International Journal of Emerging Technologies and Advanced Applications

A Survey of Anomalous Behavior Detection Techniques in Video Surveillance

Liang Ni
[1]Liaoning Dalian Tieda Comprehensive Market Management Co., Ltd., China
3672741@qq.com

***Abstract*—Anomalous behavior detection in video surveillance is a central research problem in the field of intelligent security. With the large-scale deployment of surveillance cameras in public spaces, the need for automatic identification of anomalous events—such as fighting, falling, and unattended objects—from massive video streams has become increasingly urgent. This paper systematically reviews the mainstream technical approaches in this domain, encompassing traditional methods based on optical flow and background modeling, deep learning methods based on convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), unsupervised anomaly detection methods based on autoencoders and generative adversarial networks (GANs), and the recently emerged Transformer and graph neural network (GNN) approaches. The paper further surveys widely adopted benchmark datasets—including UCF-Crime and ShanghaiTech—along with evaluation metrics such as frame-level area under the curve (AUC). The latest advances in multimodal fusion and edge deployment are discussed, and key challenges including high annotation costs, weak cross-scene generalization, and insufficient model interpretability are analyzed. Finally, the paper provides an outlook on frontier directions including vision-language large models and few-shot learning, aiming to serve as a comprehensive technical reference for researchers in this field.**

***Index Terms*—Anomalous behavior detection, Video surveillance, Deep learning, Spatiotemporal features, Multimodal fusion**

I. Introduction

Video surveillance systems have become an indispensable component of modern public safety infrastructure. Globally, the number of installed surveillance cameras has surpassed one billion, generating hundreds of petabytes of video data daily [1]. Faced with such a vast volume of video data, relying on human operators for real-time monitoring is not only inefficient but also highly susceptible to fatigue-induced miss-detections, making intelligent automated anomaly detection technology an urgent necessity [2].

The core objective of anomalous behavior detection is to automatically identify events that deviate from normal patterns in continuous video streams—such as crowd gatherings, violent confrontations, personnel falls, and suspicious unattended objects. Since the definition of an anomalous event is inherently scene-dependent and such events occur at an extremely low frequency in real-world data, this task is fundamentally a highly imbalanced open-set recognition problem [3]. Unlike standard visual tasks such as image classification, anomaly detection typically lacks sufficient annotated training samples, requiring models that can learn discriminative boundaries from normal examples alone. In recent years, the rapid advancement of deep learning has provided new solutions to this problem, with unsupervised and weakly supervised methods based on reconstruction error, prediction error, and contrastive learning being successively proposed, driving rapid progress in the field [4].

This paper aims to provide a systematic survey of technical approaches for anomalous behavior detection in video surveillance, tracing the evolution from traditional methods to deep learning approaches, analyzing mainstream datasets and evaluation frameworks, examining the latest advances in multimodal fusion and lightweight deployment, and presenting an outlook on future research directions—with the goal of offering a comprehensive reference for researchers and practitioners in this domain.

II. Classification and Development of Anomalous Behavior Detection Methods

The evolution of video anomaly detection methods has undergone a systematic transformation: from hand-crafted features to learned deep representations, from single-modal to multimodal inputs, and from fully supervised to unsupervised learning paradigms. Early research relied primarily on manually designed low-level visual features—such as optical flow fields, 3D histogram of oriented gradients (HOG3D), and background subtraction models—using statistical modeling to characterize the distributional boundary of normal behavior [5]. These methods exhibited reasonable utility in controlled settings with fixed cameras and simple backgrounds, but their robustness deteriorated noticeably under illumination changes, crowd occlusion, and complex scene dynamics. As convolutional neural networks achieved breakthroughs in image recognition, researchers began integrating deep features

ISSN:3006-2985(Print)
ISSN:3006-9297(Online)

Received: Dec 19, 2025
Revised: Jan 09, 2026
Accepted: Jan 25, 2026

into anomaly detection frameworks, progressively superseding hand-crafted features as the dominant paradigm [6].

From a methodological perspective, existing anomalous behavior detection approaches can be categorized into three major classes: reconstruction-based methods, prediction-based methods, and weakly supervised ranking-based methods. Reconstruction-based methods, exemplified by autoencoders, train a model to reconstruct normal frames and treat high reconstruction error as an anomaly signal [7]. Prediction-based methods train a model to predict future or missing frames, detecting anomalies as frames that deviate significantly from predicted values [8]. Weakly supervised ranking methods leverage coarse video-level labels and employ a multiple instance learning (MIL) framework to estimate anomaly scores at the segment level, substantially reducing the need for fine-grained annotation [9]. Additionally, GNN-based methods model human keypoints or scene objects as graph-structured nodes, capturing abnormal interactions between individuals via message-passing mechanisms [10]. In recent years, Transformer architectures have demonstrated unique advantages in temporal modeling of long video sequences, owing to their global self-attention mechanism, and have emerged as a leading technical direction [11].

Fig. 1 presents the overall taxonomy of video anomaly detection techniques, encompassing four dimensions: input modality, feature extraction, detection paradigm, and application scenario.

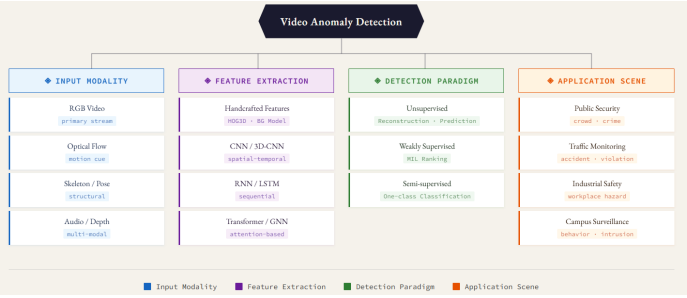

Fig. 1. Taxonomy of video anomaly detection along the dimensions of input modality, feature extraction, detection paradigm, and application scenario.

Among traditional methods, the Gaussian Mixture Model (GMM) is the most representative background modeling approach. By fitting a multi-Gaussian mixture distribution to the intensity values of each pixel, it achieves dynamic segmentation of foreground moving objects [5]. Optical flow methods compute motion vector fields between adjacent frames to characterize dynamic scene features, and are commonly combined with support vector machines (SVMs) for action classification [6]. However, these methods are sensitive to parameter tuning and struggle to capture high-level semantic information. The introduction of deep learning fundamentally transformed the feature extraction paradigm: end-to-end training frameworks enable models to automatically learn hierarchical spatiotemporal representations from raw pixels, significantly improving detection performance in complex scenes [7]. From a broader perspective, the field is converging toward three synergistic directions—unsupervised learning, multimodal integration, and model lightweighting—in response to the dual challenges of annotation scarcity and computational resource constraints in real-world deployment [9], [11].

III. Deep Learning-Based Anomalous Behavior Detection Methods

The rise of deep learning has fundamentally reshaped the technical landscape of video anomaly detection. Compared with traditional hand-crafted feature methods, deep neural networks can automatically learn hierarchical spatiotemporal representations from raw video data without manual feature engineering, exhibiting significant performance advantages in large-scale, complex scenes [12]. Based on differences in network architecture and training strategy, existing deep learning methods can be organized into five major technical lines: autoencoder-based reconstruction methods, prediction network-based sequential methods, weakly supervised ranking-based methods, Transformer-based attention methods, and GNN-based relational inference methods. Each line presents different trade-offs among detection accuracy, computational overhead, and annotation requirements.

A. Autoencoder-Based Reconstruction Methods

The autoencoder (AE) is the most widely adopted foundational architecture for unsupervised anomaly detection. Its core principle is to train an encoder-decoder network exclusively on normal samples, enabling the model to learn a compact latent representation of normal behavior. During inference, normal samples yield low reconstruction error due to their consistency with the training distribution, whereas anomalous samples produce high reconstruction error due to their deviation from the learned normal manifold, which directly serves as the anomaly score [13]. Let the input video frame be $\mathbf{x}$, the encoder mapping be f_e, and the decoder mapping be f_d; the reconstruction error is defined as:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - f_d(f_e(\mathbf{x}))\|_2^2 \tag{1}$$

The anomaly score $S(\mathbf{x})$ is typically taken as the reconstruction error directly, or after normalization, compared with a predefined threshold τ:

$$S(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \quad \text{anomaly detected if } S(\mathbf{x}) > \tau \tag{2}$$

However, standard autoencoders suffer from an "over-generalization" problem, whereby the model sometimes reconstructs anomalous samples with low error as well, reducing detection sensitivity. To address this, researchers proposed the Memory-Augmented Autoencoder (MemAE), which introduces an addressable memory module at the bottleneck layer to restrict the model's capacity to represent anomalous patterns, effectively raising the reconstruction error floor for anomalous samples [13]. The Variational Autoencoder (VAE) imposes a probabilistic distributional constraint on the latent space, encouraging the latent codes of normal samples to concentrate near a standard Gaussian prior; anomalous samples are identified by their deviation from this prior [14].

B. Prediction Network-Based Sequential Methods

Prediction-based methods reformulate anomaly detection as a video frame prediction task, training models to predict future or missing frames and using prediction error as an anomaly indicator. Such methods naturally incorporate temporal dynamic information, conferring strong detection capability for motion-pattern anomalies [15]. A frame prediction network combining U-Net with convolutional LSTM achieved state-of-the-art performance on the ShanghaiTech dataset by jointly constraining appearance reconstruction error and optical flow prediction error, enabling simultaneous detection of appearance and motion anomalies. Let $\hat{\mathbf{x}}_t$ denote the predicted value of frame t; the joint loss function is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{int}} + \lambda_2 \mathcal{L}_{\text{gd}} + \lambda_3 \mathcal{L}_{\text{flow}} \tag{3}$$

where $\mathcal{L}_{\text{int}}$ is the intensity loss, $\mathcal{L}_{\text{gd}}$ is the gradient difference loss, $\mathcal{L}_{\text{flow}}$ is the optical flow constraint loss, and $\lambda_1, \lambda_2, \lambda_3$ are balancing weights [15]. More recently, diffusion model-based prediction methods have been introduced into anomaly detection, leveraging learned denoising distributions of normal video to produce larger reconstruction deviations for anomalous frames at inference time [16].

C. Weakly Supervised Ranking-Based Methods

Weakly supervised methods exploit coarse video-level labels (indicating only whether a video contains anomalies, without frame-level annotation) to estimate segment-level anomaly scores via a Multiple Instance Learning (MIL) framework. The ranking loss proposed by Lv *et al.* is a foundational contribution in this line of work; its core assumption is that the highest-scoring segment from an anomalous video bag should substantially exceed the highest-scoring segment from a normal video bag [17]. Let $s_a^{\max}$ and $s_n^{\max}$ denote the maximum segment scores in the anomalous bag $\mathcal{B}_a$ and the normal bag $\mathcal{B}_n$, respectively; the ranking loss is:

$$\mathcal{L}_{\text{rank}} = \max(0,\ 1 - s_a^{\max} + s_n^{\max}) \tag{4}$$

Building on this foundation, subsequent work introduced temporal smoothness constraints and sparsity regularization to further improve the temporal coherence of anomaly score curves [17]. The incorporation of graph convolutional networks and attention mechanisms has enabled more precise localization of anomalous segment boundaries in time, with frame-level AUC on the UCF-Crime dataset being continuously improved [18].

D. Transformer-Based Methods

Transformer architectures have demonstrated unique advantages in video anomaly detection owing to their self-attention mechanism's powerful capacity for modeling global temporal dependencies. Spatiotemporal joint modeling methods based on the Vision Transformer (ViT) partition videos into spatiotemporal tubes and employ multi-head self-attention to simultaneously capture intra-frame spatial relationships and inter-frame temporal dependencies, effectively overcoming the limited receptive field of convolutional networks [19]. Moreover, pre-training strategies based on the Masked Autoencoder (MAE) have been introduced into anomaly detection: by randomly masking video patches and reconstructing them, models acquire more robust representations of normal video content [20]. Within weakly supervised frameworks, Transformer-encoded segment features combined with temporal attention modules can more accurately capture the abruptness and persistence of anomalous events, surpassing CNN-based baselines across multiple benchmark datasets.

E. Graph Neural Network-Based Methods

Graph Neural Network (GNN) methods model individuals or objects in a scene as graph nodes and their spatial or interaction relationships as graph edges, capturing abnormal interaction patterns between individuals via message passing [21]. Skeletal keypoint graph methods extract 17 human body keypoints using pose estimation networks, construct human skeleton graphs, and model action sequences with Spatial-Temporal Graph Convolutional Networks (ST-GCN), achieving high detection accuracy for limb-related anomalies such as falls and fights [22]. Scene graph methods construct detected multi-object scenes and their relative spatial relationships as scene graphs, employ Graph Attention Networks (GAT) to learn the graph-structural distribution of normal scenes, and identify graph structures deviating from this distribution as anomalies [23]. Compared with RGB-feature methods, skeleton-based approaches offer inherent robustness to illumination changes and background interference, though their performance is highly dependent on pose estimation accuracy and may be limited in occlusion scenarios.

Table I provides a systematic performance comparison of the five categories of mainstream deep learning methods on representative datasets.

IV. Benchmark Datasets and Evaluation Metrics

High-quality benchmark datasets constitute a critical foundation for advancing research in video anomaly detection. Because the collection and annotation of anomalous events entail extremely high costs and raise ethical concerns related to privacy protection, the number of publicly available datasets in this domain is relatively limited. Nevertheless, existing datasets differ significantly in scene complexity, anomaly type, and annotation granularity, providing complementary validation platforms for different technical approaches [24]. Early datasets were characterized by single-scene settings and low resolution; in recent years, the release of large-scale weakly supervised datasets has shifted research toward more complex scenes that better reflect real-world deployment conditions [25].

A. Major Publicly Available Datasets

The **UCSD Pedestrian** dataset is among the earliest and most widely used benchmarks for anomaly detection. It comprises two subsets—Ped1 and Ped2—captured by a fixed camera overlooking a pedestrian walkway, with anomalies consisting of non-pedestrian objects such as cyclists and skateboarders. The dataset provides pixel-level anomaly region

TABLE I
PERFORMANCE COMPARISON OF MAINSTREAM DEEP LEARNING ANOMALY DETECTION METHODS

Method	Architecture	Training	Dataset	Frame-AUC (%)	Year
MemAE [13]	Memory Autoencoder	Unsupervised	UCSD Ped2	94.1	2019
MNAD [14]	VAE + Memory	Unsupervised	ShanghaiTech	91.9	2024
SSPCAB [15]	Prediction Network	Unsupervised	Avenue	91.5	2022
LAW [17]	MIL + Transformer	Weakly Sup.	UCF-Crime	84.3	2021
VadCLIP [19]	CLIP + Transformer	Weakly Sup.	UCF-Crime	88.0	2023
STG-NF [22]	Skeleton Graph + NF	Unsupervised	ShanghaiTech	85.9	2023
MGFN [18]	Graph Feature Network	Weakly Sup.	UCF-Crime	86.7	2023
LFVAD [20]	MAE Pre-training	Unsupervised	Avenue	92.8	2024

TABLE II
COMPARISON OF MAJOR VIDEO ANOMALY DETECTION DATASETS

Dataset	Scenes	Train Videos	Test Videos	Anomaly Types	Annotation	Year
UCSD Ped2 [24]	1	16	12	5	Pixel-level	2010
CUHK Avenue [24]	1	16	21	5	Frame-level	2013
ShanghaiTech [25]	13	330	107	130+	Frame-level	2018
UCF-Crime [26]	Multiple	1610	290	13	Video-level	2018
XD-Violence [27]	Multiple	3954	800	6	Segment-level	2020
MSAD [28]	20	720	200	14	Frame-level	2023

annotations, supporting both frame-level and pixel-level evaluation; however, its single-scene setting and low resolution limit its ability to reflect the complexity of realistic surveillance [24].

The **CUHK Avenue** dataset contains 16 training videos and 21 test videos covering anomaly types such as running, throwing objects, and abnormal loitering in a campus corridor. The background is relatively clean, though moderate camera shake is present [24].

The **ShanghaiTech Campus** dataset is a large-scale, multi-scene unsupervised anomaly detection benchmark spanning 13 different camera viewpoints, comprising 330 training videos and 107 test videos with over 130 anomaly types. Its scene complexity substantially exceeds that of the preceding two datasets [25].

The **UCF-Crime** dataset is the largest weakly supervised anomaly detection benchmark to date, containing 1,900 real surveillance videos covering 13 categories of criminal behavior. Only video-level labels are provided, making it the standard platform for evaluating weakly supervised methods [26].

The **XD-Violence** dataset further extends UCF-Crime with 4,754 clips sourced from movies, surveillance footage, and online videos, providing both audio and visual bimodal information to support the evaluation of multimodal fusion methods [27].

Table II presents a systematic comparison of the core properties of these major datasets.

B. Evaluation Metrics

The *frame-level Area Under the Curve* (Frame-AUC) is the most widely used evaluation metric in this field. By plotting the receiver operating characteristic (ROC) curve at different decision thresholds, it quantifies the model's overall detection performance across all threshold settings via the area under the curve [26]. Letting TPR denote the true positive rate and FPR the false positive rate, the AUC is defined as:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR}) \tag{5}$$

An AUC value closer to 1.0 indicates that the model maintains a high detection rate across a wide range of false positive rates. The *Equal Error Rate* (EER) is another commonly used metric, corresponding to the operating point at which the false positive rate equals the false negative rate; a lower EER indicates better detection performance. For datasets that provide pixel-level annotations, *pixel-level AUC* offers a finer-grained assessment of anomaly localization accuracy, requiring the model not only to detect anomalous frames but also to accurately delineate the spatial extent of anomalous regions [24]. In weakly supervised scenarios, some works adopt *mean Average Precision* (mAP) to evaluate segment-level detection performance, which is more closely aligned with the practical requirements of real-world alarm systems [27]. It should be noted that different datasets employ varying evaluation protocols, and direct cross-dataset performance comparisons should be made with caution.

V. MULTIMODAL FUSION AND EDGE COMPUTING DEPLOYMENT

Single-modality anomaly detection methods face inherent limitations in complex real-world scenarios: pure RGB methods are sensitive to drastic illumination changes, skeleton-based methods fail under occlusion, and static appearance methods struggle to distinguish normal from anomalous events with similar visual appearance. Multimodal fusion strategies address these blind spots by integrating complementary information from different perceptual dimensions, improving the

TABLE III
PERFORMANCE COMPARISON OF MULTIMODAL FUSION VS. SINGLE-MODALITY METHODS

Method	Modalities	Fusion Strategy	Dataset	AUC/AP (%)	Year
RGB Baseline	RGB	—	XD-Violence	73.2	2020
Wu *et al.* [30]	RGB + Audio	Late Fusion	XD-Violence	75.9	2021
MTC-Net [31]	RGB + Optical Flow	Intermediate Fusion	UCF-Crime	75.23	2024
VadCLIP [19]	RGB + Text	CLIP Alignment	UCF-Crime	88.0	2023
UMIL [32]	RGB + Text	Unified MIL	XD-Violence	86.7	2023
DAKD [33]	RGB + Audio + Text	Cross-modal Transformer	XD-Violence	85.61	2025
SAA [29]	RGB	I3D	UCF-Crime	86.19	2024

robustness of detection systems in open environments [29]. In recent years, as sensor costs have continued to decline and edge computing hardware has proliferated, research into lightweight anomaly detection systems designed for practical deployment has also grown in importance.

A. Multimodal Fusion Strategies

Multimodal fusion in video anomaly detection can be categorized, from the perspective of input modality, into intra-visual multimodal fusion and cross-sensory multimodal fusion. Intra-visual multimodal fusion is epitomized by the two-stream fusion of RGB and optical flow, where the two network branches extract appearance and motion features respectively, achieving complementarity through feature-level or decision-level fusion [29]. Cross-sensory multimodal fusion further incorporates heterogeneous information sources such as audio, depth maps, or skeletal keypoints, demonstrating clear performance gains on multimodal datasets such as XD-Violence [30].

In terms of fusion mechanism, *early fusion* concatenates raw multimodal features at the input stage before jointly feeding them into the subsequent network; this is straightforward to implement but suffers from high inter-modal feature heterogeneity. *Late fusion* independently extracts features from each modality and combines them at the decision layer via weighted aggregation; this offers flexibility but ignores low-level inter-modal interactions. *Intermediate fusion* performs cross-modal attention interaction at intermediate network layers, and is currently the best-performing mainstream strategy [31]. Cross-modal Transformer-based fusion methods design inter-modal cross-attention modules to align visual and audio features at the semantic level, achieving significant improvements on violence detection tasks [30]. More recently, the introduction of vision-language pre-trained models such as CLIP has offered a new multimodal fusion paradigm: by mapping video clips and textual descriptions into a unified semantic space, zero-shot or few-shot anomaly detection becomes feasible [32].

Table III compares the performance of multimodal and single-modality methods on representative datasets.

B. Edge Computing Deployment

Deploying anomaly detection models on edge devices is a critical pathway to achieving low-latency real-time alert systems; however, a significant tension exists between the high computational complexity of deep learning models and the constrained resources of edge hardware. The mainstream lightweighting strategies encompass four technical lines: model pruning, knowledge distillation, quantization compression, and neural architecture search [33]. Model pruning reduces model parameter counts by removing redundant weights or channels; knowledge distillation uses large teacher networks to guide the training of small student networks, enabling lightweight models to dramatically reduce inference overhead while maintaining relatively high accuracy [34]. Quantization compression converts floating-point weights into low-bit integer representations; INT8 quantization incurs no more than a 1% AUC drop on most anomaly detection models while achieving a $2\times$ to $4\times$ speedup in inference [33].

In terms of practical deployment, NVIDIA Jetson series chips and Rockchip RK3588 edge AI processors have been widely adopted in intelligent surveillance terminals, supporting INT8 inference acceleration. Research demonstrates that lightweight anomaly detection models jointly optimized through knowledge distillation and INT8 quantization can achieve real-time inference at more than 25 frames per second on the Jetson Orin NX platform, meeting the latency requirements of most surveillance scenarios [34]. The integration of federated learning provides a privacy-preserving mechanism for cooperative multi-camera model training: each edge node completes model training locally and uploads only gradient updates, avoiding the centralized transmission of raw video data—which is of significant regulatory compliance value [35].

VI. CHALLENGES, TRENDS, AND FUTURE OUTLOOK

Although video anomaly detection has achieved substantial progress in recent years, a considerable gap remains between laboratory benchmark performance and real-world deployment scenarios. The challenges facing current techniques stem not only from algorithmic limitations but also from multiple dimensions including data acquisition, privacy regulations, and systems engineering; a thorough understanding of these challenges is a prerequisite for sustained progress in the field.

A. Core Technical Challenges

Scarcity of annotated data is the primary bottleneck constraining progress in this field. Anomalous events are inherently rare in reality, and fine-grained frame-level or pixel-level annotation requires substantial human effort, resulting in datasets that are generally small in scale and coarse in annotation granularity. Large-scale datasets such as UCF-Crime

provide only video-level weak labels, from which models struggle to learn precise temporal localization capabilities. Data augmentation and synthetic data generation are important strategies for alleviating this problem; anomalous sample synthesis methods based on generative adversarial networks and diffusion models have demonstrated preliminary promise in improving detector performance, though the distributional gap between synthetic and real anomalous samples remains an unsolved challenge.

Insufficient cross-scene generalization is another core challenge. Existing methods are typically trained and evaluated on datasets captured in specific scenes, such that the normal-pattern representations and anomaly-discriminative boundaries learned by the model are highly dependent on the statistical characteristics of the training scene, often leading to severe performance degradation during cross-scene transfer. Constructing a scene-adaptive universal anomaly detection framework is a key challenge for scaling the technology from laboratory settings to large-scale practical deployment. Domain adaptation and meta-learning approaches offer promising technical avenues, yet systematic exploration of these methods specifically within the video anomaly detection domain remains limited.

Insufficient model interpretability hinders the practical deployment of intelligent security systems. The black-box nature of deep neural networks makes it difficult for security operators to understand the basis for system alarms, posing compliance risks in scenarios involving legal forensics and liability determination. Interpretability methods based on attention visualization, gradient-weighted class activation mapping (Grad-CAM), and concept bottleneck models have been preliminarily introduced into anomaly detection, but providing intuitive and reliable decision explanations while maintaining high detection accuracy remains an open research problem.

B. Emerging Research Trends

The rapid development of *Vision-Language Models* (VLMs) and large-scale video understanding models introduces a new technical paradigm for anomaly detection. Large models such as GPT-4V, VideoLLaMA, and InternVideo possess powerful open-vocabulary semantic understanding capabilities, enabling the specification of anomaly types through natural language descriptions and achieving zero-shot anomaly detection and localization. This language-guided detection paradigm transcends the limitations of traditional closed-set anomaly categories, making it particularly relevant for coping with novel anomalous events that continuously emerge in real-world settings. However, the inference latency and computational overhead of large models currently remain insufficient for real-time monitoring demands; efficient large model compression and inference acceleration are necessary prerequisites for practical deployment.

Few-shot and zero-shot anomaly detection represents another important frontier. In real-world deployments, novel anomalous events typically lack sufficient training samples, requiring models with the ability to rapidly learn new anomaly patterns from very few examples. Few-shot methods based on metric learning and prototypical networks, as well as zero-shot methods leveraging pre-trained features from large foundation models, are progressively emerging as active research topics in this domain, with the potential to fundamentally overcome the dependence of traditional methods on large-scale annotated data.

Privacy protection and data compliance issues are becoming increasingly prominent as data protection regulations are enacted worldwide. How to complete model training and updating without accessing raw video data is a practical problem that must be resolved for large-scale commercial deployment of intelligent security systems. The deep integration of privacy-preserving computing techniques—including federated learning, differential privacy, and homomorphic encryption—with anomaly detection methods will constitute an important cross-disciplinary research direction in the future.

VII. Conclusion

This paper presents a systematic survey of anomalous behavior detection techniques in video surveillance, tracing the evolution from traditional hand-crafted feature methods to deep learning approaches across five mainstream technical lines: autoencoder-based reconstruction, frame prediction, weakly supervised ranking, Transformer-based attention, and graph neural network-based relational reasoning. Key benchmark datasets—including UCSD Pedestrian, CUHK Avenue, ShanghaiTech, UCF-Crime, and XD-Violence—along with standardized metrics such as frame-level AUC and mAP, have been systematically reviewed to contextualize performance comparisons across methods. Research has demonstrated that deep learning methods have achieved high detection accuracy across multiple standard benchmark datasets, with multimodal fusion strategies and lightweight edge deployment techniques accelerating the practical advancement of this field. In particular, knowledge distillation combined with INT8 quantization has proven effective for real-time inference on resource-constrained edge platforms, while federated learning offers a principled approach to privacy-compliant multi-camera collaborative training. Nevertheless, core challenges including the scarcity of annotated data, insufficient cross-scene generalization, and limited model interpretability continue to hinder large-scale real-world deployment. Looking ahead, the integration of vision-language large models, the maturation of few-shot learning frameworks, and the strengthening of privacy-preserving mechanisms will open new avenues for video anomaly detection technology, propelling intelligent surveillance systems toward greater accuracy, stronger generalization capability, and broader application coverage.

References

[1] G. Sreenu and M. A. S. Durai, "Intelligent video surveillance: A review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, no. 1, p. 48, 2019, doi: 10.1186/s40537-019-0212-5.

[2] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2293–2312, May 2022, doi: 10.1109/TPAMI.2020.3040591.

[3] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, China, 2017, doi: 10.1109/ICME.2017.8019325.

[4] Q. Yang, C. Wang, P. Liu, Z. Jiang, and J. Li, "Video anomaly detection via self-supervised and spatio-temporal proxy tasks learning," *Pattern Recognit.*, vol. 158, p. 111021, 2025, doi: 10.1016/j.patcog.2024.111021.

[5] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 3449–3456, doi: 10.1109/CVPR.2011.5995434.

[6] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5288–5301, Dec. 2015, doi: 10.1109/TIP.2015.2479561.

[7] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6536–6545, doi: 10.1109/CVPR.2018.00684.

[8] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 733–742, doi: 10.1109/CVPR.2016.86.

[9] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[10] 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 Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 200–212, Jan. 2023, doi: 10.1109/TCSVT.2021.3134410.

[11] M. Abdalla, S. Javed, M. Al Radi *et al.*, "Video anomaly detection in 10 years: A survey and outlook," *Neural Comput. Appl.*, vol. 37, pp. 26321–26364, 2025, doi: 10.1007/s00521-025-11659-8.

[12] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 4955–4966, doi: 10.1109/ICCV48922.2021.00493.

[13] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 1705–1714, doi: 10.1109/ICCV.2019.00179.

[14] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 14360–14369, doi: 10.1109/CVPR42600.2020.01438.

[15] N.-C. Ristea, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 13576–13586, doi: 10.1109/CVPR52688.2022.01321.

[16] C. Yan, S. Zhang, Y. Liu, G. Pang, and W. Wang, "Feature prediction diffusion model for video anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 5504–5514, doi: 10.1109/ICCV51070.2023.00509.

[17] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Trans. Image Process.*, vol. 30, pp. 4505–4515, 2021, doi: 10.1109/TIP.2021.3072863.

[18] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, 2023, vol. 37, no. 1, pp. 387–395, doi: 10.1609/aaai.v37i1.25112.

[19] 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 *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, 2024, vol. 38, no. 6, pp. 6074–6082, doi: 10.1609/aaai.v38i6.28423.

[20] F. A. Croitoru, N.-C. Ristea, D. Dăscălescu, R. T. Ionescu, F. S. Khan, and M. Shah, "Lightning fast video anomaly detection via multi-scale adversarial distillation," *Comput. Vis. Image Underst.*, vol. 247, p. 104074, 2024, doi: 10.1016/j.cviu.2024.104074.

[21] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021, doi: 10.1016/j.neucom.2019.12.148.

[22] O. Hirschorn and S. Avidan, "Normalizing flows for human pose anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 13499–13508, doi: 10.1109/ICCV51070.2023.01246.

[23] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 8022–8031, doi: 10.1109/CVPR52729.2023.00775.

[24] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014, doi: 10.1109/TPAMI.2013.111.

[25] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 13568–13577, doi: 10.1109/ICCV48922.2021.01333.

[26] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[27] 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 *Proc. European Conf. Comput. Vis. (ECCV)*, Glasgow, UK, 2020, pp. 322–339, doi: 10.1007/978-3-030-58577-8_20.

[28] F. Caetano, P. Carvalho, and J. S. Cardoso, "Unveiling the performance of video anomaly detection models—A benchmark-based review," *Intell. Syst. Appl.*, vol. 18, p. 200236, 2023, doi: 10.1016/j.iswa.2023.200236.

[29] Y. Fan, Y. Yu, W. Lu, and Y. Han, "Weakly-supervised video anomaly detection with snippet anomalous attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5480–5492, Jul. 2024, doi: 10.1109/TCSVT.2024.3350084.

[30] 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 *Proc. European Conf. Comput. Vis. (ECCV)*, Glasgow, UK, 2020, pp. 322–339, doi: 10.1007/978-3-030-58577-8_20.

[31] M. Xu, D. Li, R. Wu, H. Ding, K. Huang, and C. Chu, "MTC-Net: Multi-granular temporal cascade learning for weakly supervised video anomaly detection," in *Proc. Int. Conf. Image Process. Comput. Vis. Mach. Learn. (ICICML)*, Shenzhen, China, 2024, pp. 881–886, doi: 10.1109/ICICML63543.2024.10958105.

[32] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 8022–8031, doi: 10.1109/CVPR52729.2023.00775.

[33] J. Dalvi, A. Dabouei, G. Dhanuka, and M. Xu, "Distilling aggregated knowledge for weakly-supervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Tucson, AZ, USA, 2025, pp. 5439–5448, doi: 10.1109/WACV61041.2025.00531.

[34] C. Zhang, G. Li, Y. Qi, H. Ye, L. Qing, M.-H. Yang, and Q. Huang, "Dynamic erasing network with adaptive temporal modeling for weakly supervised video anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, pp. 16706–16720, 2025, doi: 10.1109/TNNLS.2025.3553556.

[35] H. T. Duong, V. T. Le, and V. T. Hoang, "Deep learning-based anomaly detection in video surveillance: A survey," *Sensors*, vol. 23, no. 11, p. 5024, 2023, doi: 10.3390/s23115024.