Frequency-Guided Diffusion Model with Perturbation Training for Skeleton-Based Video Anomaly Detection

Xiaofeng Tan[1] Hongsong Wang[1,2] Xin Geng[1,2]
[1]Department of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications
{hongsongwang, xgeng}@seu.edu.cn

Abstract

Video anomaly detection is an essential yet challenging open-set task in computer vision, often addressed by leveraging reconstruction as a proxy task. However, existing reconstruction-based methods encounter challenges in two main aspects: (1) limited model robustness for open-set scenarios, (2) and an overemphasis on, but restricted capacity for, detailed motion reconstruction. To this end, we propose a novel frequency-guided diffusion model with perturbation training, which enhances the model robustness by perturbation training and emphasizes the principal motion components guided by motion frequencies. Specifically, we first use a trainable generator to produce perturbative samples for perturbation training of the diffusion model. During the perturbation training phase, the model robustness is enhanced and the domain of the reconstructed model is broadened by training against this generator. Subsequently, perturbative samples are introduced for inference, which impacts the reconstruction of normal and abnormal motions differentially, thereby enhancing their separability. Considering that motion details originate from high-frequency information, we propose a masking method based on 2D discrete cosine transform to separate high-frequency information and low-frequency information. Guided by the high-frequency information from observed motion, the diffusion model can focus on generating low-frequency information, and thus reconstructing the motion accurately. Experimental results on five video anomaly detection datasets, including human-related and open-set benchmarks, demonstrate the effectiveness of the proposed method. Our code is available at https://github.com/Xiaofeng-Tan/FGDMAD-Code.

1. Introduction

Video anomaly detection (VAD) serves as a research topic that aims to recognize irregular events in videos [5, 22, 33, 39, 41, 43, 44, 46, 47]. Due to the rarity of abnormal events and their ambiguous definitions [1], this problem is often considered a challenging task in unsupervised learning. One promising and effective solution [7, 29, 31] is to train models to learn irregular patterns from normal motions. On this basis, unknown patterns are recognized as anomalies.

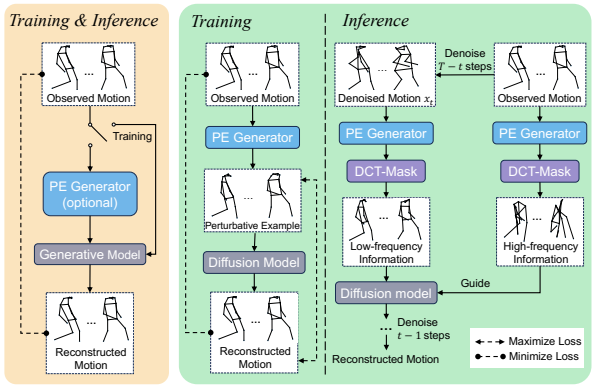


Figure 1. Comparison between existing reconstruction-based methods and the proposed method. The left (yellow) illustrates existing methods, which reconstruct motions and treat those with larger reconstructed errors as anomalies. Particularly, methods with input perturbation utilize the perturbative example (PE) generator, and only during the testing phase. The right (green) demonstrates the proposed method. The PE generator that is trained adversarially is employed in both the training and testing phases. During inference, the frequency information obtained by DCT-mask helps the diffusion model robustly reconstruct the motion.

According to the data of interest, the VAD methods [6, 13, 19, 24, 37, 42, 45] fall into two main categories: RGB-based and skeleton-based methods. Compared to RGB videos, skeletal videos are less sensitive to noise caused by illumination and background clutter [28]. Moreover, skeleton-based methods extract low-dimensional and semantic-rich features that focus on humans [29], making them particularly effective for human-related VAD. Hence,

these methods have received considerable attention in recent years [7, 27, 29, 35].

Based on the learning approaches, skeleton-based methods can be further divided into four categories [28]: reconstruction-based, prediction-based, hybrid methods, and others. Reconstruction-based methods, as one of the most classic deep-learning-based VAD approaches, train models on normal videos and assess anomalies based on reconstruction errors. However, existing reconstruction-based methods may suffer from insufficient diversity [7], overfitting [30], and limited capability along the temporal dimension [38]. As a result, these methods have underperformed compared to the prediction-based and hybrid methods, which incorporate prediction as a more challenging auxiliary task and extract semantic features along the temporal dimension. To address these issues, memory networks [10, 18] and input perturbation [38] have been integrated into autoencoders. Furthermore, recent works [7, 31] have employed advanced generative models, such as diffusion models, as backbones for VAD.

Despite satisfactory progress, the reconstruction-based methods are still inferior to the prediction-based methods, especially for open-set VAD. The reasons can be summarized in two aspects: (1) **Most existing methods do not take model robustness into account**, which may result in previously unseen normal motion being misclassified as anomalies. In fact, the reconstruction-based methods learn normal patterns using consistent inputs and outputs, which may cause the model to learn shortcuts. Consequently, when normal motion is perturbed, the model struggles to reconstruct it using the learned shortcuts, leading to misclassification. A simple yet effective method is to expand the domain of the model by training with perturbed normal motion, and thus enhancing the model's robustness. (2) **The principal and detailed information are treated equally for existing methods, however, their contributions and generation difficulty vary significantly.** Intuitively, generating approximate motions for a given actor is straightforward, but accurately reconstructing these motions is challenging due to individual variations in personal habits. From a signal processing perspective, the principal and detailed information can be represented as high-frequency and low-frequency information, respectively. Therefore, designing a model that is guided by high-frequency information and focused on low-frequency information generation is expected to address this problem.

To this end, we propose a frequency-guided diffusion model with perturbation training, as demonstrated in Fig. 1. To enhance the model robustness, the perturbative example (PE) is generated by a trainable neural network designed to maximize the objective function against the diffusion model. Through perturbation training, the diffusion model, trained on perturbed normal motions, becomes robust and enhances the separability between normal and abnormal events with perturbation. To process the detailed motion components, the discrete cosine transform (DCT) is introduced to deconstruct high-frequency information from observed motions and low-frequency information from generated motions. Guided by the DCT, the intractable high-frequency components can be obtained by observation instead of being generated exclusively by the diffusion model.

In summary, the main contributions are as follows: (1) We propose a reconstruction-based perturbation training model for VAD, which enhances the model robustness through perturbation training between PE generator and diffusion model. Subsequently, input perturbations are introduced to further enhance the separability between normal and abnormal motion during the inference phase. (2) We investigate a frequency-guided denoising process that first decomposes the frequency information of the motion, and then, utilizes the observed high-frequency information of the motion to guide the completion of motion generation. (3) Extensive experiments on five publicly available VAD datasets demonstrate that the proposed method substantially outperforms state-of-the-art (SoTA) methods.

2. Related Work

Reconstruction-Based VAD As one of the most popular VAD methods, reconstruction-based methods [12, 21] typically use generative models to learn to reconstruct the samples representing normal data with low reconstruction error. TSC [21] investigates a temporally-coherent sparse coding for a special type of stacked recurrent neural network (sRNN). To mitigate the overfitting of the reconstruction-based methods, [10, 18, 30] integrate the memory-augmented module to improve these methods. STEAL Net [3] uses a pseudo anomaly generator that synthesizes anomalies only using normal data, which trains the model by both normal data and synthesized anomalies. perturbation example assisted by the OCC technique is introduced to reconstruct only normal data [2]. A latent diffusion-based model is also investigated to generate generic spatio-temporal pseudo-anomalies by inpainting a specific portion of the given image [28].

Nevertheless, these methods leverage reconstruction as the proxy task, which may not effectively extract semantic and discriminative features, leading to inferior performance compared to prediction-based methods. Hence, we propose a DCT-Mask to guide the model in focusing on generating low-frequency information instead of the entire motion, which contains identity-irrelevant features.

Skeleton-Based VAD Due to the well-structured, semantically rich, and highly descriptive nature regarding human actions and motion, skeletal data for VAD has garnered increasing attention in recent years. MPED-RNN [29] extracts global and local features modeled by two RNN

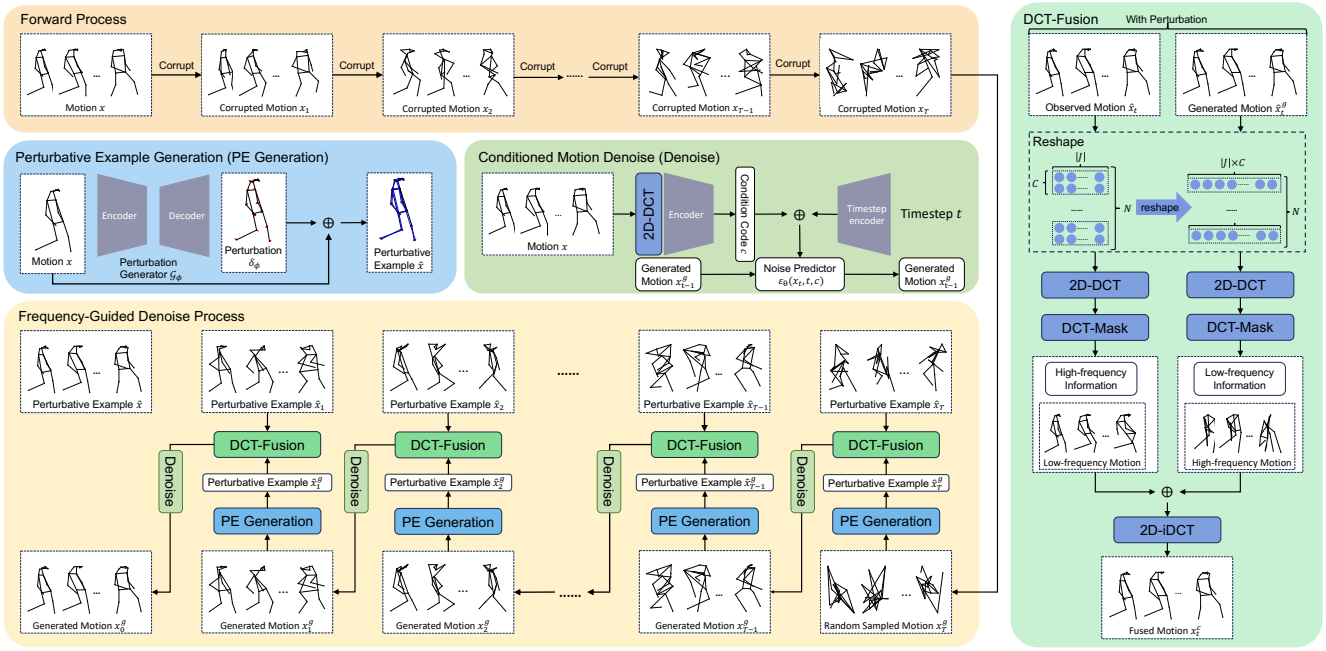


Figure 2. The framework of the proposed method. The model is trained utilizing generated perturbation examples. The training phase includes two processes: minimizing the mean square error to train the noise predictor ε_θ and maximizing this error to train the perturbation generator $\mathcal{G}_\phi$. During the testing phase, the high-frequency information of observed motions and the low-frequency information of generated motions are fused for effective anomaly detection.

branches. Considering multiple timescales, Pred. [32] trains a bi-directional framework for both past and future pose trajectories. GEPC [27] combines reconstruction and clustering, relying on temporal pose graphs to perform clustering in the latent space. Based on the diffusion model, MoCoDAD [7] generates feature motions conditioned on latent representations of past motions. From the perspective of OCC, COSKAD [8] maps normal motions to the latent space and identify anomalies. TrajREC [35], a holistic representation of trajectories, is introduced to identify anomalies in past, present, or future motions. A label-efficient method is presented for representing pose motion in video pose regularity learning [44]. Among the mentioned approaches, most works focus on predicting future frames from past frames or identifying anomalies from latent space using various techniques. Additionally, for some models [29, 32] that can extend to prediction-based and reconstruction-based variants, the prediction-based variant generally performs better. In contrast, we investigate an effective reconstruction-based method for skeleton-based VAD.

AD with Perturbations Skeleton-based VAD is a sub-field of anomaly detection, where some advanced works are performed with perturbations. Thus, we first review these methods. The concept of input perturbation is introduced in [11], which first discovers that neural networks are vulnerable to perturbative examples. By generating such samples, the loss increases with gradient ascent in the input space. Assuming that normal samples are more affected by perturbative samples, some works [14, 17] apply this technique to anomaly detection and have proved its effectiveness for enhancing the separability of anomalies. Specifically, [17] propose an effective anomaly detection method based on tiny perturbations to the input to separate the softmax score distributions between normal and abnormal samples. [14] put forward a modified input preprocessing method without tuning on anomalies. However, these methods apply input perturbations only during the testing phase and may not always be effective, as the underlying assumption may not hold in all cases. In the VAD field, only limited works [38] introduce input perturbation for RGB-based VAD.

3. Methodology

3.1. Preliminaries

For skeleton-based video anomaly detection, the data of interest is denoted as a pose sequence, referred to as motion. Consider the motion $x^{1:N} = \{x^1, x^2, \ldots, x^N\}$, which consists of T human pose from a specific actor. To simply the symbol, $x^{1:N}$ is denoted as x. For skeletal data, motion x_i is represented as a graph (J, A), comprising a body joints set J and an adjacency matrix $A \in \mathbb{R}^{|J|\times|J|}$. In this study, body joints are represented as spatial coordinates in 2D space.

Given observed motions, the diffusion model can be re-

garded as a VAD model that reconstructs motions by denoising them from random sampled noise. Trained on normal motions, the diffusion model can generate normal motions but struggles to produce anomalies, allowing its application in VAD. Given a diffusion timestep t sampled from discrete uniform distribution $\mathcal{U}_{[1,T]}$ and a variance scheduler $\alpha_t \in (0, 1)$, the noise addition process is denoted as

$$x_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\varepsilon, \tag{1}$$

where ε is a noise sampled from $\mathcal{N}(\mathbf{0}, \mathbb{I})$.

For the denoising process, the motion is generated by removing the predicted noise. Prior to this, the noise predictor ε_θ with parameter θ is optimized with the objective loss:

$$\mathcal{L}(x, \theta) = \mathbb{E}_{x,t}[\|\varepsilon - \varepsilon_\theta(x_t, t, c)\|_2^2], \tag{2}$$

where c is conditional code.

Using the trained noise predictor, the reconstructed motions can be generated from sampled noises. For each step, this process can be denoted as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha}{\sqrt{1-\bar{\alpha}}}\varepsilon_\theta(x_t, t, c)) + (1-\alpha)\varepsilon. \tag{3}$$

For a given motion x, an approximate motion $x^g \approx x$ is generated through the denoised process, and the reconstruction error between them is used to measure its anomalous degree:

$$S(x) = \|x - x^g\|_2^2. \tag{4}$$

In response to the issues mentioned in Sec. 1, we propose a frequency-guided diffusion model with perturbation training. Specifically, the motions x are first corrupted, and perturbative samples $\hat{x}$ are generated by perturbation generator $\mathcal{G}_\phi$. Subsequently, the perturbation generator $\mathcal{G}_\phi$ and the noise predictor ε_θ perform perturbation training alternately to obtain a robust model. During the inference phase, the model fuses the low-frequency information from generated motion $\hat{x}_t^g$ with the high-frequency information from observed motion $\hat{x}_t^o$. Furthermore, the frequency-guided denoise model generates motions using the low-frequency information from generated motions and supplements the detail using the high-frequency data from observed motions. The framework is illustrated in Fig. 2.

3.2. Diffusion Model with Perturbation Training

Effect of Perturbative Motion As discussed in Sec. 1, existing reconstruction-based VAD methods suffer from restricted robustness for unseen normal samples. Inspired by the concept of adversarial examples [11], the core of our method is to expand the model's domain by learning from perturbed samples, enabling it to reconstruct unseen normal samples. As shown in Fig. 3, due to the lack of abnormal training motions, reconstruction-based methods are trained on limited normal samples x_k, leading to limited robustness for unseen normal samples.

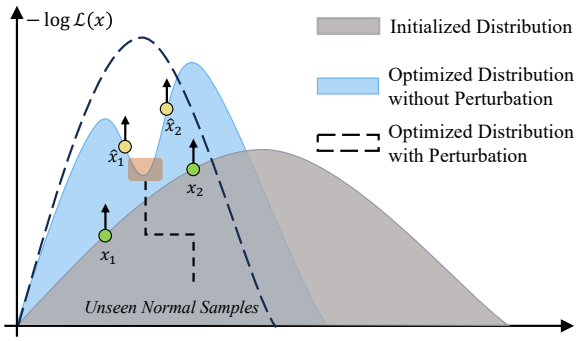


Figure 3. The illustration of perturbation training. The green and yellow points denote the original training x_k and perturbative motion $\hat{x}_k$, respectively. The red region represents the distribution of unseen normal samples.

To this end, we aim to generate samples $\hat{x}_k$ located in these unsatisfied distributions, and then train the model on them to enhance the robustness. In general, these samples $\hat{x}_k$ are similar to normal motions x_k but exhibit relatively high loss values $\mathcal{L}(x_k, \theta)$. In other words, given the network parameters θ and a normal sample x^k, we aim to find a similar but with a high loss value sample $\hat{x}_k$ to train a robust model. With this goal in mind, we consider introducing a small perturbation δ to a given normal motion x_k to generate a potential unseen motion $\hat{x}_k$.

Perturbative Motion Generation First, we focus on finding a small perturbation δ that significantly reduces the loss function, denoted as:

$$\delta = \arg\max_{\delta \in \mathcal{N}(x,\lambda)} \mathcal{L}(x + \delta, \theta), \tag{5}$$

where $\mathcal{N}(x, \lambda)$ represents the norm constraint with maximum perturbation budget λ, and θ denotes the parameter of the model.

In Eq. (5), δ can be approximatively solve by [11]:

$$\delta = \lambda \text{sign}(\nabla_\theta \mathcal{L}(x, \theta)). \tag{6}$$

Hence, the unseen motion $\hat{x}$ can be constructed as $\hat{x} = x + \delta$. In this case, its loss function value $\mathcal{L}(\hat{x}, \theta)$ is relatively larger than that of x, and $\hat{x}$ is similar to x since λ is small. Thus, the motion $\hat{x}_k$ can be used to train for a robust model.

However, computing gradients of diffusion models at each iteration requires substantial computational time and extra memory. Inspired by the perturbation training [26], we treat a lightweight network as a trainable perturbation generator $\mathcal{G}_\phi$ to indirectly predict the optimal perturbation δ_ϕ with a low computational cost:

$$\begin{aligned} \delta_\phi &= \lambda_p \text{sign}(\mathcal{G}_\phi(x)), \\ \hat{x}_\phi &= x + \delta_\phi, \end{aligned} \tag{7}$$

Algorithm 1 Perturbation training for diffusion model

Input: The motions x, the noising steps T, the maximum iterations I_{max}

Output: The noise predictor ε_θ, the perturbation generator $\mathcal{G}_\phi$

1: Encode the conditional code: $c = \text{DCT-Enc}(x)$
2: **for** i = 1,2,3,...,I_{max} **do**
3: Sample the timestep t from $\mathcal{U}_{[1,T]}$
4: Sample Gaussian noise ε from $\mathcal{N}(\mathbf{0}, \mathbb{I})$
5: Add noise ε on x using variance scheduler α_t: $x_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\varepsilon$
6: Obtain perturbative example using perturbation generator: $\hat{x}_t = x_t + \lambda_p \text{sign}(\mathcal{G}_\phi(x_t, t))$
7: Calculate noise prediction loss: $\mathcal{L}(x, \theta, \phi) = \mathbb{E}_{\hat{x},t}[\|\varepsilon - \varepsilon_\theta(\hat{x}_t, t, c)\|_2^2]$
8: Freeze parameters of $\mathcal{G}_\phi$, and update parameters of ε_θ by minimize the loss $\mathcal{L}(x, \theta, \phi)$
9: Repeat the process from line 4 to 7
10: Freeze parameters of ε_θ, and update parameters of $\mathcal{G}_\phi$ by maximize the loss $\mathcal{L}(x, \theta, \phi)$
11: **end for**

where $\mathcal{G}_\phi$ is the perturbation generator with parameters ϕ, and optimized by generating a efficient perturbation:

$$\max_{\phi} \mathcal{L}(x + \lambda_p \text{sign}(\mathcal{G}_\phi(x)), \theta). \tag{8}$$

Perturbation Training Diffusion model ε_θ focus to minimize the loss function $\mathcal{L}(x, \theta)$, while perturbation generator $\mathcal{G}_\phi$ aims to maximize the loss function $\mathcal{L}(\hat{x}, \theta)$, where x is similar to $\hat{x}$. Hence, we optimize them adversarially:

$$\min_{\theta} \max_{\phi} \mathcal{L}(x + \lambda_p \text{sign}(\mathcal{G}_\phi(x)), \theta, \phi), \tag{9}$$

where the loss function $\mathcal{L}(\hat{x}, \theta, \phi)$ is a variant of Eq. (2) that takes into account the perturbation generator $\mathcal{G}_\phi$, denoted as:

$$\mathcal{L}(x, \theta, \phi) = \mathbb{E}_{x,t}[\|\varepsilon - \varepsilon_\theta(\hat{x}_t, t, c)\|_2^2]. \tag{10}$$

Here, given x_t defined by Eq. (1), the corrupted perturbative motions $\hat{x}_t$ is denotes as:

$$\hat{x}_t = x_t + \lambda_p \text{sign}(\mathcal{G}_\phi(x_t, t)). \tag{11}$$

Note that c is the conditional code, which is obtained by selecting the top k largest DCT coefficients.

To sum up, the proposed model adversarially optimizes the perturbation generator and noise predictor during the training phase. Details for perturbation training are provided in Algorithm 1.

3.3. Frequency-Guided Motion Denoise Process

Frequency Information in Motion In signal processing [16], high-frequency information refers to rapid variations or fine details, while low-frequency information represents slower changes or broad features. Similarly, the *low-frequency information* in human motion provides basic outlines of behavior, e.g., the center of gravity, the gesture pose, and action categories. In contrast, *high-frequency information* captures details of the motion. Owing to the diversity of personal habits, high-frequency information tends to vary from person to person, such as the stride length and the extent of hand swing while walking. As shown in Fig. 4 (a), (b), and (c), the motions containing only low-dimensional information are almost identical to the original motions, except for only a few joints. A closer examination of these joints in Fig. 4 (d) reveals that most differences are derived from personal habits, such as the degree of knee bending when walking. In this case, the reconstruction quality, especially that of the joints with high-frequency information, is no longer a reliable indicator for anomaly detection.

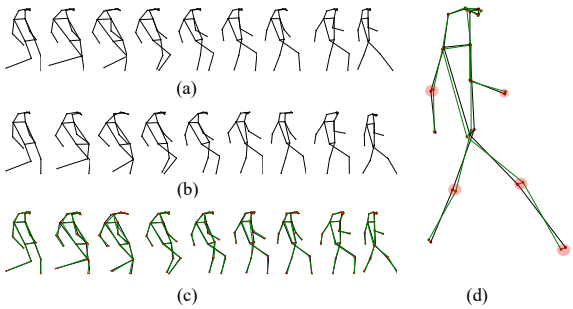


Figure 4. The visualization of human motions processed by 2D-DCT. (a) original motions; (b) motions with low-frequency information only; (c) the comparison between (a) and (b); (d) the skeletal example. Note that the red lines in (d) denote the discarded high-frequency information, and red circles represent the high-frequency joints w.r.t. temporal and spatial dimension.

Due to the diversity of generated motions, it is challenging for generative models to reconstruct details accurately. Inaccurate reconstruction for the details of high-frequency motions should not imply that the generated motions are unrealistic. On the contrary, if the model accurately reconstructs the low-frequency information from the given motions and provides rich high-frequency details, it suggests that the given motion is not an anomaly that deviates from the distribution. However, existing methods biasedly treat the reconstruction of all frequency information equally, limiting the accuracy of detection results.

To this end, we propose a frequency-guided denoise process for anomaly detection. There are three main steps in our pipeline: 1) frequency information extraction; 2) high-frequency and low-frequency information separation; 3) frequency information fusion. Specifically, the frequency-

guided motion denoising process first encodes the conditional code of the input motion using the DCT. Then, Gaussian noise is sampled and added to the motion data using the variance scheduler. Subsequently, for each time step, the motion data is corrupted by the Gaussian noise. Next, the perturbative example is generated using the perturbation generator and both the observed and generated motions are transformed into DCT space. Finally, the observed and generated motion data are combined using a DCT-Mask, and then the fused motion data are converted back to the original space by iDCT.

Frequency Information Extraction Considering both temporal and spatial aspects, we 2D-DCT utilize to extract frequency information. To deal with spatial coordinates, the original motion x is reshaped as a condensed motion $\bar{x} \in \mathbb{R}^{N \times C \cdot |J|}$. The process of 2D-DCT is denoted as [25]:

$$y = \mathrm{DCT}(\bar{x}) = D\bar{x}, \tag{12}$$

where $D \in \mathbb{R}^{N \times N}$ is a pre-defined matrix containing the DCT basis, and y denotes the DCT coefficients by projecting the motion $\bar{x}$ using matrix D. Correspondingly, the inversed Discrete Cosine Transform (iDCT) is defined as [25]:

$$\bar{x} = \mathrm{iDCT}(y) = D^T y. \tag{13}$$

Frequency Information Separation Guided by frequency, low-frequence information can be obtained by the DCT-Mask, denoted as:

$$\mathcal{M}_l(y) = \begin{cases} 1, \text{if } |y_{ij}| \geq \tau, \\ 0, \text{else}, \end{cases} \tag{14}$$

where τ is the top λ largest absolute value of DCT coefficients. Similarly, the DCT-Mask containing high-frequence information of y is denoted as $\mathcal{M}_h(y)$.

Pipeline & Frequency Information Fusion The proposed model utilizes frequency information by fixing the high-frequency components and combining them with low-frequency components to generate motions accurately. Given a motion $\bar{x}_t$ corrupted from observation, and $\bar{x}_t^d$ generated from the denoising process, the corresponding frequency information is obtained by 2D-DCT:

$$y_t^o = \mathrm{DCT}(\bar{x}_t), \quad y_t^g = \mathrm{DCT}(\bar{x}_t^g). \tag{15}$$

Then, the DCT coefficients are masked with the DCT-Mask and combined into a fused coefficient:

$$y_t^c = y_t^o \odot \mathcal{M}_h(y_t^o) + y_t^g \odot \mathcal{M}_l(y_t^g). \tag{16}$$

Subsequently, the fused coefficients are transformed into the original space by iDCT:

$$\bar{x}_t^c = \mathrm{iDCT}(y_t^c). \tag{17}$$

Finally, the coefficient $\bar{x}_t^c$ is reshaped to the motion x_t^c.

During this process, the fused motion is obtained by fusing high-frequency components of observation with low-frequency components of generation. Furthermore, the generated motions of the $t-1$ step are respected as:

$$x_{t-1}^g = \frac{1}{\sqrt{\alpha_t}}(x_t^c - \frac{1-\alpha}{\sqrt{1-\bar{\alpha}}}\varepsilon_\theta(x_t^c, t, c)) + (1-\alpha_t)\varepsilon. \tag{18}$$

4. Experiment

4.1. Experimental Setup

Here, we briefly introduce the datasets and implementation details.

Datasets Five popular benchmark datasets, i.e., Avenue, HR-Avenue [20], HR-STC [21], UBnormal [1], and HR-UBnormal, are used to evaluate the performance. Among them, the UBnormal and HR-UBnormal are open-set benchmarks. The Area Under Curve (AUC) is adopted as the evaluation metric.

Implementation Details Following previous works [7, 29], the data is pre-processed through segmentation and normalization. The model consists of a perturbation generator and a noise predictor, both of which use Graph Convolutional Networks (GCN) as their backbone. During the training phase, all mentioned networks are optimized by Adam with an exponential learning rate scheduler. The base learning rate is set to 0.01, and the multiplicative factor of learning rate decay is 0.99, consistent with the MoCoDAD. To obtain a smoothed curve, the anomaly scores of all objects in all clips inferred by the model are aggregated and padded by the post-processing technique [7, 29]. The hyper-parameter λ_p is set to 0.1 for all datasets, while λ_{dct} is set to 0.9 for the UBnormal and HR-UBnormal, and 0.1 for others.

4.2. Comparison with State-of-the-Art Methods

The performance of the proposed method and SoTA methods is presented in Table 1. We analyze the results in three aspects: comparison with reconstruction-based methods, comparison with skeleton-based methods, and performance on open-set VAD benchmarks.

Reconstruction-Based Methods The performance of all reconstruction-based methods reveals that the proposed method surpasses other reconstruction-based methods and outperforms the previous SoTA results by 1.87%, 3.54%, 4.10%, and 18.79% on four datasets. One possible explanation is that most reconstruction-based methods lack robustness and may suffer from overfitting, resulting in their limited ability to identify unseen normal samples. In contrast, the proposed method introduces perturbative examples with input perturbations to achieve a robust model, enhancing its ability to distinguish between unseen normal samples and abnormal samples that resemble normal ones. Additionally, the proposed model emphasizes low-frequency information,

Table 1. Comparison of the proposed method against other SoTA methods. The best results across all methods are in bold, the second-best ones are underlined, and the superscript ‡ denotes the best performance across all the methods under each paradigm.

Type	Method	Venue	Modality	Avenue	HR-Avenue	HR-STC	UBnormal	HR-UBnormal
Pred.	MPED-RNN-Pred. [29]	*CVPR' 2019*	Skeleton	-	-	74.5	-	-
	Multi-Time. Pred. [32]	*WACV' 2020*	Skeleton	-	88.3	77.0	-	-
	PoseCVAE [15]	*ICPR' 2021*	Skeleton	-	87.8	75.7	-	-
	AMMC [5]	*AAAI' 2021*	RGB	86.6	-	-	-	-
	F^2PN [23]	*T-PAMI' 2022*	RGB	85.7	-	-	-	-
	FPDM [40]	*ICCV' 2023*	RGB	**90.1** ‡	-	-	62.7	
	TrajREC-Ftr. [35]	*WACV' 2024*	Skeleton	-	<u>89.4</u> ‡	<u>77.9</u> ‡	68.0 ‡	68.2 ‡
Hybrid	MPED-RNN [29]	*CVPR' 2019*	Skeleton	-	86.3	75.4	60.6	61.2
	sRNN [22]	*T-PAMI' 2021*	RGB	83.5 ‡	-	-	-	-
	MoCoDAD [7]	*ICCV' 2023*	Skeleton	-	89.0 ‡	77.6 ‡	<u>68.3</u> ‡	<u>68.4</u> ‡
Others	GEPC [27]	*CVPR 2020*	Skeleton	-	58.1	74.8	53.4	55.2
	COSKAD-Hype. [8]	*PR' 2024*	Skeleton	-	87.3	75.6	64.9	65.5 ‡
	COSKAD-Eucli. [8]	*PR' 2024*	Skeleton	-	87.8 ‡	77.1 ‡	65.0 ‡	63.4
	EVAL [34]	*CVPR' 2023*	RGB	86.0	-	-	-	-
	OVVAD [39]	*CVPR' 2024*	RGB	86.5 ‡	-	-	62.9	-
Rec.	MPED-RNN-Rec. [29]	*CVPR' 2019*	Skeleton	-	-	74.4	-	-
	TrajREC-Prs. [35]	*WACV' 2024*	Skeleton	-	86.3	73.5	-	-
	TrajREC-Pst.[35]	*WACV' 2024*	Skeleton	-	87.6	75.7	-	-
	ST-PAG [31]	*CVPR' 2024*	RGB	86.5	-	-	58.0	-
	Ours	-	Skeleton	<u>88.0</u> ‡	**90.7** ‡	**78.6** ‡	**68.9** ‡	**69.0** ‡

representing the principal components of motion, thereby reducing the impact of difficult-to-generate high-frequency information. As a result, the proposed method achieves the best performance among all reconstruction-based methods.

Skeleton-Based Methods Additionally, we compare the proposed method with the skeleton-based methods, achieving the best result. Notably, among these methods, reconstruction-based methods are significantly inferior to other methods. For instance, the previous SoTA reconstruction-based method, TrajREC-Pst.[35] achieves an accuracy of only 75.7 on the HR-STC dataset, whereas TrajREC-Frt.[35] achieves 77.9 and MoCODAD [7] achieves 77.6. Similar trends are observed across other datasets. In contrast, the proposed method aims to reconstruct samples based on completing the low-frequency components from high-frequency information. As a result, the proposed method achieves AUC scores of 88.0, 90.7, 78.6, 68.9, and 69.0 on selected datasets, surpassing other skeleton-based methods. It can be attributed to certain abnormal events in Avenue that are unrelated to humans or undetectable by the pose detector [29]. By removing such frames, the proposed method achieves an AUC score of 90.7, verifying its effectiveness.

Open-Set Benchmarks Furthermore, on the open-set datasets, i.e., UBnormal and HR-UBnormal, the proposed model exceeds all methods. The results reveal that the proposed method utilizes perturbation training to extend the domain and enhance robustness, allowing it to adapt to VAD in open-set scenarios.

Table 2. Comparison with supervised and weakly supervised methods. "W.S.", "W.S.", and "S." denote unsupervised, weekly supervised, and supervised methods, respectively.

Method	Training Type	Params	UBnormal
Sultani et. al [36]	S.	-	50.3
AED-SSMTL [9]	S.	>80M	61.3
TimeSformer [4]	S.	121M	68.5
AED-SSMTL [9]	W.S.	>80M	59.3
Ours	U.S	**556K**	**68.9**

Comparison with Supervised and Weakly Supervised Methods Table 2 evaluates the proposed method with supervised and weakly supervised methods. The proposed method outperforms existing methods with fewer parameters. Even without supervision or visual information, our approach performs competitively with methods that utilize different types of supervision. Additionally, our approach boasts a significantly smaller parameter count compared to its competitors.

4.3. Ablation Studies

Ablation studies involve four models for comparison, called "baseline", "Ours w/o IP", "Ours w/ double IP", and "Ours

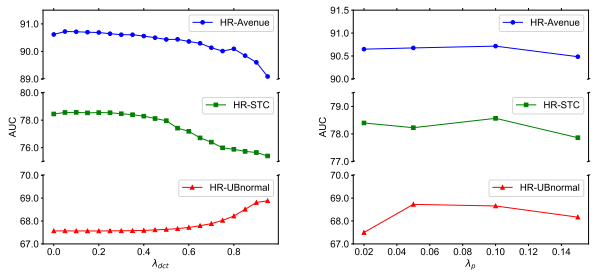
Table 3. Ablation studies of each component in the proposed method. We report the results on three HR datasets. The baseline refers to MoCoDAD-E2E, which is studied in [7]. The second row indicates the result of the model without input perturbations (IP) in the training phase, and the third row shows the result when the weight of input perturbations is 0.1 during the training phase but is doubled during the testing phase. The fourth row presents the result of the model without DCT-mask, which is replaced with the temporal-mask. The best results are highlighted in bold.

Method	HR-Avenue	HR-STC	HR-UBnormal
Baseline	87.5 (↓ 3.2)	75.2(↓ 3.4)	64.4 (↓ 4.6)
Ours w/o IP	90.4 (↓ 0.3)	77.4 (↓ 1.2)	68.7 (↓ 0.3)
Ours w/ double IP	90.7 (-)	78.5 (↓ 0.1)	68.6 (↓ 0.4)
Ours w/o DCT-Mask	89.9 (↓ 0.8)	78.0 (↓ 0.6)	68.1 (↓ 0.9)
Ours	**90.7**	**78.6**	**69.0**

w/o DCT-Mask".

Effect of DCT-Mask By examining the fourth row and the fifth row in Table 3, the results depict that the DCT-Mask is beneficial for anomaly detection, yielding improvements of 0.89%, 0.77%, and 1.17%. For generative models, it is challenging to accurately reconstruct motion details. Thanks to the DCT-Mask, the proposed method can focus on generating low-frequency information with guidance of high-frequency information, leading to satisfying results.

Effect of Perturbation Training In Table 3, the second row reports the results of the model without perturbation training, indicating its effectiveness for obtaining a robust model. To further verify this, we increased the magnitude of input perturbations only during testing, and the results are presented in the third row. The results remained unchanged on the HR-Avenue and declined by only 0.1% and 0.3% on the others, demonstrating the robustness of the model.



(a) (b)

Figure 5. Sensitivity analyses of DCT-Mask threshold and input perturbations weight. (Sensitivity analysis of the threshold of DCT-Mask. Here, $\lambda_{dct} \in \{0.05, 0.10, 0.15, \ldots, 0.90, 0.95\}$.) (b) Sensitivity analysis of the weight of input perturbations. Here, $\lambda_p \in \{0.02, 0.05, 0.10, 0.15\}$.

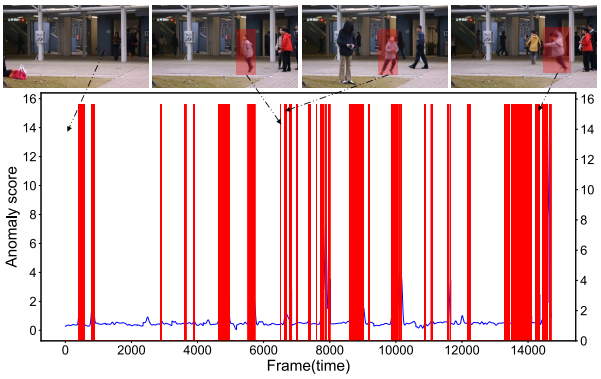


Figure 6. Anomaly score curves on HR-Avenue datasets.

4.4. Parameter Analysis and Visualizations

Parameter Sensitivity The proposed method includes two parameters, the DCT-Mask and the weight of the input perturbations. Fig. 5 (a) and 5 (b) illustrate the evaluation results of corresponding parameters on the three HR datasets. Theoretically, a smaller value of λ_{dct} indicates that the proposed model relies less on high-frequency information and is more aligned with a generative model.

For HR-Avenue and HR-STC, Fig. 5 (a) depicts that the optimal results appear around $\lambda_p = 0.1$. Interestingly, a contrasting pattern is observed for the HR-UBnormal dataset. We attribute this phenomenon to the diverse poses present in UBnormal, which is designed for supervised open-set VAD and exhibits diverse human poses. Consequently, the proposed model requires a larger threshold to effectively capture sufficient information for the accurate generation of human poses. Results in Fig. 5(b) exhibit a consistent trend, where AUC increases as the value of λ_p ranges from 0 to 0.1 and decreases for values greater than 0.1. This trend suggests that input perturbations with appropriate magnitude facilitate the training of robust models.

Visualizations Fig. 6 shows the anomaly score w.r.t. video clips across the three datasets. The results show that the proposed method is sensitive to anomalies and can effectively detect anomalous events. For example, as shown in Fig. 6, the anomaly score rises sharply when a kid jumps, and then, the anomaly curve returns to normal.

5. Conclusion

In this paper, we introduce a novel frequency-guided diffusion model with perturbation training for video anomaly detection. To enhance model robustness, we propose a perturbation training paradigm to broaden the domain of the reconstructed model. Furthermore, we utilize generated perturbative examples for inference to improve the separability between normal and abnormal motions. Focusing on motion details generation, a frequency-guide motion denoise

process is investigated. Extensive empirical results demonstrate that the proposed method outperforms other SoTA methods.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20111–20121, 2022. 1, 6

[2] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. In *British Machine Vision Conference*, 2021. 2

[3] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 207–214, 2021. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 7

[5] 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*, pages 938–946, 2021. 1, 7

[6] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20392–20401, 2023. 1

[7] Alessandro Flaborea, Luca Collorone, 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 *IEEE/CVF International Conference on Computer Vision*, pages 10284–10295, 2023. 1, 2, 3, 6, 7, 8

[8] Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Stefano D'arrigo, Marco Aurelio Sterpa, Alessio Sampieri, and Fabio Galasso. Contracting skeletal kinematics for human-related video anomaly detection. *Pattern Recognition*, page 110817, 2024. 3, 7

[9] Mariana-Iuliana Georgescu, Antonio Bărbălău, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12737–12747, 2021. 7

[10] 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 *IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019. 2

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 3, 4

[12] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016. 2

[13] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023. 1

[14] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10948–10957, 2020. 3

[15] Yashswi Jain, Ashvini Kumar Sharma, Rajbabu Velmurugan, and Biplab Banerjee. Posecvae: Anomalous human activity detection. In *International Conference on Pattern Recognition*, pages 2927–2934, 2021. 7

[16] Edmund Y Lam and Joseph W Goodman. A mathematical analysis of the dct coefficient distributions for images. *IEEE Transactions on Image Processing*, 9(10):1661–1666, 2000. 5

[17] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 3

[18] 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 *IEEE/CVF International Conference on Computer Vision*, pages 13568–13577, 2021. 2

[19] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24500–24510, 2023. 1

[20] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 6

[21] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 6

[22] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1070–1084, 2021. 1, 7

[23] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future frame prediction network for video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7505–7520, 2022. 7

[24] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for

weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2023. 1

[25] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 9488–9496, 2019. 6

[26] Yuhao Mao, Mark Müller, Marc Fischer, and Martin Vechev. Connecting certified and adversarial training. In *Advances in Neural Information Processing Systems*, pages 73422–73440. Curran Associates, Inc., 2023. 4

[27] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10536–10544, 2020. 2, 3, 7

[28] Pratik K. Mishra, Alex Mihailidis, and Shehroz S. Khan. Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(2):1073–1085, 2024. 1, 2

[29] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11988–11996, 2019. 1, 2, 3, 6, 7

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14360–14369, 2020. 2

[31] Ayush K Rai, Tarun Krishna, Feiyan Hu, Alexandru Drimbarean, Kevin McGuinness, Alan F Smeaton, and Noel E O'connor. Video anomaly detection via spatio-temporal pseudo-anomaly generation: A unified approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3887–3899, 2024. 1, 2, 7

[32] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2615–2623, 2020. 3, 7

[33] Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10330–10340, 2023. 1

[34] Ashish Singh, Michael J. Jones, and Erik G. Learned-Miller. Eval: Explainable video anomaly localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18717–18726, 2023. 7

[35] Alexandros Stergiou, Brent De Weerdt, and Nikos Deligiannis. Holistic representation learning for multitask trajectory anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6729–6739, 2024. 2, 3, 7

[36] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 7

[37] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22846–22856, 2023. 1

[38] Yizhou Wang, Can Qin, Yue Bai, Yi Xu, Xu Ma, and Yun Fu. Making reconstruction-based method great again for video anomaly detection. In *IEEE International Conference on Data Mining*, pages 1215–1220, 2022. 2, 3

[39] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024. 1, 7

[40] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *IEEE/CVF International Conference on Computer Vision*, pages 5504–5514, 2023. 7

[41] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5527–5537, 2023. 1

[42] 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. 1

[43] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18899–18908, 2024. 1

[44] Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao Gan, Cewu Lu, and Wei Wu. Regularity learning via explicit distribution modeling for skeletal video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 3

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023. 1

[46] Yuanhong Zhong, Xia Chen, Yongting Hu, Panliang Tang, and Fan Ren. Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8285–8296, 2022. 1

[47] Yuanhong Zhong, Ruyue Zhu, Ge Yan, Ping Gan, Xuerui Shen, and Dong Zhu. Inter-clip feature similarity based weakly supervised video anomaly detection via multi-scale temporal mlp. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1