Learning Causality-inspired Representation Consistency for Video Anomaly Detection

Yang Liu¹, Zhaoyang Xia^{2†}, Mengyang Zhao^{1†}, Donglai Wei^{1†}, Yuzheng Wang¹, Siao Liu¹, Bobo Ju¹, Gaoyun Fang¹, Jing Liu^{1*}, Liang Song^{1*}

¹Academy for Engineering & Technology, Fudan University, Shanghai, China

²Shanghai AI Laboratory, Shanghai China

{yang_liu20, jingliu19, songl}@fudan.edu.cn

Abstract—Video anomaly detection is an essential vet challenging task in the multimedia community, with promising applications in smart cities and secure communities. Existing methods attempt to learn abstract representations of regular events with statistical dependence to model the endogenous normality, which discriminates anomalies by measuring the deviations to the learned distribution. However, conventional representation learning is only a crude description of video normality and lacks an exploration of its underlying causality. The learned statistical dependence is unreliable for diverse regular events in the real world and may cause high false alarms due to overgeneralization. Inspired by causal representation learning, we think that there exists a causal variable capable of adequately representing the general patterns of regular events in which anomalies will present significant variations. Therefore, we design a causality-inspired representation consistency (CRC) framework to implicitly learn the unobservable causal variables of normality directly from available normal videos and detect abnormal events with the learned representation consistency. Extensive experiments show that the causality-inspired normality is robust to regular events with label-independent shifts, and the proposed CRC framework can quickly and accurately detect various complicated anomalies from real-world surveillance videos.

Index Terms—Causal Representation Learning, Video Anomaly Detection, Unsupervised Learning, Normality Learning, Deep Clustering

I. INTRODUCTION

Video anomaly detection (VAD) aims to automatically analyze the spatial-temporal patterns and contactlessly detect anomalous events of concern (e.g., traffic accidents, violent acts, and illegal operations) from surveillance videos [1], [2], which has promising applications in emerging areas such as traffic management [3], [4], security protection [5], [6], and intelligent manufacturing [7], [8]. However, *anomaly* is a vague concept, making anomalous events unbounded and difficult to predefine. Therefore, collecting all possible positive samples is impractical, making VAD remain challenging in the multimedia [9], [10], [11], [12], [13], [14], [15], [16], [17] and pattern recognition [18], [19], [20], [21], [22], [23], [24], [25], [26] communities.

To avoid the cost of collecting and labeling anomalous events, existing VAD methods typically use only normal videos to train a deep generative model (e.g., autoencoder [27], [28], [29], generative adversarial network [30], and transformer [31]) to perform reconstruction or prediction tasks, which learns the distribution dependence of regular events in an unsupervised manner and treating out-of-distribution samples as anomalies. They assume that models learned on negative samples cannot characterize unseen positive ones, leading to significant deviations from the learned normality. As shown in Figure 1(a), these methods [32], [33], [34] expect to map regular events (green cubes) to a hyperspace, while uncharacterizable anomalous events (red cubes) will fall outside. Benefiting from deep representation learning (DeepRL) [35], [36], [37], [38], [39], unsupervised VAD has achieved remarkable progress in recent years.

However, the long-overlooked problem is that regular events are also diverse. Besides, due to the high dimensionality of real-world videos and the complexity of target-scene interactions, regular events contain both shared and private semantics, i.e., prototypical and personalized features [40]. In addition, positive and negative samples captured from the same scene usually share the most appearance context, making it difficult to filter these task-irrelevant semantics under an unsupervised setting. Ultimately, an unaffordable consequence is that the learned model may represent abnormal events well due to the overgeneralization of deep neural networks [41], leading to missed detections, as sample \mathcal{V}_a^1 in Figure 1(a). Moreover, for regular events with label-independent distribution offsets (e.g., the color of pedestrians' clothes and their walking posture in crowd anomaly detection), existing unsupervised methods cannot resist such random disturbances due to insufficient robustness, resulting in high false alarms, as sample \mathcal{V}_n^1 in Figure 1(a). Therefore, existing methods are limited by DeepRL and only establish the crude statistical dependence of the normal distribution, making the learned normality unable to cope with complicated anomalies and normal events with unseen bias.

Inspired by causal representation learning (CausalRL) [42], we attempt to learn task-specific representations that contain potential causal mechanisms capable of revealing the intrinsic properties of regular events, which can mitigate the negative impact of event diversity and random label-independent bias in unsupervised normality learning [43]. In this regard, we construct a structural causal model (SCM) shown in Figure 1(b),

[†]Equal contribution.

^{*}Corresponding authors.



Fig. 1. Motivation for addressing unsupervised VAD from a causality perspective. (a) illustrates the expectation (left) and the actuality (right) of the existing methods: they expect to train a characterizer to learn the pattern boundaries for regular events. However, the learned boundaries cannot effectively detect weak anomalies and unseen normal events. The structural causal model in (b) states the shortcomings of existing methods: they try to establish the statistical association (dashed arrow) of observable normal videos X with labels Y, lacking effective exploration of causal factors. The sparse mechanism shift hypothesis in (c) suggests that label-independent domain shifts of diverse normal events ($n \rightarrow n'$) have a limited and local impact on the learned causality (marked by \mathfrak{K}). In contrast, anomalous events ($n \rightarrow a$) cause a full collapse of consistency (marked by \mathfrak{G}) learned on regular events. Inspired by the above observations, we expect to construct robust and efficient VAD models with CausalRL.

where X and Y denote observable normal videos and labels, respectively. According to the common cause principle, we consider that there are causal factors in X that can fully describe normality. We attempt to learn these unobservable causal factors with the label consistency between shared and private features of X. In addition, the sparse mechanism shift hypothesis point out that diverse normal events with significant distribution differences only vary locally in the high-level causality space. This hypothesis suggests that learning causality-inspired normality may enable the model to correctly infer shifted regular events and reduce false alarm rates. To this end, we propose an end-to-end causality-inspired representation consistency (CRC) framework to mining causal variable for unsupervised video anomaly detection. In the training phase, we optimize the CRC framework using the causal independence and the consistency of the multi-view representations for regular events. While testing, the causalityinspired characterizer learned on negative samples will not work for anomalies, making the anomalous events show significant differences in causal consistency, as shown in Figure 1(c).

Specifically, the CRC framework first utilizes a DeepRLbased feature extractor to obtain original spatial-temporal patterns containing causal and non-causal variables. Then, an iteratively updated memory pool [44] is used to record the general pattern of regular events. Distinguishing from existing methods that use features retrieved from the memory pool as representations for anomaly discrimination, we introduce a prototype decomposer to split the shared and private features [40]. The shared and private features come from the same batch of training samples, so both are representations of regular events by nature. Therefore, we use causal factorization to reduce them into a set of causal factors and capture the intrinsic causal mechanism. Finally, the features represented by causal factors are fed into a clustering algorithm [45] to obtain compact task-specific causal representations. The main contributions of this paper are summarized as follows:

- We address the unsupervised video anomaly detection from a causality perspective and propose a causalityinspired representation consistency framework to learn video normality and detect anomalous events by consistency.
- We design prototype decomposer and causal factorization to mine causal variable directly from the videos and use causal representations to characterize normal events.
- Our method can cope with label-independent shifts and amplify the deviation of subtle anomalies with causal consistency. Experimental results prove the effectiveness of CRC, which achieves superior performance on benchmarks.

II. RELATED WORK

A. Video Anomaly Detection

VAD has been extensively studied for years due to its potential applications in emerging fields such as smart cities and secure communities, and various routes have been derived with the development of DeepRL. Among them, unsupervised methods [27], [46] follow the open-world assumptions without predefining and collecting anomalies and avoiding annotation costs and data imbalance problems, becoming the preferred solutions. Early unsupervised methods [47], [48] treat VAD as a one-class classification task and use OC-SVM, OC-NN, or deep clustering to determine the boundaries of manual features, which are prone to the curse of dimensionality when dealing with real videos. Benefiting from the rise of deep generative models, researchers used autoencoders [45] and generative adversarial networks [49], [30] to extract spatialtemporal representations and introduce proxy tasks to learn the prototypical pattern of normal videos, i.e., normality. Specifically, such methods assume that generative models trained with normal videos are only effective in representing regular events. Thus, positive samples will experience significant performance degradation on the proxy task during the downstream anomaly detection phase. For example, Hasan et al. [27] propose a 2D convolutional autoencoder to reconstruct input sequences and use the reconstruction error to compute anomaly scores. Liu et al. [49] pioneered a video prediction framework to learn video normality and measure the degree of anomalies with the appearance and motion prediction errors. Following efforts lie in structure modifications (e.g., using dual-stream networks to learn appearance and motion normality separately [45], [50]) and proxy task stacking [51].

Recently, the intrinsic semantics consistency between different dimensions [46] or regions [52] is considered feasible for video normality learning. Drawing that memory networks [41], [44] can store prototype patterns of training samples, Cai et al. [46] construct two memory-enhanced autoencoders to learn appearance-motion consistency by learning relationships between RGB images and optical flow. They argue that the consistency learned on normal samples holds only for regular events. In addition, emerging object-level schemes [53], [52], [54] attempt to explore normal target-scene semantics interactions and discriminate anomalies accordingly. However, due to the high dimensionality and complexity of videos, both normal and abnormal events are diverse. Conventional DeepRL struggles to obtain representations with sufficient discrimination to describe diverse regular events. Studies show that such methods may miss-detect positive samples due to overgeneralization or fail to effectively reason about unseen negative samples due to insufficient representation ability. In contrast, our proposed CRC framework learns video normality with CausalRL, aiming to balance representation and generalization with causal mechanisms. Although causality has been proven practicable in numerous image processing tasks [55], the one-class classification setting that only negative samples are available for training in VAD makes designing reasonable causal interventions and factorization schemes for casual variable extremely challenging.

B. Causal Representation Learning

Conventional deep representation learning has flourished in recent years with the emergence of large-scale multisource datasets. However, DeepRL only learns the statistical independence between training samples and given labels, subject to the independent and identically distributed (i.i.d) assumption [42]. Causal representation learning considers statistical independence as a crude description of the physical world, which cannot perform correct inference under distribution changes and intervention conditions. Currently, causalitydriven representation models achieve leading performance in various applications such as domain generalization [55] and online recommendation systems, demonstrating great potential for learning robust representations and reusable mechanisms, which are also the key to high-performance VAED. On the one hand, real-world normal videos suffer from unpredictable bias, i.e., regular events contain personalized semantics that does not constitute anomalies. DeepRL-based characterizer can hardly accommodate such private features and may misclassify unseen regular events as anomalies. On the other hand, the unbounded nature of anomalous events makes it inevitable that their patterns intersect with the normal distribution. To summarize, robust anomaly detectors need to learn representations containing essential factors that adequately describe normality, which motivates us to learn normality with CausalRL.

III. METHODOLOGY

A. VAD in Causality Perspective

As discussed in Sec. I, we consider that the spatial-temporal features extracted by DeepRL contain both causal variable that determine the normality and label-independent non-causal variable. Existing VAD methods typically use DeepRL to learn the statistical independence of normal videos with redundant and non-helpful representations. In response, we design a structural causal model (SCM) shown in Figure 1(b) to formulate unsupervised VAD and guide our CRC framework to learn robust knowledge beyond the available training data. The following common cause principle describes the connection between statistical dependence and causality:

§1 **Common cause principle:** If two observables X and Y are statistically dependent, then there exists a variable Z that causally influences both and explains all the dependence in the sense of making them independent when conditioned on Z.

For the unsupervised VAD task, we set X and Y to denote observable regular events and normality (label 0), respectively. The causal variable Z, which has a causal effect on both the original data distribution and normality learning, is not directly observable, which may be usual targets (as opposed to unexpected objects in appearance anomalies) or regular object-scene interactions (as opposed to violations in motion anomalies). In this regard, we attempt to use the directed acyclic graph in Figure 1(b) to implicitly learn a set of causal factors $\{z_1, \dots, z_n\}$ with internal consistency for characterizing various types of normal events, as follows:

$$\boldsymbol{X} := f\left(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{P}\right), \boldsymbol{Z} \bot \boldsymbol{U} \bot \boldsymbol{P}, \tag{1}$$

$$\boldsymbol{Y} := h\left(\boldsymbol{Z}, \boldsymbol{P}\right) = h\left(g(\boldsymbol{X}), \boldsymbol{P}'\right), \boldsymbol{P} \bot \boldsymbol{P}', \qquad (2)$$

where U denotes the non-causal variable only affecting X, e.g., domain-specific information that contributes nothing to normality learning. P and P' denotes joint-independent unexplained perturbation noise. $f(\cdot, \cdot, \cdot)$, $h(\cdot, \cdot)$ and $g(\cdot)$ are regarded as unknown structural functions with causal mechanisms. According to §1 and the invariant causal mechanism, for any distribution $P(X, Y) \in \mathcal{P}$, when the causal variable Zis given, then a general conditional distribution P(Y|Z) must exist. Thus, representations that imply causality are essential for learning normality robust to diverse patterns of normal events.

However, it is impractical to observe causal variable from unstructured videos, i.e., there is no available prior to guide us to define causal representations dissuasively. Following the consensus in CausalRL, we attempt to encourage the model to learn a set of orthogonal causal factors with the following principle [56], [42]:

§2 Independent causal mechanism: The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.



Fig. 2. Overview of the proposed causally-inspired representation consistency (CRC) framework. In the training phase, CRC extracts the spatial-temporal features F of b normal sequences and stores the prototypes in the memory pool \mathcal{M} , and then uses the prototype decomposer to strip the private and shared features $\{F_p, F_s\}$, which are fed to the causally-inspired characterizer (CiC) to learn the causal variables. Inspired by §1 and §2, we exploit the independence of the causal variables to compute the correlation matrices $C_{1,2,3}$ and optimize CiC. A clustering is introduced to obtain compact task-specific causal representations.

§2 inspires us to find the unobservable causal factors $\{z_1, \dots, z_n\}$ with the separate intervening nature of Z and their independence. Corresponding to the causal factorization of the VAD representations, we know that: (1) Z that fully generalize normality are separated from U, i.e., interventions on U do not change Z and Y. (2) The causal factors $\{z_1, \dots, z_n\}$ is jointly independent, and mechanism $P(z_i|PA_i)$ does no inluence or transfer information with $P(z_j|PA_j)$ if $j \neq i$, where PA denotes the causal parents. (3) The learned task-specific causal representations are causally sufficient for normality learning to explain all statistical independence between X and Y. Therefore, we can factorize the joint distribution of causal factors into conditional as follows:

$$P(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n) = \prod_{i=1}^n P(\boldsymbol{z}_i \mid PA_i).$$
(3)

In this work, we consider both the shared and private features of regular events as a phenotypic form of normality. Due to the diversity of video events, conventional representation learning is challenging to outline the distribution of these multi-view features that point to the same causal variable. It is feasible to learn causal representations implicitly through independence, which is consistent with the following hypothesis [42]:

§3 **Sparse mechanism shift:** *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization, that is, they should usually not affect all factors simultaneously.*

which motivates us to learn stable causal variable that can respond sensitively to anomalies through intrinsic consistency while resisting label-independent shifts. For implementation, we construct a prototype decomposer to decompose the original representations into private and shared features and train the causality-inspired characterizer to further represent these features with the same causal factors. Besides, we utilize clustering and similarity constraints to explore the consistency and obtain task-specific representations for unsupervised video anomaly detection.

B. Prototype Learning and Decomposition

Inspired by the ability of the memory [44] to record the prototype of normal events and constrain the overgeneralization of the DeepRL, we use memory to construct the prototype learning module and obtain the shared and private features. As shown in Figure 2, the memory update process $\mathcal{M}_t \to \mathcal{M}_{t+1}$ illustrate the recording of the general patterns of the spatial-temporal features $F \in \mathbb{R}^{H \times W \times C}$. Specifically, the memory pool is a two-dimensional matrix, denoted as $M \in \mathbb{R}^{C \times N}$, where N denotes the number of memory entries and determines the information capacity of \mathcal{M}_t . The memory pool contains no learnable parameters but updates its memory entries to record normality through the write operation with M serving as query Q_M , as follows:

$$\mathcal{M}_{t+1} = l_2 \left(\boldsymbol{M} + \boldsymbol{V}_F \boldsymbol{\Psi} \left(\frac{\boldsymbol{K}_F^T \boldsymbol{Q}_M}{\sqrt{C}} \right) \right), \qquad (4)$$
$$\boldsymbol{V}_F = \boldsymbol{K}_F = e(\boldsymbol{F}) \in \mathbb{R}^{C \times \hat{N}},$$

where $e(\cdot)$ denotes expanding F along the spatial dimension so that $\hat{N} = H \times W$. $l_2(\cdot)$ is L2-norm to keep the data scale of \mathcal{M}_t and \mathcal{M}_{t+1} consistent, and Ψ denotes softmax. In contrast, the read operation aims to reconstruct F as prototype F' with expanded F serving as query \mathcal{Q}_F , as shown in Figure 2:

$$F' = V_M \Psi\left(\frac{K_M^T e(F)}{\sqrt{C}}\right), V_M = K_M = M \in \mathbb{R}^{C \times N}.$$
 (5)

Existing work [44], [57], [58] assumes that F' can effectively detect anomaly differences, i.e., the anomalous event will lose its own patterns and thus encounter significant errors in proxy tasks. However, over-strong memory may make the well-trained model unable to reason about normal events with shifts. Learning the distribution of prototype features is insufficient for describing diverse regular events and discriminating complex anomalies. Therefore, We use CausalRL to further explore the intrinsic connection and understand the inherent differences between positive and negative samples with causal consistency.

The raw features F contain shared prototypical semantics and unique personalized semantics, i.e., shared and private features, denoted as $\{F_s, F_p\}$. As stated in Sec. I, both F_s and F_p are statistically associated with label 1. Referring to sparse representation learning [40], we design a SE-like [59] process to strip F_s and F_p from F and F'. The details are shown in Figure 2. First, F and F' are average and max pooled, denoted as $\{f_{avg}, f'_{avg}, f_{max}, f'_{max}\} \in \mathbb{R}^C$, which are then mapped to the difference scores $\{\alpha, \beta\}$ by two multi-layer perception (MLP) with learnable parameters $\{\theta_1, \theta_2\}$, as follows:

$$\alpha = \text{MLP}(\boldsymbol{f}_{\text{avg}} - \boldsymbol{f}'_{\text{avg}}; \theta_1), \beta = \text{MLP}(\boldsymbol{f}_{\text{max}} - \boldsymbol{f}'_{\text{max}}; \theta_2).$$
(6)

Finally, we use α and β to quantitatively filter the prototypical semantics in F, as follows:

$$\boldsymbol{F}_{p} = \frac{\alpha + \beta}{2} \circledast \boldsymbol{F}, \boldsymbol{F}_{s} = \left(1 - \frac{\alpha + \beta}{2}\right) \circledast \boldsymbol{F}', \qquad (7)$$

where \circledast denotes the channel-wise multiplication.

C. Representation Consistency Learning

Inspired by principles §1 and §2, we know that there exist jointly independent causal factors capable of fully generalizing the statistical dependence from low-level normal videos to high-level normality. Furthermore, §3 indicates that the individualized features of normal events have a limited effect on the causal factors and their consistency. Therefore, we construct a causality-inspired characterizer (CiC) to learn unobservable causal variable and model the intrinsic consistency of normal events. Specifically, the spatial-temporal features of b video clips from the same batch are decomposed and fed into CiC, which maps their shared and private features into causal representations: $R = \{r_1; r_2; \cdots; r_b\} =$ $\begin{array}{l} \operatorname{CiC}(\boldsymbol{F}_s^1, \boldsymbol{F}_s^2, \cdots, \boldsymbol{F}_s^b) \in \mathbb{R}^{b \times n} \text{ and } \tilde{\boldsymbol{R}} = \{\tilde{\boldsymbol{r}}_1; \tilde{\boldsymbol{r}}_2; \cdots; \tilde{\boldsymbol{r}}_b\} = \\ \operatorname{CiC}(\boldsymbol{F}_p^1, \boldsymbol{F}_p^2, \cdots, \boldsymbol{F}_p^b) \in \mathbb{R}^{b \times n}. \text{ In practice, } n \ll H \times W \times C. \end{array}$ Unsupervised VAD attempts to learn the normality using only regular events so that r_i and \tilde{r}_i point to the same label. Then, the causal variables should remain causally invariant to the so-called decomposition intervention, i.e., the causal representations of shared and private features should remain close in the causal factor dimension:

$$\max \frac{1}{n} \sum_{i=1}^{n} \frac{f_i \tilde{f}_i}{\parallel f_i \parallel \parallel \tilde{f}_i \parallel},\tag{8}$$

where f_i and \tilde{f}_i denote the *i*-th column of R and \tilde{R} , respectively.

By maximizing the similarity of the same *n* causal factors on shared and private features, we can encourage CiC to learn causal factors that can strip label-independent non-causal variable from redundant deep spatial-temporal features. In addition, to ensure that the causal factors are jointly independent, we construct three correlation matrices on $\mathbf{R} \to \tilde{\mathbf{R}}$, $\mathbf{R} \to \mathbf{R}$, and $\tilde{\mathbf{R}} \to \tilde{\mathbf{R}}$, denoted as C_1 , C_2 , and C_3 , as shown in Figure 2. similar to Eq. 8, the non-diagonal elements of C_1 are also the cosine similarity between the corresponding columns of \mathbf{R} and $\tilde{\mathbf{R}}$. In contrast, C_2 and C_3 present the similarity within \mathbf{R} and $\tilde{\mathbf{R}}$: $C_2(i,j) = \frac{f_i f_j}{\|f_i\| \|f_j\|}$ and $C_3(i,j) = \frac{\tilde{f}_i \tilde{f}_j}{\|\tilde{f}_i\| \|\tilde{f}_j\|}$. The final optimization objective is to maximize the diagonal elements of the correlation matrix C_1 (Note that the diagonal elements of C_2 and C_3 are constant 1) and minimize the non-diagonal matrices of C_1 , C_2 , and C_3 , as follows:

$$\min \lambda \parallel C_1 - I \parallel_F^2 + \parallel C_2 - I \parallel_F^2 + \parallel C_3 - I \parallel_F^2, \quad (9)$$

where λ is a trade-off hyper-parameter, ans I denotes the identity matrix. In this way, we can ensure that the causal factors are jointly independent and invariant to the decomposition intervention. According to §3, normal events with shifts are only locally different regarding causal representations. Therefore, we follow [45] to introduce clustering to obtain a tighter causal representation R and further enhance the model to discriminate normal events with clustering effects. In addition, we introduce memory separateness and compactness loss [44] to optimize the memory pool.

D. Anomaly Detection with Causal Consistency

Due to the special setting of the anomaly detection task, only negative samples are available for training, so the welltrained CRC framework is only effective in decomposing and constructing causal representation consistency for normal events. In the testing phase, we compute anomaly score s_t by measuring the deviations to the learned causal factors in terms of consistency and representations:

$$s_t = g(\parallel \boldsymbol{C}_1 - \boldsymbol{I} \parallel_F^2 \times \boldsymbol{D}), \tag{10}$$

where $g(\cdot)$ denotes the max-min normalization over all frames. *D* is the clustering distance between the causal representations of the input video clip and the cluster center. The former part of s_t , $C_1 - I$, discriminates anomalies by the consistency in the causal variable, while the latter *D* measures the distance to the normal representation.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: We conduct extensive experiments to validate the effectiveness of the proposed CRC framework on three leading unsupervised VAD benchmarks, including UCSD Ped2 [48], CUHK Avenue [60], and ShanghaiTech [61]. All training sets are normal videos collected from the real world, while anomalous events from similar scenes are only available to the test set. UCSD Ped2 [48] is a small-scale dataset containing 16 training and 12 test videos, captured from the



Fig. 3. Quantitative performance comparison. (a)-(b) show the frame-level AUC and EER of our method (marked by pentagrams) with existing methods (marked by circles) on the three datasets, respectively, while (d) presents the inference speed on the CUHK Avenue dataset. \uparrow denotes that larger values indicate better performance, while \downarrow vice versa. Best viewed in color.

university campus. As an early unsupervised VAD benchmark, its scenario is simple, with regular samples walking normally on the sidewalk, while the anomalous events include riding bikes, skateboarding, and driving. *CUHK Avenue* [60] is a large-scale single-scene VAD dataset. The training and test sets contain 21 and 16 videos with 47 anomalous events. The collectors simulated appearance-only (e.g., the person on the lawn), motion-only (e.g., loitering and wandering), and appearance-motion anomalies (e.g., papers being scattered), making CUHK Avenue more challenging. *ShanghaiTech* [61] is the most challenging benchmark, collecting 130 anomalies from 13 scenes. The data size and cross-scene nature make it difficult for unsupervised methods to learn effective deep representations to describe diverse normal events.

2) Evaluation Metrics: In the testing phase, we measure the input samples against the learned causality-inspired normality to calculate the degree of abnormality and output a continuous anomaly score in the range [0,1]. A high score indicates that the more likely the test sequence is to be anomalous. In contrast, the given labels are binary discrete, where 0 indicates normal and 1 indicates abnormal. Following previous work [49], [54], we calculate the true-positive-rate and falsepositive-rate at multiple thresholds and plot the receiver operating characteristic curve, using the area under the curve (AUC) as the primary evaluation metric to present the effectiveness of our method for anomaly detection. In addition, the equal error rate (EER) is used as a complementary metric to demonstrate the robustness of the CRC framework, which is compared with available methods. With the same implementation platform, we report the average inference speed of our method on the CUHK Avenue [48] dataset to validate its deployment potential on resource-limited terminal devices.

3) Implementation Details: We use the PyTorch [62] framework to implement the proposed method on an Nvidia 3090 GPU. The Adam [63] optimizer is used to train the model with an initial learning rate of 8×10^{-5} . The batch size *b* is set to 8. In the initial stage, we remove the clustering constraints and optimize the characterizer without clustering. After 100 epochs, we compute the clustering centers using Kmeans and update them with an alternating optimization [45]. The video frames are resized to 224×224 pixels. The feature extractor is a 5-layer convolutional encoder. The two MLPs

 TABLE I

 Results of the frame-level AUC comparison.

Type	Method	Frame-level AUC (%)				
-71-		UCSD Ped2	CUHK Avenue	ShanghaiTech		
Traditional	MPPCA [47]	69.3	-	-		
	MPPC+SFA [47]	61.3	-	-		
	MDT [48]	82.9	-	-		
	AMDN [65]	90.8	-	-		
	Unmasking [66]	82.2	80.6	-		
	MT-FRCN [67]	92.2	-	-		
	ConvAE [27]	90.0	70.2	-		
	ConvLSTM-AE [68]	88.1	77.0	-		
	AMC [51]	96.2	86.9	-		
	FFP [49]	95.4	95.4 85.1			
	MemAE [41]	94.1	83.3	71.2		
ed	AnoPCN [69]	96.8	86.2	73.6		
oas	Mem-Guided [44]	97.0	88.5	70.5		
5	AMMC-Net [46]	96.6	96.6 86.6			
nir	Clustering [45]	96.5	86.0	73.3		
Deep Lear	TAC-Net [70]	98.1 88.8		76.5		
	STD [50]	96.7	87.1	73.7		
	STC-Net [71]	96.7	87.8	73.1		
	STM-AE [30]	98.1	89.8	73.8		
	Bi-Prediction [53]	97.4	86.7	73.6		
	HSNBM [54]	95.2	91.6	76.5		
	MAAM-Net [58]	97.7	90.9	71.3		
	CRC (ResNet-18)	97.6	90.5	77.6		
	CRC (ResNet-50)	98.7	92.5	78.3		

Bold numbers indicate the best performance.

in the prototype decomposer are three-layer fully connected neural networks with sigmoid activation in the output layer. We select ResNet-18 and ResNet-50 [64] as the backbone of CiC. The trade-off hyper-parameter λ in Eq. 9 is set to 10, 18, and 20 for UCSD Ped2 [48], CUHK Avenue [60], and ShanghaiTech [61], respectively.

B. Comparisons with State-of-the-art Methods

We perform quantitative comparisons with traditional handicraft feature-based [47], [48] and DeepRL-based methods [49], [41], [30], [69], [54] to demonstrate the effectiveness of the proposed CRC framework. The results are shown in Table I and Figure 3. Among them, Table I presents the frame-level AUCs of existing unsupervised VAD methods on three mainstream datasets, and our CRC framework achieves 98.7%, 92.5%, and 78.3% AUCs on UCSD Ped2 [48], CUHK Avenue [60] and ShanghaiTech [61] datasets, respectively,

TABLE II Results of ablation study.

ID	Component			Constraint		Frame-level AUC (%)			
	\mathcal{C}	AP	MP	$\overline{oldsymbol{C}_{1(F)}}$	$C_{2(F)}$	$C_{3(F)}$	Ped2	Avenue	S.T.
1	×	~	~	1	1	1	91.6	83.2	72.4
2	~	X	~	~	1	1	96.9	89.6	77.1
3	~	1	×	1	1	1	97.1	89.4	76.6
4	~	1	~	×	1	1	89.1	81.9	70.7
5	~	1	~	1	×	1	96.3	88.2	76.1
6	~	1	~	1	1	×	96.6	89.1	75.8
7	~	1	~	1	×	×	96.1	88.1	75.3
8	~	~	~	~	~	~	97.6	90.5	77.6

C: Clustering; AP/MP: Average/Max Pooling; $C_{i(F)} = \parallel C_i - I \parallel_F^2$, $i = \{1, 2, 3\}$.

outperforming other methods. Compared with earlier manual feature-based methods, deep learning methods achieve significant performance gains due to the powerful representational learning capability of deep neural networks. However, they fail to process the complex cross-scene ShanghaiTech [61] dataset, and the performance is limited under the unsupervised setting. The proposed CRC framework pioneers the causal representations into unsupervised normality learning, which attempts to mitigate the negative impact of diverse normal events from a causal perspective and discriminates anomalies with consistency, achieving an AUC gain of 1.8% on the ShanghaiTech dataset. For complex real-world videos, the learned normality with intrinsic causality by CausalRL is more effective than conventional representation learning.

In addition, we show the EERs and inference speed of the proposed method in Figure 3, where (a)-(c) show the AUC and EER results on three publicly available benchmarks, while (d) visually compares the performance and inference speed on the CUHK Avenue [48] dataset. In addition to those already cited in Table I, other methods involved in the comparison include DRAM [72], WTA-AE [73], stackRNN [61], DFSN [74], Street Scene [75], Trans-STR [31], and HN-MUM [76]. Our CRC framework implemented using ResNet-50 achieves EERs of 4.1%, 11.6%, and 21.1% on the UCSD Ped2 [48], CUHK Avenue [60], and ShanghaiTech [61] datasets. The average inference speed of CRC (ResNet-18) and CRC (ResNet-50) is around 46 FPS and 32 FPS, respectively, which means that they take 0.022s and 0.031s from video read-in to anomaly score output, meeting the demand of real-time detection. Although the inference speed of stackRNN [61] is faster than the proposed method, our CRC (ResNet-18) and CRC (ResNet-50) show an advantage in detection accuracy, with AUC improvements of 8.8% and 10.8% (90.5% & 92.5% vs. 81.7%), respectively.

C. Ablation Study

To validate the impact of individual components and optimization constraints on causality-inspired normality learning and their effectiveness in the VAD task, we conduct an ablation study and quantitatively compared the frame-level AUC of each model variant, as shown in Table II. Model 1, which removes the clustering module and learns the causal factors by characterizing consistency only, suffers significant



Fig. 4. Results of sensitivity analysis on N and k. For better visual effects, we interpolated the original data during the mapping. Best viewed in color.

performance degradation on all three benchmarks. Unlike the classification task, VAD has only negative samples available during the training phase and cannot construct classifiers to encourage the model to learn task-specific representations, so we introduce clustering to characterize normal events as intrinsic to causal representations. The causal characterizer is optimized to learn task-specific representations valid for video anomaly detection by updating the clustering centers, yielding remarkable improvement for the UCSD Ped2 [48], CUHK Avenue [60], and ShanghaiTech [61] datasets with AUC gains of 6.0%, 7.3% and 5.2%.

Models 2 & 3 compare the contribution of average pooling and maximum pooling in the prototype decomposer. The performance gap indicates that average pooling aggregates global information and separates shared and private features more effectively. Furthermore, both pooling strategies contribute to normality learning with cumulative gains when compared to the full framework in Model 8. As stated in Sec. III, we are inspired by §1 and §2 to use constraints on the correlation matrix to encourage the model to learn a set of independent causal factors. The impact of each constraint on performance is shown in Models 4-7. Model 4 suffers the overall and most severe AUC decline, indicating that the $m{R} o ilde{m{R}}$ correlation constraint, i.e., $m{C}_1$, is critical for causal representation learning. In contrast, the other two constraints of the representation matrix, i.e., C_2 in model 5 and C_3 in model 6, bring only minor gains for video anomaly detection. Even ignoring the corresponding constraints under limited computational resources, the impact on model performance is limited, as shown in Model 7.

D. Sensitivity Analysis

We conduct extension experiments on CUHK Avenue [60] to explore the sensitivity of model performance to the number of memory entries and clustering centers, as shown in Figure 4, where (a) is implemented with ResNet-18 while (b) is with ResNet-50. Specifically, we test the AUC performance under different parameter settings: $N = \{16, 32, 64, 128, 256, 512\}$ and $k = \{4, 8, 16, 32, 64, 128\}$. As stated in Sec. III, N determines the information capacity of the memory pool. A large N may make the recorded memory entries untypical, while a too-small N will cause the memory pool to lose prototypical features. The results in Figure (4) show that choosing an appropriate N is necessary for prototype decomposition. The



Fig. 5. $\mathbf{R} \rightarrow \mathbf{R}$ correlation matrix visualization.

clustering is used to make representations compact, and k is critical for representation learning and downstream anomaly detection. Figure 4(b) clearly demonstrates that the impact of k is twofold: too small a k may make it difficult for the CRC framework to find the appropriate cluster centers that make the causal representation tight, while too large a k may make learned cluster centers close to the anomalous events. The effects N and k on the model performance are joint. Our CRC framework achieves the best performance at $\{N = 64, k = 128\}$ and $\{N = 32, k = 128\}$ on the CUHK Avenue [60] with ResNet-18 and ResNet-50. In addition, the performance shown in Figure 4(b) is more stable than that in 4(a), indicating the superior robustness of ResNet-50 compared to ResNet-18.

E. Feasibility Analysis

The proposed CRC framework attempt to address unsupervised VAD via causal representation learning and detect anomalies with the learned consistency of the causal variables for regular events. To qualitatively present the response of causal representation consistency to anomalies, we randomly select a normal and abnormal sample from the test set of ShanghaiTech [61] dataset and partially visualized their $R \rightarrow$ R correlation matrices, as shown in Figure 5. The causal factors of the regular events in (a) show good independence, i.e., the diagonal elements of the matrix are close to 1 while the other elements are as small as possible. In contrast, many non-diagonal elements of the matrix for abnormal events in (b) are greater than 0.4, indicating that the causal factors learned on negative samples fail to represent the abnormal patterns. Therefore, we can quantitatively measure the deviation of the test samples from the learned representation consistency by calculating the Frobenius norm (F-norm) of the correlation matrix with the identity matrix, i.e., $\parallel \boldsymbol{C}_1 - \boldsymbol{I} \parallel_F^2$. The F-norm of the two selected samples are 18.3 and 22.4, respectively, which can amplify the score gap between regular and abnormal events, as presented in Eq. 10.

F. Temporal Localization

Moreover, we plot the score curves of two sample videos from the UCSD Ped2 [48] dataset to qualitatively verify the quick response of our method to abnormal events, as shown in Figure 6. For the regular interval, the anomaly scores fluctuate slightly but always remain low (generally < 0.2). When



Fig. 6. Results of temporal localization for anomalies.

anomalous events occur, the anomaly score rises rapidly and remains high (generally >0.6) until the anomaly ends or leaves the field of view, indicating that our CRC framework can quickly respond to anomalies and provide accurate temporal localization for abnormal events.

V. CONCLUSION

In this paper, we address unsupervised video anomaly detection from a causal perspective and propose causally-inspired representation consistency learning to model video normality with intrinsic causality. The proposed CRC framework exploits causal principles to mine unobservable causal factors that can fully characterize normality and discriminates anomalous events with the learned representation consistency. Extensive experimental results on three benchmarks validate the effectiveness and superiority of causal representation learning on video anomaly detection. The learned normality-specific causal variable with inherent consistency can effectively reason about regular events with label-independent bias and respond quickly and sensitively to real-world anomalies. In future work, we will further explore potential causal mechanisms in unsupervised normality learning and develop robust video anomaly detection models that can bridge the domain gaps in multi-scene and multi-view real-world videos.

ACKNOWLEDGEMENTS

This work is funded by the China Mobile Research Fund of MoE (Grant No. KEH2310029), NSFC (Grant No. 62250410368), and the Specific Research Fund of the Innovation Platform for Academicians of Hainan Province (Grant No. YSPTZX202314). This work is also supported by the Shanghai Key Research Lab of NSAI and the Joint Lab on Networked AI Edge Computing Fudan University-Changan.

REFERENCES

- Y. Liu, J. Liu, X. Zhu, D. Wei, X. Huang, and L. Song, "Learning taskspecific representation for video anomaly detection with spatial-temporal attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 2190–2194.
- [2] Y. Liu, J. Liu, K. Yang, B. Ju, S. Liu, Y. Wang, D. Yang, P. Sun, and L. Song, "Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system," *IEEE Transactions on Industrial Informatics*, pp. 1–13, 2023.

- [3] K. Yang, P. Sun, J. Lin, A. Boukerche, and L. Song, "A novel distributed task scheduling framework for supporting vehicular edge intelligence," in *IEEE International Conference on Distributed Computing Systems* (*ICDCS*), 2022, pp. 972–982.
- [4] Y. Liu, J. Liu, M. Zhao, S. Li, and L. Song, "Collaborative normality learning framework for weakly supervised video anomaly detection," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 5, pp. 2508–2512, 2022.
- [5] D. Wei, Y. Liu, X. Zhu, J. Liu, and X. Zeng, "Msaf: multimodal supervise-attention enhanced fusion for video anomaly detection," *IEEE Signal Processing Letters*, vol. 29, pp. 2178–2182, 2022.
- [6] Y. Liu, J. Liu, W. Ni, and L. Song, "Abnormal event detection with selfguiding multi-instance ranking framework," in 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022, pp. 01–07.
- [7] L. Song, X. Hu, G. Zhang, P. Spachos, K. N. Plataniotis, and H. Wu, "Networking systems of ai: on the convergence of computing and communications," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 20352–20381, 2022.
- [8] B. Ju, Y. Liu, L. Song, G. Gan, Z. Li, and L. Jiang, "A high-reliability edge-side mobile terminal shared computing architecture based on task triple-stage full-cycle monitoring," *IEEE Internet of Things Journal*, 2023.
- [9] Z. Chen, B. Li, J. Xu, S. Wu, S. Ding, and W. Zhang, "Towards practical certifiable patch defense with vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15148–15158.
- [10] Y. Liu, D. Yang, Y. Wang, J. Liu, and L. Song, "Generalized video anomaly event detection: systematic taxonomy and comparison of deep models," arXiv preprint arXiv:2302.05087, 2023.
- [11] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, and L. Zhang, "Emotion recognition for multiple context awareness," in *European Conference on Computer Vision*. Springer Nature Switzerland Cham, 2022, pp. 144–162.
- [12] K. Yang, J. Liu, D. Yang, H. Wang, P. Sun, Y. Zhang, Y. Liu, and L. Song, "A novel efficient multi-view traffic-related object detection framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [13] Z. Chen, B. Li, S. Wu, J. Xu, S. Ding, and W. Zhang, "Shape matters: deformable patch attack," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV.* Springer, 2022, pp. 529–548.
- [14] D. Yang, H. Kuang, S. Huang, and L. Zhang, "Learning modalityspecific and-agnostic representations for asynchronous multimodal language sequences," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1708–1717.
- [15] D. Yang, Z. Chen, Y. Wang, S. Wang, M. Li, S. Liu, X. Zhao, S. Huang, Z. Dong, P. Zhai et al., "Context de-confounded emotion recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19005–19015.
- [16] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge, and W. Zhang, "Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos," in *Proceedings* of the 30th ACM International Conference on Multimedia, 2022, pp. 101–110.
- [17] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83-84, pp. 19–52, 2022.
- [18] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings* of the 30th ACM International Conference on Multimedia, 2022, pp. 1642–1651.
- [19] D.-L. Wei, C.-G. Liu, Y. Liu, J. Liu, X.-G. Zhu, and X.-H. Zeng, "Look, listen and pay more attention: Fusing multi-modal information for video violence detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1980–1984.
- [20] D. Yang, Y. Liu, C. Huang, M. Li, X. Zhao, Y. Wang, K. Yang, Y. Wang, P. Zhai, and L. Zhang, "Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences," *Knowledge-Based Systems*, vol. 265, p. 110370, 2023.
- [21] B. Ju, Y. Liu, X. Hu, D. Zhao, and L. Jiang, "A novel cell contour-based instance segmentation model and its applications in her2 breast cancer discrimination," *Biomedical Signal Processing and Control*, vol. 85, p. 104941, 2023.
- [22] Z. Chen, B. Li, S. Wu, S. Ding, and W. Zhang, "Query-efficient decisionbased black-box patch attack," arXiv preprint arXiv:2307.00477, 2023.

- [23] S. Liu, Z. Chen, Y. Liu, Y. Wang, D. Yang, Z. Zhao, Z. Zhou, X. Yi, W. Li, W. Zhang, and Z. Gan, "Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation," 2023.
- [24] Z. Chen, B. Li, S. Wu, K. Jiang, S. Ding, and W. Zhang, "Content-based unrestricted adversarial attack," arXiv preprint arXiv:2305.10665, 2023.
- [25] Y. Wang, Z. Chen, J. Zhang, D. Yang, Z. Ge, Y. Liu, S. Liu, Y. Sun, W. Zhang, and L. Qi, "Sampling to distill: Knowledge transfer from open-world data," *arXiv preprint arXiv:2307.16601*, 2023.
- [26] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20922–20931.
- [27] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.
- [28] Y. Liu, J. Liu, J. Lin, M. Zhao, and L. Song, "Appearance-motion united auto-encoder framework for video anomaly detection," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 5, pp. 2498–2502, 2022.
- [29] Y. Wang, Z. Chen, D. Yang, Y. Liu, S. Liu, W. Zhang, and L. Qi, "Adversarial contrastive distillation with adaptive denoising," in *ICASSP* 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
- [30] Y. Liu, J. Liu, M. Zhao, D. Yang, X. Zhu, and L. Song, "Learning appearance-motion normality for video anomaly detection," in 2022 *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.
- [31] X. Sun, J. Chen, X. Shen, and H. Li, "Transformer with spatio-temporal representation for video anomaly detection," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2022, Montreal, QC, Canada, August 26–27, 2022, Proceedings.* Springer, 2023, pp. 213–222.
- [32] Y. Liu, S. Li, J. Liu, H. Yang, M. Zhao, X. Zeng, W. Ni, and L. Song, "Learning attention augmented spatial-temporal normality for video anomaly detection," in 2021 3rd International Symposium on Smart and Healthy Cities (ISHC). IEEE, 2021, pp. 137–144.
- [33] Y. Liu, D. Li, W. Zhu, D. Yang, J. Liu, and L. Song, "Msn-net: Multiscale normality network for video anomaly detection," in *ICASSP 2023-*2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
- [34] K. Cheng, X. Zeng, Y. Liu, M. Zhao, C. Pang, and X. Hu, "Spatialtemporal graph convolutional network boosted flow-frame prediction for video anomaly detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.
- [35] J. Liu, Y. Liu, D. Li, H. Wang, X. Huang, and L. Song, "Dsdcla: Driving style detection via hybrid cnn-lstm with multi-level attention fusion," *Applied Intelligence*, pp. 1–18, 2023.
- [36] X. Xiang, Y. Liu, G. Fang, J. Liu, and M. Zhao, "Two-stage alignments framework for unsupervised domain adaptation on time series data," *IEEE Signal Processing Letters*, 2023.
- [37] J. Liu, Y. Liu, W. Zhu, X. Zhu, and L. Song, "Distributional and spatial-temporal robust representation learning for transportation activity recognition," *Pattern Recognition*, vol. 140, p. 109568, 2023.
- [38] D. Yang, S. Huang, Z. Xu, Z. Li, S. Wang, M. Li, Y. Wang, Y. Liu, K. Yang, Z. Chen *et al.*, "Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception," *arXiv preprint arXiv:2307.13933*, 2023.
- [39] K. Yang, D. Yang, J. Zhang, M. Li, Y. Liu, J. Liu, H. Wang, P. Sun, and L. Song, "Spatio-temporal domain awareness for multi-agent collaborative perception," arXiv preprint arXiv:2307.13929, 2023.
- [40] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Selfsupervised sparse representation for video anomaly detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII.* Springer, 2022, pp. 729– 745.
- [41] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: memoryaugmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.
- [42] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

- [43] D. Li, Y. Liu, and L. Song, "Adaptive weighted losses with distribution approximation for efficient consistency-based semi-supervised learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7832–7842, 2022.
- [44] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.
- [45] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Computer Vision–ECCV* 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer, 2020, pp. 329–345.
- [46] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 938–946.
- [47] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 2921–2928.
- [48] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis* and Machine Intelligence, vol. 36, no. 1, pp. 18–32, 2013.
- [49] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection-a new baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6536–6545.
- [50] Y. Chang, Z. Tu, W. Xie, B. Luo, S. Zhang, H. Sui, and J. Yuan, "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognition*, vol. 122, p. 108213, 2022.
- [51] T.-N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1273–1283.
- [52] Y. Liu, Z. Guo, J. Liu, C. Li, and L. Song, "Osin: object-centric scene inference network for unsupervised video anomaly detection," *IEEE Signal Processing Letters*, 2023.
- [53] C. Chen, Y. Xie, S. Lin, A. Yao, G. Jiang, W. Zhang, Y. Qu, R. Qiao, B. Ren, and L. Ma, "Comprehensive regularization in a bi-directional predictive network for video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 230–238.
- [54] Q. Bao, F. Liu, Y. Liu, L. Jiao, X. Liu, and L. Li, "Hierarchical scene normality-binding modeling for anomaly detection in surveillance videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6103–6112.
- [55] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8046–8056.
- [56] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [57] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13588–13597.
- [58] L. Wang, J. Tian, S. Zhou, H. Shi, and G. Hua, "Memory-augmented appearance-motion network for video anomaly detection," *Pattern Recognition*, p. 109335, 2023.
- [59] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

- [60] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.
- [61] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.
- [62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [63] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
 [65] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events
- [65] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.
- [66] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2895–2903.
- [67] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proceedings* of the IEEE International Conference on Computer Vision, 2017, pp. 3619–3627.
- [68] Y. Lu, K. M. Kumar, S. shahabeddin Nabavi, and Y. Wang, "Future frame prediction using convolutional vrnn for anomaly detection," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.
- [69] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "Anopcn: Video anomaly detection via deep predictive coding network," in *Proceedings of the 27th* ACM International Conference on Multimedia, 2019, pp. 1805–1813.
- [70] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, and Y. Wang, "Abnormal event detection using deep contrastive learning for intelligent video surveillance system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5171–5179, 2021.
- [71] M. Zhao, Y. Liu, J. Liu, and X. Zeng, "Exploiting spatial-temporal correlations for video anomaly detection," in 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022, pp. 1727– 1733.
- [72] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," arXiv preprint arXiv:1510.01553, 2015.
- [73] H. T. Tran and D. Hogg, "Anomaly detection using a convolutional winner-take-all autoencoder," in *Proceedings of the British Machine Vision Conference 2017.* British Machine Vision Association, 2017.
- [74] B. Ramachandra, M. Jones, and R. Vatsavai, "Learning a distance function with a siamese network to localize anomalies in videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2598–2607.
- [75] B. Ramachandra and M. Jones, "Street scene: a new dataset and evaluation protocol for video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2569–2578.
- [76] H. Li, Y. Wang, M. Chen, and J. Li, "Hn-mum: heterogeneous video anomaly detection network with multi-united-memory module," *Multimedia Tools and Applications*, pp. 1–18, 2023.