# Denoising Diffusion-Augmented Hybrid Video Anomaly Detection via Reconstructing Noised Frames

**Kai Cheng**[1] , **Yaning Pan**[2] , **Yang Liu**[1,3] , **Xinhua Zeng**[1] and **Rui Feng**[2]
[1]Academy for Engineering and Technology, Fudan University
[2]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
[3]Department of Computer Science, University of Toronto
{chengkai1729, yaningpan}@163.com, {yang_liu20, fengrui, zengxh}@fudan.edu.cn

## Abstract

Video Anomaly Detection (VAD) is crucial for enhancing security and surveillance systems through automatic identification of irregular events, thereby enabling timely responses and augmenting overall situational awareness. Although existing methods have achieved decent detection performances on benchmarks, their predicted objects still remain ambiguous in terms of the semantic aspect. To overcome this limitation, we propose the Denoising diffusion-augmented Hybrid Video Anomaly Detection (DHVAD) framework. The proposed Denoising diffusion-based Reconstruction Unit (DRU) enhances the understanding of semantically accurate normality as a crucial component in DHVAD. Meanwhile, we propose a detection strategy that integrates the advantages of a prediction-based Frame Prediction Unit (FPU) with DRU by exploring the spatial-temporal consistency seamlessly. The competitive performance of DHVAD compared with state-of-the-art methods on three benchmark datasets proves the effectiveness of our framework. The extended experimental analysis demonstrates that our framework can gain a better understanding of the normality in terms of semantic accuracy for VAD and efficiently leverage the strengths of both components.

## 1 Introduction

Video Anomaly Detection (VAD) aims to detect unexpected events in various applications, such as security and surveillance, industrial safety, transportation, and healthcare [Huang *et al.*, 2022a; Liu *et al.*, 2023b; Huang *et al.*, 2022d; Cheng *et al.*, 2023a]. Due to the demand for reliable detection results and different settings of anomalies across scenarios, VAD still remains a challenging task. Additionally, it requires tremendous resources to collect all kinds of anomalies and manually annotate them. Thus, VAD is often defined as an unsupervised task in various existing methods [Gong *et al.*, 2019; Liu *et al.*, 2023c], which concentrates on learning normal patterns with solely normal examples during the training process. In the testing phase, the reconstruction or prediction of anomalous objects will differ from their ground truth due to a deviation from the learned normality, resulting in larger errors compared to the normal ones. Hence, the quality of the reconstruction or prediction of anomalous objects will significantly influence the effectiveness of VAD.

It is widely recognized that video anomaly detection based on reconstruction or prediction requires a comprehensive modeling and understanding of normality from normal videos during the training phase [Park *et al.*, 2020; Cai *et al.*, 2021]. Since normality is a complex system, it is essential to model it in multiple aspects. From one classification of anomalies, we need to comprehensively consider spatial normality, temporal normality, and their relationships, such as consistency [Liu *et al.*, 2018; Huang *et al.*, 2022c; Wu *et al.*, 2024]. Many works have endeavored to delve into this field [Chang *et al.*, 2022; Hao *et al.*, 2022; Liu *et al.*, 2023a]. However, from another perspective, there still exists untapped potential to be explored in modeling normality through understanding the semantic aspects of the object. Since prediction-based methods gradually replace reconstruction-based methods as a popular paradigm, current mainstream methods utilize the characteristics of prediction-based methods to make anomalous objects shift relative to ground truth in the predicted output. In the predicted outputs of these methods, the object is often fully preserved. These methods rely on errors caused by the shift to produce detection results. It is effective for detecting anomalies that violate the spatial-temporal normality, such as pedestrians and vehicles moving at high speeds. However, when the displacement caused by anomalies is minor, the errors produced by these methods are primarily not significant for existing at the edges of the object, which eventually have an impact on VAD. Such scenarios will be visualized in experiments. This indicates that these methods that still retain the abnormal object completely ignore some normality in the semantic perspective for VAD.

To tackle this challenge, we design the Denoising diffusion-based Reconstruction Unit (DRU) to fully explore the semantic information in normal samples. Denoising diffusion-based models as a class of deep generative models have widely been used to address various challenging real-world tasks due to their flexibility and significant representation ability [Song *et al.*, 2020; Yang *et al.*, 2023a]. Although they have strong random image generation ability, what we need in VAD tasks belongs to the conditional image generation. Specifically, DRU utilized noised frames as the conditions to complete the sampling task instead of a random noise vector. Thus, DRU can formulate the denoising diffusion-based reconstruction that is distinct from the traditional one utilizing compression and decompression in VAD. The denoising diffusion-based reconstruction provides a better understanding of semantically accurate normality as a crucial contribution to the comprehensive modelling of DHVAD, which will be discussed in experiments in detail. Specifically, Both normal and abnormal objects are disrupted by noises during the diffusion process. DRU ensures that the normal parts can be well recovered. However, the abnormal objects are reconstructed into new states that are in line with normality, resulting in significant reconstruction errors that are greatly conducive to VAD. Meanwhile, in order to further utilize the understanding of the consistency extracted by prediction-based methods, we design a Frame Prediction Unit (FPU) to enhance the capture of spatial-temporal relationships. This is because some anomalous objects are often detectable through the examination of spatial-temporal consistency by considering a video as a sequence. Given that the objective of the FPU is to uncover spatial-temporal consistency to the fullest extent possible, the FPU concurrently utilizes both video frames and optical flow for extracting spatial-temporal normality. For the triplet of ground truth, reconstructed frames, and predicted frames, along with their corresponding optical flow triplet, we devise an appropriate detection strategy to seamlessly integrate the advantages. Thus, we have formulated a framework composed of above components and strategy, which is called Denoising diffusion-augmented Hybrid Video Anomaly Detection (DHVAD). Our contributions to this paper can be summarized as follows:

- We propose a VAD framework called DHVAD, where the advantages of components DRU and FPU are complementary. DHVAD performs comparably to state-of-the-art methods on three benchmark datasets.
- We propose a denoising diffusion-based reconstruction method to better extract semantically accurate normality. The experimental analysis and visualization indicate its effectiveness.
- To leverage the multi-source information in the DHVAD framework for more efficient VAD, we also devise a detection strategy. The experimental results prove the practicality and effectiveness of it.

## 2 Related Work

### 2.1 Video Anomaly Detection

As mentioned above, the goal of unsupervised VAD is to train a model on the training datasets of only normal events and detect anomalous objects. Early one-class model-based approaches widely use reconstruction errors, which are typically produced by a pixel-level function as a measure for VAD. Among them, many deep learning techniques have been employed. For example, [Lu *et al.*, 2013] introduced the convolutional auto-encoder (ConvAE) to address VAD by capturing the spatial information of the video frames and identify anomalies by the quality of the reconstructed results. However, [Liu *et al.*, 2018] pointed out that the commonly used spatial constraint is not enough to predict a future frame with higher quality for normal events. In response, they introduced a temporal constraint in video prediction with optical flow generation, forcing the U-Net to model motion normality by enforcing the optical flow between predicted frames to be close to their optical flow ground truth. Furthermore, [Ravanbakhsh *et al.*, 2019] trained two GANs to learn the temporal and spatial distribution and used a cross-channel approach to prevent the discriminator from learning mundane constant functions, boosting the anomaly detection performance but leading to unstable training process and high training cost. In order to control the generalization capability of Auto-Encoder (AE), [Gong *et al.*, 2019] proposed a Memory-augmented Auto-Encoder (MemAE) for anomaly detection, which is encouraged to store normal patterns in the memory. Following this work, [Park *et al.*, 2020] introduced a more compact memory module to record the prototypical patterns of the items. [Hao *et al.*, 2022] developed a spatial-temporal consistency-enhanced network (STCEN) with a well-designed 3D-2D U-shape structure to focus on capturing spatial-temporal high-level features and predicting future frames. To explore better methods for mining complex spatial-temporal relationships in the video, [Yang *et al.*, 2023b] proposed a novel USTN-DSC to restore the video event based on keyframes, which focuses on temporal context relationships in the video.

### 2.2 Denoising Diffusion Models

As a category of probabilistic generative models, diffusion models [Croitoru *et al.*, 2023] have attracted increasing interest in the wide-ranging fields, spanning from computer vision, natural language processing to interdisciplinary subjects. Three predominant formulations of diffusion models are denoising diffusion probabilistic models(DDPMs) [Ho *et al.*, 2020], score-based generative models(SGMs) [Song *et al.*, 2021], and stochastic differential equations(Score SDEs) [Karras *et al.*, 2022]. The key to all these three methods is to smoothly perturb data by adding noise, then reverse this process to successively remove noise to generate new data. Subsequent researches have focused on improving these three classical diffusion models from three main perspectives: sampling speed, accuracy of likelihood estimation and consideration of data with special structures. [Pinaya *et al.*, 2022] introduced a diffusion and VQ-VAE based approach, which first encodes the brain images and then obtains the quantified latent representation from a codebook.

The use of diffusion models in DRU differs from those in image generation due to distinct input data characteristics and task objectives. In denoising, input comprises intentionally noisy video frames, while image generation introduces randomly generated Gaussian noise, aiming to explore normal patterns rather than distinct scenarios. DRU's objectives diverge from image generation, focusing not on high-quality, photorealistic images but on comprehending visual semantics. DRU identifies deviations from normal behavior, serving as indicators of anomalies in video streams. But image generation prioritizes creating visually appealing images.

# 3 Methodology

## 3.1 Overall Framework

As depicted in Figure 1 (a), the overall framework of DHVAD comprises two essential components: the DRU for enhancing the semantically accurate normality modeling and the FPU for seamlessly capturing spatial-temporal consistency normality. Under the setting of unsupervised learning, our model is trained on the dataset consisting solely of normal samples following rigorous data processing techniques. Following [Liu *et al.*, 2021], we sample $t+1$ consecutive frames $x_{n:n+t}$ of videos $\mathcal{X}$ and obtain their corresponding optical flows $f_{n:n+t}$ at time $n$. It is widely acknowledged that the normality of videos typically lies in various aspects [Wu and Liu, 2021; Huang *et al.*, 2022b; Cheng *et al.*, 2023b]. Recently, the most widely concerned types of normality are spatial and temporal normalities in the field of VAD. Especially, the interplay between them also contributes to augmenting the performance of VAD. Nonetheless, there exists untapped potential in delving into the spatial-temporal normality through understanding the semantic aspect of the objects. Most existing methods often concentrate on exploiting the prediction-based methods that will result in some pixel spatial shifts of anomalous objects in predicted frames compared with their ground truth. The shifts will further influence the prediction errors to obtain higher anomaly scores. Such efforts mostly manage to leverage the temporal normality for VAD, which particularly enhances the performance of detecting spatial-temporal anomalies. For example, scenarios including high-speed movements of people or vehicles can be captured. However, these performances are not enough to prove that they are perfect. By analyzing the predictions generated by baselines, such as FFP [Liu *et al.*, 2018] and MNAD [Park *et al.*, 2020], the anomalous objects are reconstructed rather similar to the ground truth by these models. This indicates that VAD still needs to be enhanced via understanding normality in a more semantically accurate manner. To address this, we propose DRU to explore more semantic information from normal samples.

Through denoising diffusion, DRU possesses a profounder comprehension of normal objects. We assign the reconstruction task as the proxy task of DRU, thereby emphasizing its ability to grasp the semantic information deeply. Specifically, the video frames $x_{n:n+t}$ serve as the input for DRU to initially generate the noised states of videos and produce a sequence of noisy frames $x^K_{n:n+t}$. Subsequently, the denoised samples $\hat{x}^d_{n:n+t}$ can be derived through the denoising process by predicting the noise via the denoising U-Net. After the process of adding and removing noise, the reconstructions of video frames are implemented by DRU. This process can be interpreted to be the acquisition of the ability of learning how to reconstruct anomalous images back to corresponding to the normality. This ability also helps the denoising diffusion-based reconstruction excel at output more semantically accurate results that conform to normality compared with the reconstruction that compresses images into hidden space, which will be further explained in the discussions of experiments. In order to enhance the framework's understanding of spatial-temporal consistency, we also calculate the optical flows $f^d_{n:n+t}$ corresponding to the reconstructed video frames. In addition, we leverage FPU to augment the temporal modeling. Based on the prevalent VAD paradigms, we employ an AE with the memory pool and skip connections to carry out downstream predicting proxy task. The video frames $x_{n:n+t-1}$ and corresponding optical flows $f_{n:n+t-1}$ are simultaneously input into the FPU to fully explore their relationships. During the training process, the memory pool is progressively updated to preserve normality. In the testing phase, the output of encoder $\Psi_E$ is processed through read and retrieve processes on the memory pool conditioned on the optical flow to obtain the input of decoder $\Psi_D$. The decoder generates the predicted frames $\hat{x}_{n+t}$ as the outputs of FPU. Finally, the strategy of performing VAD by employing the approach depicted in Figure 1 (b).

## 3.2 Denoising Diffusion-based Reconstruction Unit

DRU excels in reconstructing images that deviate from normality back to those that semantically conform to normality. By employing noise addition and removal, DRU also demonstrates robustness when facing unknown anomalies. Specifically, DRU includes a forward diffusion stage $\kappa(x_k|x_{k-1})$ to gradually add Gaussian noise and a backward denoising stage $\nu_\theta(x_{k-1}|x_k)$ that can remove the noise step by step to recover normal data $\kappa(x_0)$. The former, called the diffusion process, is designed to transform input data distribution into Gaussian distribution, while the latter, called the denoising process, reverses the former Markov chain with the predicted noise from a U-Net architecture.

First, considering $x_0$ sampled from a real data distribution satisfying $x_0 \sim \kappa(x_0)$, the forward process generates the noisy image $x_k$ based on the Markov property and probability rules. So the joint distribution of $x_{1:K}$ conditioned on $x_0$ is denoted as below:

$$\kappa(x_{1:K}|x_0) = \prod_{k=1}^{K} \kappa(x_k|x_{k-1}), \tag{1}$$

where $K$ denotes the total timesteps and $x_k$ represents the image generated in step $k$. We assume that real-world images obey Gaussian distribution $\mathcal{N}(\mu, \sigma)$. According to the property of Gaussian distribution, we have:

$$\kappa(x_k|x_{k-1}) = \mathcal{N}(x_k; \sqrt{1-\gamma_k}x_{k-1}, \gamma_k\mathbf{I}), \tag{2}$$

where $\gamma_k \in (0,1)$ is a hyperparameter schedule which can be defined as a linear schedule [Ho *et al.*, 2020] or a cosine schedule [Nichol and Dhariwal, 2021]. Unlike general latent variable models such as VAE, the inference process of $\kappa(x_{1:K}|x_0)$ in DRU is fixed, and the dimensions of latent variables are the same as input data.

Then for $\forall k \in (1, K)$, we can easily obtain the analytical form of $\kappa(x_k|x_0)$ from Eq. 1 with $\eta_k := 1 - \gamma_k$ and $\overline{\eta_k} := \prod_{i=1}^{k} \eta_i$, that is:

$$\kappa(x_k|x_0) = \mathcal{N}(x_k; \sqrt{\overline{\eta}_k}x_0, (1-\overline{\eta}_k)\mathbf{I}). \tag{3}$$

So, it is noted that the diffusion process admits sampling $x_k$ at any timestep in closed form, and $x_k$ can be further denoted as:

$$x_k = \sqrt{\overline{\eta}_k}x_0 + \sqrt{1-\overline{\eta}_k}\varepsilon, \tag{4}$$

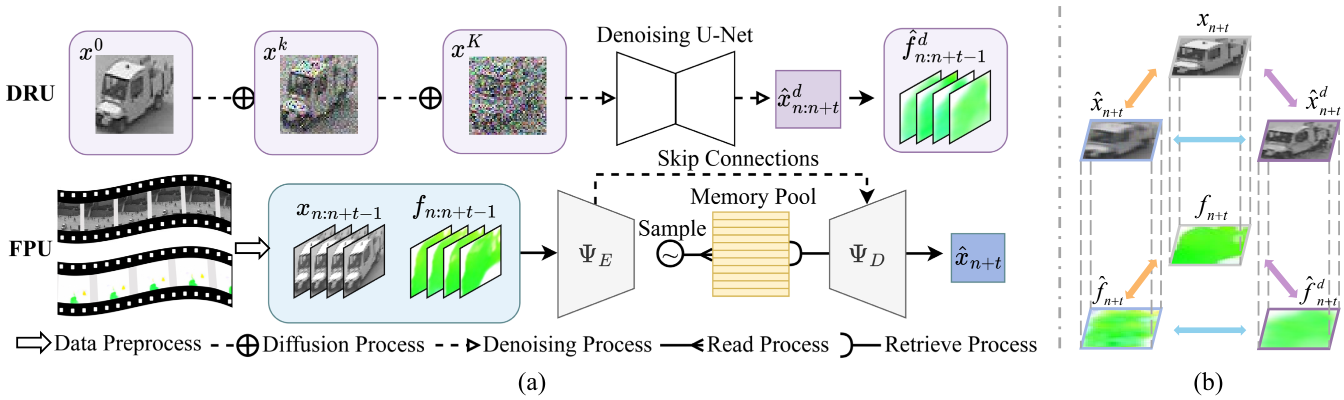

Figure 1: The overall framework of DHVAD, which shows the crucial components DRU and FPU (a) along with the detection strategy (b).

where $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ is a Gaussian noise variable.

For the denoising process, it starts from $\nu(x_K) = \mathcal{N}(x_K; 0, \mathbf{I})$ and then gradually remove noise by learning Gaussian transition $\nu_\theta(x_{k-1}|x_k)$ parameterized by $\theta$:

$$\nu_\theta(x_{k-1}|x_k) = \mathcal{N}(x_{k-1}; \mu_\theta(x_k, k), \sum_\theta (x_k, k)), \tag{5}$$

where the mean $\mu_\theta(x_k, k)$ and variance $\sum_\theta(x_k, k)$ can be parameterized by U-Net architecture.

In the case where all the conditions are represented as Gaussians with trainable mean functions and fixed constant variances, we can obtain a simplified objective as follows:

$$\mathcal{L}_{diff} = \sum_{k=1}^{K} \mathbb{E}_{x_0 \sim \kappa(x_0), \varepsilon \sim \mathcal{N}(0,\mathbf{I})} ||\varepsilon - \varepsilon_\theta(x_k, k)||^2. \tag{6}$$

## 3.3 Frame Prediction Unit

Prediction-based approaches yield better performance of VAD owing to their exceptional capacity to capture temporal related information. The video can be seen as a sequence, which makes abnormal objects in videos often detected in terms of unexpected spatial-temporal consistency. Given its role in mining spatial-temporal consistency information, FPU will simultaneously leverage video frames and optical flow to extract spatial-temporal normality. Specifically, FPU is trained to model $\Psi_D(x_{n+t}|x_{n:n+t-1}, f_{n:n+t-1})$ given previous frames $x_{n:n+t-1}$ and optical flows $f_{n:n+t-1}$. According to Conditional Variational Auto-Encoder (CVAE) [Liu *et al.*, 2024], the ELBO is:

$$\begin{aligned} &log\Psi_D(x_{n+t}|f_{n:n+t-1}) \\ &\geq \mathbb{E}_{\Psi_E} \log \frac{\Psi_D(x_{n+t}|z, f_{n:n+t-1})\Psi_D(z|f_{n:n+t-1})}{\Psi_E(z|x_{n+t}, f_{n:n+t-1})}. \end{aligned} \tag{7}$$

We assume that the distribution of $x_{n+t}$ and $x_{n:n+t-1}$ is the same, which can be determined by the hidden variables $z$. This is because the time duration is very small in real-time VAD, and their content information only contains subtle pixel-level shifts of objects. Hence, replacing $\Psi_E(z|x_{n+t}, f_{n:n+t-1})$ with $\Psi_E(z|x_{n:n+t-1}, f_{n:n+t-1})$ in Eq. 7, we have:

$$\begin{aligned} &log\Psi_D(x_{n+t}|f_{n:n+t-1}) \\ &\geq \mathbb{E}_{\Psi_E} \log \frac{\Psi_D(x_{n+t}|z, f_{n:n+t-1})\Psi_D(z|f_{n:n+t-1})}{\Psi_E(z|x_{n+t}, f_{n:n+t-1})} \\ &\approx \mathbb{E}_{\Psi_E} \log \frac{\Psi_D(x_{n+t}|z, f_{n:n+t-1})\Psi_D(z|f_{n:n+t-1})}{\Psi_E(z|x_{n:n+t-1}, f_{n:n+t-1})} \quad , \\ &= -KL[\Psi_E(z|x_{n:n+t-1}, f_{n:n+t-1}) \parallel \Psi_D(z|f_{n:n+t-1})] \\ &+ \mathbb{E}_{\Psi_E}[\log \Psi_D(x_{n+t}|z, f_{n:n+t-1})] \end{aligned} \tag{8}$$

where KL signifies Kullback-Leibler divergence.

Thus, the loss function can be designed as follows:

$$\begin{aligned} \mathcal{L}_{pred} =&KL[\Psi_E(z|x_{n:n+t-1}, f_{n:n+t-1})||\Psi_D(z|f_{n:n+t-1})] \\ &+ ||x_{n+t} - \hat{x}_{n+t}||_2^2, \end{aligned} \tag{9}$$

where the first term, KL divergence, is used to ensure that the hidden variable $z$ follows a parametric Gaussian distribution, while the latter term ensures that the decoder $\Psi_D$ can effectively restore the data from the hidden variable.

## 3.4 Detection Strategy

As shown in Figure 1 (b), with the predicted and reconstructed frames from multiple sources and their corresponding optical flows, we can design appropriate evaluation strategies for more effective VAD. In terms of video frames, a triplet $\mathbb{X} = \left\{x_{n+t}, \hat{x}^d_{n+t}, \hat{x}_{n+t}\right\}$ can be formed among the model outputs and the ground truth. Typically, we can first consider the reconstruction errors between $x_{n+t}$ and $\hat{x}^d_{n+t}$ and the prediction errors between $x_{n+t}$ and $\hat{x}_{n+t}$ for detection, which can be defined as:

$$\mathcal{E}^x_{diff} = \left\|x_{n+t} - \hat{x}^d_{n+t}\right\|_2^2, \tag{10}$$

$$\mathcal{E}^x_{pred} = \left\|x_{n+t} - \hat{x}_{n+t}\right\|_2^2. \tag{11}$$

Meanwhile, we can also consider the error between $\hat{x}^d_{n+t}$ and $\hat{x}_{n+t}$, which can be defined as:

$$\mathcal{E}^x_{mutual} = \left\|x_{n+t} - \hat{x}_{n+t}\right\|_2^2. \tag{12}$$

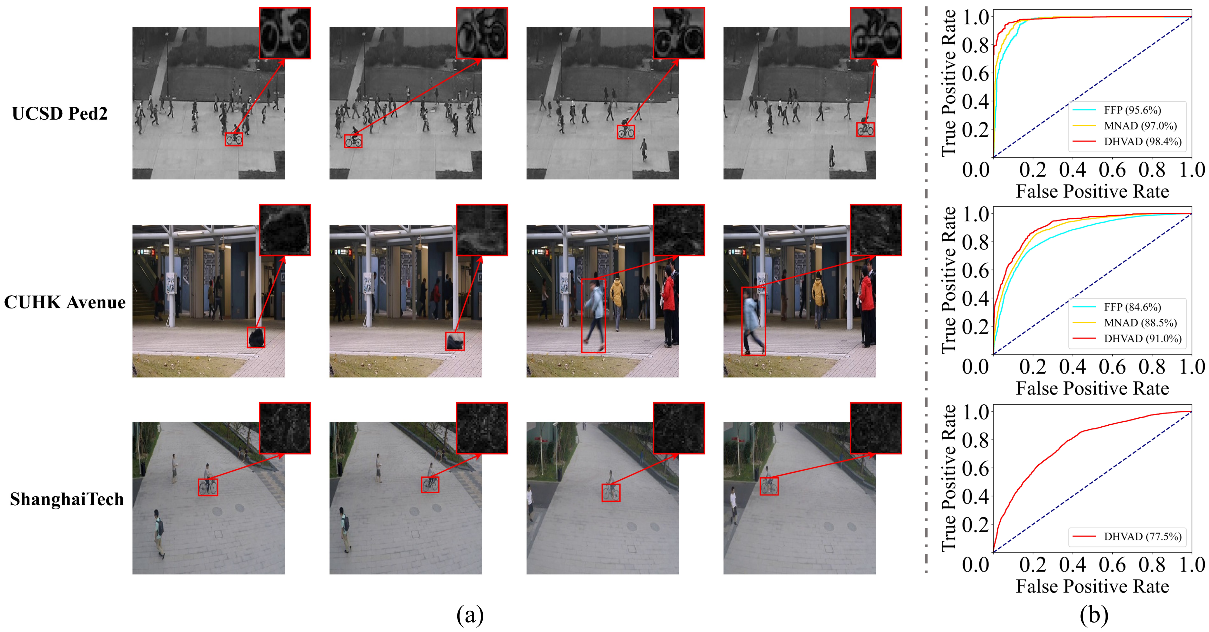

Figure 2: (a) Visualization of the error between the output of DHVAD and the ground truth. Three rows are the visualization results on UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets, respectively. (b) ROC curves of our framework DHVAD, baselines MNAD [Park *et al.*, 2020] and FFP [Liu *et al.*, 2018].

This is due to the fact that DRU and FPU will reconstruct abnormal objects in different ways. However, the normal areas that they reconstruct are quite similar. By comparing $\hat{x}^d_{n+t}$ and $\hat{x}_{n+t}$, we can more precisely identify some subtle anomalies. Additionally, optical flow can also form $\mathbb{F} = \left\{ f_{n+t}, \hat{f}^d_{n+t}, \hat{f}_{n+t} \right\}$ to assist in video anomaly detection in a similar way.

$$\mathcal{E}^f_{diff} = \left\| f_{n+t} - \hat{f}^d_{n+t} \right\|^2_2, \tag{13}$$

$$\mathcal{E}^f_{pred} = \left\| f_{n+t} - \hat{f}_{n+t} \right\|^2_2, \tag{14}$$

$$\mathcal{E}^f_{mutual} = \left\| f_{n+t} - \hat{f}_{n+t} \right\|^2_2. \tag{15}$$

We can leverage these errors to obtain the anomaly score, weighted by two hyper-parameters $\lambda_{diff}$ and $\lambda_{pred}$ called Anomaly Score Coefficient (ASC), which can be defined as:

$$\begin{aligned} \mathcal{S} = {} & \lambda_{diff} \cdot F\left(\mathcal{E}^x_{diff} + \mathcal{E}^f_{diff}\right) \\ & + \lambda_{pred} \cdot F\left(\mathcal{E}^x_{pred} + \mathcal{E}^f_{pred}\right) \\ & + (1 - \lambda_{diff} - \lambda_{pred}) \cdot F\left(\mathcal{E}^x_{mutual} + \mathcal{E}^f_{mutual}\right) \end{aligned}, \tag{16}$$

where $F$ represents the min-max normalization function.

# 4 Experiments

## 4.1 Quantitative Analysis

### Implementation Details

To demonstrate the effectiveness of our proposed DMVAD, we conduct experiments on three public benchmarks: UCSD Ped2 [Li *et al.*, 2013], CUHK Avenue [Lu *et al.*, 2013], and ShanghaiTech [Liu *et al.*, 2018] datasets. In order to save computing costs, we utilize the Cascade R-CNN [Cai and Vasconcelos, 2019] pre-trained model and FlowNet2.0 [Ilg *et al.*, 2017] pre-trained model. When training, hyperparameter timestep $K$ is set to 1200. The variance schedule $\gamma_k \in (0, 1), k = 1, ..., K$ is defined as a small linear schedule, increasing linearly from $\gamma_1 = 10^{-4}$ to $\gamma_K = 0.02$. The timestep is encoded with transformer sinusoidal positional embedding [Vaswani *et al.*, 2017]. It is noted that we set the sample distance $\lambda = 70$. We set the training batch size and testing batch size to 64 and 128, respectively. We utilize the PyTorch [Paszke *et al.*, 2019] framework on an NVIDIA GeForce RTX 3090 GPU to implement our proposed DHVAD.

### Comparison with State-of-the-Art Methods

As shown in Table 1, we compare our proposed DHVAD with previous methods on three benchmarks. Our DHVAD outperforms the state-of-the-art methods on three benchmarks. Overall, the performances of our DHVAD is 0.3%, 1.1% and 3.7% higher than the second best on UCSD ped2, CUHK Avenue, and ShanghaiTech datasets, respectively. To be specific, our DHVAD model is 1.0%, 2.5% and 3.9% higher than the best of classical methods that utilize the prevalent prediction-based tasks [Gong *et al.*, 2019; Park *et al.*, 2020; Le and Kim, 2023]. Compared with methods that model both spatial and temporal normality [Tang *et al.*, 2020; Cai *et al.*, 2021; Zhao *et al.*, 2022; Chang *et al.*, 2022; Hao *et al.*, 2022; Liu *et al.*, 2022], we obtain an improvement of 1.0–2.1%, 2.8–5.9% and 3.7–4.5% on three benchmarks. The results show that DHVAD can fully utilize complementary advantages between DRU and FPU with our strategy.

| | Methods | UCSD Ped2 | CUHK Avenue | ShanghaiTech |
|---|---|---|---|---|
| R. | MemAE [Gong *et al.*, 2019] | 94.1% | 83.3% | 71.2% |
| | MNAD-Reconstruction [Park *et al.*, 2020] | 90.2% | 82.8% | 69.8% |
| P. | FFP [Liu *et al.*, 2018] | 95.4% | 84.9% | 72.8% |
| | Multi-space [Zhang *et al.*, 2020] | 95.4% | 86.8% | 73.6% |
| | MNAD-Prediction [Park *et al.*, 2020] | 97.0% | 88.5% | 70.5% |
| | LIF [Chang *et al.*, 2020] | 96.5% | 86.0% | 73.3% |
| | AMMC-Net [Cai *et al.*, 2021] | 96.6% | 86.6% | 73.7% |
| | MPN [Lv *et al.*, 2021] | 96.9% | 89.5% | 73.8% |
| | STC-Net [Zhao *et al.*, 2022] | 96.7% | 87.8% | 73.1% |
| | STCEN [Hao *et al.*, 2022] | 96.9% | 86.6% | 73.8% |
| | AMAE [Liu *et al.*, 2022] | 97.4% | 88.2% | 73.6% |
| | Le *et al.* [Le and Kim, 2023] | 97.4% | 86.7% | 73.6% |
| O. | DDGAN [Tang *et al.*, 2020] | 96.3% | 85.1% | 73.0% |
| | STD [Chang *et al.*, 2022] | 96.7% | 87.1% | 73.7% |
| | Zhong *et al.* [Zhong *et al.*, 2022] | 97.7% | 88.9% | 70.7% |
| | USTN-DSC [Yang *et al.*, 2023b] | 98.1% | 89.9% | 73.8% |
| | **Our DHVAD** | **98.4%** | **91.0%** | **77.5%** |

Table 1: Results of quantitative frame-level AUC comparison. Numbers in bold indicate the best performance and underscored ones are the second best. ('R.', 'P.', and 'O.' indicate the reconstruction-based, prediction-based and other methods, respectively.)

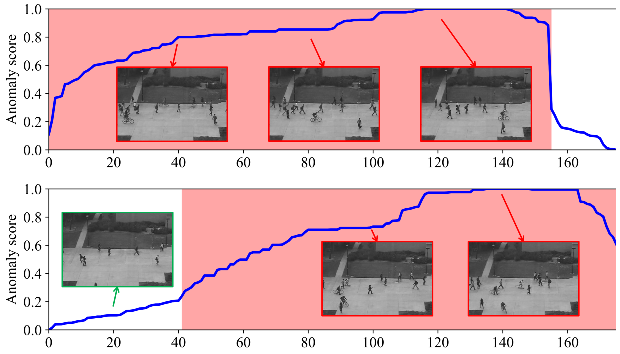

Figure 3: Temporal analysis of the anomaly score on testing videos.

## 4.2 Visualization and Qualitative Analysis

Due to DRU's understanding of semantically accurate normality and the proposed detection strategy, errors are identified significantly in the area of anomalous objects, allowing DHVAD to capture them effectively. As shown in Figure 2 (a), we demonstrate the visualization of errors between the output and the ground truth. DHVAD exhibits stability when confronted with various anomalies. For example, DHVAD can detect abnormal objects due to high-speed movement, as well as those due to inappropriate appearance in some scenes. As for anomalies that are prone to misjudgment, they can still be detected, such as objects that conform to spatial normality but deviate from temporal normality (or vice versa), whose part of compliance with normality may make detection confusing. For instance, pedestrians that adhere to spatial normality are permitted, which means running pedestrians are considered a normal mode solely from a spatial perspective. Previous methods choose to reconstruct such targets relatively completely, and their way of detection mainly relies on shifting these objects in the prediction relative to the ground truth to generate an error. But these shifts are typically slight. Once the impact of the error generated by the displacement is insufficient to outweigh the impact of the object's normal part that adheres to normality, it will lead to unreliable detection results. This will, to some extent, affect the robustness of the model. However, DHVAD successfully retains the information of the captured anomalous errors based on the reasonable proposed detection strategy.

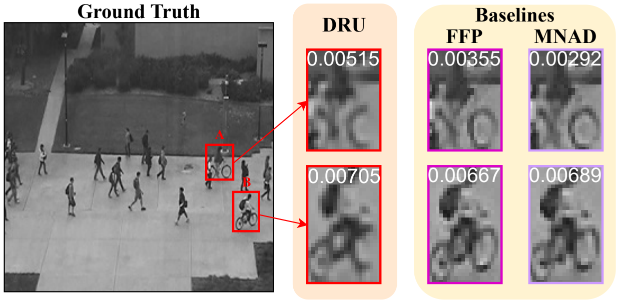

Figure 4: Examples of predictions and reconstructions of objects by DRU, MNAD [Park *et al.*, 2020] and FFP [Liu *et al.*, 2018]. Corresponding errors are labeled in the upper left.

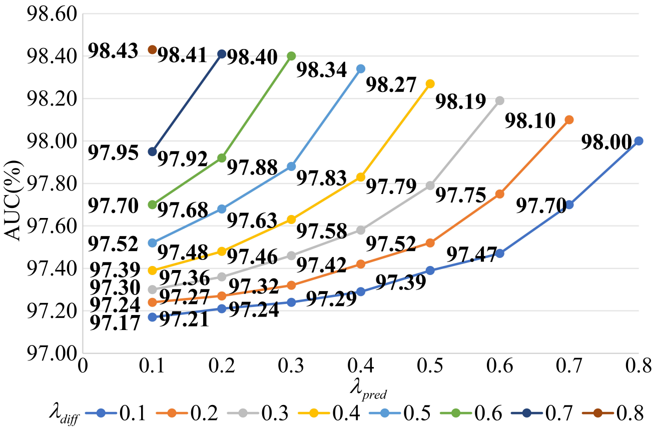

Figure 5: AUC (%) under different $\lambda_{diff}$ and $\lambda_{pred}$.

As shown in Figure 2 (b), we investigate the reliability and robustness of DHVAD based on the Receiver Operating Characteristic (ROC) curves. Due to the inaccessibility and unavailability of codes in some recent methods, we implement two baselines for the comparison and only obtain the results consistent with the claimed performance in their papers on UCSD Ped2 and CUHK Avenue. They are well explored and achieve excellent performance in the VAD task. In the figure, three red lines represent the ROC curves of DHVAD on three datasets, respectively. The yellow and green lines represent those of MNAD and FFP, respectively. In terms of the ROC curves on the UCSD Ped2 dataset, DHVAD's true positive rate is persistently higher than MNAD and FFP as the false positive rate grows from 0. This situation can be apparently observed until the false positive rate approaches around 0.2. In addition, we also conduct experiments of temporary analysis of DHVAD, as shown in Figure 3. The red background represents the existence of abnormal objects in the corresponding video segments, whereas the white background denotes that the video segments can be classified as normal ones. Some video frames within normal and abnormal segments are exhibited for clearness. Observed from the anomaly score, it is evident that when anomalies appear, DHVAD can effectively capture them, thus demonstrating the reliability of the detection outcomes.

## 4.3 Ablation Study

### Rationality of DRU

To demonstrate the effective understanding of DRU in terms of semantically accurate normality, we visualize the reconstruction of DRU as depicted in Figure 4. We also present the outputs of two baselines for comparative analysis. Two abnormal objects can be observed from the ground truth, among which object A is partially covered. Both FFP and MNAD outputs relatively complete predictions of anomalous objects, resulting in similar outcomes. Both baselines retain bicycles in the prediction of the two abnormal objects. On the contrary, DHVAD demonstrates an ability to reconstruct bicycles that do not conform to semantically accurate normality. Specifically, the visualization of object A reveals that the uncovered front wheels in object A have been partially removed visually, indicating the effectiveness of DRU. For the remaining portion of the reconstruction, it can be considered as multiple parts that conform to normality if we try to observe it in a decomposed manner. This essentially reflects another important reason why DHVAD can provide better detection performance. DRU will also try to retain parts of the anomalous objects that conform to normality to prevent misjudgment. VAD not only necessitates an accurate detection of anomalies but also requires the prevention of misjudgment of normal objects. From the perspective of generalization, the above requirements correspond to the issues of the model's too strong or too weak generalization of semantic information. For example, the dress and personal adornment of pedestrians may not be our criterion for judging anomalies. However, an overly harsh model might trigger an alert when it encounters some new dress and personal adornment to create a false detection. Therefore, the generalization of DRU is also reflected in the cases where the uncovered part of the rear wheel is reconstructed into the human leg that better conforms to the semantically accurate normality. On the one hand, DRU can remove anomalous parts from the object, while on the other, it exhibits good generalization ability in terms of semantically accurate normality. In short, DRU possesses two ways of treating anomalies by removing and reconstructing them into normal parts. It is precisely such accurate judgements of different parts of the object based on semantic information that proves the rationality of DRU. As for object B, the front wheel of the bicycle is partially removed like target A. DRU tends to reconstruct the preserved parts into legs that conform to semantically accurate normality. For the rear wheels, the remaining part is reconstructed to mimic the scenario of pedestrian legs on the crowded street from the perspective of surveillance. From the comparative experimental results with the baselines and the design of the framework, it can be seen that DHVAD has been able to implement VAD from a different perspective than existing prevalent paradigms.

### Sensitivity of ASC

We perform sensitivity analysis on ASC within DHVAD on UCSD Ped2 dataset, providing valuable insights into its impact on performance. When only considering the use of DRU for detection, its AUC drops to 96.0%. Similarly, the AUC falls to 97.5% when solely relying on FPU for detection. Compared with the detection performance of the full framework, the AUC decreases by 0.9% and 2.4%, respectively, proving the effectiveness of the detection strategy. The relationship between $\lambda_{diff}$ and $\lambda_{pred}$ satisfies the following:

$$\lambda_{diff} + \lambda_{pred} + (1 - \lambda_{diff} - \lambda_{pred}) = 1, \qquad (17)$$

$$0 < \lambda_{diff} < 1, \qquad (18)$$

$$0 < \lambda_{pred} < 1. \qquad (19)$$

Therefore, we conduct a grid search on these two parameters. As depicted in Figure 5, the different colored lines represent the changes in performance caused by changes in $\lambda_{pred}$ when $\lambda_{diff}$ remains unchanged. It can be observed that employing the detection strategy outperforms the individual detection in most instances. Even in a few cases, it is still superior to one of the performances of the individual detection. This indicates that there is indeed complementary information between DRU and FPU, thereby facilitating effective VAD. After determining the effectiveness of the detection strategy, we can see from the trend that the model performs better when $\lambda_{diff}$ rises, which means the semantically accurate normality extracted by the DRU plays a pivotal role.

## 5 Conclusion

In this paper, we propose a denoising diffusion-based reconstruction that differs from the commonly used reconstruction of compression and decompression in the prevalent VAD paradigm. The proposed framework DHVAD fully utilizes the complementary advantages between DRU and FPU in the proposed detection strategy to enhance VAD performance. The performance comparison with the state-of-the-art models on three benchmark datasets verifies the effectiveness of the framework. The diverse visual analysis further explains the rationality and role of crucial components in the model. Especially, we explained the competitive performance of DHVAD from the perspective of the generalization and understanding of semantically accurate normality using DRU. Furthermore, we conduct sensitivity analysis based on a grid search of the key parameters of the detection strategy, demonstrating its robustness. In future work, we will attempt to further explore the semantically accurate normality and establish a more sophisticated analysis.

## Acknowledgments

This work was supported by the Science and Technology Commission of Shanghai Municipality (No. 21511100500).

## Contribution Statement

Kai Cheng and Yaning Pan contributed equally to this work and should be considered co-first authors. Corresponding Authors: Xinhua Zeng and Rui Feng.

## References

[Cai and Vasconcelos, 2019] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2019.

[Cai *et al.*, 2021] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 938–946, 2021.

[Chang *et al.*, 2020] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pages 329–345. Springer, 2020.

[Chang *et al.*, 2022] Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition*, 122:108213, 2022.

[Cheng *et al.*, 2023a] Kai Cheng, Yang Liu, and Xinhua Zeng. Learning graph enhanced spatial-temporal coherence for video anomaly detection. *IEEE Signal Processing Letters*, 30:314–318, 2023.

[Cheng *et al.*, 2023b] Kai Cheng, Xinhua Zeng, Yang Liu, Mengyang Zhao, Chengxin Pang, and Xing Hu. Spatial-temporal graph convolutional network boosted flow-frame prediction for video anomaly detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023.

[Croitoru *et al.*, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[Hao *et al.*, 2022] Yi Hao, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognition*, 121:108232, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2022a] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, 2022.

[Huang *et al.*, 2022b] Chao Huang, Chengliang Liu, Zheng Zhang, Zhihao Wu, Jie Wen, Qiuping Jiang, and Yong Xu. Pixel-level anomaly detection via uncertainty-aware prototypical transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 521–530, 2022.

[Huang *et al.*, 2022c] Chao Huang, Yabo Liu, Zheng Zhang, Chengliang Liu, Jie Wen, Yong Xu, and Yaowei Wang. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 307–315, 2022.

[Huang *et al.*, 2022d] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Ilg *et al.*, 2017] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[Le and Kim, 2023] Viet-Tuan Le and Yong-Guk Kim. Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence*, 53(3):3240–3254, 2023.

[Li *et al.*, 2013] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2013.

[Liu *et al.*, 2018] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13588–13597, 2021.

[Liu *et al.*, 2022] Yang Liu, Jing Liu, Jieyu Lin, Mengyang Zhao, and Liang Song. Appearance-motion united autoencoder framework for video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(5):2498–2502, 2022.

[Liu *et al.*, 2023a] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Transactions on Industrial Informatics*, 2023.

[Liu *et al.*, 2023b] Yang Liu, Zhaoyang Xia, Mengyang Zhao, Donglai Wei, Yuzheng Wang, Liu Siao, Bobo Ju, Gaoyun Fang, Jing Liu, and Liang Song. Learning causality-inspired representation consistency for video anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 203–212, 2023.

[Liu *et al.*, 2023c] Yang Liu, Dingkang Yang, Gaoyun Fang, Yuzheng Wang, Donglai Wei, Mengyang Zhao, Kai Cheng, Jing Liu, and Liang Song. Stochastic video normality network for abnormal event detection in surveillance videos. *Knowledge-Based Systems*, 280:110986, 2023.

[Liu *et al.*, 2024] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 56(7), 2024.

[Lu *et al.*, 2013] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15425–15434, 2021.

[Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[Park *et al.*, 2020] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[Pinaya *et al.*, 2022] Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 705–714. Springer, 2022.

[Ravanbakhsh *et al.*, 2019] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1896–1904. IEEE, 2019.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Song *et al.*, 2021] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021.

[Tang *et al.*, 2020] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wu and Liu, 2021] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021.

[Wu *et al.*, 2024] Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, and Yanning Zhang. Toward video anomaly retrieval from video anomaly detection: New benchmarks and model. *IEEE Transactions on Image Processing*, 33:2213–2225, 2024.

[Yang *et al.*, 2023a] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14592–14601, 2023.

[Zhang *et al.*, 2020] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multispace for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3694–3706, 2020.

[Zhao *et al.*, 2022] Mengyang Zhao, Yang Liu, Jing Liu, and Xinhua Zeng. Exploiting spatial-temporal correlations for video anomaly detection. In *International Conference on Pattern Recognition*, pages 1727–1733. IEEE, 2022.

[Zhong *et al.*, 2022] Yuanhong Zhong, Xia Chen, Jinyang Jiang, and Fan Ren. A cascade reconstruction model with generalization ability evaluation for anomaly detection in videos. *Pattern Recognition*, 122:108336, 2022.