

A Comprehensive Review of Multimodal Weakly Supervised Video Anomaly Detection

Mr. Sreekanth K U¹, Mr. Adwaid R S²

¹ Assistant Professor, IES College of Engineering, Chittilappilly, Kerala, India

² Scholar, Government Engineering College, Idukki, Kerala, India

Abstract — Video Anomaly Detection (VAD) has emerged as a critical research area in intelligent surveillance, public safety, industrial monitoring, and smart city applications. Traditional supervised approaches rely on frame-level annotations, which are labor-intensive, time-consuming, and expensive to obtain, thereby limiting their scalability in large-scale surveillance systems. Weakly Supervised Video Anomaly Detection (WSVAD) addresses this challenge by utilizing only video-level labels while enabling effective temporal anomaly localization. In recent years, multimodal learning has significantly improved detection performance by integrating complementary information from visual, audio, and semantic modalities. This paper presents a comprehensive review of multimodal WSVAD techniques, emphasizing recent advancements in Multiple Instance Learning (MIL), vision-language models, audio-visual learning, knowledge transfer, and semantic alignment strategies. Representative methods are systematically analyzed and compared with respect to their methodologies, benchmark datasets, evaluation metrics, strengths, and limitations. Furthermore, major research challenges, including modality imbalance, limited multimodal datasets, computational complexity, and domain generalization, are discussed. Finally, emerging research directions involving foundation models, multimodal large language models (MLLMs), self-supervised learning, federated learning, and explainable artificial intelligence (XAI) are highlighted to support the development of scalable, robust, and interpretable anomaly detection systems.

Keywords — Video Anomaly Detection, Weakly Supervised Learning, Multimodal Learning, Semantic Alignment, Knowledge Transfer, Vision-Language Models.

I. INTRODUCTION

Video Anomaly Detection (VAD) has become an important research area in computer vision owing to its broad applications in intelligent surveillance, public safety management, industrial automation, healthcare monitoring, and smart city infrastructures. The rapid deployment of surveillance cameras has resulted in the generation of massive volumes of video data, making continuous manual monitoring impractical and susceptible to human error. Consequently, automated anomaly detection systems have gained considerable attention for their ability to identify unusual activities that deviate from normal behavioural patterns, thereby facilitating timely responses to security threats, accidents, and operational failures [1], [2].

Early VAD approaches primarily relied on handcrafted feature extraction combined with traditional machine learning techniques. Features such as Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), motion trajectories, optical flow patterns, and sparse reconstruction-based representations were widely employed to characterize normal and abnormal activities [3], [4]. Although these methods achieved satisfactory performance in controlled environments, their dependence on manually engineered features limited

their adaptability to complex real-world scenarios involving dynamic backgrounds, illumination variations, occlusions, and crowded scenes. Furthermore, they lacked the semantic understanding necessary for recognizing complex anomalous events.

The emergence of deep learning has significantly transformed video anomaly detection by enabling automatic learning of discriminative spatial and temporal representations. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Three-Dimensional CNNs (3D-CNNs), and Transformer-based architectures have demonstrated superior performance by effectively capturing motion dynamics and long-range temporal dependencies directly from video data [5], [6]. Despite their effectiveness, these methods generally require large-scale frame-level annotations, resulting in substantial labelling costs and limiting their practical deployment in large surveillance systems.

To mitigate the annotation burden associated with fully supervised learning, Weakly Supervised Video Anomaly Detection (WSVAD) has emerged as a promising alternative. Unlike fully supervised approaches, WSVAD utilizes only video-level labels that indicate the presence or absence of

anomalies. Multiple Instance Learning (MIL) has become the dominant learning paradigm in this setting, where each video is treated as a bag of temporal segments. Seminal methods such as MIL-Rank, RTFM, MIST, and MGFN have demonstrated that effective anomaly localization can be achieved without requiring frame-level annotations [7]–[10].

Despite significant progress, visual information alone is often insufficient for robust anomaly detection in real-world environments. Factors such as poor illumination, weather conditions, motion blur, camera occlusion, and low-resolution recordings can substantially degrade detection performance. Moreover, many anomalous events inherently contain acoustic cues that may not be visually observable. Examples include explosions, gunshots, alarms, machinery failures, vehicle collisions, and human distress signals. These limitations have motivated the development of multimodal learning frameworks that integrate information from multiple modalities, particularly visual and audio streams [11].

Recent advances in multimodal learning have demonstrated that combining heterogeneous information sources significantly improves both detection accuracy and robustness. Audio-visual learning frameworks leverage complementary temporal and semantic information to enhance anomaly localization under challenging conditions. Various multimodal fusion strategies, including early fusion, late fusion, intermediate fusion, and cross-modal attention mechanisms, have been proposed to effectively capture interactions between modalities while preserving modality-specific characteristics [12], [13].

Another notable advancement is the introduction of Vision-Language Models (VLMs), such as CLIP, which learn shared representations from large-scale image-text datasets [14]. By incorporating textual descriptions and prompt-based learning strategies, these models provide rich semantic knowledge that enhances anomaly understanding and improves generalization to previously unseen abnormal events. Nevertheless, most vision-language approaches primarily focus on visual semantics and do not fully exploit acoustic information available in surveillance scenarios.

More recently, research efforts have focused on multimodal frameworks that combine knowledge transfer and semantic alignment techniques. Knowledge transfer mechanisms facilitate information sharing across modalities, whereas semantic alignment aims to map heterogeneous representations into a common latent space to improve cross-modal consistency and anomaly discrimination. These approaches have demonstrated superior robustness against noisy supervision, modality imbalance, and domain variations, thereby advancing the development of more reliable multimodal anomaly detection systems [15], [16].

Despite these advances, several challenges remain unresolved. Existing multimodal approaches often require synchronized multimodal data, involve significant computational complexity, and exhibit limited robustness in the presence of missing or corrupted modalities. Furthermore, the scarcity of large-scale multimodal benchmark datasets, lack of standardized evaluation protocols, limited model interpretability, and real-time deployment constraints continue to hinder practical adoption.

To address these issues, this paper presents a comprehensive review of multimodal weakly supervised video anomaly detection. Unlike existing surveys that predominantly focus on visual anomaly detection, this review provides a detailed analysis of recent developments in audio-visual learning, vision-language modelling, cross-modal interaction, knowledge transfer, semantic alignment, and contrastive learning. In addition, representative methods are comparatively analysed based on their learning strategies, strengths, limitations, benchmark datasets, and evaluation metrics. Finally, current research challenges and emerging directions, including foundation models, multimodal large language models, self-supervised learning, federated learning, and explainable artificial intelligence, are discussed to provide valuable insights for future research.

The remainder of this paper is organized as follows. Section II reviews existing literature on video anomaly detection and multimodal learning techniques. Section III describes the review methodology adopted in this study. Section IV presents a comparative analysis of representative approaches and discusses recent research trends. Section V outlines major research challenges and promising future directions. Finally, Section VI concludes the paper.

II. LITERATURE SURVEY

Video Anomaly Detection (VAD) has undergone significant evolution over the past decade owing to the increasing demand for intelligent surveillance systems capable of automatically identifying abnormal events in large-scale video streams. Initial research efforts primarily focused on handcrafted feature extraction and statistical learning techniques, while recent advancements have shifted toward deep learning, weakly supervised learning, vision-language modeling, and multimodal representation learning.

This progression has been driven by the need to improve anomaly localization accuracy, reduce annotation costs, and enhance model generalization across diverse surveillance environments. Recent studies further emphasize semantic reasoning, multimodal understanding, and explainable anomaly detection as key directions for future research. Furthermore, the growing availability of large-scale datasets

and advances in computational resources have accelerated the development of robust and scalable anomaly detection frameworks. Consequently, modern research increasingly focuses on developing efficient and interpretable systems capable of understanding complex anomalous activities with minimal human intervention.

A. Traditional Video Anomaly Detection

Early video anomaly detection methods relied on handcrafted feature representations combined with conventional machine learning algorithms. Commonly employed features included Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), optical flow patterns, motion trajectories, and spatiotemporal interest points. These representations were subsequently analyzed using clustering, classification, reconstruction, or statistical modeling techniques to distinguish normal and abnormal behaviors.

Among the pioneering approaches, sparse reconstruction and dictionary learning methods assumed that normal events could be faithfully reconstructed using learned representations, whereas anomalous events generated significantly larger reconstruction errors. Similarly, statistical crowd analysis techniques, such as the Social Force Model, characterized pedestrian interactions to identify abnormal crowd behaviors.

Although these approaches demonstrated promising performance in relatively controlled environments, their dependence on manually engineered features substantially limited their adaptability to real-world surveillance scenarios. Variations in illumination, camera motion, dynamic backgrounds, occlusions, and viewpoint changes often resulted in considerable performance degradation. Moreover, the lack of semantic understanding restricted their ability to recognize complex anomalies involving multiple objects and contextual interactions.

B. Deep Learning-Based Video Anomaly Detection

The introduction of deep learning significantly transformed the field of video anomaly detection by enabling automatic extraction of hierarchical spatial and temporal representations from raw video data. Convolutional Neural Networks (CNNs) improved spatial feature learning, while Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Three-Dimensional Convolutional Neural Networks (3D-CNNs) enhanced temporal modeling capabilities.

More recently, Transformer-based architectures have attracted considerable attention due to their ability to capture long-range temporal dependencies through self-attention mechanisms. Compared with traditional handcrafted approaches, deep learning models achieve superior performance by learning discriminative feature representations directly from large-scale datasets.

Despite these advantages, most deep learning methods rely on fully supervised training paradigms requiring frame-level or temporal annotations. Since anomalous events are inherently rare and difficult to label, constructing large-scale annotated datasets remains an expensive and labor-intensive process. Consequently, the high annotation cost has become a major obstacle to deploying fully supervised anomaly detection systems in practical surveillance environments.

C. Weakly Supervised Video Anomaly Detection

To alleviate the dependency on detailed annotations, Weakly Supervised Video Anomaly Detection (WSVAD) has emerged as a promising research paradigm. Unlike fully supervised approaches, WSVAD utilizes only video-level labels indicating whether an anomaly exists within a video sequence. This significantly reduces annotation costs while maintaining competitive detection performance.

The pioneering work of Sultani et al. introduced the Multiple Instance Learning (MIL) framework for anomaly detection, wherein a video is represented as a bag of temporal segments and anomaly localization is performed without frame-level supervision. This work established the foundation for subsequent WSVAD research.

Building upon the MIL paradigm, several studies have proposed enhanced learning strategies to address temporal ambiguity and noisy supervision. Robust Temporal Feature Magnitude Learning (RTFM) improved anomaly discrimination by maximizing feature magnitude differences between normal and abnormal segments. Subsequently, methods such as Multiple-Instance Self-Training (MIST) and Magnitude-Contrastive Glance-and-Focus Network (MGFN) incorporated pseudo-label refinement, temporal attention mechanisms, graph-based reasoning, and self-training strategies to further improve temporal localization accuracy and model generalization.

Although these approaches considerably reduce annotation requirements, they predominantly rely on visual information and remain susceptible to noise, temporal uncertainty, and domain shifts encountered in practical surveillance scenarios.

D. Vision-Language Models for Video Anomaly Detection

Recent advances in Vision-Language Models (VLMs) have introduced a new direction for anomaly detection by integrating semantic knowledge derived from visual-textual representations. Models such as Contrastive Language-Image Pretraining (CLIP) learn joint embedding spaces by leveraging large-scale image-text pairs, thereby enabling richer semantic understanding compared to conventional vision-only approaches.

By incorporating textual prompts and semantic descriptions, VLM-based methods improve the recognition of complex and previously unseen anomalies. Representative approaches such as VadCLIP adapt pretrained vision-language representations to the weakly supervised anomaly detection setting through prompt learning and semantic alignment mechanisms.

Compared with traditional WSVAD methods, VLM-based approaches offer stronger semantic reasoning capabilities and improved generalization across diverse anomaly categories. Nevertheless, these methods primarily exploit visual and textual information, limiting their ability to capture acoustic cues that frequently accompany anomalous events in surveillance environments.

E. Multimodal Video Anomaly Detection

The limitations of visual-only anomaly detection systems have motivated the development of multimodal learning frameworks that integrate visual, audio, and textual information. In many real-world scenarios, abnormal events generate distinctive acoustic signatures, including alarms, gunshots, explosions, vehicle collisions, machinery failures, and distress sounds. Such cues provide valuable complementary information that may not be evident from visual observations alone.

Recent multimodal anomaly detection frameworks employ various fusion strategies to exploit cross-modal interactions effectively. These include feature fusion, cross-modal attention, multimodal contrastive learning, knowledge distillation, and semantic alignment techniques. By leveraging complementary information across modalities, these methods demonstrate improved robustness under challenging conditions such as poor lighting, visual occlusion, motion blur, and environmental noise.

Furthermore, multimodal learning facilitates a transition from conventional anomaly detection toward anomaly

understanding, where the objective extends beyond identifying abnormal events to interpreting their contextual meaning and underlying causes.

F. Knowledge Transfer in Video Anomaly Detection

Knowledge transfer has emerged as an effective mechanism for improving multimodal anomaly detection systems. The fundamental objective of knowledge transfer is to leverage information learned in one modality, task, or domain to enhance performance in another.

Teacher-student learning, knowledge distillation, and cross-modal transfer learning have been widely adopted for this purpose. These approaches enable the transfer of semantic information between visual and audio modalities, thereby improving feature representation quality and mitigating the effects of noisy supervision.

Recent studies indicate that knowledge transfer not only enhances anomaly localization performance but also contributes to better domain adaptation and cross-environment generalization. Consequently, knowledge transfer has become a key component in the design of robust multimodal anomaly detection systems.

G. Semantic Alignment and Contrastive Learning

Semantic alignment aims to project heterogeneous multimodal features into a common latent representation space where semantically related information across modalities can be effectively correlated. This mechanism plays a crucial role in ensuring consistency among visual, audio, and textual representations.

To achieve semantic alignment, recent approaches extensively employ contrastive learning frameworks, cross-modal attention mechanisms, shared embedding spaces, and prototype-based learning strategies. These methods encourage semantically similar multimodal representations to remain close in the feature space while simultaneously increasing the separation between normal and anomalous events.

Effective semantic alignment improves anomaly localization performance, strengthens multimodal consistency, and mitigates modality imbalance. Recent studies integrating foundation models and multimodal representation learning further demonstrate that semantic alignment is a critical factor in achieving reliable and scalable multimodal anomaly detection.

H. Benchmark Datasets and Evaluation Metrics

Benchmark datasets play a vital role in the development and evaluation of video anomaly detection algorithms. Several publicly available datasets have become standard benchmarks for comparing anomaly detection methods.

Among these, UCF-Crime is one of the most widely used datasets for weakly supervised anomaly detection and contains diverse real-world anomalous events. XD-Violence extends anomaly detection to multimodal scenarios by providing synchronized audio-visual data. ShanghaiTech, Avenue, and UCSD Ped2 are commonly employed for evaluating unsupervised and supervised anomaly detection approaches in controlled surveillance environments.

Performance evaluation is typically conducted using metrics such as Area Under the Receiver Operating Characteristic Curve (AUC), Average Precision (AP), Frame-Level AUC, Precision, Recall, and F1-Score. For weakly supervised approaches, temporal localization accuracy and video-level classification performance are also frequently considered. The adoption of standardized datasets and evaluation metrics enables fair comparison among competing methods and facilitates continued progress in the field.

III. REVIEW METHODOLOGY

This review adopts a structured methodology to identify, classify, evaluate, and synthesize research on Multimodal Weakly Supervised Video Anomaly Detection (WSVAD). The objective is to examine the development of video anomaly detection techniques, advances in multimodal learning, and emerging approaches involving knowledge transfer and semantic alignment. The selected literature is analysed in terms of learning paradigms, input modalities, temporal modelling strategies, evaluation practices, and limitations.

A. Literature Search Strategy

Relevant publications were sought from major scientific literature sources, including IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect, Google Scholar, and arXiv. The search period was defined as 2018–2026 to cover the development of deep learning-based WSVAD and recent multimodal approaches. Earlier publications were considered when necessary to explain foundational methods or concepts.

The search used individual keywords and combinations of the following terms:

- Video Anomaly Detection (VAD)

- Weakly Supervised Video Anomaly Detection (WSVAD)
- Multiple Instance Learning (MIL)
- Vision–Language Models (VLMs)
- Audio–Visual Learning
- Multimodal Learning
- Knowledge Transfer
- Semantic Alignment
- Contrastive Learning
- Cross-Modal Attention
- Foundation Models

The search strategy covered both foundational studies and recent methods relevant to weak supervision, multimodal fusion, temporal localization, and semantic knowledge transfer.

B. Study Selection Criteria

The studies were assessed for relevance to the review scope, methodological contribution, availability of technical details, and evaluation evidence.

1) Inclusion Criteria

Studies were considered for inclusion if they:

- Addressed video anomaly detection or a closely related video-based event understanding task.
- Investigated weak supervision, multimodal learning, vision–language modelling, audio–visual learning, or relevant knowledge-transfer techniques.
- Presented a methodological contribution or a meaningful analysis of anomaly detection.
- Were published in peer-reviewed journals or conference proceedings, or were relevant preprints with identifiable publication records.
- Provided sufficient experimental or methodological information to support critical analysis.

2) Exclusion Criteria

Studies were excluded if they:

- Focused exclusively on image anomaly detection without temporal modelling.
- Were unrelated to surveillance video or event anomaly analysis.
- Lacked sufficient technical description or experimental evaluation.
- Duplicated previously reported work without providing a distinct contribution.
- Were not available in English.

These criteria were intended to maintain the relevance and technical usefulness of the reviewed literature. Peer-reviewed studies were prioritized, while preprints were considered where they provided relevant developments not adequately represented by published work.

C. Classification of Reviewed Studies

The selected literature was organized into six categories according to its learning paradigm, representation strategy, and use of multimodal information.

1) Traditional feature-based approaches: These methods use handcrafted descriptors, statistical learning, sparse reconstruction, or other manually designed representations to characterize normal and abnormal behaviour.

2) Deep learning-based approaches: These approaches learn representations from video data using architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), 3D CNNs, and Transformers.

3) Weakly supervised video anomaly detection: This category includes multiple instance learning, temporal feature magnitude learning, self-training, and pseudo-label refinement. The principal objective is to learn segment-level anomaly scores from video-level labels.

4) Vision–language-based approaches: These methods incorporate pretrained vision–language representations, including CLIP-based features, prompt learning, and semantic guidance, to support anomaly recognition and temporal localization.

5) Multimodal learning approaches: These studies investigate the integration of audio and visual information

through early, intermediate, or late fusion, cross-modal attention, and multimodal contrastive learning.

6) Knowledge transfer and semantic alignment: These methods investigate knowledge distillation, cross-modal transfer learning, shared representation learning, and semantic alignment to transfer information between pretrained models or modalities.

This classification supports the comparison of methodological developments, from traditional feature engineering and visual representation learning to multimodal and semantically guided anomaly detection.

D. Evaluation Framework

Each selected study was examined using a common set of analytical criteria:

- Learning paradigm and supervision requirements.
- Input modalities and data representation.
- Feature extraction strategy and network architecture.
- Temporal modelling and anomaly localization mechanisms.
- Modality fusion and knowledge-transfer strategy.
- Benchmark datasets and evaluation protocols.
- Reported performance and computational requirements, where available.
- Principal strengths, limitations, and applicability.

Particular attention was given to whether multimodal components provide complementary evidence, improve temporal localization, or support generalization beyond the training environment. The analysis also considered whether reported improvements could be attributed to the multimodal strategy itself or might reflect differences in model architecture, pretrained features, datasets, or experimental protocols.

E. Benchmark Datasets and Performance Metrics

The review considers commonly used video anomaly detection benchmarks, including UCF-Crime, XD-Violence, ShanghaiTech, Avenue, and UCSD Ped2. These datasets differ in scene characteristics, anomaly categories, annotation practices, and available modalities. In particular, XD-Violence supports audio–visual violence detection, whereas the other

datasets listed here are primarily used for visual anomaly detection.

The following metrics are considered where reported by the original studies:

- Area Under the Receiver Operating Characteristic Curve (AUC): Measures the ability to distinguish anomalous from normal observations across decision thresholds.
- Average Precision (AP): Summarizes precision–recall performance.
- Precision and recall: Measure the proportion of positive predictions that are correct and the proportion of actual positive cases detected, respectively.
- F1-score: Represents the harmonic mean of precision and recall at a specified threshold.
- Frame-level AUC: Evaluates anomaly-score discrimination at the frame level.
- Video-level classification accuracy: Measures the proportion of videos classified correctly.

The precise definition and evaluation unit of each metric must be retained when reporting results. Metrics calculated at different levels, such as frames and videos, should not be treated as interchangeable. Similarly, performance figures from different datasets or experimental protocols should not be interpreted as a direct ranking unless the comparison conditions are sufficiently consistent.

F. Research Trend Analysis

In addition to comparing individual methods, the reviewed literature was examined to identify broader methodological trends:

- The transition from handcrafted features to learned spatiotemporal representations.
- The development of weakly supervised learning to reduce the need for frame-level annotations.
- The integration of audio, visual, and textual information for richer event representations.
- The adoption of pretrained foundation models and vision–language representations.

- The use of knowledge transfer and semantic alignment to support representation learning.
- The increasing importance of robustness, explainability, computational efficiency, and real-time deployment.

These trends provide the basis for identifying unresolved technical challenges and formulating future research directions.

G. Methodological Framework of the Review

Fig. 1 illustrates the overall methodology adopted in this review.



Fig. 1. Methodological framework of the review.

IV. COMPARATIVE ANALYSIS

The evolution of Video Anomaly Detection (VAD) has led to the development of diverse methodologies, ranging from traditional handcrafted feature-based techniques to advanced multimodal learning frameworks. These approaches differ substantially in their supervision requirements, feature representation capabilities, computational complexity, and anomaly localization performance. This section provides a comparative analysis of representative methods and highlights major technological trends observed in the literature.

A. Comparison of Learning Paradigms

Traditional anomaly detection methods rely on handcrafted descriptors, including optical flow, Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), motion trajectories, and sparse reconstruction features. These approaches generally require limited computational resources and perform adequately in constrained environments with relatively simple scene dynamics. However, their dependence on manually engineered features restricts adaptability to complex surveillance scenarios involving viewpoint variations, occlusions, illumination changes, and dynamic backgrounds. Furthermore, their inability to capture semantic information limits the detection of context-dependent anomalous events.

Deep learning-based approaches overcome many of these limitations through automatic feature learning. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, 3D Convolutional Neural Networks (3D-CNNs), and Transformer architectures are capable of learning discriminative spatiotemporal representations directly from video data.

These models generally achieve superior anomaly detection performance compared with traditional approaches. Nevertheless, they typically depend on fully supervised training strategies that require large amounts of finely annotated temporal data, resulting in substantial annotation costs.

Weakly Supervised Video Anomaly Detection (WSVAD) addresses this limitation by utilizing video-level labels rather than frame-level annotations. Multiple Instance Learning (MIL)-based techniques, including MIL-Rank, RTFM, MIST, and MGFN, have demonstrated that effective temporal anomaly localization can be achieved with significantly reduced annotation effort. These approaches improve scalability and practical applicability while maintaining competitive detection performance. However, most WSVAD frameworks continue to rely primarily on visual information and remain susceptible to temporal ambiguity and noisy supervision.

Vision-Language Models (VLMs) introduce semantic understanding into the anomaly detection process by leveraging large-scale visual-textual pretraining. Methods such as VadCLIP utilize pretrained CLIP representations and prompt learning to establish semantic relationships between video content and textual descriptions.

Compared with conventional WSVAD approaches, VLM-based methods exhibit improved generalization to unseen

anomaly categories. However, the absence of explicit acoustic information limits their ability to detect anomalies characterized by distinctive audio cues.

Recent multimodal approaches integrate visual, audio, and textual information to exploit complementary evidence from different modalities. Through mechanisms such as multimodal fusion, cross-modal attention, semantic alignment, and knowledge transfer, these methods achieve improved anomaly localization and robustness under challenging conditions. Although multimodal frameworks represent the current state of the art, they often require synchronized multimodal data and involve higher computational complexity than unimodal approaches.

B. Analysis of Multimodal Fusion Strategies

A critical aspect of multimodal anomaly detection is the effective integration of heterogeneous information sources. Existing multimodal approaches generally employ one of four fusion strategies.

1) Early Fusion

Early fusion combines features from different modalities before high-level representation learning. This strategy enables the model to learn joint multimodal representations from the initial stages of processing. While computationally straightforward, early fusion may not adequately preserve modality-specific characteristics and is often sensitive to noisy or incomplete inputs.

2) Late Fusion

Late fusion processes each modality independently and combines their outputs at the decision level. This approach preserves modality-specific information and provides greater flexibility when modalities exhibit different statistical properties. However, interactions between modalities are not explicitly modelled, limiting the exploitation of complementary information.

3) Intermediate Fusion

Intermediate fusion integrates information at intermediate network layers, allowing gradual interaction between modalities while preserving modality-specific representations. Compared with early and late fusion, intermediate fusion generally provides a better balance between representational richness and robustness.

4) Cross-Modal Attention

Cross-modal attention has emerged as one of the most effective multimodal fusion strategies. By dynamically assigning attention weights across modalities, these methods enable the model to focus on the most informative cues for anomaly detection. Recent audio-visual frameworks demonstrate that cross-modal attention significantly improves anomaly localization performance, particularly in scenarios involving occlusions, poor visibility, or ambiguous visual evidence.

Third, multimodal learning is gradually replacing unimodal visual analysis. Audio, visual, and textual information provide complementary evidence that enhances robustness across diverse surveillance conditions.

Fourth, knowledge transfer and semantic alignment techniques are becoming increasingly important for addressing modality imbalance, domain shifts, and weak supervision challenges.

TABLE I
COMPARISON OF REPRESENTATIVE VIDEO ANOMALY DETECTION METHODS

Method	Year	Paradigm	Modalities	Strengths	Limitations
MIL-Rank [1]	2018	WSVAD	Visual	Introduces a MIL-based ranking framework for weakly supervised anomaly detection.	Limited localization accuracy under video-level supervision.
RTFM [2]	2021	WSVAD	Visual	Robust feature magnitude learning to emphasize discriminative segments.	Relies on visual representations, without using additional modalities.
MIST [3]	2021	WSVAD	Visual	Uses self-training and pseudo-label refinement to improve temporal localization.	Performance can be affected by pseudo-label quality and error propagation.
MGFN [4]	2023	WSVAD	Visual	Uses magnitude-contrastive learning to extract discriminative temporal features.	Does not explicitly integrate audio or language modalities.
VadCLIP [5]	2024	Vision-Language	Visual + Text	Adapts pretrained vision-language knowledge to weakly supervised anomaly detection.	Depends on semantic representations and adaptation strategy; lacks audio information.
MACIL-SD [6]	2022	Audio-Visual	Audio + Visual	Uses modality-aware contrastive instance learning and self-distillation.	Increased computational complexity from multimodal learning.
AVCL [7]	2025	Audio-Visual	Audio + Visual	Improves audio-visual representation learning through cross-modal collaboration.	Dependence on synchronized audio-visual data and dataset-specific performance.
ASTRA [8]	2026	Multimodal	Audio + Visual	Uses knowledge transfer and latent alignment for enhanced anomaly detection.	High training cost and computational requirements.

C. Comparative Analysis of Representative Methods

Table I summarizes representative approaches in the development of video anomaly detection, highlighting their learning paradigms, input modalities, principal strengths, and limitations.

The comparison reveals a clear shift from visual-only anomaly detection towards multimodal representation learning and semantic understanding. Recent approaches increasingly emphasize complementary modality utilization, knowledge transfer, and semantically meaningful feature representations.

D. Current Research Trends

Several important trends emerge from the reviewed literature.

First, there is a transition from anomaly detection toward anomaly understanding. Modern systems increasingly seek not only to determine whether an anomaly exists but also to explain its semantic context and underlying cause.

Second, pretrained foundation models and Vision-Language Models are becoming central components of anomaly detection frameworks. Their ability to leverage large-scale semantic knowledge improves generalization and reduces dependence on task-specific labelled data.

Finally, explainability, efficiency, and real-time deployment are emerging as critical design considerations. Future anomaly detection systems are expected to provide interpretable results while maintaining computational efficiency suitable for edge and large-scale surveillance environments.

E. Discussion

The comparative analysis demonstrates that multimodal learning represents the most promising direction for future WSVAD research. By integrating complementary information sources and exploiting semantic knowledge acquired through foundation models, modern frameworks substantially improve anomaly localization performance and generalization capability. Nevertheless, several challenges remain unresolved, including limited multimodal benchmark datasets, computational scalability, missing-modal scenarios, and insufficient model interpretability. Addressing these challenges is essential for the development of robust and trustworthy anomaly detection systems capable of operating in complex real-world environments.

V. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Despite significant progress in Weakly Supervised Video Anomaly Detection (WSVAD), several technical and practical

challenges continue to limit the deployment of robust multimodal anomaly detection systems in real-world surveillance environments. Addressing these limitations is essential for improving reliability, scalability, interpretability, and adaptability. This section discusses the major challenges identified from the literature and highlights promising future research directions.

A. Major Research Challenges

1) Limited Availability of Multimodal Datasets

The effectiveness of multimodal learning depends heavily on the availability of large-scale datasets containing synchronized visual, audio, and semantic information. However, most benchmark datasets used in video anomaly detection, such as UCF-Crime, ShanghaiTech, Avenue, and UCSD Ped2, primarily provide visual data with limited multimodal annotations.

Although XD-Violence contains synchronized audio-visual information, the number of publicly available multimodal datasets remains relatively small. The lack of diverse benchmark datasets limits the development, evaluation, and fair comparison of multimodal anomaly detection methods across different application domains.

2) Modality Imbalance and Missing Information

Different modalities often contribute unequally to anomaly detection. In some scenarios, visual cues may provide strong evidence of anomalous behaviour, whereas in others, audio signals may be more informative. This imbalance can negatively affect model learning and lead to biased representations.

Furthermore, real-world surveillance systems frequently encounter missing or corrupted modalities due to sensor failures, environmental noise, poor lighting conditions, or communication interruptions. Designing robust models capable of maintaining detection performance under incomplete multimodal information remains a significant challenge.

3) Weak Supervision and Noisy Labels

Although weak supervision substantially reduces annotation costs, video-level labels provide only coarse information regarding anomaly occurrence. As a result,

models must infer anomalous temporal segments without explicit frame-level supervision.

This weak labelling strategy often introduces temporal ambiguity and noisy supervision, leading to inaccurate anomaly localization and reduced detection reliability. Developing more effective mechanisms for pseudo-label generation, uncertainty estimation, and noise-robust learning remains an important research problem.

4) High Computational Complexity

Modern multimodal frameworks frequently employ large neural architectures, multiple feature extractors, cross-modal attention modules, and pretrained foundation models. While these components improve representation quality and detection accuracy, they substantially increase computational and memory requirements.

Such complexity restricts deployment in resource-constrained environments, including edge devices, embedded surveillance systems, and real-time monitoring applications. Achieving a balance between accuracy and computational efficiency remains a critical requirement for practical deployment.

5) Generalization Across Domains

Many anomaly detection models perform well on benchmark datasets but experience noticeable performance degradation when applied to unseen environments. Variations in camera perspectives, scene characteristics, crowd density, weather conditions, and anomaly types introduce significant domain shifts.

The ability to generalize across different surveillance settings remains limited, particularly for weakly supervised models trained on relatively narrow data distributions. Improving cross-domain robustness is therefore a major research objective.

6) Limited Explainability and Interpretability

Most deep learning-based anomaly detection systems function as black-box models, providing anomaly scores without explaining the underlying reasoning behind their decisions.

In safety-critical applications such as public surveillance, transportation, healthcare, and industrial monitoring, explainability is essential for building user trust and supporting decision-making. The lack of interpretable

anomaly explanations remains a major barrier to widespread adoption.

B. Emerging Research Directions

1) Foundation Models for Video Anomaly Detection

Recent advances in foundation models have demonstrated remarkable capabilities in representation learning across diverse tasks and domains. Foundation models pretrained on large-scale visual, audio, and textual data possess extensive semantic knowledge that can be transferred to anomaly detection applications.

Future research may focus on adapting multimodal foundation models to weakly supervised settings, thereby reducing dependence on task-specific annotations while improving anomaly understanding and generalization.

2) Multimodal Large Language Models (MLLMs)

The emergence of Multimodal Large Language Models (MLLMs) presents new opportunities for anomaly understanding. By integrating visual, audio, and textual information with advanced reasoning capabilities, MLLMs can potentially move beyond anomaly detection toward anomaly interpretation.

Such systems may not only identify abnormal events but also generate human-readable explanations describing the nature, cause, and severity of anomalies. This capability could significantly enhance situational awareness in intelligent surveillance systems.

3) Self-Supervised and Continual Learning

Self-supervised learning has gained increasing attention as a means of reducing dependence on manually annotated datasets. By exploiting intrinsic relationships within multimodal data, self-supervised approaches can learn useful feature representations without extensive labelling.

Continual learning represents another promising direction, enabling anomaly detection systems to adapt incrementally to evolving environments without requiring complete retraining. Combining self-supervised and continual learning may significantly improve long-term adaptability and scalability.

4) Advanced Cross-Modal Knowledge Transfer

Knowledge transfer has demonstrated substantial benefits for representation learning and domain adaptation. Future research may investigate more effective mechanisms for

transferring information between visual, audio, and textual modalities.

Advanced teacher-student architectures, cross-modal distillation frameworks, and shared latent representation learning strategies have the potential to improve multimodal consistency while reducing computational complexity.

5) Adaptive Semantic Alignment

Current semantic alignment methods generally assume relatively stable relationships among modalities. However, real-world surveillance environments often involve changing contexts, noisy observations, and dynamic event structures.

Future anomaly detection systems should incorporate adaptive semantic alignment mechanisms capable of continuously adjusting cross-modal relationships according to environmental conditions. Such adaptability could improve robustness and anomaly localization accuracy.

6) Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) has emerged as a critical research area for trustworthy machine learning systems. Integrating explainability into anomaly detection frameworks can provide visual, auditory, or textual evidence supporting model decisions.

Future multimodal systems may utilize attention visualization, causal reasoning, saliency analysis, and language-based explanations to enhance transparency and user confidence in anomaly detection outcomes.

7) Federated and Privacy-Preserving Learning

Surveillance applications often involve sensitive data, raising concerns regarding privacy, confidentiality, and regulatory compliance. Federated learning offers a promising solution by allowing models to learn collaboratively across distributed devices without centralized data sharing.

Combining federated learning with multimodal anomaly detection may enable privacy-preserving surveillance systems while maintaining competitive detection performance.

8) Edge-Based Intelligent Surveillance

The growing deployment of Internet of Things (IoT) devices and edge computing platforms is driving interest in lightweight anomaly detection models capable of operating at the network edge.

Research on model compression, knowledge distillation, efficient transformers, and edge-optimized multimodal architectures will play a crucial role in enabling real-time intelligent surveillance for smart cities, transportation systems, and industrial applications.

C. Summary of Future Directions

The future of multimodal weakly supervised video anomaly detection is expected to be shaped by advancements in foundation models, multimodal large language models, self-supervised learning, knowledge transfer, semantic alignment, and explainable artificial intelligence. These technologies have the potential to transform anomaly detection systems from simple event recognition tools into intelligent anomaly understanding frameworks capable of reasoning about complex events across multiple modalities.

Addressing challenges related to dataset availability, modality imbalance, computational efficiency, domain generalization, and interpretability will be essential for developing scalable, reliable, and trustworthy anomaly detection systems suitable for real-world deployment.

VI. CONCLUSION

This review presented a comprehensive overview of recent advances in Multimodal Weakly Supervised Video Anomaly Detection (WSVAD). The evolution of the field from traditional handcrafted feature-based methods to deep learning, weakly supervised learning, vision-language models, and multimodal frameworks was examined. The analysis showed that multimodal approaches, which integrate visual, audio, and semantic information, offer improved anomaly localization, robustness, and generalization compared with unimodal methods.

Despite these advancements, challenges such as limited multimodal datasets, noisy weak supervision, modality imbalance, computational complexity, and limited interpretability continue to restrict practical deployment. Emerging research directions, including foundation models, multimodal large language models, self-supervised learning, knowledge transfer, semantic alignment, and explainable AI, provide promising opportunities to address these limitations.

Overall, multimodal WSVAD represents a promising direction for next-generation intelligent surveillance systems. Continued research toward efficient, scalable, and interpretable multimodal frameworks is expected to enable more reliable anomaly detection and understanding across diverse real-world applications.

VII. REFERENCES

- [1] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6479-6488.
- [2] 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)*, 2021, pp. 4975-4986.
- [3] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14009-14018.
- [4] Y. Chen, Z. Liu, B. Zhang, W. T. 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.*, vol. 37, no. 1, pp. 387-395, 2023.
- [5] J. Yu, J. Liu, Y. Cheng, R. Feng, and Y. Zhang, "Modality-Aware Contrastive Instance Learning With Self-Distillation for Weakly-Supervised Audio-Visual Violence Detection," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2022.
- [6] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748-8763.
- [7] 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," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 6, pp. 6074-6082, 2024.
- [8] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep Learning for Anomaly Detection: A Review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1-38, 2021.
- [9] R. Mehran, A. Oyama, and M. Shah, "Abnormal Crowd Behavior Detection Using Social Force Model," in *CVPR*, 2009, pp. 935-942.
- [10] Y. Cong, J. Yuan, and J. Liu, "Sparse Reconstruction Cost for Abnormal Event Detection," in *CVPR*, 2011, pp. 3449-3456.
- [11] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in *ICCV*, 2013, pp. 2720-2727.
- [12] W. Luo, W. Liu, and S. Gao, "A Revisit of Sparse Coding-Based Anomaly Detection in Stacked RNN Framework," in *ICCV*, 2017.
- [13] W. Luo, W. Liu, and S. Gao, "Remembering History With Convolutional LSTM for Anomaly Detection," in *ICME*, 2017, pp. 439-444.
- [14] W. Liu, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection: A New Baseline," in *CVPR*, 2018.
- [15] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training Adversarial Discriminators for Cross-Channel Abnormal Event Detection in Crowds," in *WACV*, 2019, pp. 1896-1904.
- [16] S. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Deep-Anomaly: Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 88-92, 2017.

- [17] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in *CVPR*, 2017, pp. 4724-4733.
- [18] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” in *NeurIPS*, 2014.
- [19] A. Dosovitskiy *et al.*, “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in *ICLR*, 2021.
- [20] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, “Video Action Transformer Network,” in *CVPR*, 2019, pp. 244-253.
- [21] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting the Unreasonable Effectiveness of Data in Video Recognition,” in *ICCV*, 2017.
- [22] S. Hershey *et al.*, “CNN Architectures for Large-Scale Audio Classification,” in *ICASSP*, 2017.
- [23] A. Alayrac *et al.*, “Self-Supervised Multimodal Versatile Networks,” in *NeurIPS*, vol. 33, 2020.
- [24] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, “HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips,” in *ICCV*, 2019, pp. 2630-2640.
- [25] J. Meng, H. Tian, G. Lin, J.-F. Hu, and W.-S. Zheng, “Audio-Visual Collaborative Learning for Weakly Supervised Video Anomaly Detection,” *IEEE Trans. Multimedia*, 2025.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *CVPR*, 2016, pp. 770-778.
- [27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification With Deep Convolutional Neural Networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84-90, 2017.
- [28] F. Chollet, “Xception: Deep Learning With Depthwise Separable Convolutions,” in *CVPR*, 2017.
- [29] A. Howard *et al.*, “Searching for MobileNetV3,” in *ICCV*, 2019.
- [30] A. Vaswani *et al.*, “Attention Is All You Need,” in *NeurIPS*, vol. 30, 2017.
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL-HLT*, 2019, pp. 4171-4186.
- [32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” OpenAI Technical Report, 2019.
- [33] M. Zaheer, A. Mahmood, M. H. Khan, M. Astrid, and S.-I. Lee, “CLAWS: Clustering Assisted Weakly Supervised Learning With Normalcy Suppression for Anomalous Event Detection,” in *ECCV*, 2022.
- [34] D. Purwanto, Y.-T. Chen, and W.-H. Fang, “Dance With Self-Attention: A New Look of Conditional Random Fields on Video Anomaly Detection,” *IEEE Trans. Multimedia*, vol. 24, pp. 2991-3004, 2022.
- [35] H. Lv, Z. Yue, Q. Sun, B. Luo, and F. Shen, “BN-WVAD: Bag-Level Normalcy-Guided Weakly Supervised Video Anomaly Detection,” in *ICME*, 2023.
- [36] B. Xu, Y. Wang, J. Wang, Y. Long, and Z. Wang, “Self-Guided Learning for Weakly Supervised Video Anomaly Detection,” in *AAAI*, 2023.
- [37] J. Georgescu, A. Barbalau, R. T. Ionescu, and M. Shah, “Anomaly Detection in Video via Self-Supervised and Multi-Task Learning,” in *CVPR*, 2021.
- [38] T. Wang, H. Snoussi, and F. Smach, “Video Anomaly Detection Based on Deep Prediction Networks,” *Pattern Recognit. Lett.*, vol. 132, pp. 345-352, 2020.
- [39] UCF-Crime Dataset Paper (Sultani *et al.*, CVPR 2018).
- [40] XD-Violence Dataset: J. Wu *et al.*, “Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision,” in *ECCV*, 2020.
- [41] ShanghaiTech Campus Dataset: W. Luo *et al.*, 2017.
- [42] H. Zhong, J. Li, Z. Zhang, Y. Jiao, and H. Shan, “Open-Vocabulary Video Anomaly Detection,” 2024.
- [43] H. Zhong, J. Li, Z. Zhang, and H. Shan, “Uni-VAD: A Unified Framework for Weakly Supervised Video Anomaly Detection,” 2024.
- [44] M. Kim, J. Lee, C. Bang, and M. Cho, “CLUE-VAD: Structured Semantic Clues for Explainable Video Anomaly Detection,” 2026.
- [45] Y. Kim, T. Kim, W. Lee, and M. Cho, “ASTRA: Joint Latent Representation and Knowledge Transfer for Multimodal Anomaly Detection,” *IEEE Access*, 2026.