

# Safeguarding Sustainable Cities: Unsupervised Video Anomaly Detection through Diffusion-based Latent Pattern Learning

Menghao Zhang\*, Jingyu Wang\*, Qi Qi, Pengfei Ren, Haifeng Sun, Zirui Zhuang<sup>†</sup>, Lei Zhang and Jianxin Liao<sup>†</sup>

State Key Laboratory of Networking and Switching Technology,  
Beijing University of Posts and Telecommunications

{zhangmenghao, wangjingyu, qiqi8266, rpf, hfsun, zhuangzirui, zhanglei, liaojx}@bupt.edu.cn

## Abstract

Sustainable cities requires high-quality community management and surveillance analytics, which are supported by video anomaly detection techniques. However, mainstream video anomaly detection techniques still require manually labeled data and do not apply to real-world massive videos. Without labeling, unsupervised video anomaly detection (UVAD) is challenged by the problem of pseudo-labeled noise and the openness of anomaly detection. In response, a diffusion-based latent pattern learning UVAD framework is proposed, called DiffVAD. The method learns potential patterns by generating different patterns of the same event through diffusion models. The detection of anomalies is realized by evaluating the pattern distribution. The different patterns of normal events are diverse but correlated, while the different patterns of abnormal events are more diffuse. This manner of detection is equally effective for unseen normal events in the training set. In addition, we design a refinement strategy for pseudo-labels to mitigate the effects of the noise problem. Extensive experiments on six benchmark datasets demonstrate the design's promising generalization ability and high efficiency. Specifically, DiffVAD obtains an AUC score of 81.9% on the ShanghaiTech dataset.

## 1 Introduction

The development of sustainable cities requires high-quality community management, which is strongly supported by surveillance technology. Through surveillance systems, community managers can monitor activities in their neighborhoods and effectively ensure that normal activities are taking place. To reduce the cost of human supervision, Video Anomaly Detection (VAD), which is designed to automatically detect abnormal activities [Wu *et al.*, 2021], has gained increasing attention.

In the real world, anomalous events are rare and unconstrained, which makes collecting and labeling enough anomalous events nearly impossible. Accordingly, most previous

\*Equal Contribution.

<sup>†</sup>Corresponding Author.

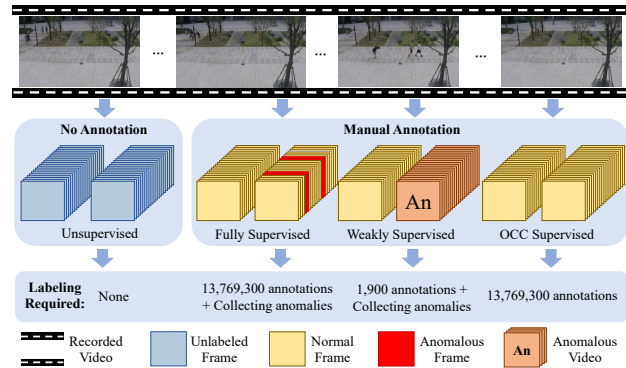


Figure 1: Different supervised settings for video anomaly detection. From left to right: unsupervised setting without any training data annotation, fully supervised setting requiring frame-level annotation in the training data, weakly supervised setting requiring video-level annotation, and OCC supervised setting requiring only normal data (frame-level). The labeling requirements are illustrated with the real-world dataset UCF-Crime.

methods detect anomalies in a weakly supervised setting that learns with video-level labels [Sultani *et al.*, 2018; Tian *et al.*, 2021], or One-Class Classification (OCC) supervised setting that only learns from normal data [Wang *et al.*, 2022; Chen *et al.*, 2022]. The OCC supervised VAD refers to model the distribution of normal pattern with normal data, labeling the events that deviate from the model as anomalous events. However, labeling videos or labeling training sets as purely normal events is still labor-intensive and costly, especially when dealing with massive amounts of surveillance video in the real world. To alleviate this problem, a natural idea is to perform **Unsupervised VAD (UVAD)** which aims to detect anomalies directly from completely unlabeled videos as shown in Figure 1. Several recent works [Giorno *et al.*, 2016; Pang *et al.*, 2020; Yu *et al.*, 2022; Zaheer *et al.*, 2022] have explored this approach. Figure 1 illustrates the different supervised settings in VAD.

Despite the progress made, two prominent limitations constrain further improvements in the performance of existing UVAD methods. (1) Representative methods [Yu *et al.*, 2022; Zaheer *et al.*, 2022] generate pseudo-labels by reconstruction or simple classification, and the model learns the distribution

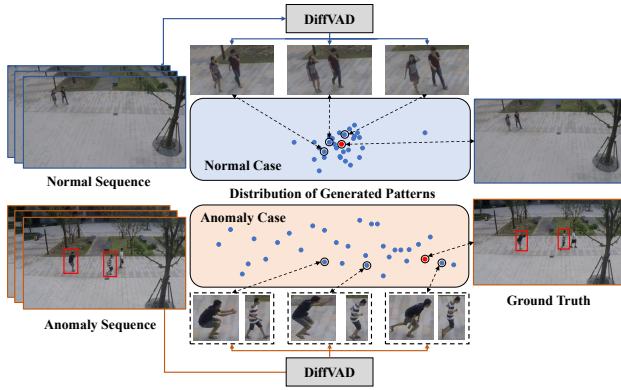


Figure 2: DiffVAD detects anomalies by generating and statistically summarizing latent patterns. Conditioned on history frames (left frames). The blue (top) and red (bottom) distributions represent normal and abnormal generations (mapped by t-SNE 2d). In the distribution of patterns, the red dots are the actual real futures (right frames) corresponding to the history frame condition. In the normal case, the real future lies within a dominant distribution pattern and the generated latent patterns are biased toward ground truth. In the anomalous case (red boxes mark running and jumping anomalies), the real future lies at the edge of the distribution pattern, generating diverse but scattered latent patterns.

of normal patterns under the supervision of pseudo-labels. However, the noise in the pseudo-labels (incorrect pseudo-labels) limits the effectiveness of model. (2) Existing UVAD methods attempt to overfit the training set for learning normal patterns, ignoring the fact that unsupervised anomaly detection is essentially an open-set problem. In the unsupervised setting, the training set cannot include all normal patterns, thus only fitting the training set will make the model treat the unseen normal events as abnormal events.

Inspired by recent studies [de Moraes *et al.*, 2019; Chen *et al.*, 2022; Flaborea *et al.*, 2023] of the distribution of features of abnormal skeletal anomalies, we explore the distribution of latent patterns of abnormal and normal events. As shown in Figure 2, we find distributional difference in the latent patterns. In the case of normality, different patterns of the same event are varied but correlated, i.e., they are biased towards the ground truth of what actually happens. In the case of abnormality, different patterns of the same event are diverse but not correlated. The distinctions outlined above offer the potential to identify diverse normality and abnormality within the real world. Nevertheless, exploiting this discrepancy requires the generation of multiple diverse patterns for the same occurrence. Introducing diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020] into UVAD might fulfill this requirement. Diffusion models are characterized by minor modifications and corrections to the generated samples at each step, thus having the ability to learn multiple latent patterns or forms of the same action or event.

To this end, we propose a latent pattern learning framework for UVAD based on diffusion models, called DiffVAD, which comprehensively consider the diversity of normal and abnormal patterns. Given a video sequence that is either normal or abnormal, the video frames are corrupted to random noise,

after which multiple patterns of the corrupted frames are reconstructed based on the clean input frame in the past. DiffVAD distinguishes between normal and abnormal by comparing the distribution of different patterns, with different normal patterns for the same event tending to be more correlated. To enhance this correlation, we condition the reconstruction of damaged frames on past motion features. Furthermore, we refine the pseudo-labels based on knowledge aggregation from neighbours sample to reduce the effect of noise. Since noise in pseudo-labels is unavoidable, we evaluate the reliability of the refined pseudo-labels to penalize unreliable pseudo-labels to guide learning through reliable pseudo-labels.

The main contributions of the article are as follows:

- A diffusion-based latent pattern UVAD learning framework is proposed for sustainable cities. We exploits the motion conditioned diffusion model to generate multiple patterns of the same event to account for the openness of anomaly detection.
- A pseudo-label refinement strategy based on knowledge aggregation is proposed to alleviate the effect of noise. We weight the loss based on the reliability of pseudo-labels to train the model with reliable pseudo-labels.
- Extensive experiments on six benchmark datasets show that our method achieves state-of-the-art performance and even outperforms the OCC methods.

## 2 Related Work

### 2.1 Video Anomaly Detection

**VAD as One-Class Classification.** Most previous work considers VAD as one class classification task due to the difficulty of collecting and labeling abnormal samples. The detection model learns the feature distribution of the normal pattern only from the normal samples during training, and deviation points that do not conform to the model representation are labeled as anomalies during detection. Reconstruction-based methods [Lv *et al.*, 2021] and prediction-based methods [Liu *et al.*, 2018a; Wang *et al.*, 2022] are the two mainstream methods. These two methods typically use autoencoders (AEs) [Hasan *et al.*, 2016], memory-augmented AEs [Gong *et al.*, 2019; Lv *et al.*, 2021] to reconstruct predict frames so that frames with large reconstruction or prediction errors are recognized as anomalies.

**Weakly Supervised VAD.** Unlike OCC methods that require frame-level annotations, some work [Sultani *et al.*, 2018; Feng *et al.*, 2021; Tian *et al.*, 2021; Wu *et al.*, 2021; Zhang *et al.*, 2023] has attempted to train models with video-level annotations. Video-level labels are provided in such a way that normal labeled videos are completely normal, while abnormal labeled videos contain both normal and abnormal content. Mainstream weakly supervised methods [Feng *et al.*, 2021; Lv *et al.*, 2023] detect video-level anomalies through multi-instance learning.

**Unsupervised VAD.** To address the need for detection of large amounts of unlabeled surveillance video in the real world, limited recent work [Giorno *et al.*, 2016; Pang *et al.*, 2020; Yu *et al.*, 2022; Zaheer *et al.*, 2022] has explored UVAD. For instance, Del *et al.* [Giorno *et al.*, 2016] pioneer

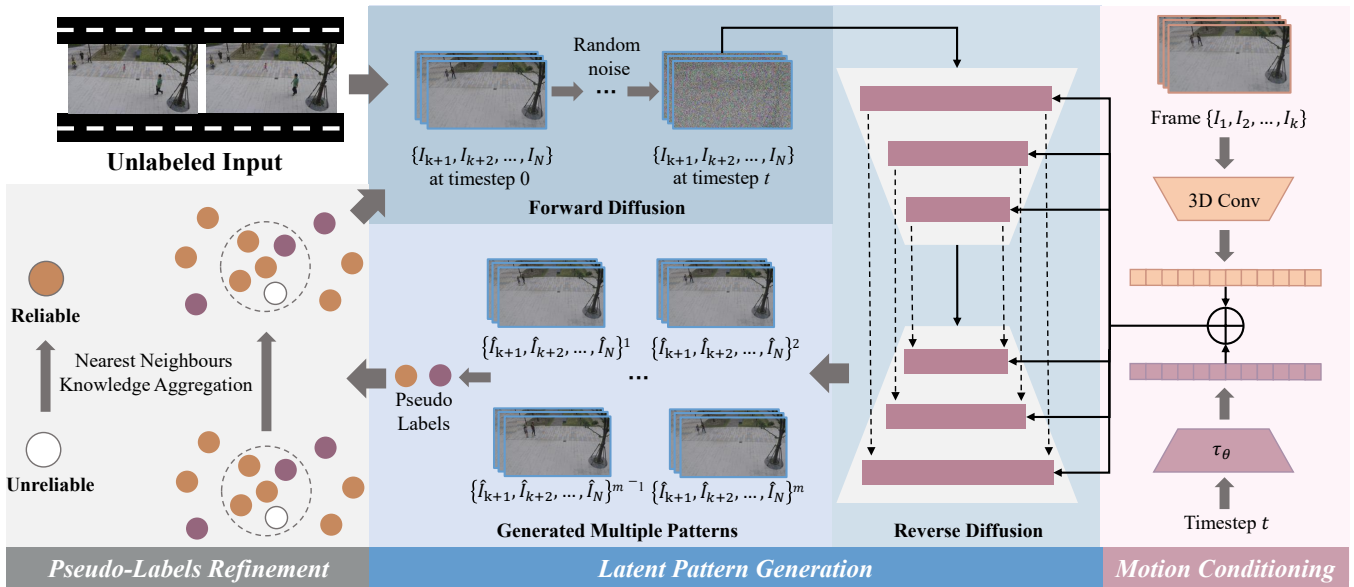


Figure 3: DiffVAD framework overview. The unlabeled input frames are first added with random noise step by step in a forward diffusion process and then the noisy video frames are fed into the generator to perform the reverse process. In the reverse diffusion process, the generator denoises the noisy video and generates  $m$  latent patterns based on the motion features of the history frames. For the multiple patterns generated, we compute the distance and mean of each pattern relative to the ground truth. Pseudo-labels are assigned based on the distribution of latent patterns, with more discrete distributions representing that the frame is more likely to be anomalous. Finally before starting the next round of training, we refine the generated pseudo-labels to make them more reliable. Best viewed in color.

the exploration of UVAD by detecting dramatic changes in the anomalies. Pang *et al.* [Pang *et al.*, 2020] obtain preliminary results by using pre-trained DNNs then refine the results by performing two classes of ordered regression in a self-training manner. Yu *et al.* [Yu *et al.*, 2022] propose the normality advantage and extend the reconstruction from OCC supervised VAD to UVAD. Zaheer *et al.* [Zaheer *et al.*, 2022] proposes a cooperative learning framework to generate pseudo-labels for self-training of detection models.

Although the above methods can detect anomalies from unlabeled data through self-training and pseudo-labeling, the noise in the pseudo-labels limits the improvement of performance. Meanwhile, these methods fit the distribution of the training set without considering the unseen normal patterns. In contrast to the above methods, we refine the pseudo-labels and estimate their uncertainty to eliminate noise. Furthermore, diffusion model [Ho *et al.*, 2020] is introduced to account for the diversity of normal and abnormal patterns.

## 2.2 Diffusion Model

Recently, diffusion models [Ho *et al.*, 2020] have achieved superior performance in many generative tasks, outperforming GANs [Kaneko and Harada, 2020] and AEs [Kim, 2022]. Diffusion models have also found many applications in computer vision tasks such as image restoration [Song *et al.*, 2021], image super-resolution [Metzger *et al.*, 2023], and image editing [Yang *et al.*, 2023a]. For VAD, Yan *et al.* [Yan *et al.*, 2023] introduces diffusion models to improve anomaly detection. They utilizes the robustness of the diffusion model to noise to predict and reconstruct video frames in feature space without resorting to other auxiliary networks. How-

ever, this OCC work is only a single-pattern reconstruction of events and is not applicable to UVAD with openness. The latest work [Flaborea *et al.*, 2023] utilizes diffusion models to generate unbounded skeletal motion to detect anomalies, but its applicability is limited to human-related anomalies.

Inspired by generating unbounded data [Flaborea *et al.*, 2023], for unlabeled videos, our work covers multiple patterns of normal and anomalies by latent pattern learning of diffusion models. Differently, our approach learns unbounded patterns rather than being limited to skeletal motion. In particular, to enable the model to accurately learn abnormal and normal from unlabeled data, we propose a pseudo-label refinement strategy to alleviate the effect of noise.

## 3 Methodology

### 3.1 Overall Framework

Figure 3 depicts the overall architecture of proposed framework. The diffusion-based latent pattern learning framework for UVAD consists of a diffusion-based generator (performing forward diffusion process and reverse diffusion process) and pseudo-label refinement module. In addition, we introduce motion conditions in the generator network to help the network generate more relevant multiple patterns for the input frames. Specifically, we condition on the motion features of the history frame sequence that preceded the input frame sequence. Pseudo-labels for input frames can be generated by observing the distribution of generated patterns (counting the distances of potential patterns from the ground truth); discrete distributions are more likely to be anomalous. The whole framework is trained under the supervision of pseudo-labels.

### 3.2 Generation of Latent Patterns

We define a diffusion-based generative network that learns to reconstruct multiple latent patterns for corrupted future frame sequences under the condition of clean past frame sequences.

Let  $I = \{i_1, \dots, i_N\}$  be a sequence belonging to  $N$  time-continuous frames. We divide  $X$  into two parts: a sequence of past history frames  $I_{1:k}$  and a future sequence  $I_{k+1:N}$  with  $k \in \{1, \dots, N\}$ .

In the forward process  $q$ , we corrupt the sequence of future video frames by adding random noise. We sample a random noise map  $\varepsilon_{k+1:N}$  from the distribution  $\mathcal{N}(0, \mathbf{I})$  and add it to  $I_{k+1:N}$  to randomly destroy its pixels.

The extent of added noise depends on the variance scheduler  $\beta_t \in (0, 1)$  and the diffusion time step  $t$ . As a result,  $q$  at each diffusion time step  $t$  makes  $i_n^{t=T}$  indistinguishable from a noisy frame with random sampling.

The reverse process  $p_\theta$  performs denoising of the noisy frames and estimates the noise map  $\varepsilon_{k+1:N}$  through the U-net structure  $\varepsilon_\theta$ . The architecture of the main diffusion model is the neural network represented by the purple blocks in Figure 3, tasked with estimating the noise in the sequence of input frames and consequently generating multiple latent patterns. Our diffusion network of latent patterns gradually shrinks and then expands (reconstructs) the spatial dimension of the input frame sequence. To take into account the temporal dimension of the input sequence, we construct a U-Net with the spatio-temporally separable GCN (STS-GCN) layer proposed in [Sofianos *et al.*, 2021]. In detail, the U-Net receives the input  $I_{k+1:N}$  and the motion time conditioning signal  $h + \tau_\theta(t)$ , which provides the network with the diffusion time step and encoded features of  $I_{1:k}$ . Furthermore, to align the dimension of this conditioning signal with the dimension of the network layer, the former is fed to the embedding layer, which projects it to the correct vector space. This embedded conditioning signal is then fed to each STS-GCN layer.

Formally, to obtain an approximation of  $\varepsilon_{k+1:N}$ , we define the approximation objective conditioned on the diffusion time step  $t$  and the feature embedding  $h$  of  $I_{1:k}$ . We define the approximation objective as [Flaborea *et al.*, 2023]:

$$\mathcal{L}_{\text{appr}} = \mathbb{E}_{t, I, \varepsilon} [|\varepsilon - \varepsilon_\theta(I_{1:k}^t, t, h)|]. \quad (1)$$

We smooth  $\mathcal{L}_{\text{appr}}$  as follows:

$$\mathcal{L}_{\text{smooth}} = \begin{cases} 0.5 \cdot (\mathcal{L}_{\text{appr}})^2 & \text{if } |\mathcal{L}_{\text{appr}}| < 1 \\ |\mathcal{L}_{\text{appr}}| - 0.5 & \text{otherwise} \end{cases} \quad (2)$$

Given a random noise sample  $z$  and consider it as the start for generating latent patterns  $I_{k+1:N}^{P=m}$  of the input frame sequence.  $P = m$  represents  $m_{th}$  pattern with  $m \in \{1, \dots, M\}$ . We perform the generation by [Flaborea *et al.*, 2023]:

$$I_{k+1:N}^{m,t-1} = \frac{1}{\sqrt{\alpha_t}} \left( I_{k+1:N}^{m,t-1} - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \varepsilon_\theta(I, t, h) \right) + z \sqrt{\beta_t}. \quad (3)$$

We generate  $M$  distinct latent patterns  $I^1, \dots, I^M$ . For each  $I^m$ , we compute the reconstruction error by means of the smoothing loss  $\mathcal{L}_{\text{smooth}}(|I - I^m|)$  used in training.

**Conditions of Generation.** The conditioning strategy is a

key element of the diffusion model we designed, as it directly affects the quality of the output. Depending on the features of the history  $k$  frames, the latent patterns of the output can be more relevant to the ground truth, helping to distinguish between normal and anomalous.

The right part in Figure 3 illustrates our conditioning strategy, where we use 3D convolution on  $I_{1:k}$  to extract features embedding  $h$ . 3D convolution to extract features as conditioning signals is simple and effective. Since the diffusion models benefits from being conditional on the time step  $t$ , we add the embedding  $\tau_\theta(t)$  to the embedding  $h$  and feed the generated motion time signal to each layer of our network  $\varepsilon_\theta$ .  $\tau_\theta(t)$  is implemented with a MLP.

### 3.3 Generation of Pseudo-labels

In our proposed UVAD framework, the pseudo-labels generated from this round of training are used for the next round of training. The pseudo-labels are generated on the basis of observing the distribution of the generated latent patterns.

$M$  different latent patterns  $I^1, \dots, I^M$  is generated by  $\varepsilon_\theta$ , we compute the mean of the distances of the generated latent patterns relative to the ground truth by smoothing the loss:

$$\mathcal{D}_I = \frac{1}{M} \sum_{i=1}^M \mathcal{L}_{\text{smooth}}(|I - I^m|). \quad (4)$$

We use  $D^{th}$  as the thresholds for generator  $\varepsilon_\theta$ ,  $\hat{y}_I$  as the pseudo-label of  $I$ :

$$\begin{cases} \hat{y}_I = 1, & \text{if } \mathcal{D}_I \geq D^{th} \\ \hat{y}_I = 0, & \text{otherwise} \end{cases} \quad (5)$$

Inspired by [Zaheer *et al.*, 2022], we choose to generate thresholds by a data-driven approach. For the output of generator,  $\mu_D$  and  $\sigma_D$  is the mean and standard deviation of distance  $\mathcal{D}$  for each batch. The  $D^{th}$  is as follow:

$$D^{th} = \mu_D + 0.1\sigma_D. \quad (6)$$

This approach of using a fixed percentage of the maximum distance of the generated latent patterns relative to the ground truth as the threshold is simple and effective. In addition we have tried other implementations, such as continuously adjusting the thresholds with training. These methods have the same effect as the approach of Eq. (9), which shows that our method is robust to the selection of thresholds.

### 3.4 Refinement Strategy of Pseudo-Labels

The refinement of pseudo-labels is done by aggregating the knowledge of the nearest neighbor samples. We assume that features of semantically similar images should be neighboring in the feature space. This assumption is satisfied by extracting similar features from the samples using contrastive learning. The strategy is shown on the left side of Figure 3.

Formally, given a video frame  $i_n$  and a feature extractor  $f_\theta$ , we obtain a feature vector  $z = f_\theta(i_n)$  from the video frame  $i_n$ . The feature  $z$  is then used to search for neighbors of the sample  $i_n$  in the feature space. Thus, the pseudo-label of  $i_n$



is refined by aggregating knowledge from the selected neighbors. For this purpose, the anomaly probabilities  $\hat{p}_i$  outputs of the selected neighbors are averaged to perform a soft voting:

$$\hat{p}_i = \frac{1}{K} \sum_{i \in \mathcal{I}} p'_i, \quad (7)$$

where  $\mathcal{I}$  is the index set of the selected neighbors, and  $\hat{p}_i$  is the average score of vector  $z_i$ . To obtain fine pseudo-labels, we use the argmax operation on  $\hat{p}_i$ :

$$\hat{y}_i = \arg \max \hat{p}_i. \quad (8)$$

The above refinement process entails searching a representation of the feature space in which neighboring samples are searched. This is allowed by the bank, which stores the features obtained from the samples and the original anomaly probabilities. Neighbors are then selected by calculating the cosine distance between the features of  $i_n$  and the features stored in the bank. According to [He *et al.*, 2020], in order to keep the information stored in the bank more stable, we use the slowly varying momentum model  $g'_i(\cdot) = h'_i(f'_i(\cdot))$  to update the feature  $z$  and the anomaly probabilities  $p'$ .

The refined pseudo-label  $\hat{y}_i$  is used as a self-supervised signal for loss calculation during training. Since refinement is an iterative process, the refined pseudo-labels obtained through knowledge aggregation of neighbors still contain some noise. To address this problem, we reweight the training loss by estimating the uncertainty of the pseudo-label refinement.

Following [Litrico *et al.*, 2023], we exploit an entropy-based uncertainty estimation method that exploits the consistency between neighboring predictions. The basic idea is that if the network predicts the same kind of neighboring samples, we can consider the derived pseudo-labels to be reliable (low uncertainty). The average score vector  $\hat{p}_i$ , obtained by averaging the probabilistic outputs of the neighbors, has low entropy in the former case while high entropy in the latter. Therefore, we reweight the training loss by computing the weight  $w$ . The weight  $w$  is more important for pseudo-labels obtained from  $\hat{p}_i$  with low entropy and less important for pseudo-labels obtained from  $\hat{p}_i$  with high entropy.

Formally, given the frame  $i_n$ , we obtain the average score vector  $\hat{p}_i$  from the soft-voting strategy. note that  $\hat{p}_i$  is obtained by averaging over a probability distribution, thus it is still a probability distribution. We compute the entropy of  $\hat{p}_i$  as:

$$\mathcal{H}(\hat{p}_i) = \mathbb{E}[I(\hat{p}_i)] = -\hat{p}_i \log_2 \hat{p}_i. \quad (9)$$

Based on the normalized entropy value  $\mathcal{H}(\hat{p}_i)$ , we obtain the weight  $w$  of the sample  $i_n$  as:

$$w_{i_n} = \exp\left(-\hat{\mathcal{H}}(\hat{p}_i)\right) \quad (10)$$

### 3.5 Training and Inference

**Training.** Based on the pseudo-labels, we generate latent patterns for input data and use  $\mathcal{L}_{\text{smooth}}$  weighted according to uncertainty in order to encourage to generate as more relevant patterns as possible for pseudo-labeled normal data and as more irrelevant patterns as possible for pseudo-labeled abnormal data. After updating the parameters of  $\varepsilon_\theta$ , we generate new pseudo-labels for this batch of data. Before starting a

new round of training, the pseudo-labels need to be refined and uncertainty estimated.

**Inference.** For the input sequence of unlabeled frames, we generate  $M$  latent patterns for each frame as trained and count their distance  $\mathcal{D}$  from the ground truth. We consider the mean as aggregation statistics for the distance  $\mathcal{D}$  between the generated latent patterns and the ground truth. Latent patterns for normal events are more correlated, and the distribution of latent patterns for anomalous events is more discrete.

## 4 Experiments

### 4.1 Implementation Details

#### Datasets

Six real-world benchmark datasets are used to evaluate our approach, including the ShanghaiTech dataset [Luo *et al.*, 2017], CUHK Avenue dataset [Lu *et al.*, 2013], UCSD dataset [Mahadevan *et al.*, 2010], Subway dataset [Adam *et al.*, 2008], UMN dataset [Mehran *et al.*, 2009], and UCF-Crime dataset [Sultani *et al.*, 2018]. These datasets correspond to different scenarios, including scenarios such as schools, pedestrian streets, subway entrances and exits, and city streets. In particular, the UCF-Crime dataset contains some illegal activities that harm sustainable cities.

Note that the training set in these datasets contains only normal events, with abnormal activity occurring only in the test set. To perform UVAD, we employ two types of UVAD settings based on previous UVAD work: (i) **Partial mode** [Giorno *et al.*, 2016; Liu *et al.*, 2018b; Yu *et al.*, 2022]: only the original test set of the dataset is used for learning, while the original training set is discarded. (ii) **Merged mode** [Pang *et al.*, 2020; Zaheer *et al.*, 2022]: the original training and test sets are merged into one unlabeled set for learning. For both modes, labels are strictly not used in learning. Performance evaluation is performed only on the original test set to allow comparison with existing VAD methods.

#### Evaluation Metrics

Following the existing method [Pang *et al.*, 2020; Lin *et al.*, 2022; Zaheer *et al.*, 2022], the area under the ROC curve (AUC) is used for evaluation and comparison. The AUC is calculated based on the frame-level annotations of the test videos in each dataset.

#### Training Details

We resize each video frame to  $256 \times 256$  and normalize it to the range  $[-1, 1]$ . We set the numbers of generated latent patterns  $M$  to 10. Both the generator network and the condition extractor network are trained with AdamW-based stochastic gradient descent method. The learning rate, and weight decay were set to 0.0002, and 0.0001, respectively. Training epochs are set to 60; 60; 10; 25; 30; 15 on Ped1, Ped2, Avenue, ShanghaiTech, Subway, UMN and UCF-Crime, respectively. We use pre-trained Resnext101 [Xie *et al.*, 2017] as the extractor for motion features in experiments.

### 4.2 Results

#### Quantitative Results

We compare our method with state-of-the-art OCC supervised VAD methods and UVAD methods on six benchmark

Supervision	Method	ShTech	Avenue	UCSD		Subway		UMN				UCF Crime
				Ped1	Ped2	Entrance	Exit	Scene1	Scene2	Scene3	All Scenes	
OCC Supervised	SRC [Cong <i>et al.</i> , 2011]	-	-	-	-	80.0	83.0	99.5	97.5	96.4	97.8	-
	LSHF [Zhang <i>et al.</i> , 2016]	-	-	87.0	91.0	-	-	99.2	98.3	99.9	99.7	-
	GNG [Sun <i>et al.</i> , 2017]	-	-	93.8	94.1	-	-	99.8	99.3	99.9	99.7	-
	FFP [Liu <i>et al.</i> , 2018a]	72.8	85.1	83.1	95.4	-	-	-	-	-	-	-
	MemAE [Gong <i>et al.</i> , 2019]	71.2	83.8	-	94.1	-	-	-	-	-	-	-
	OCAA [Ionescu <i>et al.</i> , 2019a]	78.7	90.4	-	97.8	-	-	-	-	-	-	-
	DAE [Ionescu <i>et al.</i> , 2019b]	-	-	-	-	93.5	95.1	99.9	98.2	99.8	99.3	-
	MLEP [Liu <i>et al.</i> , 2019]	76.8	92.8	-	-	-	-	-	-	-	-	-
	PMem [Park <i>et al.</i> , 2020]	70.5	88.5	-	97.0	-	-	-	-	-	-	-
	SSMT [Georgescu <i>et al.</i> , 2021]	82.4	92.8	85.1	96.9	-	-	-	-	-	-	-
	DPU [Lv <i>et al.</i> , 2021]	73.8	89.5	-	99.8	-	-	-	-	-	-	-
	HF <sup>2</sup> -VAD [Liu <i>et al.</i> , 2021]	76.2	91.1	-	99.3	-	-	-	-	-	-	-
	ROADMAP [Wang <i>et al.</i> , 2022]	76.6	88.3	83.4	96.3	95.2	95.5	-	-	-	99.1	72.9
	BDPN [Chen <i>et al.</i> , 2022]	78.1	90.3	-	98.3	-	-	-	-	-	-	-
	DLAN [Yang <i>et al.</i> , 2022]	74.7	89.9	-	97.6	-	-	-	-	-	-	-
	SCAE [Cao <i>et al.</i> , 2023]	79.2	86.8	-	-	-	-	-	-	-	-	-
	USTN-DSC [Yang <i>et al.</i> , 2023b]	73.8	89.9	-	98.1	-	-	-	-	-	-	-
	FPDM [Yan <i>et al.</i> , 2023]	78.6	90.1	-	-	-	-	-	-	-	-	74.7
	<b>DiffVAD<sub>OCC</sub> (Ours)</b>	<b>82.9</b>	<b>90.3</b>	<b>90.1</b>	<b>99.3</b>	<b>96.0</b>	<b>97.2</b>	<b>99.9</b>	<b>99.9</b>	<b>99.9</b>	<b>99.9</b>	<b>77.1</b>
Unsupervised	ADF [Giorno <i>et al.</i> , 2016]	-	-	59.6	57.6	74.6	87.2	80.2	88.3	77.1	84.8	-
	Unmask [Ionescu <i>et al.</i> , 2017]	-	-	68.4	82.2	70.6	85.7	99.3	87.7	98.2	95.1	-
	CTS [Liu <i>et al.</i> , 2018b]	-	81.1	69.0	87.5	71.6	93.1	-	-	-	95.2	-
	DAW [Wang <i>et al.</i> , 2018]	-	85.3	77.8	96.4	-	84.5	-	-	-	-	-
	STDOR [Pang <i>et al.</i> , 2020]	-	-	71.7	83.2	88.1	92.7	99.9	99.9	99.7	97.4	-
	CIL-UVAD [Lin <i>et al.</i> , 2022]	-	90.3	84.9	98.7	91.3	97.6	99.9	99.9	99.8	99.2	-
	NASP [Yu <i>et al.</i> , 2022]	71.9	89.7	79.4	97.0	-	-	-	-	-	-	-
	GCL [Zaheer <i>et al.</i> , 2022]	78.9	85.2	-	97.9	-	-	-	-	-	99.3	71.0
	<b>DiffVAD<sup>-</sup> (Ours)</b>	<b>83.1</b>	<b>91.1</b>	<b>90.2</b>	<b>99.2</b>	<b>93.2</b>	<b>99.0</b>	<b>99.9</b>	<b>99.9</b>	<b>99.9</b>	<b>99.9</b>	<b>77.0</b>
	<b>DiffVAD<sup>+</sup> (Ours)</b>	<b>81.9</b>	<b>90.3</b>	<b>87.6</b>	<b>98.9</b>	<b>92.1</b>	<b>98.2</b>	<b>99.9</b>	<b>99.9</b>	<b>99.9</b>	<b>99.9</b>	<b>75.2</b>

Table 1: Comparisons with state-of-the-art methods including OCC supervised methods that require labeled normal data in the upper blocks and unsupervised methods that do not require labeled data in the bottom block. The numbers are the frame-level AUC(%) performance.

datasets. For a fair comparison, we provide three versions of DiffVAD under different training settings. DiffVAD<sub>OCC</sub> represents the OCC supervised version trained with normal data only. DiffVAD<sup>-</sup> and DiffVAD<sup>+</sup> represent the unsupervised versions in partial mode and merged mode, respectively. The results are summarized in Table 1. Generally, our proposed method is significantly better than all previous UVAD methods, and even higher than most OCC supervised methods. This demonstrates the proposed method is perfectly suited to the need for anomaly detection in real-world scenarios. In the unsupervised setting, DiffVAD<sup>-</sup> outperforms DiffVAD<sup>+</sup> on all datasets. This is because partial mode contains less data, i.e., fewer unseen events, than merged mode.

### Qualitative Results

The examples in Figure 4 show anomaly curves of the testing video from the ShanghaiTech dataset compared among MemAE [Gong *et al.*, 2019], PMem [Park *et al.*, 2020], GCL [Zaheer *et al.*, 2022] and DiffVAD. MemAE and PMem are both OCC supervised VAD methods based on memory modules, while GCL is an unsupervised VAD method. All three methods perform single-pattern reconstruction of the input data. The anomaly curve shows the anomaly scores of all frames in the video in turn, allowing a more intuitive comparison of the performance of different methods. It can be seen that DiffVAD performs significantly better than the other methods. The anomaly scores of DiffVAD are lower and more stable on normal segments. On abnormal segments, the anomaly score curves of the other three methods have obvious errors whilst DiffVAD has longer abnormal durations and more accurate anomaly scores. The curves shown in the Fig-

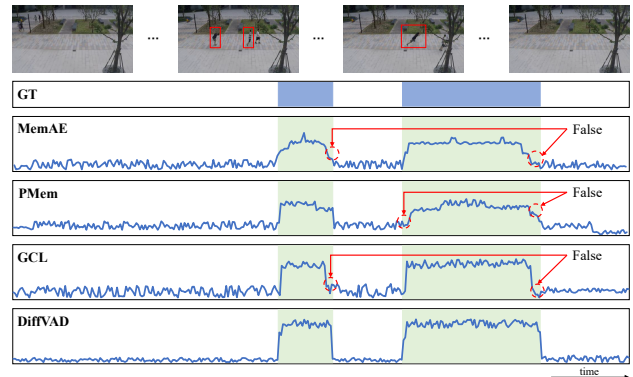


Figure 4: The example of anomaly detection comparison on ShanghaiTech. From top to bottom, we show the sampled video frames, ground-truth abnormal sections (blue regions are abnormal), result of MemAE, result of PMem, result of GCL and result of DiffVAD.

ure 4 are consistent with the results in Table 1.

### 4.3 Ablation Study and Discussion

As shown in Table 2, to validate each component in the proposed DiffVAD effectiveness in detail, the ablation study is conducted on the ShanghaiTech dataset. Not using any components represents a binary classification of abnormal and normal. It can be seen that both training the model through the classification task and performing single pattern reconstruction for training perform poorly. A significant improvement in AUC scores was obtained by designing multiple latent pattern generation and diffusion models to form a latent

Generator	Latent Patterns Generation	Single Pattern Reconstruction	Pseudo-Labels Refinement	AUC(%)
Diffusion model	✓			61.4
	✓	✓	✓	78.1 72.4 <b>81.9</b>
VAE	✓			60.1
	✓	✓	✓	74.7 70.1 <b>76.3</b>
GANs	✓			59.6
	✓	✓	✓	74.2 68.9 <b>77.0</b>

Table 2: Ablation study results on ShanghaiTech dataset. Anomaly detection performance is reported by AUC.

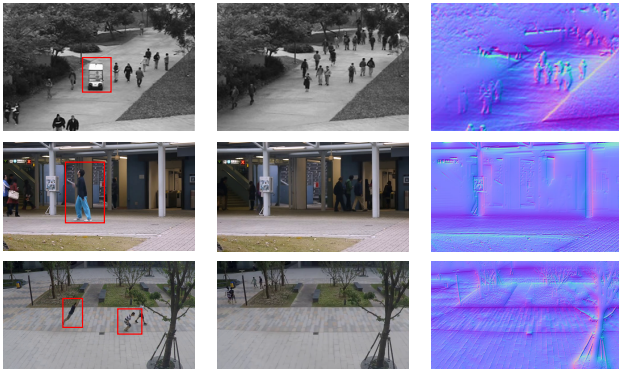


Figure 5: Visualization examples of normalcy maps of the multiple latent patterns on (top to bottom) UCSD Ped2, CUHK Avenue, and ShanghaiTech: the abnormal frames (left); the normal frames in the same scene (middle); visualization of normalcy maps of the multiple latent patterns (right).

pattern learning framework.

Latent pattern learning improves 16.7% and 5.7% compared to classification and single pattern reconstruction, respectively. The effectiveness of the pseudo-labels refinement is demonstrated by the 3.8% increase in score after introducing it during training. This is explained by the fact that more reliable pseudo-labels help the model to learn the feature representations more correctly and also help to produce the correct training loss weights. The results in Table 2 illustrate that the components of our design fit each other and justify our design’s rationality.

In addition to the ablation experiments with diffusion models as the generator, we also experiment with VAE [Kim, 2022] and GANs [Kaneko and Harada, 2020] as generators in the same setup. As shown in Table 2, both VAEs and GANs performed less than 80% as generators, while the diffusion model achieved a score of 81.9%. The superior performance of the diffusion-based generator demonstrates the suitability of the diffusion model for the UVAD task. It is notable that the latent pattern learning and pseudo-labels refinement also improve the score by about 14% and 3% when other models as the generator.

## Evaluation of Latent Patterns

Unlike previous work [Gong *et al.*, 2019; Park *et al.*, 2020; Liu *et al.*, 2021] that performs single pattern reconstruction, our proposed DiffVAD generates multiple latent patterns of the same event. In the case of normality, different patterns of the same event are varied but correlated, i.e., they are biased towards the ground truth of what actually happens. In the case of abnormality, different patterns of the same event are diverse but not correlated. Figure 5 shows the normal frames, the abnormal frames in the same scene and the visualization of normalcy maps of the multiple latent patterns on UCSD Ped2, Avenue and ShanghaiTech. We can see that (1) the normalcy maps of the multiple latent patterns is more similar to the normal frame than abnormal frames, indicating that the latent patterns of normal events is more biased toward the ground truth of what really happens. (2) There is no anomalous events in the normalcy map, such as vehicles, jumping and running. Anomalous features in anomalous frames are not reconstructed due to the smaller similarity. Therefore, even in the same scene, the latent pattern of normal events would not be similar to the anomalous one.

## Analysis of Pseudo-Labels Refinement Design

The pseudo-label refinement strategy we devise not only produces finer pseudo-labels, but also weights the training loss according to its uncertainty to avoid noise in the early stages of training. The results in Table 2 demonstrate the benefits of pseudo-labels refinement designed in this way. The proposed design functions similarly to the teacher-student model. Differently, we do not need the teacher model only to aggregate the information of the nearest neighbors in the feature space. Meanwhile, it is also the refinement design based on aggregated nearest neighbors that allows us to train with reliable pseudo-labels without the need for auxiliary networks or deeper models, keeping the lightweight feature. DiffVAD meets the need for fast detection of real-world scenarios.

## 5 Conclusion

UVAD technology is an important safeguard for sustainable cities. Existing UVAD methods are limited by the openness problem to learn sufficient patterns, we address this challenge by introducing diffusion models into the video anomaly detection task. Specifically, we propose a latent pattern-based learning framework that learns unseen event representations in training by generating multiple latent patterns of the same event, with fast inference speed. The design of pseudo-labels refinement and uncertainty estimation minimize the noise problem. The comprehensive experiments on six benchmarks and exhaustive ablation studies validate the effectiveness of proposed framework.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants ( 62201072, 62101064, 62171057, U23B2001, 62001054, 62071067), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), and the BUPT Excellent Ph.D. Students Foundation (Grant CX20241007).

## References

- [Adam *et al.*, 2008] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):555–560, 2008.
- [Cao *et al.*, 2023] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *CVPR*, pages 20392–20401, 2023.
- [Chen *et al.*, 2022] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In *AAAI*, pages 230–238, 2022.
- [Cong *et al.*, 2011] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pages 3449–3456, 2011.
- [de Moraes *et al.*, 2019] Romero F. A. B. de Moraes, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Reda Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *CVPR*, pages 11996–12004, 2019.
- [Feng *et al.*, 2021] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. MIST: multiple instance self-training framework for video anomaly detection. In *CVPR*, pages 14009–14018, 2021.
- [Flaborea *et al.*, 2023] Alessandro Flaborea, Luca Collocone, Guido Maria D’Amely di Melendugno, Stefano D’Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *ICCV*, pages 10318–10329, 2023.
- [Georgescu *et al.*, 2021] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *CVPR*, pages 12742–12752, 2021.
- [Giorno *et al.*, 2016] Allison Del Giorno, J. Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *ECCV*, pages 334–349, 2016.
- [Gong *et al.*, 2019] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, pages 1705–1714, 2019.
- [Hasan *et al.*, 2016] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *CVPR*, pages 733–742, 2016.
- [He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- [Ionescu *et al.*, 2017] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *ICCV*, pages 2914–2922, 2017.
- [Ionescu *et al.*, 2019a] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *CVPR*, pages 7842–7851, 2019.
- [Ionescu *et al.*, 2019b] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using narrowed normality clusters. In *WACV*, pages 1951–1960, 2019.
- [Kaneko and Harada, 2020] Takuhiro Kaneko and Tatsuya Harada. Noise robust generative adversarial networks. In *CVPR*, pages 8401–8411, 2020.
- [Kim, 2022] Minyoung Kim. Gaussian process modeling of approximate inference errors for variational autoencoders. In *CVPR*, pages 244–253, 2022.
- [Lin *et al.*, 2022] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. A causal inference look at unsupervised video anomaly detection. In *AAAI*, pages 1620–1629, 2022.
- [Litrico *et al.*, 2023] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *(CVPR)*, pages 7640–7650, 2023.
- [Liu *et al.*, 2018a] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *CVPR*, pages 6536–6545, 2018.
- [Liu *et al.*, 2018b] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *BMVC*, page 71, 2018.
- [Liu *et al.*, 2019] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, and Shenghua Gao. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, pages 3023–3030, 2019.
- [Liu *et al.*, 2021] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *ICCV*, pages 13588–13597, 2021.
- [Lu *et al.*, 2013] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 FPS in MATLAB. In *ICCV*, pages 2720–2727, 2013.



- [Luo *et al.*, 2017] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *ICCV*, pages 341–349, 2017.
- [Lv *et al.*, 2021] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021.
- [Lv *et al.*, 2023] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, pages 8022–8031, 2023.
- [Mahadevan *et al.*, 2010] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981, 2010.
- [Mehran *et al.*, 2009] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, 2009.
- [Metzger *et al.*, 2023] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In (*CVPR*), pages 18237–18246, 2023.
- [Pang *et al.*, 2020] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *CVPR*, pages 12170–12179, 2020.
- [Park *et al.*, 2020] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, pages 14372–14381, 2020.
- [Sofianos *et al.*, 2021] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *ICCV*, pages 11209–11218, 2021.
- [Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015.
- [Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.
- [Sultani *et al.*, 2018] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018.
- [Sun *et al.*, 2017] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognit.*, 64:187–201, 2017.
- [Tian *et al.*, 2021] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, pages 4955–4966, 2021.
- [Wang *et al.*, 2018] Siqi Wang, Yijie Zeng, Qiang Liu, Chengzhang Zhu, En Zhu, and Jianping Yin. Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection. In *ACM Multimedia*, pages 636–644, 2018.
- [Wang *et al.*, 2022] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Trans. Neural Networks Learn. Syst.*, 33(6):2301–2312, 2022.
- [Wu *et al.*, 2021] Jie Wu, Wei Zhang, Guanbin Li, Wenhao Wu, Xiao Tan, Yingying Li, Errui Ding, and Liang Lin. Weakly-supervised spatio-temporal anomaly detection in surveillance video. In *IJCAI*, pages 1172–1178, 2021.
- [Xie *et al.*, 2017] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017.
- [Yan *et al.*, 2023] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In (*ICCV*), pages 5527–5537, 2023.
- [Yang *et al.*, 2022] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *ECCV*, pages 404–421, 2022.
- [Yang *et al.*, 2023a] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In (*CVPR*), pages 18381–18391, 2023.
- [Yang *et al.*, 2023b] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, pages 14592–14601, 2023.
- [Yu *et al.*, 2022] Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, and Chengkun Wu. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In *CVPR*, pages 13967–13978, 2022.
- [Zaheer *et al.*, 2022] Muhammad Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segù, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, pages 14724–14734, 2022.
- [Zhang *et al.*, 2016] Ying Zhang, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Shun Sakai. Video anomaly detection based on locality sensitive hashing filters. *Pattern Recognit.*, 59:302–311, 2016.
- [Zhang *et al.*, 2023] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *CVPR*, pages 16271–16280, 2023.