological University, Singapore. He is currently an Associate Professor with Ningbo University. He received the Distinguished Youth Scholar Funding from the Zhejiang Natural Science Foundation. His research interests include image processing, visual perception, and computer vision.

Dr. Jiang received the Best Paper Honorable Mention Award of the *Journal of Visual Communication and Image Representation* and the Excellent Doctoral Dissertation Award of Zhejiang Province. He serves as the Area Chair/Session Chair/PC Member for International Joint Conference on Artificial Intelligence/the Association for the Advance of Artificial Intelligence/ACM International Conference on Multimedia/IEEE International Conference on Multimedia and Expo/International Conference on Image Processing/APSIPA-ASC. He serves as an Associate Editor for the *IET Image Processing*, the *Journal of Electronic Imaging*, and the *APSIPA Transactions on Signal and Information Processing*.

Jian Yang (Member, IEEE) received the B.S. degree in mathematics from Jiangsu Normal University, Xuzhou, China, in 1995, the M.S. degree in applied mathematics from Changsha Railway University, Changsha, China, in 1998, and the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002.

In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre of Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the School of Computer Science and Technology, NUST. He has authored more than 200 scientific articles in pattern recognition and computer vision. His research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang is a fellow of IAPR. He is/was an Associate Editor of *Pattern Recognition*, *Pattern Recognition Letters*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*. His articles have been cited more than 26 000 times in Google Scholar.

Yaowei Wang (Member, IEEE) received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2005.

He was with the Department of Electronics Engineering, Beijing Institute of Technology, Beijing, from 2005 to 2019. He was a Professor with the National Engineering Laboratory for Video Technology Shenzhen, Peking University Shenzhen Graduate School, Shenzhen, in 2019. From 2014 to 2015, he was an Academic Visitor with the Vision Laboratory, Queen Mary University of London, London, U.K. He is currently an Associate Professor with the Peng Cheng Laboratory, Shenzhen, China. His research interests include machine learning and multimedia content analysis and understanding. He has authored or coauthored over 70 refereed journals and conference papers.

Dr. Wang is a member of CIE. His team was ranked one of the Best Performers at the TRECVID CCD/SED tasks from 2009 to 2012 and PETS 2012.

David Zhang (Life Fellow, IEEE) received the B.S. degree in computer science from Peking University, Beijing, China, in 1974, the M.Sc. and Ph.D. degrees in computer science from Harbin Institute of Technology (HIT), Shenzhen, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, and an Associate Professor with Academia Sinica, Beijing. He has been the Chair Professor with The Hong Kong Polytechnic University, Hong Kong, since 2005, where he is currently the Founding Director of Biometrics Research Centre (UGC/CRC), supported by the Hong Kong Government. He is also the Presidential Chair Professor with The Chinese University of Hong Kong, Shenzhen. He also serves as a Visiting Chair Professor with Tsinghua University and HIT, and an Adjunct Professor with Shanghai Jiao Tong University, Shanghai, China, Peking University, the National University of Defense Technology, Changsha, China, and the University of Waterloo. He has been working on pattern recognition, image processing, and biometrics, over the past 30 years, where many research results have been awarded and some created directions, including palmprint recognition, computerized TCM, and facial beauty analysis are famous in the world. He has edited the book *Springer International Series on Biometrics* (KISB). He is an Associate Editor of more than ten international journals, including the IEEE TRANSACTIONS. He has authored or coauthored over 20 monographs, 500 international journal papers, and 40 patents from the USA, Japan, and China.

Prof. Zhang is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and IAPR/AAIA. He has organized the first International Conference on Biometrics Authentication. He is the Founder and the Editor-in-Chief of the *International Journal of Image and Graphics*. His articles have been cited more than 73 000 times in Google Scholar. He has been continuously listed as a Highly Cited Researchers in Engineering by Clarivate Analytics (formerly known as Thomson Reuters) for eight consecutive years from 2014 to 2021. He is also ranked about 85 with an h-index of 123 at Top 1000 Scientists for international Computer Science and Electronics.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on November 13,2023 at 05:08:10 UTC from IEEE Xplore. Restrictions apply.