2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Video Anomaly Detection: A Systematic Taxonomy and Analysis of Deep Models

Yitong Yuan[1] and Jingxin Cao[2]

[1] Tianjin Normal University, Tianjin 300387, China
16622750626@163.com

Abstract. Video anomaly detection (VAD), a foundational pillar of modern computer vision, has garnered significant attention due to its wide-ranging real-world applications. Despite considerable progress, VAD persistently grapples with formidable challenges, particularly the scarcity of anomalies in real-world datasets and the complexities of accurate anomaly annotation. This survey investigates state-of-the-art VAD methodologies, synthesizing their core challenges and elucidating tailored solutions. Addressing the shortcomings of prior reviews, we propose a unified taxonomy that classifies methods according to input modalities: raw video, mid-level visual features, and high-level visual-semantic representations, providing a perspicuous framework to discern their unique attributes. To enrich comprehension, we present rigorous comparisons and analyses across established benchmark datasets. Anticipating future developments, we delineate promising research trajectories, such as semantic context learning enabled by contrastive language-image pre-training (CLIP) and multi-modal large language models (MLLMs), to drive transformative advancements in the field. Furthermore, we meticulously examine the practical impediments existing approaches encounter in deployment in real-world environments.

Keywords: Video Anomaly Detection, Anomaly Detection, Computer Vision.

1 Introduction

Previous endeavors in manual monitoring have predominantly relied on human observation, necessitating the continuous examination of multiple surveillance feeds. This approach is inherently susceptible to operator fatigue and declining attentiveness, thereby incrementally increasing the probability of overlooking atypical incidents. VAD, a foundational component of contemporary computer vision research, addresses this limitation by facilitating the autonomous identification and accurate spatio-temporal localization of anomalies within video sequences. In this context, anomalies are characterized as events that significantly deviate from established behavioral norms, thus offering a transformative resolution to the constraints inherent in conventional monitoring methodologies.

In recent years, the extensive rollout of surveillance technologies has markedly increased the importance of VAD in practical settings. Its utility includes intelligent surveillance to enhance public safety, real-time traffic oversight, detection of violent human behavior, and interpretation of medical imaging. Despite significant progress in recent years, current VAD methods still confront two formidable challenges: (1) the inherent scarcity of anomalous events in real-world settings, which introduces pronounced data imbalance and complicates model training, and (2) the substantial difficulty in annotating anomalous segments, driven by the labor-intensive nature of human annotation, the diversity of abnormal behaviors, and their contextual reliance on specific scenes. To address these issues, state-of-the-art approaches predominantly adopt semi-supervised VAD (SS-VAD) and weakly supervised VAD (WS-VAD) frameworks, which offer promising avenues for balancing efficacy and annotation efficiency in practical deployments. To address the problem of data imbalance, where collecting adequate anomalous training samples is impractical, semi-supervised SS-VAD, such as ConvAE [1], employs only normal samples for training. This method learns to distinguish normal from abnormal patterns, flagging deviations from normality as anomalies. To address the difficulty of obtaining fine-grained labels, WS-VAD requires only coarse video-level annotations following the extraction of pre-trained features, achieving strong performance in complex and crowded scenes.

Beyond the core challenges, numerous studies have identified additional critical factors impacting VAD performance. For instance, while convolutional neural networks (CNNs) are highly effective at extracting spatial features, their integration with long short-term memory (LSTM) [2] bolsters temporal modeling by adeptly capturing sequential dependencies, resulting in enhanced accuracy. Moreover, using middle-level features like skeleton trajectory [3] for training provides a robust approach to mitigate noise stemming from dynamic background variations. Additionally, current innovative multi-modal frameworks like CLIP [4], [5] can learn from high-level semantic signals, which enable generalization to unseen datasets via natural language prompting, surpassing traditional pre-trained models dependent on fixed class labels and exhibiting greater adaptability across diverse real-world contexts.

The discussion above clearly demonstrates that multifaceted challenges in VAD require resolution. To ensure robust and generalizable conclusions, it is imperative to conduct a comprehensive analysis and systematic comparison of different types of existing methods across diverse datasets, enabling a thorough understanding of their characteristics. Therefore, there is an urgent necessity for a systematic overview of existing works to serve as a guide for newcomers and provide valuable references for established researchers.

In this survey, we focus on the progress of deep learning-based VAD, offering a systematic and rigorous analysis to illuminate the multifaceted challenges faced by VAD across diverse settings, alongside their respective solutions. Our objectives extend to exposing the limitations of current methodologies and charting potential avenues for future research. We commence by curating a collection of seminal VAD reviews from recent years, distilling their insights in Table 1. For instance, Ramachandra et al. [6] confined their scope to single-scene VAD, omitting a broader taxonomic synthesis; Ren et al. [7] examined VAD challenges and opportunities from a systems perspective, yet fell short of detailed methodological comparisons; Ericsson et al. [8] honed in on representation learning for self-supervised VAD, without probing its broader methodological ramifications; Berroukham et al. [9] surveyed deep learning approaches within existing frameworks, but overlooked the latest advancements; Cao et al. [10] targeted static image anomaly detection, leaving video-specific contexts underexplored; and Wu et al. [11] tracked emerging trends like large-scale pre-trained models and explainability, yet offered limited depth on multi-modal fusion strategies. In contrast, this survey not only synthesizes these foundational contributions and incorporates cutting-edge developments, but also proposes a novel classification framework. By integrating multi-modal approaches with current research frontiers, this framework addresses critical gaps in the VAD landscape, providing a forward-looking guidance for future investigations.

Table 1. A summary of recent reviews.

Paper	Year	Main Contribution	Focus
Ramachandra et al. [6]	2020	A profound analysis of single-scene VAD methodology.	Single-Scene VAD
Ren et al. [7]	2021	Summarized opportunities and challenges in various applications of VAD.	Intelligent VAD Systems
Ericsson et al. [8]	2022	Classified and explained self-supervised representation learning, discussing its applications in various domains.	Self-Supervised Representation Learning
Berroukham et al. [9]	2023	A classification and summary of deep learning-based methods of VAD.	Deep Learning-based Methods of VAD
Cao et al. [10]	2024	Emphasized the latest advancements in visual anomaly detection.	Visual Anomaly Detection
Wu et al. [11]	2024	A review of deep learning-based methods for VAD within the framework of existing classification schemes.	Deep Learning for VAD

Overall, this survey makes the following key contributions:

(1) We provided a novel taxonomy that systematically classified the existing approaches into three categories based on their model inputs: raw-video input, middle-level visual input and high-level visual-semantic input. Our classification framework is illustrated in Figure 1, using the following abbreviations: OCC: One-Class Classifier, SS-VAD: Semi-Supervised, SF-VAD: Self-Supervised, FS-VAD: Fully-Supervised, US-VAD: Unsupervised, WS-VAD: Weakly-Supervised. We hope to resolve the limitations of existing categorization.

(2) We proposed a detailed overview of representative methods within each category and compared their performance across various benchmark datasets. In addition, we discussed the pros and cons of these approaches, allowing readers to intuitively grasp their applicability to various real-world scenarios.

(3) We overview the latest VAD research trends, highlighting the potential of CLIP and MLLMs in the VAD domain, providing future research directions.

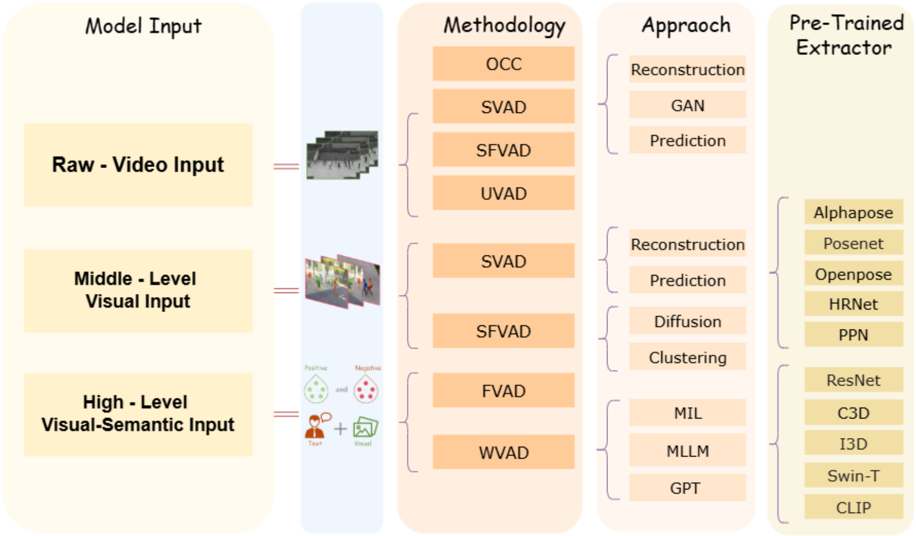

Fig. 1. The basic framework of our taxonomy.

2 A Taxonomy Of VAD Approaches

2.1 Related Benchmark Datasets

Benchmark datasets are essential for effectively evaluating VAD approaches. However, due to the rarity of anomalous events, collecting comprehensive datasets is inherently challenging. This difficulty makes publicly available benchmark datasets particularly valuable for their standardized evaluation frameworks and fair comparisons between diverse algorithms. Despite of great challenges, many prominent datasets are continuously updated and expanded through tireless efforts, which offer profound resources in diverse abnormal detection in different scenarios. We have illustrated some widely used datasets in Table 2.

UCSD Ped1 & Ped2 [12]: Designed for the purpose of VAD in crowded pedestrian scenes, the datasets use a single stationary camera to capture naturally occurring events like non-pedestrian entities (e.g., bicycles, skateboards) and unusual motions (e.g., walking on grass), with frame-level annotations and a subset offering pixel-level binary masks.

CUHK Avenue [13]: Built for abnormal event detection in surveillance, CUHK Avenue uses a single fixed camera to capture 30,652 frames across 15 sequences, including anomalies like running, throwing objects, and loitering with pixel-level annotations.

ShanghaiTech [16]. The large dataset is created for the purpose of complex real-world VAD, employing multiple surveillance cameras across a university campus to record dynamic anomalies (e.g., chasing and brawling) with frame-level annotations.

UCF-Crime [15]. Focused on detecting realistic illegal anomalies, UCF-Crime gathers 128 hours of CCTV footage from YouTube and LiveLeak, featuring 13 anomalies (e.g., fighting, burglary, robbery) with weak labels, checked via strict criteria to ensure authenticity.

XD Violence [14]. Designed for violence detection, this dataset collects untrimmed videos from movies and YouTube, featuring violence-related anomalies with weak labels averaged from multiple annotators. Additionally, it emphasis on background consistency to prevent discrimination.

IITB-Corridor [18]. Designed at IIT Bombay for abnormal human activity detection, this dataset uses a single camera in a realistic corridor setting to capture anomalies like loitering, running, fighting, and chasing, including both individual and group behaviors with frame-level annotations, ensuring a profound resource for analyzing multiple types of abnormal behaviors.

UBnormal [17]. Created as a benchmark for supervised open-set VAD, UBnormal uses Cinema4D to generate virtual scenes with 22 anomaly types (e.g., running, fighting, car crashes) across 29 real-world backgrounds, offering pixel-level annotations for training, which has better generalization to various abnormal activities.

Table 2. Characteristics of VAD datasets.

Dataset	Total Frames	Training Frames	Testing Frames	Abnormal Events
UCSD Ped1 [12]	14000	6800	7200	40
UCSD Ped2 [12]	4560	2550	2010	12
CUHK Avenue [13]	30652	15328	15324	47
XD-Violence [14]	4754 (videos)	3954 (videos)	800 (videos)	6
UCF-Crime [15]	1900 (videos)	1610 (videos)	290 (videos)	950 (videos)
ShanghaiTech [16]	317398	274575	42883	130
Ubnormal [17]	236902	116087	92640	660
IITB-Corridor [18]	482566	301999	181567	108278

2.2 Raw-Video Input

Raw-video inputs, such as RGB images and optical flow, are used directly for model training. The approach preserves full data integrity and primarily supports an end-to-end learning framework based on fully CNNs avoiding spatial information loss from fully connected layers. OCC approaches usually learn a hypersphere to enclose normal data in feature space and labeling the points outside it as anomalies to achieve classification within a unified framework. SS-VAD approaches leverage CNNs to learn regular patterns and flag the anomalous frames that deviate from these patterns. Therefore, this end-to-end learning framework is suitable for both of them in detecting mechanisms. We have summarized representative approaches in this category as follows and compared their performance in Table 3.

Table 3. The performances of raw-video input-based approaches on public datasets.

Approach	**Publication**	**Learning Type**	**Methodology**	**UCSD Ped1**	**UCSD Ped2**	**CUHK Avenue**	**ShanghaiTech**	**Subway Exit**
				Frame-Level AUC (%)				
AMDN [19]	BMVC 2015	Semi-Supervised	OCC	92.1	90.8	-	-	-
ConvAE [1]	CVPR 2016	Semi-Supervised	Reconstruction	81.8	82.9	78.3	-	80.7
ConvLSTM [2]	ICME 2017	Semi-Supervised	Reconstruction	75.5	88.1	77.0	-	87.7
GANs [20]	CVPR 2017	Semi-Supervised	Reconstruction	97.4	93.5	-	-	-
FuturePred [21]	CVPR 2018	Semi-Supervised	Prediction	83.1	95.4	85.1	72.8	-
MNAD [22]	CVPR 2020	Semi-Supervised	Reconstruction	97.0	88.5	70.5	-	-
SSMTL [23]	CVPR 2021	Self-Supervised	Multiple Tasks	99.8	92.8	90.2	-	-
GCL [24]	CVPR 2022	Unsupervised	Reconstruction	-	94.1	78.9	71.2	-
AED-MAE [25]	CVPR 2024	Semi-Supervised	Reconstruction	-	95.4	91.3	79.1	-

One-Class Classification (OCC)-based methods. To overcome the challenge of learning effective representations under limited labeled data conditions, Lu et al. [19] proposed the appearance and motion deepnet (AMDN), which integrates stacked denoising autoencoders (SDAEs) to extract compact and robust features from both appearance and motion information. By training multiple one-class SVMs (OC-SVMs) on the learned representations, they constructed a unified VAD model capable of classifying normal patterns. This approach enhances the ability to jointly learn features and classifiers in an end-to-end fashion, reducing the reliance on hand- crafted features. However, as dimensionality increases, the computational cost of the kernel matrix escalates, resulting in overfitting and limiting the performances of OCC methods in high-dimensional scenarios.

Reconstruction-based methods. To address the challenge of modeling sparse and irregular events through supervised learning, Hasan et al. [1] introduced an SS-VAD approach utilizing convolutional autoencoders (ConvAE) to directly learn temporal regularities in video sequences and flag anomalies based on reconstruction errors. However, due to its poor ability to capture motion dynamics, ConvAE often misclassifies irregular motion, leading to a high false positive rate.

To address this problem, Luo et al. [2] proposed the ConvLSTM framework, integrating CNNs with LSTM to better encode the change of appearance and motion for normal events, respectively, to reduce the possibility of false positives. However,it remains limited by the assumption that anomalies will always produce higher reconstruction errors, which may not hold if the model fails to generalize to unseen abnormal data.

To address the limitation of traditional autoencoder-based models, i.e., the tendency to produce blurry reconstructions due to reliance on pixel-level reconstruction losses, which compromises VAD accuracy in high-dimensional settings, Ravanbakhsh et al. [20] proposed a generative adversarial nets (GANs) approach. By training a generator-discriminator framework to learn the distribution of normal frames, their method enhances sensitivity to anomalies, which enhances the ability to identify anomalies in high-dimensional situations.

To improve feature discrimination in SS-VAD, Park et al. [22] introduced a memory module that employs feature compactness loss and feature separateness loss to ensure that normal data representations remain tightly clustered, while maintaining clear separation between different memory units. Therefore, this method achieves a larger reconstruction error gap between normal and abnormal frames, which improves the discrimination capability.

In an effort to operate without labeled video data to ensure high generalization and scalability, Zaheer et al. [24] leveraged the low frequency of anomaly occurrences to introduce a US-VAD framework called generative cooperative learning (GCL), where the generator and discriminator engage in cross-supervised training by iteratively exchanging pseudo-labels. Additionally, the generator employs negative learning (NL) to prevent the reconstruction of anomalous data, enhancing robustness of VAD in complex surveillance scenarios without label annotation.

In recent studies, patch-based reconstruction strategies, which operate by inferring missing data from surrounding regions to enhance reconstruction quality, have garnered significant attention. To address the challenges of detecting rare and context-dependent abnormal events in video surveillance, Ristea et al. [25] proposed a lightweight self-distilled masked autoencoder model. By leveraging motion gradient-weighted tokens, the model shifts focus from static backgrounds to dynamic foreground objects. Additionally, the integration of a teacher-student decoder framework and synthetic anomaly augmentation during training enables the model to limit the model's ability to reconstruct anomalies, thereby enhancing the balance between efficiency and accuracy in VAD. This approach tackles the issues of insufficient abnormal event data and excessive generalization of traditional models to anomalies, achieving a trade-off between speed and detection performance.

Prediction-based methods. A major limitation of reconstruction-based models lies in their tendency to overfit to normal patterns, which makes it challenging to ensure consistently higher reconstruction errors for anomalous inputs. To solve the problem, FuturePred [21] is the first work making use of U-Net architectures with GAN-based video frame prediction to forecast future. By leveraging temporal consistency, predictive models offer superior robustness against normal variations while effectively capturing deviations caused by anomalous events.

Multi-task learning methods. To address the detecting challenge of limited supervision due to the scarcity of labeled anomalous data, Georgescu et al. [23] proposed a multi-task SF-VAD framework, which integrates pretext tasks (temporal arrow prediction, motion irregularity estimation, intermediate frame reconstruction and model distillation) to jointly learn data representations, creating pseudo-labels as supervisory signals. This work converts the traditional single-objective detection task into a multi-task learning paradigm, effectively addressing the issue of suboptimal performance stemming from single-task approaches.

2.3 Middle-Level Visual Input

Middle-level visual input primarily focuses on skeleton-based VAD, leveraging human pose estimation and skeletal motion trajectory modeling to detect abnormal behaviors. While raw-video input relying on pixel information is vulnerable to noise and occlusion in complex backgrounds, leveraging compact skeletal feature representations as input enables models to alleviate the background noise interference, which is particularly effective for VAD in crowded scenes. Middle-level visual input remains a promising prospect for human-centric VAD, offering a balance between compact representation and effective motion analysis.

Compared to raw-video input, the additional pose estimation techniques(e.g. OpenPose [29], AlphaPose [30] and PoseNet [31] and HRNet [32]) enable real-time extraction of skeletal keypoints, motion trajectories, and pose maps from video data, making models adaptable to various situations. Combining with deep learning architectures like the graph convolutional network (GCN) [33], [34] and LSTM [35] network, middle-level visual-based models can effectively capture spatio-temporal relationships. However, the performance relying on the accuracy of pose estimation algorithms also makes it particularly challenging in multi-object scenes with significant occlusions. In the following sections, we analyze existing methods based on skeleton-level visual input, which is summarized on Table 4.

Table 4. The performances of middle-level visual input-based approaches on public datasets.

Approach	**Publication**	**Methodology**	**Pose Estimation**	**HR-Avenue**	**HR-ShanghaiTech**	**IITB-Corridor**	**UBnormal**
				AUC (%)			
GCAE [26]	CVPR 2020	Reconstruction + Clustering	Alphapose, Openpose	-	76.10	-	-
HSTGCNN [27]	TCSVT 2021	Prediction	HRNet	88.65	83.40	70.50	-
Multi-timescale [18]	TNNLS 2021	Prediction	Openpose	88.33	77.04	67.10	-
MPED-RNN [3]	CVPR 2021	Reconstruction + Prediction	Alphapose	86.30	75.40	-	-
MoCoDAD [28]	ICCV 2023	Reconstruction + Diffusion	Alphapose, Openpose	89.00	-	-	68.30

Reconstruction + Prediction-based approaches. To address the challenge of detecting human-related anomalies in surveillance videos with noisy and high-dimensional pixel- based features, Morais et al. [3] proposed a skeleton trajectory- based method, decomposing skeletal movements into global and local components. They employed a message-passing encoder-decoder recurrent neural network (MPED-RNN) to construct a spatio-temporal model, aiming to enhance the interpretability compared to traditional appearance-based approaches.

Prediction-based approaches. To effectively capture abnormal activities across different temporal scales, Rodrigues et al. [18] proposed a multi-timescale model to detect human anomalies at varying time scales. By performing bidirectional (past and future) multi-scale predictions on skeleton trajectory inputs and combining a hierarchical supervised approach, the model is able to learn effective representations at different temporal scales.

Reconstruction + Clustering-based approaches. To address the limitations of existing methods when dealing with complex scenes and diverse actions, Markovitz et al. [26] proposed a graph-based approach, using GCN to directly model dynamic body posture graphs and reduce background noise interference. They employed deep embedding clustering to effectively capture action diversity, followed by a Dirichlet process mixture model (DPMM) scoring mechanism to detect anomalies.

Diffusion-based approaches. To address the challenges posed by the complexity of human behaviors and the multi-modality of skeletal motion, Flaborea et al. [28] proposed a diffusion-based technique in the skeleton framework. They utilized the multi-modal generative capabilities of diffusion models to design a motion conditioned diffusion anomaly detection (MoCoDAD) model. This model leverages the coverage ability of the diffusion process to capture normal diversity and detects anomalies by aggregating future sequences, improving anomaly identification in multi-modal skeletal motion scenarios.

2.4 High-Level Visual- Semantic Input

The low level representations derived from raw-video inputs are often coarse and inadequate for modeling complex events in VAD. To address this limitation, an growing number of studies have adopted a two-stage VAD framework, using pre- trained models (such as ResNet [36], C3D [37] and Swin-T [38]) to extract high-level features. Unlike raw video data, high-level representations inherently capture more abstract and hybrid information cues such as appearance, motion and semantic context. This framework not only improves generalization across diverse scenarios but also enables the model to detect subtle and complex anomalies, even only given the coarse-grained labels. For a constructive understanding, a comparative summary of various high-level feature extraction methods is presented in Table 5.

Table 5. The performances of high-level visual-semantic input-based approaches on datasets.

Approach	**Publication**	**Feature**	**UCF-Crime**	**XD-Violence**	**ShanghaiTech**	**UCSD Ped2**
			AUC(%)	**AP (%)**	**AUC(%)**	**AUC(%)**
DeepMIL [37]	CVPR 2018	C3D^{RGB}	75.40	-	-	-
GCN [43]	CVPR 2019	TSN^{RGB}	82.12	-	84.44	-
HLNet [39]	ECCV 2020	I3D^{RGB}	82.44	75.41	-	-
RTFM [40]	ICCV 2021	I3D^{RGB}	84.30	77.81	97.21	98.60
MSL [41]	AAAI 2022	VideoSwin^{RGB}	85.62	78.59	97.32	-
TEVAD [44]	CVPR 2023	I3D^{RGB}	84.90	79.80	98.10	98.70
PE-MIL [45]	CVPR 2024	I3D^{RGB}	86.83	88.05	98.35	-
VadCLIP [4]	AAAI 2024	CLIP	88.02	84.51	-	-
STPrompt [46]	ACMMM 2024	CLIP	88.08	-	97.81	-
SUVAD [47]	ICASSP 2025	MLLM	82.3	80.1	76.8	96.2

Weakly Supervised Video Anomaly Detection (WS-VAD). For large-scale video data, providing clip-level annotations is extremely challenging. In addition, clip-level annotations are more prone to inconsistent labeling criteria due to the lack of a standardized definition of anomalies, which can negatively affect model performance. To solve these problems, many researches have shifted their focus toward WS-VAD, which utilizes weakly labeled training videos. Therefore, WS-VAD methods are typically formulated within the multiple instance learning (MIL) framework, where videos containing normal or abnormal clips were respectively represented as negative or positive bags.

Sultani et al. [37] was the first to employ this framework to construct a deep ranking model. This paradigm enables more effective identification of anomalous events in ambiguous patterns. However, despite significant progress in WS-VAD, several major challenges remain: (1) It is difficult to learn discriminative features of normal and anomalous clips through coarse-grained labels due to data imbalance. (2) WS-VAD framework usually assumes that each instance is independent, which challenges capturing long-range dependencies and local representations between clips. (3) Most WS-VAD methods can only provide clip-level anomalies, failing to localize anomalous objects. To resolve these challenges, many methods have been proposed to effectively optimize the traditional WS-VAD framework, increasing the probability of selecting abnormal clips from abnormal videos.

To address the limitations of WS-VAD for violence detection, where reliance on single-modal features and poor modeling of inter-clip relationships limit the effective localization of events. Wu et al. [39] proposed a method named holistic and localized network (HL-Net), leveraging multi-modal (audio-visual) fusion to enhance feature richness. They also employed holistic and localized branches with similarity and proximity priors to capture long and short-range clip dependencies. This method enhances the capability to localize violent events accurately.

To improve the discriminability learning within a video, i.e., the dominance of normal clips and subtle differences between normal and abnormal events often bias existing methods, Tian et al. [40] proposed a method called robust temporal feature magnitude learning (RTFM), which leverages a feature magnitude learning function to maximize the disparity between normal and abnormal features to improve discriminability. They also utilized dilated convolutions and self-attention to capture subtle anomalies. This approach improves the robustness to negative instances, alleviating the impact of data imbalance.

To reduce the probability of selection errors caused by ignoring the temporal continuity of abnormal events across multiple clips, Li et al. [41] proposed a multi-sequence learning (MSL) method, which uses a sequence composed of multiple clips as the optimization unit instead of single instance. Furthermore, they incorporated a transformer-based network architecture via self-attention to dynamically model complex inter-clip relationships and local perception, improving the accuracy of abnormal selection. Su et al. [42] integrates object detection and hierarchical disentanglement strategies. Through heterogeneous cross-scale correlation learning and a position-scale awareness inference mechanism, it simultaneously generates clip-level anomaly scores and spatially localizes anomalous targets.

Semantic + Visual Framework. The definition of anomaly is scene-dependent. Due to the diversity of visual features and their sensitivity to variations in feature extractors, object characteristic (e.g., density, scale and shooting viewing angle), a significant visual-semantic gap often arises. Moreover, existing WS-VAD methods predominantly follow a classification paradigm that relies heavily on visual features while neglecting the semantic alignment between video content and textual labels. To overcome these challenges, recent research has increasingly focused on the potential of multi-modal models where visual features are fused with rich semantic context. These approaches usually incorporate static knowledge bases or contrastive learning techniques, which can help models capture nuanced relationships between visual inputs and natural language semantics. This not only bridges the visual-semantic gap but also enhances the understanding of high-level concepts, paving the way for broader applications in the future.

To address the challenge of relying solely on visual features struggles to capture the deep semantic meanings in complex scenarios, Chen et al. [44] proposed a novel framework called text empowered VAD (TEVAD), leveraging video captions generated by the SwinBERT [48] model to obtain high-level semantic text features. The model fuses visual and textual features to address the semantic gap inherent in visual representations, which can interpret deep and rich semantic meanings. However, it remains dependent on video-level labeled training samples and high computational cost of SwinBERT, which limited the applicability in complex scenarios.

Su et al. [49] utilized a cross-modal detection network and dual consistency learning at the semantic-to-target and target-to-snippet levels to enhance discriminative representations of anomalies, enabling comprehensive identification, localization, and recognition of abnormal events in diverse scenarios.

To address the limitations of single-modal approaches in VAD, particularly their inability to perform effective cross-modal inference, Cao et al. [50] pioneered the application of GPT-4V for VAD. GPT-4V exhibits strong semantic understanding and reasoning abilities, enabling cross-modal inference through prompt engineering and exhibiting notable adaptability and generalization. They validated its great effec- tiveness in pedestrian and traffic scenes using the UCF-Crime [15] dataset. However, current research on GPT-4V relies on carefully designed prompts. Moreover, the high computational requirements also presents significant challenges for real-world deployment.

While traditional pre-trained models are dependent on fixed class datasets and ignore textual information, CLIP [51], however, uses natural language supervision on a large number of internet-sourced (image, text) pairs by contrastive learning, enabling multimodal modeling and zero-shot transfer to diverse tasks without labeled data. Wu et al. [4] was the first to employ CLIP into VAD. They leveraged frozen CLIP without additional fine-tuning. This method utilizes a local-global temporal adapter (LGT-Adapter) and a dual-branch framework with learnable prompts to harness vision-language associations. It has outperformed state-of-the-art methods on benchmarks like UCF-Crime [15] and XD-Violence [14], showing greater generalization across multiple contexts.

Although CLIP-based VAD methods have demonstrated strong image-text alignment capability, they are limited in analyzing global temporal context. To address the limitation, Wu et al. [46] proposed a spatio-temporal prompt learning approach (STPrompt) based on pre-trained vision-language models (VLMs) that enables temporal context modeling. Furthermore, they introduced a Spatial Attention Aggregation (SA^2) mechanism to suppress background noise, which improves VAD accuracy.

To address the challenge that CLIP-based VAD methods struggle to align video content with high-level semantic cues, Gao et al. [47] proposed a framework called semantic-guided unified VAD (SUVAD). By incorporating a semantic completion module (SCM) and a visual prompt generator (VPG), the framework integrates multi-modal knowledge from MLLMs to enrich temporal features with semantic information and generate prompts from scene- and object-level semantics. This design enables better guidance of visual features toward alignment with the semantic space, enhancing the capability of the model to detect semantic shifts and improving robustness in open-world scenarios.

3 Discussion

In this paper, we have reviewed recent advancements in VAD, demonstrating significant improvements in model performance on benchmark datasets. However, the practical effectiveness of these methods in real-world scenarios remains an open challenge. For instance, SS-VAD, while effective in leveraging unlabeled data, often fail to incorporate high-level semantic prior knowledge about anomalies. This limitation confines their focus to low-level pixel reconstruction or prediction tasks, leading to high false positive and false negative rates. On the other hand, WS-VAD methods, which rely on coarse annotations, have shown promise in improving detection performance. Nevertheless, these methods still face critical limitations, such as the inability to precisely localize anomalous regions and the reliance on two-stage optimization frameworks involving pre-trained feature extractors and MIL. These constraints significantly hinder their scalability and generalization.

Beyond these technical challenges, several critical issues in VAD remain underexplored. First, privacy preservation in surveillance videos has become increasingly important due to stringent regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). These regulations require the development of VAD models capable of protecting individual privacy, making distributed or privacy-aware learning frameworks a promising research direction [52]. Second, the heterogeneity of surveillance scenes—ranging from indoor environments to crowded public spaces—poses significant challenges for model generalization. Current methods often struggle to adapt to diverse scenarios, highlighting the need for more robust and adaptive architectures.

Furthermore, the real-time performance of VAD models is crucial for practical deployment in surveillance systems. To address this, research on model compression techniques, such as quantization and pruning, is essential to reduce computational overhead while maintaining detection accuracy. Finally, the integration of multi-modal data (e.g., combining visual and audio cues) and the exploration of SF-VAD learning paradigms could further enhance the robustness and efficiency of VAD systems. These directions not only address existing limitations but also open new avenues for future research in the field.

4 Conclusion

In this survey, we have conducted a comprehensive review of VAD approaches based on deep learning. We began by highlighting the predominant challenges in VAD, emphasizing the key limitations that hinder progress in the field. To provide a structured understanding, we introduced a novel taxonomy that categorizes existing methods into three levels based on their model inputs: raw-video input, middle-level visual input, and high-level visual-semantic input. For each category, we systematically analyzed representative methods, comparing their performances on various benchmark datasets. Furthermore, we provided an in-depth discussion on the pros and cons of these approaches, offering insights into their applicability across different real-world scenarios. Finally, we outlined promising research directions that can address current challenges and drive future advancements in VAD, hoping to inspire novel methodologies that push the boundaries of this field.

References

1. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 733–742.
2. W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in Proceedings of the IEEE international conference on eultimedia and expo, 2017, pp. 439–444.
3. R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11 996–12 004.
4. P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 6, 2024, pp. 6074–6082.
5. Y. Su, Y. Tan, M. Xing, and S. An, "Vpe-wsvad: Visual prompt exemplars for weakly-supervised video anomaly detection," Knowledge- Based Systems, vol. 299, p. 111978, 2024.
6. B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single- scene video anomaly detection," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 5, pp. 2293–2312, 2020.
7. J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep video anomaly detection: Opportunities and challenges," in Proceedings of the international conference on data mining workshops, 2021, pp. 959–966.
8. L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 42–62, 2022.
9. A. Berroukham, K. Housni, M. Lahraichi, and I. Boulfrifi, "Deep learning-based methods for anomaly detection in video surveillance: a review," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 1, pp. 314–327, 2023.
10. Y. Cao, X. Xu, J. Zhang, Y. Cheng, X. Huang, G. Pang, and W. Shen, "A survey on visual anomaly detection: Challenge, approach, and prospect," arXiv preprint arXiv:2401.16402, 2024.
11. P. Wu, C. Pan, Y. Yan, G. Pang, P. Wang, and Y. Zhang, "Deep learning for video anomaly detection: A review," arXiv preprint arXiv:2409.05383, 2024.

12. W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 1, pp. 18–32, 2013.
13. C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 2720–2727.
14. P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in Proceedings of the European Conference on Computer Vision, 2020, pp. 322–339.
15. W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.
16. W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in Proceedings of the IEEE International Conference on Multimedia and Expo. IEEE, 2017, pp. 439–444.
17. A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "Ubnormal: New benchmark for supervised open-set video anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 20 143–20 153.
18. R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi- timescale trajectory prediction for abnormal human activity detection," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 2626–2634.
19. D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep represen- tations of appearance and motion for anomalous event detection," arXiv preprint arXiv:1510.01553, 2015.
20. M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in Proceedings of the IEEE international conference on image processing, 2017, pp. 1577–1581.
21. W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6536– 6545.
22. H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14 372–14 381.
23. M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi- task learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12 742–12 752.
24. M. Z. Zaheer et al., "Generative cooperative learning for unsupervised video anomaly detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14 744–14 754.
25. N.-C. Ristea, F.-A. Croitoru, R. T. Ionescu, M. Popescu, F. S. Khan, M. Shah et al., "Self-distilled masked auto-encoders are efficient video anomaly detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15 984–15 995.
26. A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 539–10 547.
27. X. Zeng, Y. Jiang, W. Ding, H. Li, Y. Hao, and Z. Qiu, "A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 1, pp. 200–212, 2021.

28. A. Flaborea, L. Collorone, G. M. D. Di Melendugno, S. D'Arrigo, B. Prenkaj, and F. Galasso, "Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 10 318–10 329.
29. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291– 7299.
30. H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2334–2343.
31. G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4903–4911.
32. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5693–5703.
33. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.
34. M. Xing, Z. Feng, Y. Su, Y. Zhang, C. Oh, V. Gribova, V. F. Filaretoy, and D. Huang, "Spatio-temporal graph-based self-labeling for video anomaly detection," Neurocomputing, p. 129576, 2025.
35. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," Advances in neural information processing systems, 2015.
36. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
37. W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.
38. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10 012–10 022.
39. P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in Proceedings of the European Conference on computer vision, 2020, pp. 322–339.
40. Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4975–4986.
41. S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, 2022, pp. 1395–1403.
42. Y. Su, Y. Tan, S. An, and M. Xing, "Anomalies cannot materialize or vanish out of thin air: A hierarchical multiple instance learning with position-scale awareness for video anomaly detection," Expert Systems with Applications, vol. 254, p. 124392, 2024.
43. J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1237 – 1246.

44. W. Chen, K. T. Ma, Z. J. Yew, M. Hur, and D. A.-A. Khoo, “Tevad: Improved video anomaly detection with captions,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5549–5559.
45. J. Chen, L. Li, L. Su, Z.-J. Zha, and Q. Huang, “Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18 319–18 329.
46. P. Wu, X. Zhou, G. Pang, Z. Yang, Q. Yan, P. Wang, and Y. Zhang, “Weakly supervised video anomaly detection and localization with spatio-temporal prompts,” in Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 9301–9310.
47. S. Gao, P. Yang, and L. Huang, “Suvad: Semantic understanding based video anomaly detection using mllm,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.
48. K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, “Swinbert: End-to-end transformers with sparse attention for video captioning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17 949–17 958.
49. Y. Su, Y. Tan, S. An, M. Xing, and Z. Feng, “Semantic-driven dual consistency learning for weakly supervised video anomaly detection,” Pattern Recognition, vol. 157, p. 110898, 2025.
50. Y. Cao, X. Xu, C. Sun, X. Huang, and W. Shen, “Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead,” arXiv preprint arXiv:2311.02782, 2023.
51. A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proceedings of the International Conference on Machine Learning, 2021, pp. 8748–8763.
52. Y. Su, H. Zhu, Y. Tan, S. An, and M. Xing, “Prime: privacy-preserving video anomaly detection via motion exemplar guidance,” Knowledge- Based Systems, vol. 278, p. 110872, 2023.