Scene-Dependent Video Anomaly Detection via Discriminative-Contrastive Learning from Intrinsic Scene Labels

Xichen Tan
National University of Defense Technology
Changsha, Hunan, China
tanxc23@nudt.edu.cn

Guang Yu*
National University of Defense Technology
Changsha, Hunan, China
guangyu@nudt.edu.cn

Yuanjing Luo
Central South University of Forestry and Technology
Changsha, Hunan, China
luoyuanjing@csuft.edu.cn

Chang Liu
National University of Defense Technology
Changsha, Hunan, China
liudawn@nudt.edu.cn

Siqi Wang
National University of Defense Technology
Changsha, Hunan, China
wangsiqi10c@nudt.edu.cn

Fang Liu
Hunan University
Changsha, Hunan, China
fangl@hnu.edu.cn

Zhiping Cai
National University of Defense Technology
Changsha, Hunan, China
zpcai@nudt.edu.cn

Abstract

Scene-dependent video anomaly detection (SDVAD) is a recently emerging and valuable topic. Existing methods typically rely on scene-conditioned generative learning, which implicitly embeds scene context without explicitly considering the intrinsic video scene labels and the connection between foreground events and their corresponding scenes. To this end, we propose a novel discriminative-contrastive learning framework, which can explicitly exploit scene labels and bridge foreground events and scenes for effective SDVAD. First, we deploy a discriminative learning process by directly taking scene class labels from different surveillance cameras as learning targets. It trains deep neural networks (DNNs) to classify events from different scenes, and anomalies are determined based on scene-aware classification probabilities in inference. Second, to fully establish the connection between events and scenes, we develop a contrastive learning strategy, which pulls events closer to their corresponding scenes while pushing them away from other scenes. During inference, the distance between events and scenes is utilized to perform anomaly scoring for further enhancing performance. Finally, we design a joint discriminative-contrastive DNN, where the discriminative module embeds scene modulation signals into multiple layers with learnable weights to properly incorporate scene information, and the contrastive module utilizes a memory bank to augment contrastive samples, thus strengthening the connection between events and scenes. Extensive experiments on commonly-used SDVAD datasets demonstrate that our method achieves state-of-the-art performance with up to 8.5% AUC improvement over existing methods.

CCS Concepts

• **Computing methodologies** → **Scene anomaly detection**.

Keywords

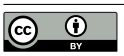
Video Anomaly Detection, Discriminative Learning, Contrastive Learning

ACM Reference Format:
Xichen Tan, Guang Yu, Yuanjing Luo, Chang Liu, Siqi Wang, Fang Liu, and Zhiping Cai. 2026. Scene-Dependent Video Anomaly Detection via Discriminative-Contrastive Learning from Intrinsic Scene Labels. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835493

1 Introduction

As a critical enabling technology for public security and intelligent monitoring, video anomaly detection (VAD) has attracted sustained attention in multimedia research and responsible AI [15, 38, 39, 44, 46, 48]. Most existing VAD methods adopt a semi-supervised setting, using large collections of normal videos to learn normal patterns [8]. They typically employ generative objectives, such as reconstruction [6, 13, 25, 41, 49] or prediction [5, 7, 24, 32, 50], and identify anomalies based on high generative errors. However, these methods mainly characterize deviations from learned appearance and motion patterns, providing limited evidence for why an event is anomalous in a particular scene. This limitation is especially relevant to scene-dependent settings, where the same behavior can be normal in one scene but abnormal in another.

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835493

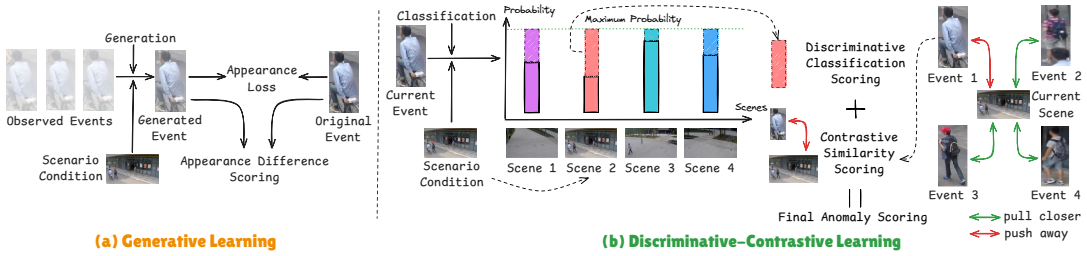


Figure 1: Comparison of different SDVAD paradigms. (a) illustrates generative learning paradigms, which implicitly embed scene information and derive anomaly scores from appearance-level differences between the original event and its generated counterpart. In contrast, our discriminative-contrastive paradigm in (b) explicitly combines scene-conditioned event classification with contrastive event-scene similarity, offering a more direct and semantically grounded basis for anomaly scoring.

Early methods typically focus on single-scene VAD [31], where a video dataset from a fixed surveillance camera is collected to construct a normality model tailored to that scenario. Since multiple surveillance cameras are usually deployed at different locations and angles in public area, subsequent methods have gradually shifted from single-scene to multi-scene VAD [12, 23, 27, 35]. They collect video data from multiple scenes to build models capable of detecting abnormal events across various scenes. It is important to note that the manifestations of anomalies vary across different scenes, i.e. anomalies exhibit ***scene dependence*** [5], which means that an event is normal in one scene but anomalous in another. For example, running on a playground is normal, but running in a bank constitutes a notable abnormal event. Obviously, scene-dependent anomalies are prevalent in daily life, and detecting such anomalies holds significant practical value. However, most methods fail to consider the scene dependence of anomalies. On the one hand, these methods do not explicitly incorporate scene information during training, thus tending to classify the same type of event (e.g. running) as anomalous across different scenes. On the other hand, most existing VAD datasets (e.g. ShanghaiTech [22]) do not include scene-dependent anomalies, as the abnormality of the same event remains consistent across different scenes. This is also the reason why most methods overlook scene-dependent anomalies. They are not penalized for failing to detect such anomalies on these datasets.

To advance the research on scene-dependent video anomaly detection (SDVAD), recently Cao et al. propose the first SDVAD dataset named NWPU Campus [5]. It is currently the largest and only semi-supervised VAD dataset that includes scene-dependent anomalies. Subsequently, a few SDVAD methods [5, 7, 50] have been proposed and validated on this dataset. Their key idea is to embed scene information into the model to influence model training, thereby integrating scene information for VAD. For instance, Cao et al. [5] propose to embed scene images into a conditional variational autoencoder when training a generative model for video frame prediction learning.

Despite certain progress, we have observed some valuable insights: Current SDVAD methods usually do not explicitly consider ***intrinsic video scene labels***, which provide crucial information for distinguishing scene-dependent anomalies. Specifically, *since video events from various scenes are captured by different surveillance cameras, they inherently carry natural scene class labels from different cameras (i.e. camera IDs [50]), which contain direct information about which events belong to which scenes.* However, as shown in Fig. 1(a), current SDVAD methods typically follow classic generative learning, which implicitly embeds scene information (e.g. scene images) into generative models without utilizing scene labels in generative learning objectives (e.g. mean square error minimization). This may result in the limited impact of scene information. Besides, existing SDVAD methods do not explicitly establish the connection between foreground events and scene context, as generative learning objectives usually do not include direct relationships between foregrounds and scenes.

To alleviate the aforementioned issues, we propose a novel discriminative-contrastive learning framework that can achieve effective SDVAD by explicitly leveraging scene labels and the connection between foreground events and scene context. As shown in Fig. 1(b), the core idea is to perform discriminative learning on foreground events from different scenes based on scene class labels, while conducting contrastive learning using the affiliation relationships between foreground events and scenes. During inference, events that cannot be correctly discriminated or establish correct connection with the corresponding scenes are detected as abnormal events. This dual-objective design enables the model to capture both categorical event-scene relationships and continuous event-scene semantic similarity, yielding more discriminative and robust representations for anomaly detection. Specifically, our contributions are summarized as follows:

- To explicitly utilize scene labels, we propose a discriminative learning-based SDVAD method. It directly takes scene class labels of different cameras as learning targets to train deep neural networks (DNNs) for classifying foreground events from different scenes. During inference, we propose to use scene-aware classification probabilities based on scene labels to score events for SDVAD.
- To fully consider the connection between foreground events and scenes, we develop a contrastive learning strategy. It reduces the distance between foreground events and their corresponding scenes while increasing the distance from other scenes. The distance is also used for anomaly scoring in inference, thus leveraging the connection between foregrounds and scenes to enhance SDVAD.
- We further design a joint discriminative-contrastive DNN for effective learning, where the discriminative module embeds scene modulation information into multiple network layers with learnable weights to balance foreground events and scene information, and the contrastive module introduces

a memory bank to store and update contrastive samples in learning process, thereby strengthening the connection between events and scenes.

To verify the effectiveness of the proposed method, we conduct experiments on commonly-used SDVAD datasets as well as the scene-independent ShanghaiTech dataset. Experimental results demonstrate that our method achieves state-of-the-art performance on scene-dependent datasets with an obvious AUC improvement (up to 8.5%), while also exhibiting strong generalizability on scene-independent settings.

2 Related Work

2.1 Video Anomaly Detection Methods

Video anomaly detection (VAD) is inherently challenging due to the extreme scarcity of anomalous samples and the assumption that training data contain only normal behaviors. Consequently, most existing methods adopt a semi-supervised learning paradigm, where the model is trained exclusively on normal samples and detects anomalies as deviations from learned patterns. Current semi-supervised VAD approaches can be broadly categorized into two types: reconstruction-based [3, 6, 12–14, 25, 37, 41, 49], and prediction-based [4, 9, 11, 16, 17, 19, 23, 24, 27, 32, 34, 35] methods. These methods are generative in nature, relying on the assumption that anomalies are deviations from the learned distribution of normal data. Despite their success, these approaches are scene-agnostic and do not explicitly account for the fact that the same behavior may be normal in one scene but abnormal in another. As a result, they fail to capture the contextual dependency between scene semantics and anomalous behaviors, which is essential for reliable detection in real-world multi-scene environments.

The release of the NWPU Campus dataset [5] has drawn increasing attention to scene-dependent video anomaly detection (SDVAD). Cao et al. [5, 7] and subsequent works [50] proposed a series of scene-aware generative methods, which incorporate scene context or scene embeddings into reconstruction- or prediction-based frameworks. Specifically, Cao et al. [5] integrated the encoded scene image as a conditional input into a Conditional Variational Auto-Encoder (CVAE) to generate scene-related latent features and decode them into predicted frames. Later, Cao et al. [7] employed a Hierarchical Variational Auto-Encoder (HVAE) to extract scene-generic features and a Tiny Auto-Encoder to capture scene-specific latent representations, which were then fused within a Score-based Diffusion Model (SDM) and decoded through the HVAE decoder for scene-dependent prediction. Building upon this line of work, Zhang et al. [50] proposed a Noise-Conditioned Score Transformer (NCST), which adds noise to input video frames, divides them into patches, and jointly models scene information to predict noise-conditioned patches, with anomaly scores computed via Autoregressive Denoising Score Matching (ADSM). While these methods effectively integrate scene information into generative frameworks, they primarily embed scene context implicitly, rather than fully exploiting the explicit scene labels.

2.2 Video Anomaly Detection Datasets

This work mainly focuses on semi-supervised video anomaly detection, where models are trained solely on normal data without video-level annotations. Consequently, weakly-supervised datasets such as UCF-Crime [36] and XD-Violence [43] that rely on video-level labels are not considered. Representative datasets commonly used for semi-supervised VAD include UCSD Ped1 & Ped2 [42], Subway Entrance & Exit [2], UMN [40], CUHK Avenue [20], ShanghaiTech [22], Street Scene [30], IITB Corridor [34], UBnormal [1], and NWPU Campus [5].

Among them, only UMN, ShanghaiTech, UBnormal, and NWPU Campus contain multiple scene types, while the others are single-scene datasets that lack diversity in spatial or contextual conditions. The UMN dataset does not provide an official train-test split, which limits its consistency for benchmarking. The UBnormal dataset is synthetically generated using animated characters, leading to a domain gap between synthetic and real-world videos. Among the remaining real-world multi-scene datasets, ShanghaiTech represents a multi-scene but *scene-independent* dataset, where the definition of anomaly is consistent across all scenes, while NWPU Campus represents a multi-scene and *scene-dependent* dataset, where the same event may be normal in one scene but anomalous in another.

Specifically, NWPU Campus includes typical scene-dependent anomalies such as *cycling on roadway (normal)* versus *cycling on square (abnormal)*. Building upon ShanghaiTech, the authors of NWPU Campus further constructed a scene-dependent extension named ShanghaiTech-sd, which includes four distinct scenes where *cycling* is defined as a scene-dependent anomaly. These datasets collectively highlight the evolution of VAD research from single-scene, context-agnostic settings toward multi-scene, context-aware anomaly detection, motivating the development of our label-guided scene-dependent framework. In this work, we focus on NWPU Campus and ShanghaiTech-sd for scene-dependent evaluation, and additionally include ShanghaiTech to verify the generalizability of our method on scene-independent multi-scene settings.

3 The Proposed Discriminative-Contrastive Learning Framework for SDVAD

As shown in Fig. 2, our proposed SDVAD framework integrates two complementary learning paradigms (i.e. discriminative learning and contrastive learning) via a joint discriminative-contrastive DNN. Following the classic semi-supervised VAD setting, all training videos contain only normal events, without exposure to anomalies or manually annotated behavior labels. Accordingly, our framework uses intrinsic camera-based scene labels to learn normal event-scene matching relations and identifies test events that violate these learned relations as anomalies. Unlike generative approaches that implicitly exploit scene information, our framework jointly leverages scene labels and event-scene relationships. Within this framework, discriminative learning provides scene-aware supervisory signals by classifying each event into its corresponding scene, while contrastive learning enhances the semantic alignment between events and scenes through similarity-driven objectives. These two learning objectives are cohesively implemented through a dedicated scene-conditioned DNN, which serves as the architectural backbone enabling hierarchical fusion of event and scene features.

It should be noted that our framework does *not* assume a one-to-one mapping between event types and scenes. The same event (e.g., walking, cycling) can occur across multiple scenes and may

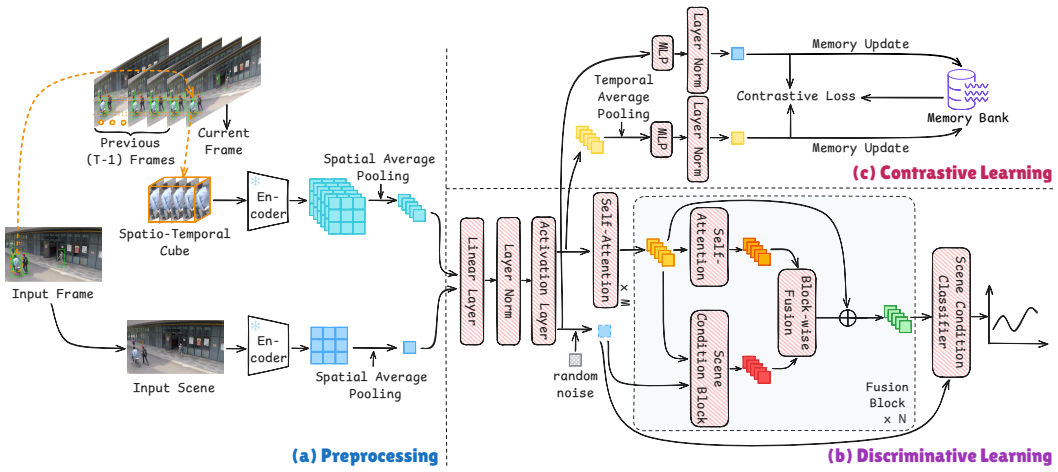


Figure 2: Overview of the proposed discriminative-contrastive framework for scene-dependent video anomaly detection.

be normal in all of them, which is consistent with the used datasets. The key insight is that *the same event exhibits distinct contextual patterns in different scenes*, and scene information is integrated to capture these scene-conditioned patterns rather than to bind events rigidly to a single scene identity. For scene-invariant anomalies (e.g., fighting, shooting), they are absent from all normal training data and thus unmatched to any scene, causing both the discriminative and contrastive objectives to consistently assign them high anomaly scores at test time.

3.1 Discriminative Learning-Based SDVAD

Motivation. The core nature of SDVAD lies in that anomalies are *defined with respect to a specific scene*, making it essentially an *event-scene mismatch* problem. Although prior generative paradigms incorporate scene information during training or inference as conditional guidance, their underlying motivation remains the same: normal events are easy to reconstruct or predict, whereas abnormal ones are not. As a result, these methods primarily capture appearance-level discrepancies rather than the semantic matching between an event and its scene. In contrast, the motivation behind a discriminative learning paradigm is more aligned with the essence of SDVAD: if a model is unable to classify a foreground event into the correct scene, the event is likely inconsistent with that scene and therefore potentially anomalous. This provides a natural and direct formulation of scene-dependent anomaly detection.

Learning Objective. To capture the semantic matching between an event and the scene in which it occurs, we formulate the discriminative learning objective as a scene-conditioned event classification problem. Let the data distribution over event-scene pairs and their scene labels be $(e_i, s_j, y) \sim p(e, s, y)$, where $y \in \{1, \ldots, K\}$ denotes the scene identity. The objective is to learn a scene-conditioned classifier g_θ that minimizes the expected negative log-likelihood of the correct scene label:

$$\min_\theta \mathbb{E}_{(e_i,s_j,y)\sim p(e,s,y)} \left[-\log p_\theta\big(y \mid f_{e,i}, f_{s,j}\big) \right]. \tag{1}$$

Concretely, given the event and the corresponding scene embeddings $f_{e,i}, f_{s,j} \in \mathbb{R}^D$, the classifier predicts scene logits:

$$z = g_\theta(f_{e,i}, f_{s,j}), \tag{2}$$

where $z \in \mathbb{R}^K$. The conditional distribution over scenes is obtained using the softmax:

$$\hat{y}_k = p_\theta(y = k \mid f_{e,i}, f_{s,j}) = \frac{\exp(z_k)}{\sum_{m=1}^{K} \exp(z_m)}, \tag{3}$$

where $k = 1, \ldots, K$. Given the one-hot ground-truth label $y \in \{0, 1\}^K$ and its index y_{scene}, the instantiated negative log-likelihood corresponds to the cross-entropy loss:

$$\mathcal{L}_{\text{classify}} = -\log \hat{y}_{y_{\text{scene}}}. \tag{4}$$

During inference, anomaly scoring is scene-aware: a mismatch between an event and its scene lowers the classifier's confidence in the correct scene. Thus, the anomaly score is derived from the model's uncertainty regarding the ground-truth scene:

$$S_{\text{classify}} = 1 - \hat{y}_{y_{\text{scene}}}. \tag{5}$$

3.2 Contrastive Learning-Based SDVAD

Motivation. Although scene-label classification provides a strong discriminative signal, the supervision it offers lacks fine-grained relational constraints. It only learns to associate an event with a single categorical label, without explicitly modeling *how strongly* an event corresponds to its scene or *how different* it is from other scenes. To this end, we introduce a symmetric contrastive learning objective that quantifies the alignment between event and scene embeddings, thereby improving the model's ability to distinguish abnormal behaviors that deviate from scene context.

Learning Objective. The contrastive learning objective is formulated under the same probabilistic setting as discriminative learning. Given event-scene pairs $(e_i, s_j, y) \sim p(e, s, y)$, the goal is to learn event-scene alignment by pulling together embeddings sharing the same scene label while pushing apart those from different scenes. Formally, the objective is to minimize the expected contrastive loss:

$$\min_\theta \mathbb{E}_{(e_i,s_j,y)\sim p(e,s,y)} \left[\mathcal{L}_{\text{contrast}}(e_i, s_j, y; \theta) \right], \tag{6}$$

where $\mathcal{L}_{\text{contrast}}$ is a bidirectional event-scene contrastive loss defined below. For the given event embeddings $f_{e,i} \in \mathbb{R}^D$ and scene

embeddings $f_{s,j} \in \mathbb{R}^D$, we construct positive and negative pairs based on their scene labels $y_{e,i}$ and $y_{s,j}$:

$$m_{i,j} = \begin{cases} 1, & \text{if } y_{e,i} = y_{s,j}, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $m_{i,j}$ indicates whether the i_{th} event and the j_{th} scene share the same scene label. We use cosine similarity with temperature τ:

$$\text{sim}(f_{e,i}, f_{s,j}) = \frac{f_{e,i}^\top f_{s,j}}{\tau ||f_{e,i}|| ||f_{s,j}||}. \tag{8}$$

The event→scene contrastive loss is defined as:

$$\mathcal{L}_{e2s} = -\sum_i \log \frac{\sum_j m_{i,j} \exp(\text{sim}(f_{e,i}, f_{s,j}))}{\sum_j \exp(\text{sim}(f_{e,i}, f_{s,j}))}. \tag{9}$$

Symmetrically, the scene→event loss is:

$$\mathcal{L}_{s2e} = -\sum_j \log \frac{\sum_i m_{i,j} \exp(\text{sim}(f_{s,j}, f_{e,i}))}{\sum_i \exp(\text{sim}(f_{s,j}, f_{e,i}))}. \tag{10}$$

It is worth noting that although cosine similarity itself is symmetric, $\mathcal{L}_{e2s}$ and $\mathcal{L}_{s2e}$ are *not* redundant: $\mathcal{L}_{e2s}$ sums over events (index i), contrasting each event against all scenes in the batch, while $\mathcal{L}_{s2e}$ sums over scenes (index j), contrasting each scene against all events. Their comparison targets therefore differ, and combining both directions ensures that the alignment is enforced from both the event and scene perspectives. The full contrastive objective combines both directions:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{2}(\mathcal{L}_{e2s} + \mathcal{L}_{s2e}). \tag{11}$$

During inference, the same similarity function is used to assess event-scene alignment. A foreground event that poorly aligns with its scene yields a lower similarity value and is thus more likely to be abnormal. The contrastive anomaly score is defined as:

$$S_{\text{contrast}} = 1 - \text{sim}(f_{e,i}, f_{s,j}), \tag{12}$$

which provides a complementary measure of event-scene discrepancy beyond classification uncertainty.

3.3 Joint Discriminative-Contrastive DNN

To realize the above learning objectives, we design a joint DNN integrating both discriminative and contrastive modules, detailed as follows: **Preprocessing**. To accurately localize and represent foreground events, we follow the event construction pipeline of [47] to perform object localization. For each localized foreground object, we retrieve the identically positioned patches from the previous $(T-1)$ frames to form a spatio-temporal cube, which allows DNN to incorporate short-term temporal dynamics. Each cube thus corresponds to a single foreground event, capturing its localized motion and appearance evolution over time. Further details of the cube construction pipeline are provided in the supplementary material.

To obtain discriminative and robust representations, we extract features using a frozen DINOv2 encoder [26], ensuring stable and semantically rich scene and event embeddings without additional parameter updates. For the scene frame, the encoder first produces a set of spatial tokens, which we average to obtain a global scene embedding $f_{s,j} \in \mathbb{R}^D$. For each foreground event, the spatial tokens of its spatio-temporal cube are also mean-pooled per frame, yielding an event representation $f_{e,i} \in \mathbb{R}^{T \times D}$, where T is the number of patches in the cube. When computing the contrastive similarity in Eq. 8, the event embedding $f_{e,i}$ is further averaged along the temporal dimension T to obtain a single vector representing its overall spatio-temporal semantics. Note that this temporal averaging is applied exclusively for the contrastive module. In the discriminative module (e.g., Eq. 14), the full spatio-temporal representation $f_{e,i} \in \mathbb{R}^{T \times D}$ is retained and processed along the temporal axis to preserve fine-grained temporal dynamics.

Discriminative DNN Module. To prevent the classifier from overfitting to static scene appearance, we perturb the scene embedding $f_{s,j}$ using random multiplicative noise sampled per sample before injecting it into the fusion blocks. Given the original scene representation $f_{s,j}$, we draw a random scaling factor $\alpha \sim \mathcal{U}(0,1)$, and broadcast it across the feature dimension. The perturbed scene embedding is then computed as:

$$\tilde{f}_{s,j} = \alpha f_{s,j}, \tag{13}$$

which effectively reduces the dominance of raw scene appearance and encourages the model to consider both the scene context and event dynamics during subsequent fusion. Meanwhile, before event-scene fusion, each event embedding $f_{e,i}$ is processed using M layers of temporal self-attention. These pre-fusion layers capture intra-event temporal dynamics and produce a temporally contextualized representation of its spatio-temporal cube. Formally,

$$f_{e,i} = \text{Self-Attn}_{\text{temporal}}(f_{e,i}), \tag{14}$$

where attention is applied along the temporal axis T, allowing the model to capture motion-consistent temporal dependencies for the i_{th} foreground event. For each of the N Fusion Blocks l, we apply a separate temporal self-attention operation inside the block to the current event representation, yielding the updated event features $f_{e,i}^{att}$ immediately before event-scene fusion. Unlike the M pre-fusion layers, this block-level operation repeatedly sharpens event features at each fusion stage, helping preserve event-specific cues and preventing scene information from overwhelming the event representation. We then adopt an AdaLN-Zero–style modulation [29], which takes the scene embedding as a conditional input to achieve hierarchical and adaptive fusion between event and scene features. Given the perturbed scene embedding $\tilde{f}_{s,j}$, a single MLP outputs six modulation parameters for the attention and FFN branches:

$$\gamma_{\text{pre}}^{\text{att}}, \beta_{\text{pre}}^{\text{att}}, \gamma_{\text{post}}^{\text{att}}, \gamma_{\text{pre}}^{\text{ffn}}, \beta_{\text{pre}}^{\text{ffn}}, \gamma_{\text{post}}^{\text{ffn}} = \text{MLP}(\tilde{f}_{s,j}), \tag{15}$$

The attention branch is computed as:

$$f'_{e,i} = f_{e,i} + \gamma_{\text{post}}^{\text{att}} \odot \text{Attn}\left(\gamma_{\text{pre}}^{\text{att}} \odot \text{LN}(f_{e,i}) + \beta_{\text{pre}}^{\text{att}}\right), \tag{16}$$

and the FFN branch further yields:

$$f''_{e,i} = f'_{e,i} + \gamma_{\text{post}}^{\text{ffn}} \odot \text{FFN}\left(\gamma_{\text{pre}}^{\text{ffn}} \odot \text{LN}(f'_{e,i}) + \beta_{\text{pre}}^{\text{ffn}}\right). \tag{17}$$

In all expressions, Attn denotes the self-attention operator, FFN refers to the feed-forward network, and LN indicates the normalization layer. The symbol $\odot$ denotes element-wise multiplication with broadcasting along the temporal dimension.

To balance event and scene information, each block is equipped with a learnable fusion weight w_l, initialized jointly over the temporal dimension T and feature dimension D to ensure sample awareness and maintain diversity:

$$h_{e,i} = w_l f_{e,i}^{att} + (1 - w_l) f_{e,i}''. \tag{18}$$

The fused features are added residually to the event representation:

$$f_{e,i} = f_{e,i} + h_{e,i}. \tag{19}$$

For final classification, we follow the formulation in Eq. 15, 16, and 17 and design an AdaLN-Zero–style scene-conditioned classification head that further injects scene information into $f_{e,i}$. Given the perturbed scene embedding $\tilde{f}_{s,j}$, an MLP generates the modulation parameters:

$$\gamma^{\text{cls}}, \beta^{\text{cls}} = \text{MLP}(\tilde{f}_{s,j}), \tag{20}$$

where $\gamma^{\text{cls}}, \beta^{\text{cls}} \in \mathbb{R}^D$ are broadcast along the temporal dimension. The classification logits are obtained by modulating the normalized event embedding:

$$z = \text{Linear}\left((1 + \gamma^{\text{cls}}) \odot \text{LN}(f_{e,i}) + \beta^{\text{cls}}\right), \tag{21}$$

here $z \in \mathbb{R}^K$ denotes the predicted logits over the K scene categories. The discriminative loss $\mathcal{L}_{\text{classify}}$ is computed following the cross-entropy formulation in Eq. 3-4.

Contrastive DNN Module. Scene labels alone only guide the model to learn (event, scene) $\rightarrow$ label relations. To explicitly model the event & scene semantic associations, we introduce a dedicated contrastive learning module.

Within each training batch, foreground event features and scene belonging to the same scene label form positive pairs, while those belonging to different scene labels form negative pairs. To enrich the pool of contrastive samples, we maintain a *memory bank* that stores previously observed event and scene embeddings together with their labels. At each iteration, embeddings from the current batch are concatenated with stored entries in the memory bank, and contrastive learning is performed between the current batch features and this augmented set, following the formulations in Eq. 9, 10, and 11. After computing the loss, the memory bank is updated in a FIFO manner. Let $\mathcal{M}_n$ denote the memory bank at iteration n, and $\mathcal{B}_n$ the current batch of embeddings. The queue update is:

$$\mathcal{M}_{n+1} = \text{Dequeue}\big(\text{Enqueue}(\mathcal{M}_n, \mathcal{B}_n), C\big), \tag{22}$$

here C is the fixed maximum capacity of the memory bank. This queue-based mechanism ensures that the memory bank is continuously refreshed with recent samples.

In this way, by combining the symmetric contrastive objectives in Eq. 9-10 with a large and dynamically refreshed memory bank, both scenes and events are exposed to significantly richer positive and negative examples. This enables the model to learn more discriminative representations, thereby improving semantic alignment and enhancing overall robustness in scene-dependent anomaly detection during training.

3.4 Training Objective

The overall learning objective integrates both the discriminative loss and the event-scene contrastive loss. The discriminative loss in Eq. 4 optimizes the model to correctly assign each event to its ground-truth scene label. Complementarily, the contrastive loss in Eq. 11 enforces semantic alignment between events and their corresponding scenes while increasing separation from mismatched scenes. To jointly optimize these two complementary objectives, we formulate the total training loss as a weighted combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{classify}} + \lambda \mathcal{L}_{\text{contrast}}. \tag{23}$$

The balancing coefficient λ controls the relative contribution of the contrastive component.

3.5 Anomaly Inference

During inference, two complementary anomaly scores are computed for each detected event: the discriminative classification score S_{classify} in Eq. 5 and the contrastive similarity score S_{contrast} in Eq. 12. These scores are combined through a scalar weighted fusion:

$$S_{\text{fused}}^{(i)} = (1 - w) \cdot S_{\text{classify}}^{(i)} + w \cdot S_{\text{contrast}}^{(i)}, \tag{24}$$

where w represents the scalar weight applied to the contrastive score. This simple linear combination allows the two complementary anomaly cues to be balanced in a straightforward and interpretable manner. A sensitivity analysis of w is provided in the supplementary material. For a single frame, the final anomaly score is the maximum fused event-level score over all detected events:

$$S_{\text{frame}} = \max_i S_{\text{fused}}^{(i)}. \tag{25}$$

Afterward, a median-filter–based sliding window is applied to smooth the resulting score sequence.

4 Experiments and Analysis

4.1 Experimental Settings

Datasets. Experiments are primarily conducted on two SDVAD-specific datasets: NWPU Campus [5] and ShanghaiTech-sd [5], denoted as "NWPU" and "ST-sd" respectively. To further verify the generalizability of our method beyond scene-dependent settings, we also evaluate on ShanghaiTech [22], a widely used multi-scene but scene-independent VAD dataset containing 13 scenes where anomaly definitions are consistent across all scenes.

Implementation Details. As introduced in Sec. 3.3, we employ a frozen DINOv2 encoder to extract event and scene embeddings, with a feature dimension of $D = 768$. The temporal span for constructing each event's spatio-temporal cube is set to $T = 9$. For discriminative DNN module, the number of pre-fusion self-attention layers is $M = 4$, and the number of Fusion Blocks is $N = 4$. As for contrastive DNN module, the memory bank has a capacity of $C = 8{,}192$, storing up to C event embeddings, C scene embeddings, and their associated scene labels. The weight of the contrastive loss is set to $\lambda = 0.5$. All models are trained using the SGD [33] optimizer with a batch size of 512 and a learning rate of 0.01. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU using the PyTorch [28] framework. More implementation details are provided in supplementary material.

Evaluation Metric. Following the widely adopted evaluation protocol in VAD, we use the Area Under the Curve (AUC) of the Receiver Operating Characteristic as the evaluation metric. All test frames are concatenated to compute an overall frame-level AUC.

Table 1: Comparison on scene-dependent VAD datasets.

Method	Venue	NWPU	ST-sd
Generative Learning-Based			
OG-Net [49]	CVPR 20	62.5	69.6
MemAE [12]	ICCV 19	61.9	67.4
MNAD [27]	CVPR 20	62.5	68.2
HF^2-VAD [19]	ICCV 21	63.7	70.8
MPN [24]	CVPR 21	64.4	76.9
AMMC-Net [4]	AAAI 21	64.5	64.9
FBAE [5]	CVPR 23	68.2	82.7
SSAE [7]	TPAMI 24	75.6	83.6
MS-VAD [51]	CVPR 24	70.1	-
SMDR [14]	MM 25	73.8	-
ADSM [50]	ICCV 25	76.9	-
Other Existing Methods			
LLSH [21]	TCSVT 22	62.2	-
CRCL [18]	TIP 25	73.8	-
Discriminative-Contrastive Learning-Based			
Ours	MM 26	**77.3**	**92.1**

Table 2: Comparison on ShanghaiTech dataset.

Method	Venue	ShanghaiTech
FFP [17]	CVPR 18	72.8
MemAE [12]	ICCV 19	72.8
CAE [15]	CVPR 19	78.7
MNAD [27]	CVPR 20	70.5
MTP [34]	WACV 20	76.0
AMMC-Net [4]	AAAI 21	73.7
HF^2-VAD [19]	ICCV 21	76.2
VABD [16]	TIP 21	78.2
MPN [24]	CVPR 21	73.8
SSMTL [11]	CVPR 21	82.4
HSNBM [3]	MM 22	76.5
LLSH [21]	TCSVT 22	77.6
MoCoDAD [10]	ICCV 23	77.9
FBAE [5]	CVPR 23	79.2
FPDM [45]	ICCV 23	78.6
Two-stream [6]	TIP 24	83.7
SDMAE [32]	CVPR 24	79.1
SSAE [7]	TPAMI 24	80.5
ADSM [50]	ICCV 25	**84.5**
Ours	MM 26	81.3

4.2 Main Results

Tab. 1 reports the AUC comparison on the scene-dependent NWPU and ShanghaiTech-sd datasets, while Tab. 2 additionally reports the comparison on the scene-independent ShanghaiTech dataset.

Results on Scene-Dependent Datasets. As shown in Tab. 1, our method achieves an AUC of 92.1% on ShanghaiTech-sd, significantly outperforming all existing methods including SSAE [7] (83.6%) and FBAE [5] (82.7%) by a large margin of +8.5% and +9.4%, respectively. On the more challenging NWPU Campus dataset with 43 diverse scenes, our method achieves 77.3% AUC, surpassing the previous best competing methods ADSM [50] (76.9%) and SSAE [7] (75.6%). These results demonstrate that explicitly exploiting scene labels and modeling event-scene correlations through discriminative-contrastive learning is highly effective for SDVAD.

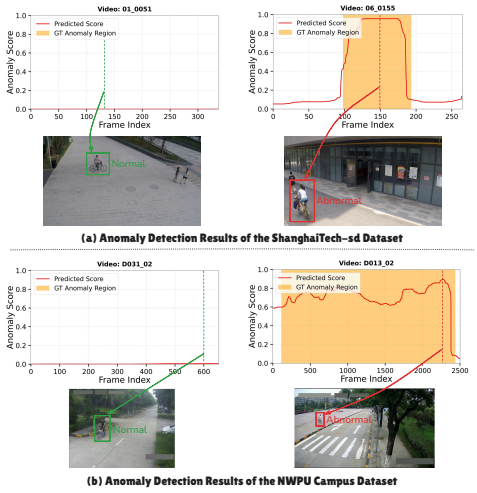


Figure 3: Qualitative visualization of scene-dependent anomaly detection results.

Table 3: Ablation results of the discriminative and contrastive learning components.

Learning Components		ST-sd
Discriminative Module	Contrastive Module	
✓		88.9
	✓	89.0
✓	✓	**92.1**

Visualization on Scene-Dependent Datasets. Fig. 3 shows qualitative results on two scene-dependent datasets. In (a) ShanghaiTech-sd, "cycling" is manually defined as a normal event in Scene 1 but abnormal in other scenes. In (b) NWPU Campus, cycling on roadway is considered normal, while cycling on square is abnormal. In both datasets, our model accurately distinguishes these scene-dependent normal and abnormal cases, confirming its ability to model event-scene consistency and perform context-aware anomaly detection.

Generalizability on Scene-Independent Datasets. To further evaluate the generalizability of our framework, we report results on the scene-independent ShanghaiTech dataset in Tab. 2. Although our method is primarily designed for scene-dependent anomaly detection, it still achieves a competitive AUC of 81.3% on ShanghaiTech, outperforming most generative methods such as FBAE [5] (79.2%), SSAE [7] (80.5%), and SDMAE [32] (79.1%). This demonstrates that our discriminative-contrastive framework possesses strong generalizability: even on multi-scene datasets where anomalies are not scene-dependent, the learned event-scene representations still provide effective cues for anomaly detection. The scene-conditioned classification and contrastive alignment mechanisms do not overfit to scene-dependent patterns, but rather learn broadly useful event-scene semantic correspondences that transfer well across different evaluation settings.

4.3 Discussion

To validate the effectiveness of our proposed framework, we conduct extensive experiments on the ShanghaiTech-sd dataset, analyzing the impact of individual learning modules, generalizability

Table 4: Generalization to unseen scene labels.

Train Scenes	Test Scenes	AUC (%)
{1, 10}	{6, 12}	81.7
{6, 12}	{6, 12}	**84.3**

Table 5: Ablation on key architectural design choices.

Random Noise	Fusion Methods		ST-sd
	Scalar Fusion	Sample-Aware Fusion	
		✓	90.9
✓	✓		90.8
✓		✓	**92.1**

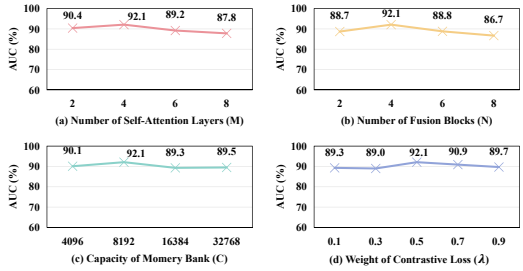


Figure 4: Hyperparameter sensitivity analysis.

to unseen scene labels, critical architectural design choices, and key hyperparameters. More results, including complementary analyses of baselines, generalization, and model behavior, as well as computational cost and comparisons with multimodal large language models (MLLMs), are reported in the supplementary material.

Effect of Discriminative & Contrastive Learning. Tab. 3 presents the ablation results of the two key learning components. When only the discriminative module is used, the model achieves an AUC of 88.9%, indicating that explicit scene-label supervision effectively guides the model to distinguish scene-dependent anomalies. When trained solely with the contrastive module, module attains a comparable AUC of 89.0%, indicating that contrastive learning alone provides useful relational cues but limited semantic guidance. Notably, each single component already surpasses all baselines in Tab. 1, further demonstrating the effectiveness of our learning design. By jointly optimizing both components, the proposed framework reaches 92.1%, demonstrating a clear +3.1% improvement over the contrastive-only variant. This result verifies that discriminative learning and contrastive learning are mutually complementary.The former provides explicit semantic matching between events and scenes, while the latter reinforces event-scene alignment and strengthens the model's ability to capture scene-dependent correlations.

Generalization to Unseen Scene Labels. A natural concern is whether the discriminative classifier, trained on a fixed set of scene labels, can handle a new or previously unseen camera without retraining. To investigate this, we evaluate on ShanghaiTech-sd by training on scenes {1, 10} and performing inference on the entirely unseen scenes {6, 12}, comparing against the in-domain baseline that trains and tests on {6, 12}. As shown in Tab. 4, the cross-scene setting achieves 81.7% AUC, only 2.6% below the in-domain result of 84.3%. This gap is modest and demonstrates that the model retains meaningful discriminative capability even on scene labels it has never encountered during training. The underlying reason is twofold. First, although the scene class labels differ, the *visual feature distributions* of normal events (e.g., walking, standing) are largely shared across scenes, and our model learns to exploit these cross-scene commonalities during training. Second, the contrastive module learns generic event-scene alignment rather than label-specific boundaries, so the similarity-based anomaly score remains informative even when the exact scene identity is unknown. Together, these properties allow our framework to generalize beyond its training scene set, addressing the unseen-scene challenge without retraining.

Effect of Scene Perturbation and Fusion Strategy. As discussed in Eq. 13 and Eq. 18, our framework introduces random perturbation to scene embeddings and a sample-aware fusion mechanism to balance event and scene information. Tab. 5 evaluates their respective contributions by removing the random perturbation and replacing the sample-aware fusion with a simple scalar fusion. Without applying random noise, the model tends to overfit to static scene appearance, achieving an AUC of 90.9%. When random noise is preserved but the fusion module is replaced with a simple learnable scalar weight, performance remains limited (90.8%). In contrast, combining scene perturbation with our proposed sample-aware fusion yields the highest AUC of 92.1%, demonstrating that both stochastic scene perturbation and the sample-aware fusion mechanism are essential for effectively balancing event-scene information in context-aware anomaly detection.

Hyperparameter Sensitivity Analysis. Fig. 4 visualizes how the key hyperparameters affect model performance. For self-attention layers (M) and fusion blocks (N) in (a) and (b), we vary one while fixing the other at 4. Both curves peak at 4 layers, suggesting an optimal balance between event-level feature learning and event-scene fusion. Too few layers underfit event representations, whereas too many may cause oversmoothing or scene dominance. In (c), a moderate memory bank capacity (C) of 8192 yields the best performance. Smaller capacities limit the diversity of positive and negative pairs, while excessively large banks introduce outdated embeddings, leading to noisy contrastive gradients. In (d), the optimal contrastive loss weight is $\lambda = 0.5$. Smaller weights reduce the contribution of contrastive supervision, whereas larger ones overemphasize feature alignment and weaken discriminative classification learning. Despite these variations, our model consistently outperforms all competing methods in Tab. 1, demonstrating strong robustness to hyperparameter choices.

5 Conclusion

In this paper, we propose a discriminative-contrastive joint learning framework for SDVAD. Unlike the classic scene-conditioned generative learning paradigm, our framework explicitly leverages natural scene labels and event-scene relationships through discriminative and contrastive learning. Building upon these two complementary objectives, we design a scene-informed DNN that achieves effective scene-dependent classification and event-scene contrastive alignment via proper scene information integration. Extensive experiments on SDVAD datasets demonstrate state-of-the-art performance, while results on the scene-independent ShanghaiTech dataset further confirm that the learned event-scene representations generalize well beyond scene-dependent settings.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62506374, 62472434, and 62402171), the Key Program of NSFC Hunan (Grant No. 2026JJ30028), and the Postgraduate Research and Innovation Project of Hunan Province (Grant No. XJJC2025012).

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20143–20153.

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence* 30, 3 (2008), 555–560.

[3] Qianyue Bao, Fang Liu, Yang Liu, Licheng Jiao, Xu Liu, and Lingling Li. 2022. Hierarchical scene normality-binding modeling for anomaly detection in surveillance videos. In *Proceedings of the 30th ACM international conference on multimedia*. 6103–6112.

[4] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. 2021. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 938–946.

[5] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. 2023. 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*. 20392–20401.

[6] Congqi Cao, Yue Lu, and Yanning Zhang. 2024. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Transactions on Image Processing* 33 (2024), 1810–1825.

[7] Congqi Cao, Hanwen Zhang, Yue Lu, Peng Wang, and Yanning Zhang. 2024. Scene-dependent prediction in latent space for video anomaly detection and anticipation. *IEEE transactions on pattern analysis and machine intelligence* (2024).

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–58.

[9] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and Feng Yang. 2020. Anomaly detection with bidirectional consistency in videos. *IEEE transactions on neural networks and learning systems* 33, 3 (2020), 1079–1092.

[10] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. 2023. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10318–10329.

[11] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12742–12752.

[12] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1705–1714.

[13] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742.

[14] Han Hu, Wenli Du, and Bing Wang. 2025. Efficient Video Anomaly Detection via Scene-Dependent Memory Assisted Inter-Frame RGB Difference Reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 43–51.

[15] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. 2019. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7842–7851.

[16] Jing Li, Qingwang Huang, Yingjun Du, Xiantong Zhen, Shengyong Chen, and Ling Shao. 2021. Variational abnormal behavior detection with motion consistency. *IEEE Transactions on Image Processing* 31 (2021), 275–286.

[17] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6536–6545.

[18] Yang Liu, Hongjin Wang, Zepu Wang, Xiaoguang Zhu, Jing Liu, Peng Sun, Rui Tang, Jianwei Du, Victor CM Leung, and Liang Song. 2025. Crcl: Causal representation consistency learning for anomaly detection in surveillance videos. *IEEE Transactions on Image Processing* (2025).

[19] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. 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*. 13588–13597.

[20] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*. 2720–2727.

[21] Yue Lu, Congqi Cao, Yifan Zhang, and Yanning Zhang. 2022. Learnable locality-sensitive hashing for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 2 (2022), 963–976.

[22] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*. 341–349.

[23] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. 2021. Future frame prediction network for video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence* 44, 11 (2021), 7505–7520.

[24] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. 2021. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15425–15434.

[25] Trong-Nguyen Nguyen and Jean Meunier. 2019. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1273–1283.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[27] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14372–14381.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[29] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

[30] Bharathkumar Ramachandra and Michael Jones. 2020. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2569–2578.

[31] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. 2020. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence* 44, 5 (2020), 2293–2312.

[32] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. 2024. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15984–15995.

[33] Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics* (1951), 400–407.

[34] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. 2020. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2626–2634.

[35] Hao Song, Che Sun, Xinxiao Wu, Mei Chen, and Yunde Jia. 2019. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. *IEEE Transactions on Multimedia* 22, 8 (2019), 2138–2148.

[36] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.

[37] Shengyang Sun and Xiaojin Gong. 2023. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22846–22856.

[38] Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. 2025. ALLVB: All-in-One Long Video Understanding Benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7211–7219.

[39] Xichen Tan, Yuanjing Luo, Yunfan Ye, Guang Yu, Chengyu Wang, Fang Liu, and Zhiping Cai. 2026. Video Plot Segmentation. *Expert Systems with Applications* (2026), 132422.

[40] University of Minnesota. 2006. Unusual Crowd Activity Dataset. http://mha.cs.umn.edu/proj_events.shtml#crowd.

[41] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. 2022. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*. Springer, 494–511.

[42] Shu Wang and Zhenjiang Miao. 2010. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*. IEEE, 1220–1223.

[43] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*. Springer, 322–339.

[44] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 2024. Vadclip: Adapting vision-language models for

weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 6074–6082.

[45] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. 2023. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5527–5537.

[46] Guoqing Yang, Zhiming Luo, Jianzhe Gao, Yingxin Lai, Kun Yang, Yifan He, and Shaozi Li. 2024. A multilevel guidance-exploration network and behavior-scene matching method for human behavior anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5865–5873.

[47] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. 2020. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM international conference on multimedia*. 583–591.

[48] Tongtong Yuan, Xuange Zhang, Bo Liu, Kun Liu, Jian Jin, and Zhenzhen Jiao. 2024. Surveillance video-and-language understanding: from small to large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 1 (2024), 300–314.

[49] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. 2020. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14183–14193.

[50] Hanwen Zhang, Congqi Cao, Qinyi Lv, Lingtong Min, and Yanning Zhang. 2025. Autoregressive Denoising Score Matching is a Good Video Anomaly Detector. *arXiv preprint arXiv:2506.23282* (2025).

[51] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. 2024. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17385–17394.