A Causal Inference Look at Unsupervised Video Anomaly Detection

Xiangru Lin[1,2*], **Yuyang Chen**[1*], **Guanbin Li**[1†], **Yizhou Yu**[2†]

[1]Sun Yat-sen University
[2]The University of Hong Kong
lin0111@connect.hku.hk, chenyy567@mail2.sysu.edu.cn,
liguanbin@mail.sysu.edu.cn, yizhouy@acm.org

Abstract

Unsupervised video anomaly detection, a task that requires no labeled normal/abnormal training data in any form, is challenging yet of great importance to both industrial applications and academic research. Existing methods typically follow an iterative pseudo label generation process. However, they lack a principled analysis of the impact of such pseudo label generation on training. Furthermore, the long-range temporal dependencies also has been overlooked, which is unreasonable since the definition of an abnormal event depends on the long-range temporal context. To this end, first, we propose a causal graph to analyze the confounding effect of the pseudo label generation process. Then, we introduce a simple yet effective causal inference based framework to disentangle the noisy pseudo label's impact. Finally, we perform counterfactual based model ensemble that blends long-range temporal context with local image context in inference to make final anomaly detection. Extensive experiments on six standard benchmark datasets show that our proposed method significantly outperforms previous state-of-the-art methods, demonstrating our framework's effectiveness.

Introduction

Video anomaly detection (VAD) refers to the task of detecting anomalous events, such as unusual pedestrian motion patterns, traffic accidents, and thrown objects, etc., in frames of a video, that divert significantly from the observed normal routine. The practical importance of the task has attracted extensive research both from industry and academia. The majority of such research share a typical setting that either a set of labeled abnormal events in the dataset is available or the training dataset must contain normal videos only, limiting the wide application of such research. Instead, another line of research focuses on designing algorithms following a completely unsupervised setting where no labeled normal/abnormal training data is provided in any form. In this paper, we focus on such unsupervised video anomaly detection (UVAD). To supervise the training, self-training with iterative pseudo label generation is usually adopted, a technique that has been extensively studied and used in unsupervised learning (Giorno, Bagnell, and Hebert 2016; Ionescu et al. 2017; Wang et al. 2018; Pang et al. 2020). The key working principle behind such a pipeline is two-fold: First, the learned representations are biased towards normal events since abnormal events in real world are rare, making the representations of abnormal events more discriminative; Second, the majority of the pseudo labels generated by heuristic designs are accurate enough, for example auto-encoder based reconstruction in (Wang et al. 2018) and Sp+iForest in (Pang et al. 2020) etc.

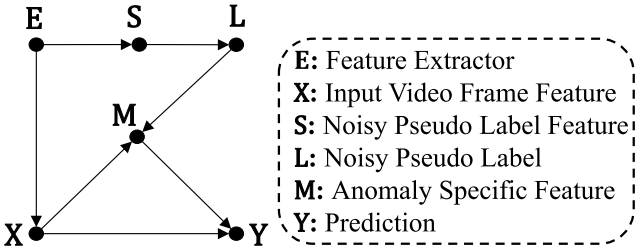

Figure 1: The proposed causal graph explaining the causal effect of noisy pseudo label.

Although an anomaly detection model trained with the aforementioned pseudo labels exhibits competitive performance, its performance gain primarily comes from the correct pseudo labels. Without a principled analysis of the negative impact introduced by the incorrect pseudo labels, further performance improvement is limited. To better understand the noisy pseudo label's impact and obtain insights about this phenomenon, we approach this problem from a causal inference perspective. According to Fig.1, the UVAD task is to learn a model that could estimate $P(Y|X, M)$. The pseudo label generation process ($E \rightarrow S \rightarrow L$) produces noisy pseudo label set L that supervises the training of anomaly specific feature representation M in $P(Y|X, M)$. On the one hand, the correct pseudo labels benefit the anomaly specific feature representation learning M, leading to a significant performance gain. This is denoted as the mediation causal path ($X \rightarrow M \rightarrow Y$) in Fig. 1. On the other hand, the incorrect pseudo labels confound X and Y via the backdoor path ($X \leftarrow E \rightarrow S \rightarrow L \rightarrow M \rightarrow Y$). A backdoor path is defined as one end of the path has an arrow into X and the other end into Y, making X and Y spuri-

*The first two authors have equal contribution. †Corresponding authors are Guanbin Li and Yizhou Yu.

ously dependent. In other words, this causal path has a detrimental effect that spuriously correlates some abnormal/normal events with normal/abnormal labels, which misleads the classifier biased towards making false predictions. Therefore, we conjecture that the mixed causal effects from the above mentioned two causal paths are one of the main causes of the performance bottleneck. Furthermore, the interactions between the long-range temporal context and image frame appearance itself are essential to discriminate anomalous video frames. Existing methods perform such interactions by collecting a small range of neighboring video frames as inputs, which lack sufficient abilities to exploit long-range temporal context in video activities since a short range of temporal context involves inconsistent temporal context information (Yu et al. 2020; Pang et al. 2020; Ionescu et al. 2017; Giorno, Bagnell, and Hebert 2016; Wang et al. 2018).

According to the above analysis, we propose a new two-stage causal inference based pipeline that aims to disentangle the noisy pseudo label's impact and incorporate long-range temporal context. Concretely, at the first stage, we conduct de-confounded training by keeping the beneficial mediation path while removing the backdoor path as shown in Fig. 3. Then, we perform a counterfactual based model ensemble by summing up predictions from the model trained in the first stage and from the same model with inputs replaced by a long-range sliding window based context feature while keeping the mediator M unchanged as shown in Fig. 4. It is important to note that the second stage does not require extra training, meaning that we just need to perform inference twice to obtain the counterfactual model ensemble prediction, which is lightweight and cost-free. The pipeline is shown in Fig. 5.

To sum up, this paper has the following contributions:

- To our best knowledge, we are the first to investigate the impact of the noisy pseudo label in UVAD from the perspective of causal inference and identify that the pseudo label generation contains a confounding effect that limits further performance improvement.
- We introduce an iterative two-stage causal inference based framework to disentangle the noisy pseudo label's impact. Specifically, we adopt de-confounded training with causal intervention to remove the detrimental backdoor causal path and perform counterfactual based long-range temporal context ensemble with the trained model.
- Our method outperforms all previous methods by a clear margin, achieving new state-of-the-art performance on six standard datasets.

Related Works

Unsupervised Video Anomaly Detection. Giorno et al. (Giorno, Bagnell, and Hebert 2016) introduced the UVAD problem and proposed to use permutation tests to detect changes on a frame sequence to see which frames are distinguishable from all the previous frames. Ionescu et al. (Ionescu et al. 2017) removed the permutation tests and instead applied unmasking to measure the abnormality according to the changes of the classification accuracy. Wang et al. (Wang et al. 2018) approached the problem from an auto-encoder perspective to solve the UVAD task. The most similar work to ours is (Pang et al. 2020) where it first adopted Sp+iForest (Liu, Ting, and Zhou 2012) to generate pseudo labels of video frames and then trained a self-supervised deep ordinal regression model iteratively in an end-to-end manner. However, our work differs in the following aspect: (1) We analyze the role of the pseudo label generation in (Pang et al. 2020) from a causal inference perspective and identified that it has a confounding effect on $P(Y|X, M)$. (2) We propose to utilize the backdoor adjustment approach to eliminate the confounding effect explicitly by stratifying (intervening) the pseudo label feature S and blocking the $X \leftarrow E \rightarrow S \rightarrow L \rightarrow M \rightarrow Y$. (3) We inject long-range temporal context prior information to the model prediction through counterfactual based model ensemble with negligible computational cost.

Causal Inference. It is a statistical tool that empowers the model to reason the casual effect between variables of interest. It has been extensively studied and applied in statistics, psychology, economics, and sociology (Morgan and Winship 2014; Chernozhukov, Fernández-Val, and Melly 2009; Rubin 2005; Petersen, Sinisi, and van der Laan 2006; Pearl 2001). Recent years have witnessed an increasing number of active research in applying causal inference to many computer related problems, including natural language processing (Liu et al. 2021; Keith, Jensen, and O'Connor 2020), computer vision (Tang et al. 2020; Zhang et al. 2020; Wang et al. 2020), Robotics (Ahmed et al. 2021), etc. We follow the same graphical notation in Pearl's graphical model (Pearl 2009). However, we proposed a tailored causal graph for the UVAD task, which, to our best knowledge, is the first attempt to investigate the confounding effect of the pseudo label generation process in UVAD. Moreover, we model the interactions between long-range temporal context and local image frame appearance via counterfactual based feature replacement.

Method

Problem Formulation

General Settings. Given a set of video frames $\mathbb{I} = \{I_i\}_{i=1}^{K}$ where K is the total number of video frames, the extracted feature set is represented as $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{K}$, where $\boldsymbol{x}_i \in \mathbb{R}^{D_b}$. We define the overall noisy pseudo label set as $L = A \cup N = \{l_i | l_i = c, c \in \mathcal{C} = \{0, 1\}\}_{i=1}^{K}$, where the pseudo anomaly label set to A and the pseudo normal label set to N. $\mathcal{C}$ represents the label set, 0 for the normal event and 1 for the abnormal event. We formulate the UVAD task as,

$$\mathcal{F} = \arg\min_{\Theta} \sum_{I \in \mathbb{I}} \mathcal{L}_{foc}(\hat{y} = \phi(\boldsymbol{m} = \varphi(\boldsymbol{x} = f(I))), l) \quad (1)$$

where we aim to learn an anomaly detector $\mathcal{F}$ via a convolutional neural network which consists of a backbone network $f(\cdot; \Theta_b) : \mathbb{R}^{H \times W \times 3} \mapsto \mathbb{R}^{D_b}$ that transforms an input video frame I to feature $\boldsymbol{x}$, an anomaly representation learning block $\varphi(\cdot; \Theta_a) : \mathbb{R}^{D_b} \mapsto \mathbb{R}^{D_s}$ that converts $\boldsymbol{x}$ to an anomaly specific representation $\boldsymbol{m}$, and an anomaly score regression layer $\phi(\cdot; \Theta_s) : \mathbb{R}^{D_s} \mapsto \mathbb{R}$ that learns to predict $\boldsymbol{m}$ to an anomaly score y. The overall parameters $\Theta = \{\Theta_b, \Theta_a, \Theta_s\}$

are optimized by the focal loss (Lin et al. 2017) $\mathcal{L}_{foc}$,

$$\begin{aligned}\mathcal{L}_{foc}(\hat{y}, l) = \alpha_1 l(1-\sigma(\hat{y}))^2 \log \sigma(\hat{y}) \\ +\alpha_2(1-l)\sigma(\hat{y})^2 \log\left(1-\sigma(\hat{y})\right)\end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is the standard sigmoid function. α_1 and α_2 are hyperparameters.

A Strong Baseline. We then introduce the training of a strong baseline anomaly detector $\mathcal{F}$ following the same logics in (Pang et al. 2020).

Round 0: Initial pseudo label set generation L_0. We use a ResNet-50 CNN (He et al. 2016) pre-trained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) as $f(\cdot)$ to extract $\boldsymbol{X}$. Then, an unsupervised algorithm is adopted to perform initial pseudo labels generation for L_0. It is true that many algorithms can be chosen for this task, such as auto-encoder network (Wang et al. 2018). However, for fair comparisons with (Pang et al. 2020), isolation forest algorithm (Liu, Ting, and Zhou 2012) is adopted. It isolates anomalous events by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. This is equivalent to building up a forest of random trees where the feature and cut-point at each tree node are randomly selected. The number of splittings required to isolate a sample equals to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality. Concretely, given a random subset $\boldsymbol{R} \subset \boldsymbol{X}$ and $\boldsymbol{x} \in \boldsymbol{R}$, the anomaly score of $\boldsymbol{x}$ is defined as,

$$\begin{aligned}\boldsymbol{z} &= PCA(\boldsymbol{x}) \\ score(\boldsymbol{z}) &= 2^{-\frac{E(h(\boldsymbol{z}))}{\tau(|\boldsymbol{R}|)}} \\ \tau(n) &= 2Har(n-1) - (2(n-1)/n)\end{aligned} \tag{3}$$

where $PCA(\cdot)$ is principle component analysis function that retains 99% of the amount of explained variance. $h(\boldsymbol{z})$ represents the path length of $\boldsymbol{z}$ that is measured by the number of edges it traverses an isolation tree from the root to a leaf node by $\boldsymbol{z}$. $E(h(\boldsymbol{z}))$ is the average of $h(\boldsymbol{z})$ from a collection of isolation trees. $|\boldsymbol{R}|$ denotes the total samples in $\boldsymbol{R}$. $Har(\cdot)$ is the harmonic number. $\tau(\cdot)$ is a normalization term.

Round 1: Learning with L_0. With L_0 computed at previous Round, we perform learning with equation (1) to obtain $\mathcal{F}_1$. Then, we re-sample the pseudo label set L_1 with $\mathcal{F}_1$.

Round 2 to T: Self-supervised pseudo label learning process. Self-training with iterative pseudo label generation is performed to gradually refine the quality of L. Specifically, new pseudo label set L_t generated with the trained $\mathcal{F}_t$ is used to train a new $\mathcal{F}_{t+1}$. This process is iterated for T Rounds until the performance plateaus.

A Causal Inference Look At UVAD

Analysis. We propose a causal graph shown in Fig. 1 to analyze the problem of the aforementioned training of $\mathcal{F}$. Here, we briefly introduce the definition of the causal graph. The causal graph in Fig. 1 consists of six variables of interest: feature extractor (E), noisy pseudo label feature (S), noisy pseudo label (L), input video frame feature (X), anomaly specific feature representation (M), and model prediction (Y). It basically contains two parts: (1) the pseudo label generation part via ***Link*** $E \rightarrow S \rightarrow L$, which represents the pseudo label generation in the Round 0 and following Rounds; (2) the model training part via ***Link*** $E \rightarrow X$ denoting $\boldsymbol{x} = f(I)$ in equation (1), ***Link*** $X \rightarrow M \rightarrow Y$ denoting $\hat{y} = \phi(\boldsymbol{m} = \varphi(\boldsymbol{x} = f(I)))$, and ***Link*** $L \rightarrow M \leftarrow X$ denoting $\mathcal{L}_{foc}(\hat{y}, l)$. Besides, ***Link*** $X \rightarrow Y$ is the direct causal effect between X and Y which we aim to achieve.

As discussed in previous section, the performance of the learned model $\mathcal{F}$ cannot imply the direct causal effect between X and Y because an apparent backdoor path $X \leftarrow E \rightarrow S \rightarrow L \rightarrow M \rightarrow Y$ makes X and Y spuriously dependent. Correct pseudo labels help $\mathcal{F}$ learn better anomaly specific representation space via $X \rightarrow M \rightarrow Y$ while incorrect pseudo labels distort the space through the backdoor path. Therefore, this provides a potential to further improve the performance.

De-confounded Training with Causal Intervention. To address the aforementioned problem, we propose an intervened causal graph to solve the confounding bias of the pseudo label generation process as shown in Fig. 3. The adjusted causal graph blocks the confounding path by blocking the causal link the $X \leftarrow E \rightarrow S \rightarrow L \rightarrow M \rightarrow Y$, which makes the pseudo label generation process not spuriously correlated with the model learning. Thus, learning with this causal graph produces the direct causal effect between X and Y denoted as $P(Y|do(X), M) = \sum_s P(Y|X, M, S=s)\, P(s)$. This technique is termed as backdoor adjustment (Pearl 2001), which amounts to partitioning the population into groups that are homogeneously relative to S, assessing the effect of X on Y in each homogeneous group, and then averaging the results. Note that we choose S here because it is the only feasible variable that can be partitioned to perform backdoor adjustment, while the feature extractor E and noisy pseudo label L are intractable to be partitioned. To this end, we define the learned model with $P(Y|do(X), M)$ as $\mathcal{F}^*$ and the implementation of $P(Y|do(X), M)$ is

$$\begin{aligned}P(Y=c|do(X=\boldsymbol{x}), M=\boldsymbol{m}) &= \mathbb{E}_{\boldsymbol{s}}\left[\sigma(\mathcal{F}^*(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{s}))\right] \\ &\approx \sigma(\mathbb{E}_{\boldsymbol{s}}\left[\mathcal{F}^*(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{s})\right])\end{aligned} \tag{4}$$

where $\mathcal{F}^*$ outputs the unbiased prediction logit of $\boldsymbol{x}$ for class c. Since $\mathbb{E}_s[\cdot]$ requires computationally expensive sampling, we thus perform approximation shown in equation (7).

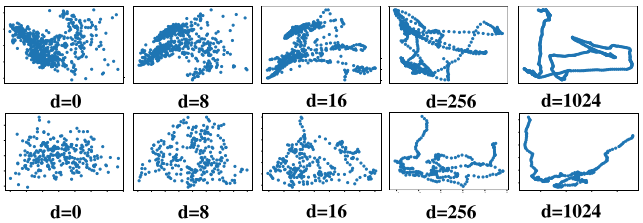

Figure 2: Temporal Context of UCSD Ped2 (first row) and Avenue (second row) from short-range (d=0) to long-range (d=1024).

Counterfactual Based Long-range Temporal Context

Ensemble. With the model trained in aforementioned de-confounded training, we further increase the model capacity by injecting long-range temporal context prior to the model prediction. In VAD, extracting robust temporal context is critical for the determination of an abnormal event. Existing methods typically model temporal context as a short-range of neighboring video frames while ignoring long-range temporal context. Different short-range temporal context representations may differ significantly from each other and have large variations, which is not conducive to capture robust temporal context representation. On the contrary, long-range temporal context representations are more stable and change slightly as the video plays. This phenomenon is illustrated in Fig. 2, where we plot the change of temporal context features as the neighboring frames are increased: (1) the left most column denotes short-range (0 neighboring frames) temporal context feature projected to 2D image plane and (2) the right most column shows long-range (1024 neighboring frames) temporal context feature projected to 2D image plane. It is clear that short-range temporal context feature representation is much noisier than the long-range counterpart and the long-range temporal context shows smoother and clearer pattern. To this end, we propose to model the long-range temporal context via counterfactual feature replacement shown in the second part of Fig 4. Due to the fact that the magnitude of normal prediction logit and that of the long-range temporal context prediction logit are different, we normalize the prediction logits from $\mathcal{F}^*$ for normal prediction and long-range prediction before summing up them together for model ensemble. The final anomaly prediction score $O(\cdot)$ for class c is defined as follows:

$$O(Y=c) = \sigma(Norm(\mathbb{E}_{\boldsymbol{s}}\left[\mathcal{F}^*(\boldsymbol{x},\boldsymbol{m},\boldsymbol{s})\right]) + Norm(\mathbb{E}_{\boldsymbol{s}}\left[\mathcal{F}^*(\boldsymbol{x}_a,\boldsymbol{m},\boldsymbol{s})\right])) \tag{5}$$

where $\boldsymbol{x}_a = \left(\sum_{i=-d}^{d}\boldsymbol{x}_i\right)/2d$ is the mean feature of a sliding window centered at $\boldsymbol{x}$ with window size d. $Norm(logit) = (logit-\mu)/\delta$ where μ is the mean value of all logits of all frames and δ is the standard deviation of all logits of all frames.

Overall Formulation. The overall formulation of the anomaly prediction problem is defined as the measurement of the $O(\cdot)$ score.

$$\arg\max_{c\in\mathcal{C}} O(Y=c) \tag{6}$$

De-confounded Training

As discussed in the previous subsection, we propose to use backdoor adjustment to derive the de-confounded model. The key idea is to stratify (intervene) one of the variables E, S or L to block the backdoor path. However, the stratification of the pseudo label generation process can be implemented as stratifying the pseudo label feature S because L is determined by the feature set S generated by E only and stratifying L or E is intractable. Therefore, we define the stratification of S as $\boldsymbol{S} = \{\boldsymbol{s}_i\}_{i=1}^{N_s}$ where $\boldsymbol{s}_i \in \mathbb{R}^{D_b}$ and N_s is a hyperparameter representing the size of the confounder set $\boldsymbol{S}$. Since the number of the noisy pseudo label features is large in reality, in the implementation, we utilize K-Means

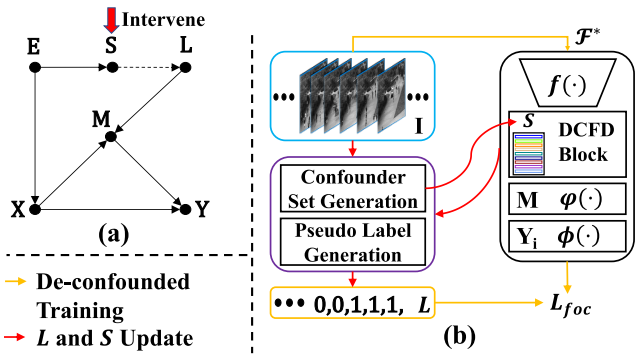

Figure 3: Overview of the proposed de-confounded training. (a) represents the intervened causal graph. The detailed implementation is illustrated in (b). In particular, the pseudo label set L and the confounder set $\boldsymbol{S}$ are iteratively generated via the red lines; then, the de-confounded training process is performed following the orange lines. The DCFD Block is the de-confound process illustrated in equation (8).

with $PCA(\cdot)$ to learn the confounder set $\boldsymbol{S}$. Therefore, the overall formulation for equation (4) is,

$$\begin{aligned} P\left(Y|do\left(X\right)\right) &= \sum_{\boldsymbol{s}} P\left(Y|X=\boldsymbol{x},M=\boldsymbol{m},S=\boldsymbol{s}\right)P\left(\boldsymbol{s}\right) \\ &\approx P\left(Y|X,\boldsymbol{m}=\sum_{\boldsymbol{s}} g\left(\boldsymbol{x}=f(I),\boldsymbol{s}\right)P\left(\boldsymbol{s}\right)\right) \end{aligned} \tag{7}$$

where the approximation is achieved by the Normalized Weighted Geometric Mean (Xu et al. 2015b)(See Supplementary Document). Blocking the backdoor path makes X have a fair opportunity to incorporate every $\boldsymbol{s}$ into Y's prediction, subject to a prior $P(\boldsymbol{s})$. $g(\cdot)$ is defined as follows,

$$\begin{aligned} \boldsymbol{m} = g(\boldsymbol{x},\boldsymbol{S})P(\boldsymbol{S}) &= \sum_{\boldsymbol{s}} g(\boldsymbol{x},\boldsymbol{s})P(\boldsymbol{s}) \\ &= softmax(\frac{(\boldsymbol{W}_1\boldsymbol{x})^T(\boldsymbol{W}_2\boldsymbol{S})}{\sqrt{D_h}})\boldsymbol{S} \end{aligned} \tag{8}$$

where $P(\boldsymbol{s}_i) = \frac{|\boldsymbol{s}_i|}{\sum_j|\boldsymbol{s}_j|}$ and $|\boldsymbol{s}_i|$ is the number of samples in cluster $\boldsymbol{s}_i$. $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{D_h\times D_b}$ are learnable parameters to project $\boldsymbol{x}$ and $\boldsymbol{s}_i$ into a joint space. $\sqrt{D_h}$ is a constant scaling factor for feature normalization. In the implementation, to better represent the anomaly specific feature, we further set $M = \boldsymbol{m}^{\oplus}$ where $\boldsymbol{m}^{\oplus} = concat(\boldsymbol{x},\boldsymbol{m})$.

Finally, the model $\mathcal{F}^*$ defined in this section is trained with the $\mathcal{L}_{foc}$.

Counterfactual Temporal Context Ensemble

With the model $\mathcal{F}^*$ trained in previous subsection, we aim to inject long-range temporal context prior to the model prediction. Given an input video frame I, the first term in equation (5) is obtained via a normal inference taking as input $\boldsymbol{x} = f(I)$ and thus $\boldsymbol{m}^{\oplus} = concat(\boldsymbol{x},\boldsymbol{m})$. The second term in equation (5) is implemented as a counterfactual feature replacement. In other words, we set $\boldsymbol{m}_a^{\oplus} = concat(\boldsymbol{x}_a,\boldsymbol{m})$

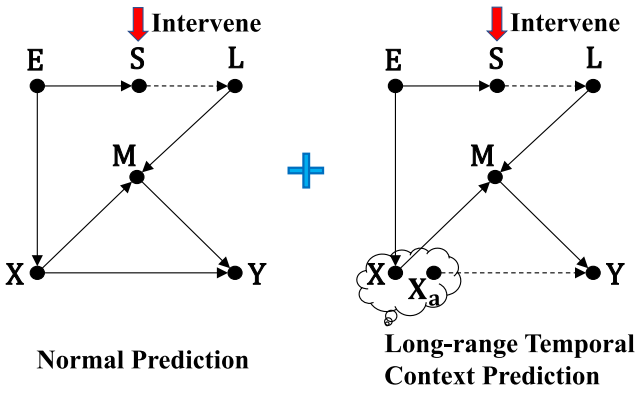

Figure 4: The counterfactual long-range temporal context model ensemble to calculate final output anomaly prediction logits. The left part is $P(Y = c|do(X = \boldsymbol{x}), M = \boldsymbol{m}^{\oplus})$ and the right represents $P(Y = c|do(X = \boldsymbol{x}_a), M = \boldsymbol{m}_a^{\oplus})$.

followed by later fusion layers. That is setting the input to the mean feature $\boldsymbol{x}_a$ of a sliding window centered at I and at the same time keeping everything else unchanged. This implementation mimics the interactions between a long-range temporal context $\boldsymbol{x}_a$ and local image context $\boldsymbol{m}$. With such a disentangled design, the first term maintains the de-confounded anomaly prediction and the second term incorporates interactions between long-range temporal context and local image context. Summing up them together resembles a model ensemble. The implementation of equation (5) is defined as,

$$O(Y = c) = \sigma(Norm(\mathbb{E}_{\boldsymbol{s}}\left[\mathcal{F}^*(\boldsymbol{x}, \boldsymbol{m}^{\oplus}, \boldsymbol{s})\right]) + Norm(\mathbb{E}_{\boldsymbol{s}}\left[\mathcal{F}^*(\boldsymbol{x}_a, \boldsymbol{m}_a^{\oplus}, \boldsymbol{s})\right])) \quad (9)$$

Self-supervised Pseudo Label Learning

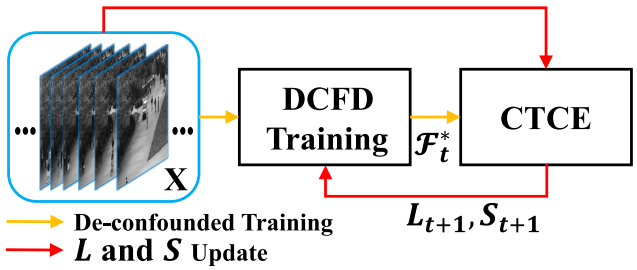

Figure 5: Overall pipeline of the two-stage iteration. DCFD Training is illustrated in Fig. 3 and CTCE is the counterfactual Temporal Context Ensemble shown in Fig. 4.

To this end, we have presented the de-confounded training module and the counterfactual based long-range temporal context ensemble module. Following (Pang et al. 2020), we follow the same self-supervised pseudo label learning setting as that of the strong baseline. Concretely, at Round 0, we use the aforementioned isolation forest algorithm to initialize the pseudo labels L as L_0. The confounder set $\boldsymbol{S}$ is first initialized as $\boldsymbol{S}_0$ using the backbone $f(\cdot)$. Then, at Round 1, we perform de-confounded training to obtain the optimized model parameter $\mathcal{F}_1^*$, which is then used to update $\boldsymbol{S}$ to $\boldsymbol{S}_1$ and $\boldsymbol{L}_0$ to $\boldsymbol{L}_1$ via counterfactual temporal context ensemble module. At Round 2 and onward, this self-supervised pseudo label learning procedure is repeated for T Rounds. In general, although our framework falls into the self-supervised pseudo learning paradigm, our contributions are that we explicitly removes the confounding bias caused by pseudo label generation process and incorporates long-range temporal context prior in a counterfactual manner. Experiments in the later section further show that our model performance surpasses previous SOTA significantly.

Experiments

Implementation Details

Training and Evaluation. Since anomalies are rare events in real-world applications and it is violated if only the test set of these datasets are used, following (Pang et al. 2020), we merge the training and test sets to construct a full dataset. We train the model on a sampled training set and evaluate our model on the full dataset with the ground truth used in the evaluation only. To obtain reliable pseudo labels for training, we construct the pseudo anomaly label set A by keeping frames with top $a\%$ anomaly scores as anomalous frames and build up the pseudo normal label set N by selecting $b\%$ most normal frames based on the anomaly scores. a and b are typically set to 5 and 20 respectively. These two cutoff thresholds are used by default as they consistently obtain substantially improved performance on datasets with diverse anomaly rates. Setting b to a higher value can always help to achieve a high-quality N because of the overwhelming presence of normal frames in the real-word datasets. For **more implementation details**, please refer to **the supplementary document**.

Evaluation Datasets and Metric

We evaluate our method on four benchmark datasets, namely UCSD dataset (Mahadevan et al. 2010), Subway surveillance dataset (Adam et al. 2008), UMN dataset (Mehran, Oyama, and Shah 2009), and Avenue dataset (Lu, Shi, and Jia 2013). Following (Sugiyama and Borgwardt 2013; Giorno, Bagnell, and Hebert 2016; Ionescu et al. 2017; Luo, Liu, and Gao 2017; Sultani, Chen, and Shah 2018; Wang et al. 2018; Liu, W. Luo, and Gao 2018; Pang et al. 2020), we employ ROC curves and the corresponding area under the curve (AUC) as the evaluation metric, which is computed with respect to the ground truth frame-level annotations. For more datasets' details, please refer to **the supplementary document**.

Ablation Study

We conduct extensive experiments to verify the effectiveness of our model in the following aspects: (1) Component Effectiveness; (2) Variants of Counterfactual Ensemble; (3) Loss Design; (4) Backbone Robustness; (5) Hyperparameter Tuning. For fair comparisons, we choose the baseline model that consists of a ResNet-50 as $f(\cdot)$ and two consecutive FC-BN-ReLU as $\varphi(\cdot)$ followed by a single FC as

Sup.	Method	ID	UCSD		Subway		UMN				Avenue
			Ped1	Ped2	Entrance	Exit	Scene1	Scene2	Scene3	All Scenes	
Labeled Data	MPPCA (Kim and Grauman 2009)	1	59.0%	69.3%	-	-	-	-	-	-	-
	SFM (Mehran, Oyama, and Shah 2009)	2	67.5%	55.6%	-	-	-	-	-	96.0%	-
	MDT (Mahadevan et al. 2010)	3	81.8%	82.9%	-	-	-	-	-	-	-
	SRC (Cong, Yuan, and Liu 2011)	4	-	-	80.0%	83.0%	99.5%	97.5%	96.4%	97.8%	-
	AMDN (Xu et al. 2015a)	5	92.1%	90.8%	-	-	-	-	-	-	-
	LSHF (Zhang et al. 2016)	6	87.0%	91.0%	-	-	99.2%	98.3%	**99.9%**	**99.7%**	-
	GNG (Sun, Liu, and Harada 2017)	7	**93.8%**	94.1%	-	-	99.8%	**99.3%**	**99.9%**	**99.7%**	-
	FFP (Liu, W. Luo, and Gao 2018)	8	83.1%	95.4%	-	-	-	-	-	-	-
	AMC (Nguyen and Meunier 2019)	9	-	96.2%	-	-	-	-	-	-	-
	MemAE (Gong et al. 2019)	10	-	94.1%	-	-	-	-	-	-	-
	OCAA (Ionescu et al. 2019)	11	-	**97.8%**	-	-	-	-	-	99.6%	-
	DAE (R.T. et al. 2019)	12	-	-	**93.5%**	**95.1%**	**99.9%**	98.2%	99.8%	99.3%	-
	MLEP (Liu et al. 2019)	13	-	-	-	-	-	-	-	-	**92.8%**
	PMem (Park, Noh, and Ham 2020)	14	-	97.0%	-	-	-	-	-	-	88.5%
	CDAE (Chang et al. 2020)	15	-	96.5%	-	-	-	-	-	-	86.0%
	CTH (Yu et al. 2020)	16	-	97.3%	-	-	-	-	-	-	89.6%
No Labeled Data	ADF (Giorno, Bagnell, and Hebert 2016)	17	59.6%	57.6%	74.6%	87.2%	80.2%	88.3%	77.1%	84.8%	-
	Unmask (Ionescu et al. 2017)	18	68.4%	82.2%	70.6%	85.7%	99.3%	87.7%	98.2%	95.1%	80.6%
	CTS (Liu, Li, and Póczos 2018)	19	69.0%	87.5%	71.6%	**93.1%**	-	-	-	95.2%	81.1%
	DAW (Wang et al. 2018)	20	**77.8%**	**96.4%**	-	84.5%	-	-	-	-	**85.3%**
	STDOR (Pang et al. 2020)	21	71.7%	83.2%	**88.1%**	92.7%	**99.9%**	**99.9%**	**99.7%**	**97.4%**	-
	iForest(ResNet-50)	22	64.5%	68.8%	80.5%	90.9%	87.7%	88.1%	90.9%	83.2%	76.9%
	Strong Baseline(ResNet-50)	23	70.7%	81.2%	81.9%	94.3%	98.8%	100%	98.4%	97.7%	84.7%
	Ours(ResNet-50) + DCFD	24	73.9%	97.9%	85.1%	96.1%	99.7%	100%	99.7%	99.5%	85.9%
	Ours(ResNet-50) + DCFD + CTCE	25	**84.9%**	**99.4%**	**89.0%**	**97.2%**	**100%**	**100%**	**99.8%**	**100%**	**87.3%**
	Ours(I3D-RGB) + DCFD + CTCE	26	**84.9%**	**98.7%**	**91.3%**	**97.6%**	**99.9%**	**100%**	**99.8** %	**99.2%**	**90.3%**

Table 1: Comparisons with state-of-the-art methods including 16 methods that require labeled normal data in the upper block and 7 methods that require no labeled data in the bottom block. The numbers are Frame-level AUC performance. The best performing result in each block is marked as bold.

$\phi(\cdot)$ for all experiments. In (4), we further test C3D (Tran et al. 2015), I3D (Carreira and Zisserman 2017) (taking RGB image as input only), and VGG (Simonyan and Zisserman 2015) for $f(\cdot)$. All experiments are performed on the challenging UCSD datasets and are conducted with self-supervised pseudo label learning. We set the default setting on the UCSD datasets to: $N_s = 16$, $d = 1024$, $(a\%, b\%) = (5\%, 20\%)$ to balance the computational cost and performance. The ablations are performed by changing each parameter at a time.

Component Effectiveness. According to Tab. 2, we perform experiments 1,2,3 to verify the effectiveness of each proposed component. Experiment 1 is the proposed strong baseline model. With experiments $2, 3$, by adding DCFD Training only, it outperforms the strong baseline by 3.2% on UCSD Ped1 and 16.7% on Ped2. By adding DCFD Training and Counterfactual TCE, it further beats the DCFD Training only model by 11% and 1.5% on UCSD Ped1 and Ped2.

Variants of Counterfactual Ensemble. We perform experiments 3,4,5 in Tab. 2 to verify the design choice of CTCE. Concretely, we construct two variants: (1) CTCE V1: the overall effect of X towards Y without further counterfactual intervention on the mediator M. We abandon the counterfactual feature replacement design and set the model inputs to $\boldsymbol{x}_a$. The mediator M in equation (5) is no longer a fixed value and it is calculated on the fly with $\boldsymbol{x}_a$ as inputs. (2) CTCE V2: change the sliding window mean feature design to zero feature design without sliding window. We further verify the usage of $\boldsymbol{x}_a$ as the counterfactual input by replacing it with $\boldsymbol{x}_0 \in \mathbb{R}^{D_b}$, a zero feature vector. The results reveals that counterfactual feature replacement with $\boldsymbol{x}_a$ performs the best, showing the superiority of our design.

Loss Design. We use two other losses to verify the effectiveness of using focal loss $\mathcal{L}_{foc}$, namely Mean Squared Error loss $\mathcal{L}_{mse}$ and Binary Cross Entropy loss $\mathcal{L}_{bce}$, respectively. The performance in experiments 3, 6, 7 in Tab. 2 demonstrates that focal loss automatically penalizes the well-learned samples and focuses on the hard ones, achieving the best performance among the three losses.

Backbone Robustness. Experiments 3,8,9,10 in Tab. 2 show that the performance of our method increases as more advanced backbone network used, which indicates that our method is not dependent on a careful choice of $f(\cdot)$.

Hyperparameter Tuning. In general, there are four types of hyperparameters in our model: (1) $a\%$ and $b\%$ in the construction of the pseudo label set; (2) the size N_s of the confounder set $\boldsymbol{S}$; (3) the sliding window size d; (4) the training Rounds T. Like most work in VAD, we report the evaluation results of all hyperparameter settings. For (1), according to Tab. 3, we set the abnormal:normal sampling ratio to $1:2$, $1:3$, $1:4$, $1:5$, $1:6$ in experiments 1,2,3,4,5 and the evaluation results suggest that setting the ratio to a

Ablation Task Name	ID	Backbone	DCFD Training	CTCE	CTCE V1	CTCE V2	$\mathcal{L}_{foc}$	$\mathcal{L}_{mse}$	$\mathcal{L}_{bce}$	UCSD	
										Ped1	Ped2
Component Effectiveness	1	ResNet-50					√			70.7%	81.2%
	2	ResNet-50	√				√			73.9%	97.9%
	3	ResNet-50	√	√			√			**84.9%**	**99.4%**
CTCE Design	4	ResNet-50	√		√		√			76.8%	96.4%
	5	ResNet-50	√			√	√			73.3%	94.4%
Loss Design	6	ResNet-50	√	√				√		79.0%	97.3%
	7	ResNet-50	√	√					√	82.5%	97.6%
Backbone Robustness	8	I3D	√	√			√			84.9%	98.7%
	9	VGG	√	√			√			82.0%	94.6%
	10	C3D	√	√			√			82.7%	95.6%

Table 2: Ablation experiments to verify the effectiveness of our method. The default setting is marked in bold.

Task	Settings	ID	UCSD	
			Ped1	Ped2
Pseudo Label Set (a%, b%)	(10%, 20%)	1	84.0%	96.7%
	(5%, 15%)	2	84.1%	98.2%
	(5%, 20%)	3	**84.9%**	**99.4%**
	(5%, 25%)	4	85.9%	98.4%
	(5%, 30%)	5	86.1%	99%
Confounder Set Size N_s	4	6	82.4%	95.8%
	16	7	**84.9%**	**99.4%**
	64	8	85.4%	99.0%
	128	9	82.0%	98.3%
Sliding Window Size d	8	10	75.5%	93.5%
	64	11	75.6%	93.6%
	512	12	80.9%	98.5%
	1024	13	**84.9%**	**99.4%**
	2048	14	86.3%	97.4%

Table 3: Ablation experiments of all hyperparameter settings. The default setting is marked in bold.

smaller value produces better performance since abnormal events in real world is rare. For (2), to justify the size of the confounder set N_s, we set N_s to 4, 16, 64, and 128 in experiments 6,7,8,9, the results of which suggest that the granularity of the confounder set $\boldsymbol{S}$ matters. Setting N_s to a value that could well represent the distribution of $\boldsymbol{X}$ benefits the de-confounded training. For (3), we set the window size from short-range to long-range values and the results show that larger window size consistently outperforms smaller counterparts, suggesting that long-range temporal context is essential for robust context representation. For (4), we plot the AUC performance in Fig. 6 to show the overall trend of the self-supervised pseudo label learning process. Concretely, the results suggest that the AUC progressively improves from $t = 0 - 8$ and typically plateaus at $t = 8$. We set $T = 8$ to balance the computational cost and performance. Furthermore, the learning process from initialization to Round 1 represents learning from traditional unsupervised method to deep neural network model. The sharp performance improvement reveals the fact that deep neural networks tend to learn simple patterns first before fitting the pseudo label noise as proved in (Li, Socher, and Hoi 2020; Arpit et al. 2017). From Round 2 onward, the learning process turns to a representation refinement procedure since our model fine-tunes from the trained model of previous Round, resulting in the less sharp increase.

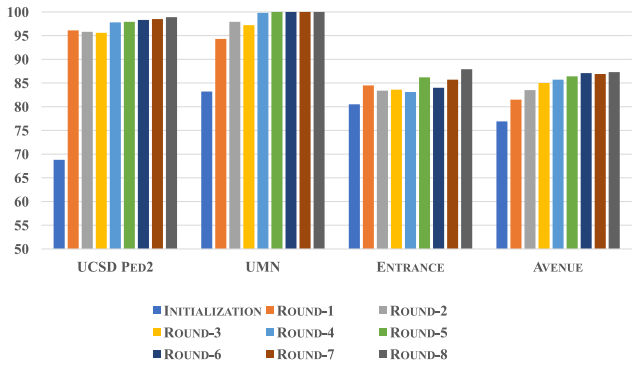

Figure 6: The AUC performance of our model at each round.

Compared to previous state-of-the-art results

Quantitative results. According to Tab. 1, we compare our method with 16 VAD methods that require labeled normal data in training and 7 UVAD methods that do not require labeled data in any form on four standard benchmark datasets. We present five models with different configurations: (1) iForest with image feature extracted from ResNet-50 in experiment 22 that serves as a simple baseline; (2) Strong Baseline $\mathcal{F}$ with $f(\cdot)$ set to ResNet-50, which is almost identical to (Pang et al. 2020) except that the training loss is replaced with $\mathcal{L}_{foc}$; (3) Ours(ResNet-50) + DCFD that differs from (2) by adding a de-confounded training module; (4) Ours(ResNet-50) + DCFD + CTCE that differs from (3) by adding a Counterfactual temporal context ensemble module; (5) Ours(I3D) + DCFD + CTCE that differs from (4) by changing $f(\cdot)$ to I3D. In general, our method outperforms all previous UVAD methods significantly and is even higher than some of the VAD methods. Concretely, we analyze the performance gain in each dataset respectively. **UCSD**: our method surpasses all previous UVAD methods significantly, 7.1% higher than the top performing model (Wang et al.

2018) for Ped1 and 3.0% higher than (Wang et al. 2018) for Ped2 by comparing experiments 25 and 20. For VAD, our method outperforms all VAD methods for Ped2 and reaches a high rank in Ped1, demonstrating the competitiveness of our d method. **Subway**: our result surpasses all previous UVAD methods in both Entrance and Exit benchmarks. **UMN**: it is clear that the performance of our method is higher than all previous UVAD methods and is competitive to the supervised methods. Specifically, we achieved 100% in All Scenes (2.6% higher than (Pang et al. 2020)). **Avenue**: our method outperforms the previous UVAD SOTA (Wang et al. 2018) by 5% by comparing experiments 26 and 20. The performance is also higher than most VAD methods, ranking the second.

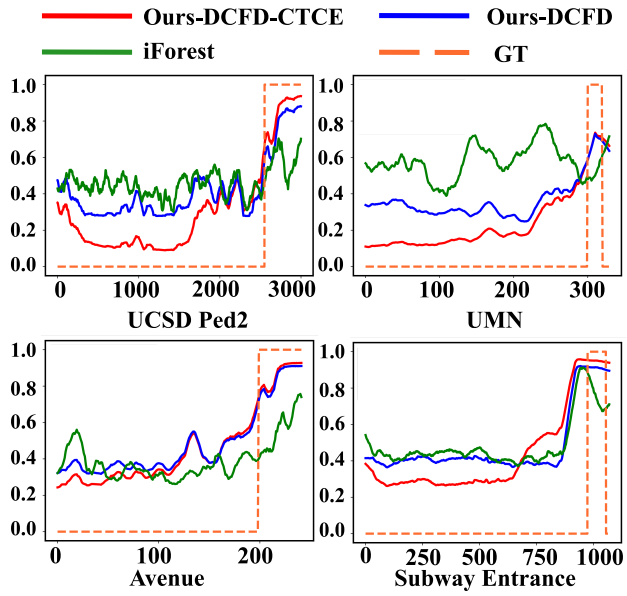

Figure 7: The qualitative results of 4 example benchmark datasets. The horizontal axis denotes the frame number of a video snapshot and the vertical axis represents the anomaly score generated by our model.

Qualitative results. Fig. 7 shows the quantitative results of our model in 4 example benchmark datasets. Compared to the iForest baseline, it is clear that our proposed method can produce better anomaly score when the event is anomalous. Extensive experiments demonstrate that our method could progressively improve the pseudo label quality (calculated by dividing the number of correct normal/abnormal frames by the number of total normal/abnormal frames) as in Fig. 8.

Conclusion

We analyzed the impact of the noisy pseudo label and long-range temporal context in unsupervised video anomaly detection from a causal inference perspective. Then, we proposed the de-confounded training and counterfactual temporal context ensemble to enhance the commonly used self-supervised pseudo label learning process in UVAD. The overall framework is simple, computationally lightweight,

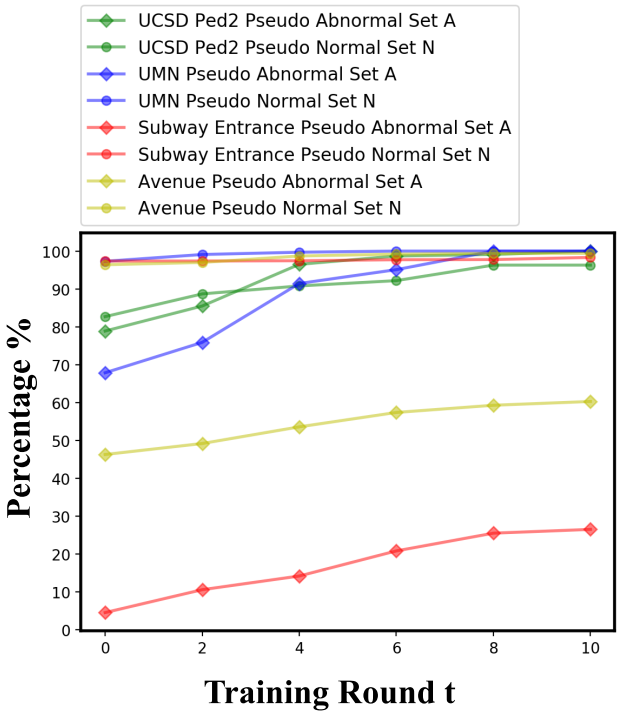

Figure 8: The progress of quality of the pseudo labels sets A and N w.r.t GT.

and robust to the noisy pseudo label. We extensively verified the effectiveness of our proposed pipeline, and experiment results on six benchmark datasets show that our method outperforms all previous methods significantly, demonstrating the superiority of our approach. Nevertheless, designing better causal graphs or feature disentanglement methods may further improve the model's performance in UVAD.

Acknowledgements

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048 and No.2020A1515010423, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463, and in part by the Guangzhou Science and technology project under Grant No.202102020633. This work was also supported by the Guangdong Provincial Key Laboratory of Big Data Computing, the Chinese University of Hong Kong, Shenzhen.

References

Adam, A.; Rivlin, E.; Shimshoni, I.; and Reinitz, D. 2008. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. *TPAMI*.

Ahmed, O.; Trauble, F.; Goyal, A.; Neitz, A.; Wuthrich, M.; Bengio, Y.; Scholkopf, B.; and Bauer, S. 2021. CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. In *ICLR*.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A. C.;

Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In *ICML*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

Chang, Y.; Tu, Z.; Xie, W.; and Yuan, J. 2020. Clustering Driven Deep Autoencoder for Video Anomaly Detection. In *ECCV*.

Chernozhukov, V.; Fernández-Val, I.; and Melly, B. 2009. Inference on counterfactual distributions. *Econometrica*.

Cong, Y.; Yuan, J.; and Liu, J. 2011. Sparse reconstruction cost for abnormal event detection. In *CVPR*.

Giorno, A. D.; Bagnell, J.; and Hebert, M. 2016. A Discriminative Framework for Anomaly Detection in Large Videos. In *ECCV*.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.; Venkatesh, S.; and Hengel, A. 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ionescu, R.; Khan, F.; Georgescu, M.; and Shao, L. 2019. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *CVPR*.

Ionescu, R.; Smeureanu, S.; Alexe, B.; and Popescu, M. 2017. Unmasking the Abnormal Events in Video. In *ICCV*.

Keith, K. A.; Jensen, D. D.; and O'Connor, B. 2020. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *ACL*.

Kim, J.; and Grauman, K. 2009. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *CVPR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Li, J.; Socher, R.; and Hoi, S. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *ICLR*.

Lin, T.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

Liu, F.; Ting, K.; and Zhou, Z. 2012. Isolation-Based Anomaly Detection. *ACM Transaction Knowledge Discovery Data*.

Liu, W.; Luo, W.; Li, Z.; Zhao, P.; and Gao, S. 2019. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies. In *IJCAI*.

Liu, W.; W. Luo, D. L.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection – A New Baseline. In *CVPR*.

Liu, X.; Yin, D.; Feng, Y.; Wu, Y.; and Zhao, D. 2021. Everything Has a Cause: Leveraging Causal Inference in Legal Text Analysis. In *NAACL*.

Liu, Y.; Li, C.; and Póczos, B. 2018. Classifier Two Sample Test for Video Anomaly Detections. In *BMVC*.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal Event Detection at 150 FPS in Matlab. In *ICCV*.

Luo, W.; Liu, W.; and Gao, S. 2017. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In *ICCV*.

Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *CVPR*.

Mehran, R.; Oyama, A.; and Shah, M. 2009. Abnormal crowd behavior detection using social force model. In *CVPR*.

Morgan, S. L.; and Winship, C. 2014. Counterfactuals and Causal Inference. In *Cambridge University Press*.

Nguyen, T.; and Meunier, J. 2019. Anomaly Detection in Video Sequence With Appearance-Motion Correspondence. In *ICCV*.

Pang, G.; Yan, C.; Shen, C.; Hengel, A.; and Bai, X. 2020. Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection. In *CVPR*.

Park, H.; Noh, J.; and Ham, B. 2020. Learning Memory-guided Normality for Anomaly Detection. In *CVPR*.

Pearl, J. 2001. Direct and Indirect Effects. In *UAI*.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Petersen, M.; Sinisi, S.; and van der Laan, M. J. 2006. Estimation of Direct Causal Effects. *Epidemiology*.

R.T.; Smeureanu, S.; Popescu, M.; and Alexe, B. 2019. Detecting Abnormal Events in Video Using Narrowed Normality Clusters. In *WACV*.

Rubin, D. 2005. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Sugiyama, M.; and Borgwardt, K. 2013. Rapid Distance-Based Outlier Detection via Sampling. In *NIPS*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*.

Sun, Q.; Liu, H.; and Harada, T. 2017. Online growing neural gas for anomaly detection in changing surveillance scenes. *PR*.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation From Biased Training. In *CVPR*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*.

Wang, S.; Zeng, Y.; Liu, Q.; Zhu, C.; Zhu, E.; and Yin, J. 2018. Detecting Abnormality without Knowing Normality: A Two-Stage Approach for Unsupervised Video Abnormal Event Detection. In *ACM MM*.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual Commonsense R-CNN. In *CVPR*.

Xu, D.; Ricci, E.; Yan, Y.; Song, J.; and Sebe, N. 2015a. Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. In *BMVC*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015b. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*.

Yu, G.; Wang, S.; Cai, Z.; Zhu, E.; Xu, C.; Yin, J.; and Kloft, M. 2020. Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. In *ACM MM*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.; and Sun, Q. 2020. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *NeurIPS*.

Zhang, Y.; Lu, H.; Zhang, L.; Ruan, X.; and Sakai, S. 2016. Video anomaly detection based on locality sensitive hashing filters. *PR*.