Learning Event-Relevant Factors for Video Anomaly Detection

Che Sun[1], **Chenrui Shi**[1], **Yunde Jia**[2,1], **Yuwei Wu**[1,2*]

[1]Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
{sunche, shichenrui, jiayunde, wuyuwei}@bit.edu.cn

Abstract

Most video anomaly detection methods discriminate *events* that deviate from normal patterns as anomalies. However, these methods are prone to interferences from *event-irrelevant* factors, such as background textures and object scale variations, incurring an increased false detection rate. In this paper, we propose to explicitly learn *event-relevant* factors to eliminate the interferences from event-irrelevant factors on anomaly predictions. To this end, we introduce a causal generative model to separate the event-relevant factors and event-irrelevant ones in videos, and learn the prototypes of event-relevant factors in a memory augmentation module. We design a causal objective function to optimize the causal generative model and develop a counterfactual learning strategy to guide anomaly predictions, which increases the influence of the event-relevant factors. The extensive experiments show the effectiveness of our method for video anomaly detection.

Introduction

Video anomaly detection aims to automatically discriminate abnormal events that do not conform to normal patterns (Chandola et al. 2009; Luo et al. 2021). Most current methods (Chen et al. 2022; Yao et al. 2022; Chang et al. 2022) tackle the task by using deep neural networks to discriminate events that deviate from the learned patterns as anomalies, and have achieved a good performance. However, these methods are prone to interferences from *event-irrelevant* factors in videos, which tends to increase the false detection rate. For example, event-irrelevant factors, such as background textures and object scale variations shown in Figure 1, may interfere with the baseline video anomaly detection method HF2-VAD (Liu et al. 2021b). *Event-relevant* factors, e.g., the appearances and motions of objects in Figures 1a and 1b, are the causes of anomalies. Existing methods unconsciously mix up event-relevant factors with event-irrelevant factors into a single abstract representation of an input video. They tend to extract an overly simplistic representation which favors event-irrelevant factors due to "shortcut learning" (Geirhos et al. 2020), and thus event-irrelevant factors dominate anomaly predictions. Furthermore, the lack of annotated anomalies poses increasing challenges on capturing the concept of event-relevant factors for existing methods.

In this paper, we propose to explicitly learn event-relevant factors to eliminate the interferences from event-irrelevant factors on anomaly predictions. The basic idea is to increase the influence of event-relevant factors on anomaly predictions and decrease the influence of event-irrelevant factors. Specifically, we introduce a causal generative model to separate the event-relevant factors and event-irrelevant

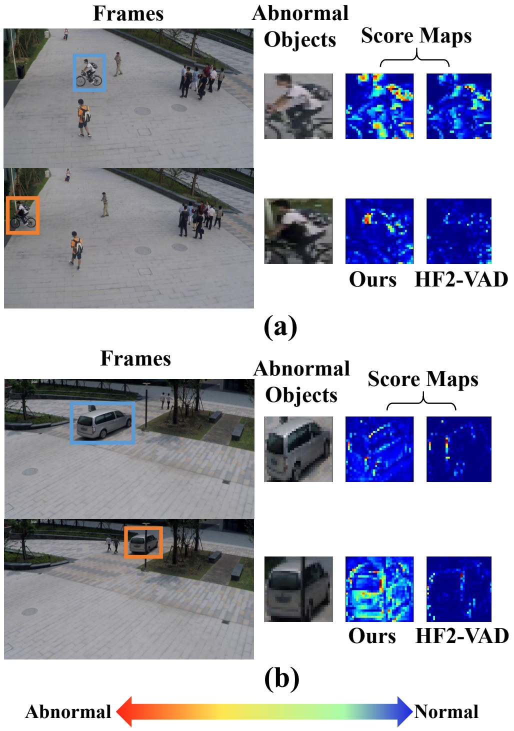

Figure 1: Examples of interferences from event-irrelevant factors. Frames with abnormal objects in orange boxes have large interferences of (a) background textures and (b) object scale variations. The score maps are normalized frame prediction errors in our method and the method HF2-VAD.

[*] Corresponding author

factors in videos, and use a memory augmentation module to learn the prototypes of event-relevant factors. We design a causal objective function to optimize the causal generative model and develop a counterfactual learning strategy to guide anomaly predictions, which increases the influence of event-relevant factors on anomaly predictions. During counterfactual learning, we generate counterfactual samples as pseudo-anomalies which help the anomaly predictions benefit from the learned event-relevant factors without extra annotated anomalies.

We evaluate our method on three datasets: ShanghaiTech (Luo, Liu, and Gao 2017), CUHK Avenue (Lu, Shi, and Jia 2013), and UCSD Ped2 (Mahadevan et al. 2009). We extend the evaluation metrics to small data scenarios, where only a small portion of training samples is available during training. In traditional big data scenarios, existing methods are likely to benefit from a performance gain by capturing inherent dataset biases without learning the causes (i.e., event-relevant factors) of anomalies. Their results are unstable, because the learned dataset biases seldom hold in an open world (e.g. previously unseen scenes). To this end, we perform additional evaluations in small data scenarios to simulate the use of anomaly detection methods in an open world. In big data scenarios, where all of the training samples are available during training, our method outperforms the baseline method HF2-VAD (Liu et al. 2021b) by learning event-relevant factors. In small data scenarios, where only 10% of the training samples are available during training, our method performs comparatively or even surpasses the baseline method HF2-VAD (Liu et al. 2021b) which uses all of the training samples. The experimental results verify the effectiveness of our method.

Our contributions are summarized as follows. (1) To our best knowledge, we are the first to explicitly learn event-relevant factors to eliminate the interferences from event-irrelevant factors on anomaly predictions. (2) We introduce a causal generative model to separate the event-relevant factors and event-irrelevant ones in videos and develop a counterfactual learning strategy to guide anomaly predictions without annotated anomalies. (3) The improvements of experimental results in both big and small data scenarios validate the effectiveness of our method.

Related Work

Video Anomaly Detection. This paper focuses on unsupervised anomaly detection where the training set provides only normal events, so we review the related deep methods in this section. Most existing methods can be categorized into two classes, namely, reconstruction-based methods and classification-based methods.

Reconstruction-based methods assume that reconstruction models trained only on normal events fail to reconstruct anomalies well so that anomalies are discriminated via larger reconstruction errors (Hou et al. 2021). Hasan *et al.* (Hasan et al. 2016) first introduced deep convolutional autoencoders to reconstruct normal events for video anomaly detection. A series of works on improving autoencoder structures were subsequently proposed for reconstructing video events better, such as convolutional LSTM autoencoders (Luo et al. 2017; Song et al. 2020), 3D convolutional autoencoders (Sun et al. 2021), two-stream autoencoders (Li, Chang, and Liu 2021), and so on. These methods studied overparameterized autoencoders and may wrongly generalize their reconstruction capacity to anomalies. To address the problem, some works proposed to use additional loss terms (e.g., adversarial losses (Abati et al. 2019) and contrastive losses (Huang et al. 2021)) to constrain the latent space of autoencoders to limit their capacity. Other works introduced a restricted memory space to directly replace the latent space, such as the memory-augmented deep autoencoder (MemAE) (Gong et al. 2019), multi-level memory modules in an autoencoder with skip connections (ML-MemAE-SC) (Liu et al. 2021b), and so on.

Classification-based methods assume that classifiers trained only on normal events fail to classify anomalies into any known class so that anomalies are discriminated via lower classification confidence scores (Pang et al. 2021). Early works argued that all normal events come from one class, and used one or more one-class classifiers for anomaly detection. For example, Sabokrou *et al.* (Sabokrou et al. 2018) applied an end-to-end generative architecture to construct a one-class classifier for discriminating anomalies. Xu *et al.* (Xu et al. 2017) extracted deep appearance and motion features and combined three one-class SVMs to compute anomaly scores. Recent methods believed that normal events in complex scenarios come from multiple classes, and a one-versus-rest SVM (Ionescu et al. 2019) was introduced to classify normal samples into multiple classes for discriminating anomalies that do not conform to any class. Furthermore, some works (Georgescu et al. 2021) collected some anomaly examples to construct binary classifiers for enhancing the performance of anomaly detection.

Although both reconstruction-based and classification-based methods perform well on existing datasets, they are prone to interferences from event-irrelevant factors, because the extracted deep representations for reconstruction or classification in these methods inevitably mix up event-relevant factors with event-irrelevant factors. In contrast, our method separates the event-relevant factors and the event-irrelevant factors in videos and increases the influence of event-relevant factors on anomaly predictions to eliminate the interferences. We apply our method to the baseline method HF2-VAD (Liu et al. 2021b) to improve the performance of video anomaly detection by explicitly learning event-relevant factors.

Causal Generative Models. Recent works have introduced causality into generative models to form causal generative models for learning the causal relationships by representation learning (Ding et al. 2022). These methods disentangle task-relevant and task-irrelevant factors, and establish causal relationships between task-relevant factors and model predictions for domain adaptations (Zhang et al. 2021; Yuan et al. 2022), out-of-distribution predictions (Liu et al. 2021a), and causal explanations (O'Shaughnessy et al. 2020; Holzinger et al. 2022), etc. Most existing methods learn task-relevant and task-irrelevant factors under the same constraints, and are likely to lose information or capture misinformation, because different types of factors have dif-

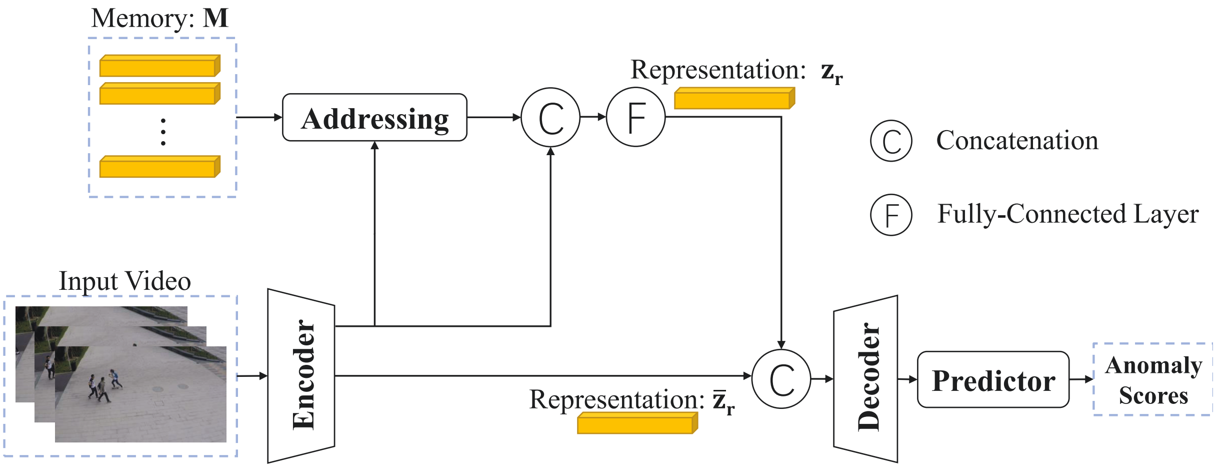

Figure 2: The architecture of our causal generative model.

ferent characteristics. Task-relevant factors are often known and typical while task-irrelevant factors are unknown and diverse. Therefore, we present a causal generative model to learn them under different constraints. We augment task-relevant (i.e., event-relevant) factors by using a memory module and leave the task-irrelevant (i.e., event-irrelevant) factors unconstrained. Our model provides fitting representations for different types of factors.

Method

We propose a causal generative model to separately model the event-relevant factors and event-irrelevant ones, and design a causal objective function to optimize the causal generative model. We further perform a counterfactual learning strategy to guide anomaly predictions.

Causal Generative Model

Given an input video sample $\mathbf{X}$, we use a predictor f to compute its anomaly score $y = f(\mathbf{X}) \in [0, 1]$ for anomaly detection. A causal generative model is used to learn two vector representations $\mathbf{z}_\mathrm{r}$ and $\overline{\mathbf{z}}_\mathrm{r}$ of $\mathbf{X}$, and we expect to capture event-relevant and event-irrelevant factors into the two representations, as shown in Figure 2. The representation $\mathbf{z}_\mathrm{r}$ is augmented by using a memory module to record prototypical event-relevant factors for reducing the noise influence. Due to the rare and unbounded natures of abnormal events, the memory module with finite elements only records prototypical factors of normal events, and abnormal events are discriminated according to the deviations from the prototypes (Gong et al. 2019; Liu et al. 2021b). We use memory variables $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_N]^\top$ and a corresponding address variable $a \in \{1, 2, \cdots, N\}$ for augmentation. We also use an additional independent representation $\overline{\mathbf{z}}_\mathrm{r}$ to represent event-irrelevant factors. $(\mathbf{z}_\mathrm{r}, \overline{\mathbf{z}}_\mathrm{r})$ constitutes the low-dimensional representations of the data distribution $p(\mathbf{X})$ through a generative mapping g, where $p(g(\mathbf{z}_\mathrm{r}, \overline{\mathbf{z}}_\mathrm{r})) \approx p(\mathbf{X})$. We construct a variational auto-encoder (VAE) to encode event-relevant and event-irrelevant factors independently from the input sample $\mathbf{X}$ and model its data distribution $p(\mathbf{X})$.

Variational Auto-Encoder. The standard VAE (Kingma and Welling 2013) assumes that each sample $\mathbf{X}$ corresponds to a low-dimensional latent variable $\mathbf{z}$ that is sampled from a Gaussian distribution, and forms a generative model as

$$p(\mathbf{X}) = \int_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{X}|\mathbf{z}) d\mathbf{z}. \tag{1}$$

The directed acyclic graph (DAG) describing the standard VAE is shown in Figure 3a. We modify the standard VAE by adding a memory module. As illustrated in Figure 3b, the VAE with a memory module forms a generative model

$$p(\mathbf{X}|\mathbf{M}) = \sum_a p(a|\mathbf{M}) \int_{\mathbf{z}} p(\mathbf{z}|\mathbf{m}_a) p(\mathbf{X}|\mathbf{z}, \mathbf{m}_a) d\mathbf{z}, \tag{2}$$

where a is the address variable, $\mathrm{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_N]^\top$ denotes the memory variables, and $\mathbf{m}_a$ is the a-th memory variable. As shown in Figure 3c, we further modify the VAE with a memory module by decoupling the latent variable $\mathbf{z}$ into two representations $\overline{\mathbf{z}}_\mathrm{r}$ and $\mathbf{z}_\mathrm{r}$ to form a generative model

$$\begin{aligned} p(\mathbf{X}|\mathbf{M}) = & \sum_a p(a|\mathbf{M}) \int_{\overline{\mathbf{z}}_\mathrm{r}} \int_{\mathbf{z}_\mathrm{r}} p(\overline{\mathbf{z}}_\mathrm{r}) p(\mathbf{z}_\mathrm{r}|\mathbf{m}_a) \\ & \times p(\mathbf{X}|\mathbf{z}_\mathrm{r}, \overline{\mathbf{z}}_\mathrm{r}, \mathbf{m}_a) d\overline{\mathbf{z}}_\mathrm{r} d\mathbf{z}_\mathrm{r}. \end{aligned} \tag{3}$$

The variational lower bound of our VAE model is

$$\begin{aligned} \log p(\mathbf{X}|\mathbf{M}) \geq & \mathbb{E}_{a, \overline{\mathbf{z}}_\mathrm{r}, \mathbf{z}_\mathrm{r} \sim q(\cdot|\mathbf{M}, \mathbf{X})} [\log p(\mathbf{X}, \overline{\mathbf{z}}_\mathrm{r}, \mathbf{z}_\mathrm{r}, a|\mathbf{M}) \\ & - \log q(a, \overline{\mathbf{z}}_\mathrm{r}, \mathbf{z}_\mathrm{r}|\mathbf{X}, \mathbf{M})], \\ q(a, \overline{\mathbf{z}}_\mathrm{r}, \mathbf{z}_\mathrm{r}|\mathbf{X}, \mathbf{M}) = & q(a|\mathbf{M}, \mathbf{X}) q(\mathbf{z}_\mathrm{r}|\mathbf{m}_a, \mathbf{X}) q(\overline{\mathbf{z}}_\mathrm{r}|\mathbf{X}). \end{aligned} \tag{4}$$

The components of our VAE include the memory variables $\mathbf{M}$, the address variable a, and the two representations $\mathbf{z}_\mathrm{r}$ and $\overline{\mathbf{z}}_\mathrm{r}$. The components and their variational posteriors in Eq. (4) are described as follows.

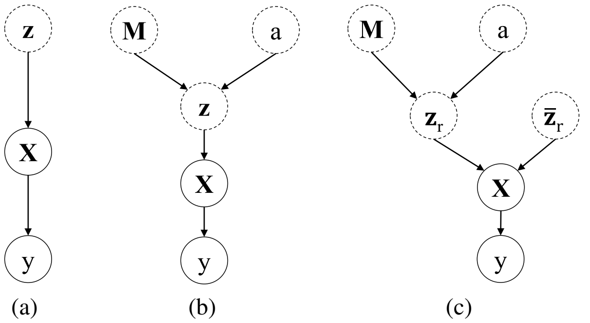

Figure 3: (a) A directed acyclic graph (DAG) describing the standard Variational Auto-Encoder (VAE). (b) A DAG describing the memory-augmented VAE. (c) A DAG describing our VAE.

The memory variables $\mathbf{M}$ are treated as model parameters, and are randomly initialized. Their values are updated by using the gradient of the objective. The address variable a is used to select a memory variable $\mathbf{m}_a$ from the memory. a is sampled from a variational posterior distribution $q(a|\mathbf{X}, \mathbf{M})$. We use the encoder of our VAE to encode $\mathbf{X}$ as a vector $\mathbf{e}$, and compute the similarity S between the encoded vector $\mathbf{e}$ and memory variables. $q(a|\mathbf{X}, \mathbf{M})$ is parameterized as a categorical distribution of the softmax over S, given by

$$q(a|\mathbf{X}, \mathbf{M}) = \frac{\exp(S(\mathbf{e}, \mathbf{m}_a)}{\sum_{j=1}^{N} \exp(S(\mathbf{e}, \mathbf{m}_j))}, \quad \text{where } S(\mathbf{e}, \mathbf{m}_a) = \frac{\mathbf{e}\mathbf{m}_a^\top}{\|\mathbf{e}\|\|\mathbf{m}_a\|}. \tag{5}$$

Once the posterior $q(a|\mathbf{X}, \mathbf{M})$ is computed, we sample a and retrieve $\mathbf{m}_a$ from $\mathbf{M}$ in a purely deterministic way. The prior of a is considered as a flat categorical distribution $p(a) = 1/N$.

We make assumptions about the priors and posteriors of the representations $\mathbf{z}_\mathrm{r}$ and $\overline{\mathbf{z}}_\mathrm{r}$ as follows: (1) For the representation $\mathbf{z}_\mathrm{r}$, we model a conditional Gaussian prior $p(\mathbf{z}_\mathrm{r}|\mathbf{m}_a)$ and an approximate conditional posterior $q(\mathbf{z}_\mathrm{r}|\mathbf{X}, \mathbf{M})$. (2) For the representation $\overline{\mathbf{z}}_\mathrm{r}$, we model a Gaussian prior $p(\overline{\mathbf{z}}_\mathrm{r})$ and an approximate conditional posterior $q(\overline{\mathbf{z}}_\mathrm{r}|\mathbf{X})$.

Optimization Strategy. The objective of our VAE is to faithfully represent the data distribution of $\mathbf{X}$. The approximated conditional posterior q in the VAE is learned from the encoder to generate $(a, \mathbf{z}_\mathrm{r}, \overline{\mathbf{z}}_\mathrm{r})$ with the inputs of $\mathbf{X}$ and $\mathbf{M}$. A generative mapping from $(a, \mathbf{z}_\mathrm{r}, \overline{\mathbf{z}}_\mathrm{r})$ to $\mathbf{X}$ is established in the decoder. We maximize the evidence lower bound (ELBO) to train the encoder and decoder. According to the log-likelihood bound in Eq. (4), the loss function of the VAE has four terms:

$$\begin{aligned}\mathcal{L}^{\mathrm{vae}} =& \mathbb{E}_{a,\mathbf{z}_\mathrm{r},\overline{\mathbf{z}}_\mathrm{r} \sim q}\left[\log p(\mathbf{X}|a, \mathbf{z}_\mathrm{r}, \overline{\mathbf{z}}_\mathrm{r})\right] \\ &+ \mathbb{E}_{a\sim q}[\mathrm{KL}\left(q(a|\mathbf{M}, \mathbf{X})\|p(a)\right)] \\ &+ \mathbb{E}_{a\sim q}[\mathrm{KL}\left(q(\mathbf{z}_\mathrm{r}|\mathbf{X}, \mathbf{m}_a)\|p(\mathbf{z}_\mathrm{r}|\mathbf{m}_a)\right)] \\ &+ \mathrm{KL}\left(q(\overline{\mathbf{z}}_\mathrm{r}|\mathbf{X})\|p(\overline{\mathbf{z}}_\mathrm{r})\right),\end{aligned} \tag{6}$$

where the first term is the expected reconstruction errors, and the last three $\mathrm{KL}(\cdot\|\cdot)$ terms are used to approximate the posteriors of all latent variables in the VAE. We compute backpropagation gradients and use the gradient descent algorithm to jointly optimize network parameters and memory variables. Different from standard VAEs, our method introduces a discrete latent variable a, which means that simply backpropagating gradients does not work well. Hence, we use the Gumbel-max relaxation-based approach (Maddison, Mnih, and Teh 2017; Jang, Gu, and Poole 2017) to compute the gradients of a in Eq. (5).

Causal Objective Function

We define the causal relationships among VAE components $\mathbf{M}$, a, $\mathbf{z}_\mathrm{r}$, $\overline{\mathbf{z}}_\mathrm{r}$, $\mathbf{X}$ and y, and introduce the causal objective function to increase the causal influence of $\mathbf{z}_\mathrm{r}$ on y. Since we focus on unsupervised anomaly detection without any annotated anomalies during training, we measure the causal influence on the predicted anomaly score y instead of the ground-truth anomaly label. The anomaly score y is acquired from a pre-trained anomaly predictor f.

The directed acyclic graph (DAG) in Figure 3c describes the causal relationships among $\mathbf{M}$, a, $\mathbf{z}_\mathrm{r}$, $\overline{\mathbf{z}}_\mathrm{r}$, $\mathbf{X}$ and y. Causal links from $\mathbf{M}$, $\overline{\mathbf{z}}_\mathrm{r}$ and a to $\mathbf{X}$ denote our memory-augmented generative process, and the causal link from $\mathbf{X}$ to y denotes the anomaly prediction process. The roles of the two representations $\mathbf{z}_\mathrm{r}$ and $\overline{\mathbf{z}}_\mathrm{r}$ are different: the causal link $\mathbf{X} \to y$ only uses features that are controlled by $\mathbf{z}_\mathrm{r}$. This means that interventions on both $\mathbf{z}_\mathrm{r}$ and $\overline{\mathbf{z}}_\mathrm{r}$ will raise changes of $\mathbf{X}$ but only interventions on $\mathbf{z}_\mathrm{r}$ will raise changes of y. We increase the causal influence of $\mathbf{z}_\mathrm{r}$ on y to achieve this goal. Recent works (Pearl 2009; O'Shaughnessy et al. 2020) introduce the information flow to measure the causal influence of the learned representation on the output of the predictor. Inspired by these works, we increase the causal influence by maximizing the information flow from $\mathbf{z}_\mathrm{r}$ to y.

Definition 1 [Information flow from U to V in a directed acyclic graph (Ay and Polani 2008)]. *Let U and V be disjoint subsets of nodes. The information flow from U to V is given by*

$$\begin{aligned}I(U \to V) := \int_U p(u) \int_V p(v|\mathrm{do}(u)) \\ \times \log \frac{p(v|\mathrm{do}(u))}{\int_{u'} p(u')p(v|\mathrm{do}(u'))du'} dV dU,\end{aligned} \tag{7}$$

where $\mathrm{do}(u)$ *denotes an intervention that fixes u to a value regardless of the values of its parents in the causal model.*

Since $\mathbf{z}_\mathrm{r}$ and $\overline{\mathbf{z}}_\mathrm{r}$ are independent when satisfying properties of the VAE evidence lower bound, the information flow from $\mathbf{z}_\mathrm{r}$ to y coincides with the mutual information between $\mathbf{z}_\mathrm{r}$ and y. That is

$$I(\mathbf{z}_\mathrm{r} \to y) = I(\mathbf{z}_\mathrm{r}; y) = \mathbb{E}_{\mathbf{z}_\mathrm{r},y}\left[\log \frac{p(\mathbf{z}_\mathrm{r}, y)}{p(\mathbf{z}_\mathrm{r})p(y)}\right]. \tag{8}$$

We refer readers to *supplementary materials* for the proof of Eq. (8). As we expect to maximize the causal influence of $\mathbf{z}_\mathrm{r}$ on y, the causal objective function is converted to a loss function $\mathcal{L}^{\mathrm{ce}} = -I(\mathbf{z}_\mathrm{r}; y)$. We optimize it together with the learning of our VAE model, and the loss function is

$$\mathcal{L}^{\mathrm{gen}} = \mathcal{L}^{\mathrm{vae}} + \lambda^{\mathrm{ce}}\mathcal{L}^{\mathrm{ce}}, \tag{9}$$

where λ^{ce} is a trade-off parameter and is set to 0.001.

Counterfactual Learning Strategy

The event-relevant factors are captured into the representation $\mathbf{z}_{\mathrm{r}}$ by using the causal generative model with the pre-trained anomaly predictor. We introduce a counterfactual learning strategy to finetune the anomaly predictor based on the learned factors. We generate counterfactual samples as pseudo-anomalies to improve the predictor so that it can pay more attention to event-relevant factors to correct the false predictions caused by event-irrelevant factors.

Given an input sample $\mathbf{X}$, we use an encoder of our VAE to obtain its representations $\mathbf{z}_{\mathrm{r}}$ and $\overline{\mathbf{z}}_{\mathrm{r}}$, and reconstruct $\mathbf{X}$ as $g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})$ via a decoder. We perform interventions on $\mathbf{z}_{\mathrm{r}}$ to generate counterfactual samples $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$. Inspired by the works of (Mothilal, Sharma, and Tan 2020; Haldar, John, and Saha 2021), a counterfactual sample $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ for a normal sample $\mathbf{X}$ is defined as such a sample that $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ is not predicted to be normal through the predictor f (i.e., $f(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})) > \epsilon$ for some threshold ϵ of the anomaly score), and $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ is sufficiently close to $g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})$ (i.e., $d(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}}), g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})) \leq \kappa$ for some threshold κ of the distance). The optimization objective is

$$\begin{aligned} &\min d(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}}), g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})), \\ &\quad \text{s.t. } f(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})) > \epsilon, \end{aligned} \tag{10}$$

where $d(\cdot, \cdot)$ is a distance measure function. In addition to generating counterfactual samples, we also perform interventions on the representation $\overline{\mathbf{z}}_{\mathrm{r}}$ to generate some normal samples for finetuning. We relax the optimization objective in Eq. (10), and design a loss function $\mathcal{L}^{\mathrm{csg}}$ for generating counterfactual samples and finetuning the predictor, given by

$$\mathcal{L}^{\mathrm{csg}} = \mathcal{L}^{\mathrm{pre}} + \lambda^{\mathrm{dis}} \mathcal{L}^{\mathrm{dis}}, \tag{11}$$

where λ^{dis} is a trade-off parameter and is set to 0.5. In Eq. (11), $\mathcal{L}^{\mathrm{pre}}$ is the anomaly prediction loss function, ensuring that the generated counterfactual samples $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ are predicted as abnormal (i.e., $f(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})) > \epsilon$) and the generated samples $g(\mathbf{z}_{\mathrm{r}}, \mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}}))$ are predicted as normal (i.e., $f(g(\mathbf{z}_{\mathrm{r}}, \mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}}))) \leq \epsilon$). $\mathcal{L}^{\mathrm{dis}}$ is the distance loss function for minimizing the distance between $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ and $g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})$.

Anomaly Prediction Loss. We generate triplet samples $\{g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}}), g(\mathbf{z}_{\mathrm{r}}, \mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}})), g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})\}$ to construct the anomaly prediction loss function, where $g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})$ is an anchor (normal) sample, $g(\mathbf{z}_{\mathrm{r}}, \mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}}))$ is the positive (normal) one and $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ is the negative (abnormal) one. The anomaly prediction loss $\mathcal{L}^{\mathrm{pre}}$ is

$$\begin{aligned} \mathcal{L}^{\mathrm{pre}} =& f\big(g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})\big) \\ &+ \max\bigg(0, \, \Big\| f\big(g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})\big) - f\big(g(\mathbf{z}_{\mathrm{r}}, \mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}}))\big) \Big\|_2^2 \\ &- \Big\| f\big(g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})\big) - f\big(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})\big) \Big\|_2^2 + \epsilon \bigg). \end{aligned} \tag{12}$$

where ϵ is a margin parameter. The first term $f(g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}}))$ is used to penalize the wrong prediction of reconstructed normal samples. The second term is a triplet loss function for closing the gap between $f(g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}}))$ and $f(g(\mathbf{z}_{\mathrm{r}}, \mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}})))$ as well as pushing $f(g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}}))$ and $f(g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}}))$ away.

Distance Loss. The distance between $g(\mathrm{do}(\mathbf{z}_{\mathrm{r}}), \overline{\mathbf{z}}_{\mathrm{r}})$ and $g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})$ can be measured in the representation space, as the abstract representation $\mathbf{z}_{\mathrm{r}}$ captures the event-relevant factors. We simply use an L2-norm to form the distance loss function

$$\mathcal{L}^{\mathrm{dis}} = \|\mathrm{do}(\mathbf{z}_{\mathrm{r}}) - \mathbf{z}_r\|_2^2. \tag{13}$$

Experiments

Datasets. We conduct experiments on three common benchmark datasets, including ShanghaiTech (Luo, Liu, and Gao 2017), CUHK Avenue (Lu, Shi, and Jia 2013), and UCSD Ped2 (Mahadevan et al. 2009). The *ShanghaiTech* dataset collects over $270k$ normal training frames and 130 abnormal events. It has 13 scenes with complex light conditions and camera angles. The *CUHK Avenue* dataset contains $35k$ frames in a single scene, and collects a total of 47 abnormal events, including throwing objects, loitering, and running. The *UCSD Ped2* dataset collects about $5k$ frames of a pedestrian walkway. It contains 12 abnormal events.

Implementation Details. We select the baseline anomaly detection model HF2-VAD (Liu et al. 2021b) as our predictor, and modify the predictor slightly to make it applicable to our method. Since the predictor should generate a normalized anomaly score $y \in [0, 1]$, we adopt the z-score normalization strategy and clamp the outputs of HF2-VAD into the range $[0, 1]$. HF2-VAD takes both RGB and optical-flow frames cropped by object bounding boxes as inputs. Since we focus on visual factors captured from RGB frames, we only use RGB frames as the input $\mathbf{X} \in \mathbb{R}^{32\times32\times15}$ to our causal generative model. We follow HF2-VAD to set both the height and width of $\mathbf{X}$ to 32 and set the channel number to 15 (i.e., 5 RGB frames). We keep the hyperparameters in our anomaly predictor the same as the baseline method HF2-VAD to make fair comparisons.

In our causal generative model, we use the ResNet-18 (He et al. 2016) as the backbone to construct the encoder and decoder. The output of the encoder is flattened, and two independent fully-connected layers are used to generate the encoded vector $\mathbf{e} \in \mathbb{R}^{32}$ and the parameters $\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n \in \mathbb{R}^{128}$ of the conditional posterior $q(\overline{\mathbf{z}}_{\mathrm{r}}|\mathbf{X})$. The memory bank $\mathbf{M}$ has 64, 64 and 256 memory variables for the CUHK Avenue, UCSD Ped2 and ShanghaiTech datasets, respectively. We concatenate the retrieved memory variable $\mathbf{m}_a \in \mathbb{R}^{32}$ and the encoded vector $\mathbf{e}$, and use a fully-connected layer to generate the parameters $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c \in \mathbb{R}^{32}$ of the posterior $q(\mathbf{z}_{\mathrm{r}}|\mathbf{X}, \mathbf{m}_a)$. $\mathbf{z}_{\mathrm{r}}$ and $\overline{\mathbf{z}}_{\mathrm{r}}$ sampled from their posteriors are concatenated, and then fed into the decoder to generate $g(\mathbf{z}_{\mathrm{r}}, \overline{\mathbf{z}}_{\mathrm{r}})$. The decoder has a symmetrical structure of the encoder. For counterfactual sample generation, we randomly initialize two intervening vectors $\Delta\mathbf{z}_{\mathrm{r}} \in \mathbb{R}^{32}$ and $\Delta\overline{\mathbf{z}}_{\mathrm{r}} \in \mathbb{R}^{32}$ for each input sample $\mathbf{X}$, and add them to the representations $\mathbf{z}_{\mathrm{r}}$ and $\overline{\mathbf{z}}_{\mathrm{r}}$ as interventions $\mathrm{do}(\mathbf{z}_{\mathrm{r}}) = \mathbf{z}_{\mathrm{r}} + \Delta\mathbf{z}_{\mathrm{r}}$ and $\mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}}) = \overline{\mathbf{z}}_{\mathrm{r}} + \Delta\overline{\mathbf{z}}_{\mathrm{r}}$. The vector $\Delta\mathbf{z}_{\mathrm{r}}$ is updated together with network parameters in a similar way as the memory variables. The intervening vector $\Delta\overline{\mathbf{z}}_{\mathrm{r}}$ is not updated. The margin parameter ϵ is set to 1 in the ShanghaiTech dataset and is set to 0.5 in the CUHK Avenue and UCSD Ped2

Method	SHTech	CUHK Avenue	UCSD Ped2
Conv-AE (Hasan et al. 2016)	-	70.2	90.0
ConvLSTM-AE (Luo et al. 2017)	-	77.0	99.1
Frame-Pred. (Liu et al. 2018)	72.8	85.1	95.4
MemAE (Gong et al. 2019)	71.2	83.3	94.1
Object-Centric (Ionescu et al. 2019)	**78.7**	87.4	94.3
MNAD-P (Park et al. 2020)	70.5	88.5	97.0
Any-Shot (Doshi and Yilmaz 2020)	71.6	86.4	97.8
HF2-VAD (Liu et al. 2021b)	76.2	91.1	99.3
STCEN (Hao et al. 2022)	73.8	86.6	96.9
BDPN (Chen et al. 2022)	78.1	90.3	98.3
Ours	78.6	**91.5**	**99.4**

Table 1: Comparisons of frame-level performance (AUROC ↑, %) under Setting-A (big data scenarios). A higher AUROC value indicates a better performance.

datasets.

We use PyTorch (Paszke et al. 2017) to train our model and adopt the Adam optimizer (Kingma and Ba 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize it. The batch size, epoch number and initialized learning rate are set to $(128, 80, 1e\text{-}4)$ and $(128, 40, 8e\text{-}5)$ for training the causal generative model and finetuning the predictor, respectively. The learning rate is decayed by 0.8 after every 40 epochs.

Evaluation Metric. We compute frame-level anomaly scores, and plot the receiver operating characteristic (ROC) curve according to the scores. The area under the receiver operating characteristic (AUROC) is computed for evaluation. A higher AUROC indicates a better performance. We calculate the AUROC based on all the frames in each dataset rather than the averaging video-level AUROC.

We train and test our model under two settings, big and small data scenarios. (1) Setting-A (big data scenarios): the model is trained and tested on the specific target dataset. Most existing methods follow this setting, where all training samples are used for training. (2) Setting-B (small data scenarios): the model is trained on a sub-dataset and is tested on the whole test dataset. We perform evaluations under this setting to simulate the use of anomaly detection methods in an open world, where only a small portion of training samples is used for training. We randomly select 10% of the training samples to form sub-datasets of the three datasets, and all samples in each sub-dataset come from one scene in each dataset.

Results. Table 1 reports the AUROC performance of our method compared with state-of-the-art methods under Setting-A. We do not compare our method with existing top-performing methods (Georgescu et al. 2021; Acsintoae et al. 2022) because they use extra annotated anomalies while we focus on deep unsupervised anomaly detection. The performance of all compared methods is taken from their original paper or from the work of (Georgescu et al. 2021). From Table 1, we can see that our method outperforms the baseline method HF2-VAD (Liu et al. 2021b) on the three datasets, gaining improvements of 2.4%, 0.4% and 0.1% in terms of the AUROC evaluation. Compared with the method Object-Centric (Ionescu et al. 2019), our method has achieved a slightly worse result on the ShanghaiTech dataset. The probable reason is that our causal generative model focuses on visual factors captured from RGB frames without considering the optical-flow images used in Object-Centric. Nevertheless, our method outperforms Object-Centric on the other two datasets with large AUROC improvements of 4.1% and 5.1%, which verifies the effectiveness of our method.

Method	SHTech	CUHK Avenue	UCSD Ped2
Conv-AE (Hasan et al. 2016)	50.7	69.4	67.0
MemAE (Gong et al. 2019)	51.5	81.5	74.3
Object-Centric (Ionescu et al. 2019)	72.7	80.6	92.9
HF2-VAD (Liu et al. 2021b)	73.8	82.2	98.4
Ours	**76.4**	**87.8**	**99.2**

Table 2: Comparisons of frame-level performance (AUROC ↑, %) under Setting-B (small data scenarios). A higher AUROC value indicates a better performance.

Method		SHTech	CUHK Avenue
B0	Baseline (HF2-VAD)	73.8	82.2
B1	+VAE+CSG	74.2	85.3
B2	+MVAE+CSG	72.6	83.5
B3	+VAE-T+CSG	75.0	85.0
B4	+MVAE-T+CSG-CE	74.1	85.8
B5	+MVAE-T+CSG2	75.2	85.9
Ours	+MVAE-T+CSG	**76.4**	**87.8**

Table 3: Ablation study performance (AUROC ↑, %) on the ShanghaiTech and CUHK Avenue datasets.

Table 2 reports the AUROC performance under Setting-B. We re-train models of all compared methods for evaluation in the small data scenarios. In Table 2, our method obtains significant improvements of 2.6%, 5.6% and 0.8% compared with the state-of-the-art baseline method HF2-VAD (Liu et al. 2021b) on the three datasets. The performance of all compared methods drops significantly from big data scenarios in Table 1 to small data scenarios in Table 2. Differently, our method achieves comparable AUROC results of 76.4% and 99.2% compared with the results of 76.2% and 99.3% of the baseline method HF2-VAD trained on the full ShanghaiTech and UCSD Ped2 datasets. On the CUHK Avenue dataset, our method achieves 96% of the AUROC performance of the baseline method ($96\% = 87.8\%/91.1\%$). This demonstrates that learning event-relevant factors can work stably in small data scenarios.

We show two qualitative results in Figure 4. Blue windows show ground-truth labels of anomalies, and yellow/red curves represent the anomaly scores computed by the pretrained/finetuned predictor under Setting-B (small data scenarios). The scores are normalized through the min-max normalization. The finetuned predictor's scores (red) match better with the anomaly annotations even under the interferences of event-irrelevant factors of the occlusion and ob-

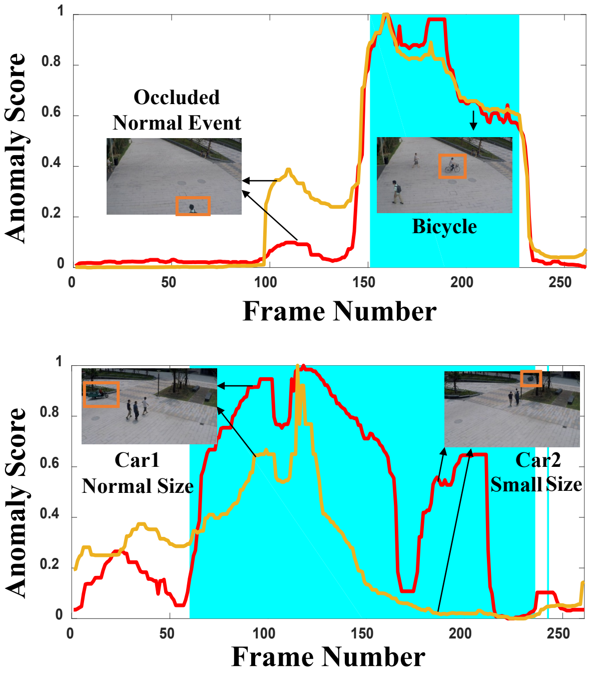

Figure 4: Qualitative results on the ShanghaiTech datasets. Frames in white/blue windows are the ground-truth normal/anomaly events. Yellow/red curves represent the anomaly scores computed by the pre-trained/finetuned predictor. Anomaly scores in the finetuned predictor match well with the ground-truth annotations under the interferences of event-irrelevant factors of the occlusion and object scale variation.

ject scale variation, indicating good discrimination of our method. More Visualization results of event-irrelevant factors can be found in *supplementary materials*.

Ablation Study. We conduct an ablation study on the ShanghaiTech and CUHK Avenue datasets to compare the contributions of different components in our method. The experimental results under Setting-B are shown in Table 3. Our baseline method is the pre-trained HF2-VAD (Liu et al. 2021b). "+VAE", "+MVAE", "+VAE-T" and "+MVAE-T" denote that we use the standard VAE shown in Figure 3a, the memory-augmented VAE shown in Figure 3b, the standard VAE with two representations $\mathbf{z}_{\mathrm{r}}$ and $\overline{\mathbf{z}}_{\mathrm{r}}$, and the memory-augmented VAE with two representations $\mathbf{z}_{\mathrm{r}}$ and $\overline{\mathbf{z}}_{\mathrm{r}}$ shown in Figure 3c, respectively, as the causal generative model. "+CSG" means that we use the counterfactual sample generation for finetuning the predictor. When the causal generative model only has one representation (i.e., "+VAE", "+MVAE"), the L2-norms containing the interventions $\mathrm{do}(\overline{\mathbf{z}}_{\mathrm{r}})$ in Eq. (12) and Eq. (13) are set to 0 during finetuning. "-CE" indicates removing the loss function $\mathcal{L}^{\mathrm{CE}}$. "+CSG2" means that we replace the intervention type $\mathrm{do}(\mathbf{z}) = \mathbf{z} + \Delta(\mathbf{z})$ with $\mathrm{do}(\mathbf{z}) = \mathrm{dropout}(\mathbf{z}, p)$ in the counterfactual sample generation, where the dropout probability p is computed by using a two-layer fully-connected network $p = \mathrm{softmax}(\mathrm{MLP}(\Delta(\mathbf{z})))$.

Setting	Time (h)	Speed (fps)	AUROC↑ (%)
Setting-A	48.0	270.3	78.6
Setting-B	6.1	270.3	76.4

Table 4: The training time (Time), inference speed (Speed) and AUROC values of our method on the ShanghaiTech dataset under Setting-A and Setting-B.

From Table 3, we can see that: (1) The causal relationships established by the memory $\mathbf{M}$, representation $\mathbf{z}_{\mathrm{r}}$ and representation $\overline{\mathbf{z}}_{\mathrm{r}}$ benefit learning event-relevant factors, because when discarding any one of them (B1, B2 and B3), we see a drop of the AUROC performance. (2) When removing the loss function $\mathcal{L}^{\mathrm{CE}}$ for capturing event-relevant factors (B4), a decrease of the AUROC from 76.4% and 87.8% to 74.1% and 85.8% is obtained on the two datasets. The performance drops show that maximizing the information flow can increase the causal influence. (3) The intervention type $\mathrm{do}(\mathbf{z}) = \mathbf{z} + \Delta(\mathbf{z})$ is more beneficial for generating counterfactual samples for finetuning the predictor compared with the intervention type in "+CSG2" (B5).

Computation Efficiency. As shown in Table 4, we list the training time, inference speed and AUROC values on the ShanghaiTech dataset under Setting-A and Setting-B. Our method achieves a comparable result in less time under setting-B compared with the result under setting-A. Our method also achieves the real-time inference speed of video anomaly detection. The experiment results are obtained on a single NVIDIA RTX3090 GPU and an Intel i9-10900X CPU, and we do not consider the pre-processing time of the object detection and optical flow estimation.

Conclusion

We have presented a novel method that can eliminate the interferences from event-irrelevant factors on anomaly predictions by explicitly learning event-relevant factors in videos. We design a variational auto-encoder with two independent representations as the causal generative model. The causal generative can separately model the event-relevant factors and event-irrelevant ones. A causal objective function is used to optimize the causal generative model, which can maximize the causal influence of event-relevant factors on anomaly predictions. We further develop a counterfactual learning strategy that can help the anomaly predictions benefit from the learned event-relevant factors without annotated anomalies by generating counterfactual samples as pseudo-anomalies. Experimental results in different settings demonstrate the effectiveness of our method.

A future extension of our work is to introduce an extra explainer with generic knowledge to explain the real causes of anomalies from the learned event-relevant factors, therefore improving the interpretability of our model.

Acknowledgments

This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 62172041 and No. 62176021.

References

Abati, D.; Porrello, A.; Calderara, S.; and Cucchiara, R. 2019. Latent space autoregression for novelty detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 481–490.

Acsintoae, A.; Florescu, A.; Georgescu, M.-I.; Mare, T.; Sumedrea, P.; Ionescu, R. T.; Khan, F. S.; and Shah, M. 2022. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 20143–20153.

Ay, N.; and Polani, D. 2008. Information Flows in Causal Networks. *Adv. Complex Syst.*, 11(1): 17–41.

Chandola, V.; Banerjee, A.; Kumar, V.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3): 15:1–15:58.

Chang, Y.; Tu, Z.; Xie, W.; Luo, B.; Zhang, S.; Sui, H.; and Yuan, J. 2022. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognit.*, 122: 108213.

Chen, C.; Xie, Y.; Lin, S.; Yao, A.; Jiang, G.; Zhang, W.; Qu, Y.; Qiao, R.; Ren, B.; and Ma, L. 2022. Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection. In *Proc. Conf. Artif. Intell.*

Ding, W.; Lin, H.; Li, B.; and Zhao, D. 2022. CausalAF: Causal Autoregressive Flow for Goal-Directed Safety-Critical Scenes Generation. In *Proc. Conf. Robot Learn.*

Doshi, K.; and Yilmaz, Y. 2020. Any-shot sequential anomaly detection in surveillance videos. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit Workshops*, 934–935.

Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11): 665–673.

Georgescu, M.-I.; Ionescu, R. T.; Khan, F. S.; Popescu, M.; and Shah, M. 2021. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Trans. Pattern Anal. Mach. Intell.*

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and van den Hengel, A. 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, 1705–1714.

Haldar, S.; John, P. G.; and Saha, D. 2021. Reliable Counterfactual Explanations for Autoencoder based Anomalies. In Haritsa, J. R.; Roy, S.; Gupta, M.; Mehrotra, S.; Srinivasan, B. V.; and Simmhan, Y., eds., *8th ACM IKDD CODS and 26th COMAD*, 83–91.

Hao, Y.; Li, J.; Wang, N.; Wang, X.; and Gao, X. 2022. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognit.*, 121: 108232.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning Temporal Regularity in Video Sequences. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 733–742.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 770–778.

Holzinger, A.; Dehmer, M.; Emmert-Streib, F.; Cucchiara, R.; Augenstein, I.; Ser, J. D.; Samek, W.; Jurisica, I.; and Rodríguez, N. D. 2022. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Inf. Fusion*, 79: 263–278.

Hou, J.; Zhang, Y.; Zhong, Q.; Xie, D.; Pu, S.; and Zhou, H. 2021. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, 8791–8800.

Huang, C.; Wu, Z.; Wen, J.; Xu, Y.; Jiang, Q.; and Wang, Y. 2021. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Trans. Ind. Informatics*.

Ionescu, R. T.; Khan, F. S.; Georgescu, M.; and Shao, L. 2019. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 7842–7851.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *Proc. Int. Conf. Learn. Res.*

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *Proc. Int. Conf. Learn. Repren.*

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, N.; Chang, F.; and Liu, C. 2021. Spatial-Temporal Cascade Autoencoder for Video Anomaly Detection in Crowded Scenes. *IEEE Trans. Multim.*, 23: 203–215.

Liu, C.; Sun, X.; Wang, J.; Tang, H.; Li, T.; Qin, T.; Chen, W.; and Liu, T.-Y. 2021a. Learning causal semantic representation for out-of-distribution prediction. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 6536–6545.

Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021b. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *Proc. IEEE Int. Conf. Comput. Vis.*, 13588–13597.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2720–2727.

Luo, W.; Liu, W.; and Gao, S. 2017. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In *Proc. IEEE Int. Conf. Comput. Vis.*, 341–349.

Luo, W.; Liu, W.; Gao, S.; and Gao, S. 2017. Remembering history with convolutional lstm for anomaly detection. In *IEEE Int. Conf. Multimedia Expo*, 439–444. IEEE.

Luo, W.; Liu, W.; Lian, D.; and Gao, S. 2021. Future Frame Prediction Network for Video Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *Proc. Int. Conf. Learn. Res.*

Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2009. Anomaly detection in crowded scenes. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 1975–1981.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In Hildebrandt, M.; Castillo, C.; Celis, L. E.; Ruggieri, S.; Taylor, L.; and Zanfir-Fortuna, G., eds., *Proc. Conf. Fairness Account. Trans.*, 607–617.

O'Shaughnessy, M. R.; Canal, G.; Connor, M.; Rozell, C.; and Davenport, M. A. 2020. Generative causal explanations of black-box classifiers. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Proc. Adv. Neural Inf. Process. Syst.*

Pang, G.; Shen, C.; Cao, L.; and van den Hengel, A. 2021. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.*, 54(2): 38:1–38:38.

Park, H.; Noh, J.; Ham, B.; and Ham, B. 2020. Learning Memory-Guided Normality for Anomaly Detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 14360–14369.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *Proc. Adv. Neural Inf. Process. Syst. Workshops*, 8024–8035.

Pearl, J. 2009. *Causality*. Cambridge university press.

Sabokrou, M.; Khalooei, M.; Fathy, M.; and Adeli, E. 2018. Adversarially Learned One-Class Classifier for Novelty Detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.*, 3379–3388.

Song, H.; Sun, C.; Wu, X.; Chen, M.; and Jia, Y. 2020. Learning Normal Patterns via Adversarial Attention-Based Autoencoder for Abnormal Event Detection in Videos. *IEEE Trans. Multim.*, 22(8): 2138–2148.

Sun, C.; Jia, Y.; Song, H.; and Wu, Y. 2021. Adversarial 3D Convolutional Auto-Encoder for Abnormal Event Detection in Videos. *IEEE Trans. Multim.*, 23: 3292–3305.

Xu, D.; Yan, Y.; Ricci, E.; and Sebe, N. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.*, 156: 117–127.

Yao, Y.; Wang, X.; Xu, M.; Pu, Z.; Wang, Y.; Atkins, E.; and Crandall, D. 2022. DoTA: unsupervised detection of traffic anomaly in driving videos. *IEEE Trans. Pattern Anal. Mach. Intell.*

Yuan, B.; Zhao, D.; Shao, S.; Yuan, Z.; and Wang, C. 2022. Birds of a Feather Flock Together: Category-Divergence Guidance for Domain Adaptive Segmentation. *IEEE Trans. Image Process.*, 31: 2878–2892.

Zhang, X.; Wong, Y.; Wu, X.; Lu, J.; Kankanhalli, M.; Li, X.; and Geng, W. 2021. Learning Causal Representation for Training Cross-Domain Pose Estimator via Generative Interventions. In *Proc. IEEE Int. Conf. Comput. Vis.*, 11270–11280.