Metric-Wise Comparative Analysis of Hybrid CNN–SRU/LSTM and Lightweight CNN–MIL Frameworks for Deployment-Oriented Video Anomaly Detection

Rajat Gupta*[1,2], Nidhi Tyagi[1]
E-mail: rajjatgupta2@gmail.com, nidhi.tyagi@shobhitunversity.ac.in
[1]School of Computational Sciences and Engineering, Shobhit Institute of Engineering and Technology, Meerut, India
[2]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
*Corresponding author

Keywords: Video anomaly detection, comparative analysis, hybrid deep learning frameworks, multiple instance learning, temporal modeling, evaluation metrics, cross-domain robustness, deployment-oriented analysis

Received: February 17, 2026

Video anomaly detection is a critical component of intelligent surveillance systems, where detection accuracy, temporal stability, computational efficiency, and real-world deployment feasibility must be jointly considered. Existing studies frequently rely on ROC–AUC as the primary evaluation metric, providing limited insight into practical system performance. This study presents a structured metric-wise comparative analysis of hybrid CNN–SRU/LSTM architectures and lightweight CNN-based multiple instance learning (MIL) frameworks, based on systematically collected benchmark results from datasets such as UCF-Crime and ShanghaiTech. The analysis follows a literature-driven methodology and evaluates models across multiple dimensions, including AUC, false alarm rate (FAR), temporal stability, inference speed (FPS), computational footprint, and calibration reliability using expected calibration error (ECE). Deployment-oriented factors such as latency–performance trade-offs, cross-domain robustness, and scalability under limited labeled data are also examined. Results indicate that hybrid CNN–SRU/LSTM frameworks achieve approximately 85.9% AUC across benchmark datasets with strong temporal consistency, while CNN–MIL approaches maintain competitive accuracy (≈82–84.7%) with significantly higher efficiency (up to 72 FPS) and improved calibration (ECE reduced from ~0.17 to ~0.10). Transformer-based and vision–language models achieve slightly higher accuracy (>86% AUC) but operate at substantially lower frame rates (<12 FPS) and higher memory requirements (>800 MB). These findings highlight that marginal accuracy gains often incur substantial computational cost, emphasizing multi-metric evaluation and hardware-aware model selection for practical video anomaly detection systems.

Povzetek: Študija primerjalno ovrednoti modele za zaznavanje anomalij v videu po več praktičnih metrikah, ne le AUC, ter pokaže kompromis med natančnejšimi, a zahtevnejšimi Transformerji in učinkovitejšimi CNN–MIL/CNN–SRU-LSTM pristopi za realno uvedbo.

1 Introduction

Video anomaly detection has become a fundamental component of intelligent surveillance systems, enabling the automatic identification of abnormal events such as violence, accidents, theft, and unusual crowd behavior in untrimmed video streams. With the rapid deployment of large-scale camera networks across smart cities, transportation hubs, healthcare facilities, and industrial environments, automated anomaly detection systems are increasingly required to enhance situational awareness while reducing reliance on manual monitoring. In real-world operational settings, effective anomaly detection frameworks must not only achieve high detection accuracy but also maintain temporal stability, low false alarm rates, computational efficiency, and scalability under real-time constraints [5,6,14].

Despite significant progress in deep learning-based approaches, video anomaly detection remains challenging due to the rarity, diversity, and context-dependent nature of anomalous events. These events often lack precise temporal boundaries, making frame-level annotation expensive and difficult to obtain at scale. To address this limitation, many recent approaches adopt weakly supervised or semi-supervised learning paradigms, where models are trained using coarse video-level labels or only normal samples [8,10]. Furthermore, real-world deployment introduces domain shifts caused by variations in illumination, camera viewpoints, scene layouts, and motion dynamics. Such variations can significantly degrade model performance when applied to unseen environments, highlighting the need for robust and generalizable detection frameworks [3,26].

To address these challenges, recent research has explored multiple architectural paradigms. Hybrid

frameworks combining convolutional neural networks (CNNs) with recurrent architectures such as long short-term memory (LSTM) and simple recurrent units (SRU) have demonstrated improved temporal modeling by capturing sequential dependencies in evolving anomalies [1,7]. In parallel, lightweight CNN-based multiple instance learning (MIL) approaches have gained attention due to their ability to operate under weak supervision while maintaining computational efficiency and scalability [9,11]. More recently, transformer-based architectures and vision–language models have emerged as promising alternatives, leveraging global attention mechanisms and multimodal representations to improve contextual understanding and detection accuracy [29,38].

Although these approaches achieve competitive performance on benchmark datasets, they differ substantially in computational complexity, latency, temporal modeling capability, and robustness to domain variation. These differences make it difficult to identify the most suitable framework for practical deployment scenarios, where hardware constraints, inference speed, and reliability play a critical role [26,35].

Another important limitation lies in the current evaluation landscape. Most studies evaluate models independently and emphasize ROC–AUC as the primary performance metric. While ROC–AUC provides a useful measure of ranking performance, it does not fully capture deployment-critical characteristics such as false alarm behavior, temporal consistency, calibration reliability, and inference latency [19,20]. Moreover, heterogeneous experimental protocols—including differences in datasets, preprocessing pipelines, backbone architectures, and hardware configurations—further complicate direct comparison across studies and reduce the interpretability of reported results [21,25].

To address these limitations, this study presents a structured, metric-wise comparative analysis of hybrid CNN–SRU/LSTM frameworks and lightweight CNN–MIL architectures. Rather than reimplementing models, the study synthesizes systematically selected results from representative research works using predefined inclusion criteria. This approach enables consistent framework-level comparison under heterogeneous experimental settings. The comparative analysis evaluates models across multiple deployment-relevant dimensions, including detection accuracy, false alarm rate, temporal stability, computational efficiency, calibration reliability, cross-domain robustness, and latency–performance trade-offs. This multi-dimensional evaluation provides practical insights for selecting suitable anomaly detection frameworks for real-world surveillance applications [32,36].

1.1 Novelty of the study

The novelty of this work lies in its deployment-oriented, multi-dimensional evaluation framework that extends beyond conventional accuracy-based comparisons. Unlike traditional studies that focus primarily on ROC–AUC performance, this work integrates multiple operational metrics, including false alarm rate, temporal stability, inference speed, computational footprint, and calibration error. This comprehensive evaluation provides a more realistic understanding of performance trade-offs relevant to practical deployment scenarios [19,37].

Additionally, the study synthesizes results from heterogeneous experimental environments, enabling comparative insights into scalability, robustness, and efficiency–accuracy trade-offs. This structured comparative approach differs from conventional survey papers, which typically provide descriptive overviews without systematic multi-metric evaluation. By linking architectural characteristics with deployment constraints, the proposed analysis provides actionable guidance for selecting anomaly detection frameworks in real-world applications [3,26,29].

1.2 Research questions

The study is guided by the following research questions:

- **RQ1:** What are the performance trade-offs between hybrid CNN–SRU/LSTM and lightweight CNN–MIL frameworks in terms of detection accuracy, reliability, and temporal modeling capability?
- **RQ2:** How do these frameworks compare in computational efficiency and scalability for real-time and edge-based deployment scenarios?
- **RQ3:** How do calibration reliability and cross-domain robustness affect the practical performance of video anomaly detection systems?
- **RQ4:** Which evaluation metrics provide the most informative basis for deployment-oriented framework selection beyond ROC–AUC?

1.3 Contributions

In response to the above research questions, the main contributions of this work are summarized as follows:

- A structured metric-wise comparative analysis of hybrid CNN–SRU/LSTM and lightweight CNN–MIL frameworks for video anomaly detection.
- A multi-dimensional evaluation framework incorporating detection accuracy, reliability, temporal behavior, computational efficiency, calibration performance, and cross-domain robustness.
- A systematic synthesis of reported results from representative studies, enabling framework-level comparison without model reimplementation.
- Deployment-oriented insights for framework selection under practical constraints such as real-time processing, hardware limitations, and domain variability [32,38].

Unlike conventional survey studies, this work provides a unified, deployment-oriented perspective that explicitly connects architectural design choices with system-level performance trade-offs. The findings aim to support informed decision-making in real-world surveillance environments, where efficiency, robustness, and reliability are as critical as detection accuracy.

In practical surveillance environments, anomaly detection systems must balance accuracy, computational efficiency, and robustness across diverse operational conditions. Therefore, deployment-oriented evaluation considering multiple performance metrics is essential for selecting suitable frameworks for real-world applications.

2 Background and related work

Video anomaly detection aims to identify events that deviate from learned normal behavior in untrimmed video streams. With the rapid expansion of surveillance systems across public infrastructure, transportation hubs, and industrial environments, automated anomaly detection has become essential for improving monitoring efficiency and reducing reliance on manual supervision [5,6]. However, anomalous events are inherently rare, diverse, and context-dependent, making precise annotation expensive and limiting the availability of labeled datasets [14]. As a result, many practical anomaly detection systems adopt unsupervised, semi-supervised, or weakly supervised learning paradigms to address limited annotation availability [8,10].

Over the past decade, research in video anomaly detection has evolved across multiple methodological directions, including reconstruction-based learning, temporal modeling frameworks, weakly supervised learning approaches, graph-based modeling, and transformer-based architectures [3,26]. These approaches differ in their temporal modeling capability, computational complexity, and deployment feasibility, leading to varying performance trade-offs in real-world surveillance applications.

2.1 Reconstruction and prediction-based methods

Early deep learning approaches focused on modeling normal behavior using reconstruction-based learning. Autoencoder-based architectures learn compact representations of normal spatiotemporal patterns and detect anomalies based on reconstruction errors [4,15]. Memory-augmented reconstruction models further improved detection reliability by storing representative normal patterns and identifying deviations during inference [16].

Prediction-based methods extend this paradigm by forecasting future frames or motion patterns. Anomalies are detected when predicted frames deviate significantly from actual observations [29,31]. These approaches enable unsupervised learning and reduce dependency on anomalous samples. However, reconstruction and prediction-based methods often struggle to detect high-level semantic anomalies, particularly in complex scenes with dynamic backgrounds or multiple interacting objects [18].

2.2 Temporal Modeling with CNN–RNN Architectures

To improve temporal modeling, hybrid architectures combining convolutional neural networks with recurrent networks have been widely explored. Long short-term memory (LSTM) networks enable modeling of sequential dependencies and improve detection of anomalies that evolve over time [1,7]. Similarly, simple recurrent unit (SRU)-based architectures enhance temporal consistency while reducing training complexity compared to traditional recurrent models.

These temporal modeling frameworks improve anomaly score stability and enable detection of long-duration anomalies such as accidents, violent behavior, and abnormal crowd movement. However, recurrent processing introduces additional computational overhead and latency, which may limit deployment in real-time surveillance environments.

2.3 Weakly supervised and MIL-based frameworks

Weakly supervised anomaly detection has emerged as a practical solution to annotation scarcity. Multiple instance learning (MIL) frameworks treat videos as bags of segments and identify anomalous segments using video-level labels [8,9]. This formulation enables scalable training across large surveillance datasets while reducing annotation costs.

Recent studies have incorporated attention mechanisms to improve segment-level anomaly localization and enhance representation quality [10,13]. Feature enhancement modules further improve anomaly separability and detection robustness [12]. In addition, lightweight convolutional neural network-based MIL frameworks have been proposed to improve computational efficiency and reduce memory requirements, making them suitable for real-time deployment scenarios [11,24].

Despite these advantages, MIL-based methods may face challenges in modeling long-range temporal dependencies due to the absence of explicit sequential modeling. Furthermore, segment-level aggregation can introduce sensitivity to short-duration noise or abrupt scene changes.

2.4 Graph-based and feature enhancement approaches

Graph-based approaches have recently been introduced to model relationships between video segments and capture contextual dependencies. These frameworks improve anomaly detection performance by incorporating relational reasoning across temporal regions and interacting objects [26]. Graph-based architectures are particularly effective in complex scenes involving multiple entities and interactions.

Feature enhancement frameworks further improve anomaly separability by refining learned representations [12]. Multi-stage architectures also enhance detection reliability by progressively refining anomaly scores.

However, these approaches typically increase computational complexity and memory requirements, which may limit their applicability in resource-constrained environments.

2.5 Transformer and attention-based models

Transformer-based architectures have gained attention due to their ability to model long-range dependencies using self-attention mechanisms [30]. These models capture global contextual relationships across video segments and improve anomaly detection performance.

More recently, vision–language models have extended this paradigm by incorporating multimodal representations and semantic reasoning capabilities [2,38]. Although these architectures achieve strong performance on benchmark datasets, they typically require large-scale training data and substantial computational resources. High memory consumption and reduced inference speed may limit their suitability for real-time surveillance deployment.

Recent survey work further highlights the growing adoption of transformer-based architectures for video anomaly detection, emphasizing improved temporal modeling, contextual representation, and robustness across diverse surveillance datasets [39]. In addition, lightweight real-time anomaly detection frameworks designed for edge computing environments have been proposed to support deployment in resource-constrained surveillance systems. These approaches aim to achieve efficient inference, reduced computational complexity, and improved scalability while maintaining competitive anomaly detection performance, making them suitable for real-time surveillance applications [40].

2.6 Limitations of existing evaluation practices

Despite significant architectural advancements, evaluation practices remain largely model-centric. Most studies primarily report ROC–AUC as the dominant performance metric, while deployment-relevant factors such as false alarm rate, temporal stability, inference speed, calibration reliability, and computational footprint are often underreported [9,19,20].

Furthermore, heterogeneous experimental settings—including differences in datasets, preprocessing pipelines, backbone architectures, and hardware configurations—make direct comparison across studies challenging and reduce the interpretability of reported results [21,25]. These limitations highlight the need for structured, multi-metric comparisons that reflect real-world deployment requirements.

To address this gap, Table 1 presents a quantitative comparison of representative video anomaly detection frameworks across accuracy, efficiency, calibration, and deployment-oriented metrics.

2.7 Quantitative comparison with state-of-the-art methods

Table 1 presents a structured comparison of representative video anomaly detection frameworks, including weakly supervised methods, MIL-based architectures, graph-based models, transformer-based approaches, and recent lightweight and hybrid frameworks. The comparison incorporates key evaluation dimensions such as detection accuracy, inference speed, memory requirements, hardware configuration, and calibration reporting.

The comparison indicates that transformer-based and vision–language architectures achieve the highest detection accuracy. However, these models typically require substantial computational resources, large memory capacity, and high-end hardware, which may limit their practical deployment in real-time surveillance systems [30,38].

Hybrid CNN–SRU/LSTM frameworks provide stable temporal modeling with moderate computational requirements, offering a balance between accuracy and efficiency [23]. In contrast, lightweight CNN–MIL approaches achieve competitive detection performance while maintaining higher inference speed and lower memory consumption, making them suitable for large-scale and real-time applications [24].

These findings suggest that marginal improvements in detection accuracy often come at the cost of increased computational complexity. Therefore, accuracy alone may not be sufficient for selecting anomaly detection frameworks for real-world deployment.

Table 1 summarizes the key characteristics and performance metrics of representative video anomaly detection methods reported in prior studies [6,8,24,30,38]. The comparison highlights variations in accuracy, efficiency, memory usage, hardware requirements, and calibration reporting across different architectural paradigms.

Despite the progress achieved by existing approaches, several limitations remain. First, most studies rely predominantly on ROC–AUC while overlooking deployment-oriented metrics such as false alarm rate, calibration reliability, and computational efficiency [9,19,20]. Second, cross-domain robustness is often evaluated qualitatively rather than through standardized quantitative metrics, making generalization performance difficult to assess [3,26]. Third, unified multi-metric comparisons across diverse architectures remain limited, reducing the interpretability of reported performance trends [21,25].

These limitations highlight the need for a structured, deployment-oriented comparative framework that integrates accuracy, efficiency, reliability, and robustness into a unified evaluation strategy. The present study addresses these gaps by providing a metric-wise, hardware-aware, and cross-domain focused comparison of representative video anomaly detection frameworks [32,36].

Table 1: Metric-based comparison of representative video anomaly detection methods [6], [8], [12], [24], [30]

Method	Year	Framework	Dataset	AUC (%)	FPS	Memory	Hardware	Calibration	Key Limitation
Weakly Supervised MIL [6]	2018	Weakly supervised	UCF-Crime	75.4	—	—	GPU	No	Limited temporal modelling
RTFM [8]	2021	MIL-based	UCF-Crime	84.3	~20	High	GPU	No	Computational overhead
Graph-based Model [26]	2022	Graph-based	ShanghaiTech	83.8	~18	High	GPU	No	Complex architecture
Feature-Enhanced MIL [12]	2023	MIL-based	UCF-Crime	85.1	~22	Moderate	GPU	Partial	Increased complexity
Transformer Model [30]	2023	Attention	UCF-Crime	86.5	12	~850MB	High-end GPU	No	Low efficiency
Vision-Language Model [2,38]	2024	Multimodal	UCF-Crime	89.0	8	~1200MB	Multi-GPU	No	High computational cost
Lightweight CNN-MIL [24]	2025	Lightweight	UCF-Crime, ShanghaiTech	82–84.7	45-72	Low	Single GPU	Yes	Limited long-range modeling
CNN-SRU/LSTM [23]	2026	Hybrid	UCF-Crime	85.9	24	Moderate	Single GPU	Partial	Sequential overhead

3 Methodology of comparative analysis

This study adopts a structured, literature-driven comparative methodology to evaluate representative video anomaly detection frameworks. Unlike experimental studies that involve model training, fine-tuning, or implementation, the proposed analysis synthesizes performance results reported in prior research. This approach enables systematic comparison across heterogeneous experimental settings while avoiding variability introduced by differences in training procedures, dataset splits, and evaluation protocols [1–3].

The comparative analysis focuses on major categories of video anomaly detection frameworks, including hybrid temporal architectures, weakly supervised multiple instance learning (MIL) methods, reconstruction-based approaches, and attention-based models. These frameworks represent complementary design paradigms with distinct trade-offs in temporal modeling capability, computational efficiency, and deployment feasibility [4–6]. The objective of this study is to analyze how these architectural differences influence performance across deployment-oriented evaluation dimensions. Figure 1 illustrates the structured workflow adopted in this study. The methodology consists of five main stages: literature identification, study selection, data extraction, multi-dimensional evaluation, and framework-level synthesis. Such structured comparative pipelines have been widely adopted to improve reproducibility and transparency in anomaly detection research [7,8].

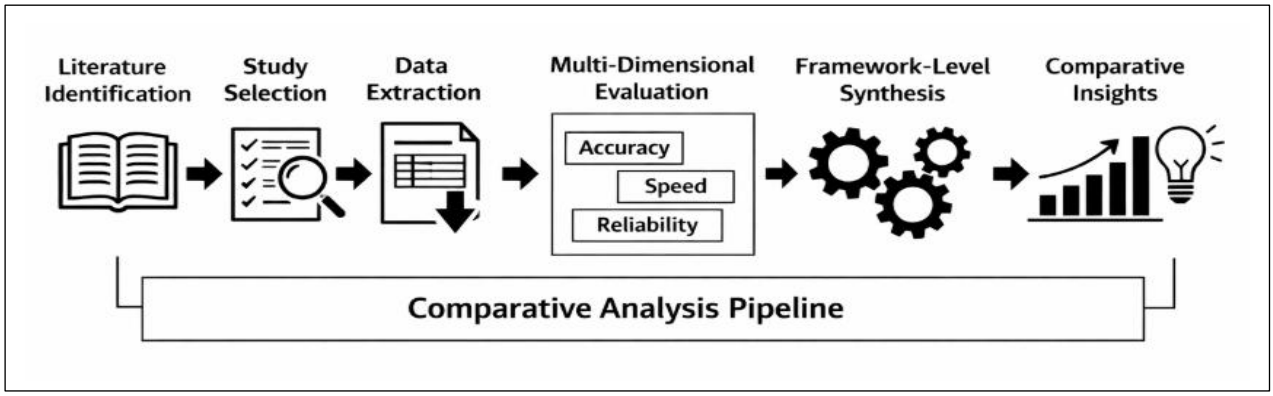

Figure 1: Overview of the methodology pipeline for comparative analysis of video anomaly detection frameworks [7,8]

3.1 Comparative pipeline and study selection

The comparative methodology begins with systematic identification of relevant research studies in video anomaly detection. Emphasis is placed on recent deep learning-based approaches designed for surveillance applications. Selected studies represent diverse architectural paradigms and deployment characteristics to ensure comprehensive comparison [9–12].

To ensure methodological consistency, studies are selected based on predefined inclusion criteria:

- **Domain relevance:** Studies must address video anomaly detection in surveillance environments involving untrimmed video streams [13].
- **Methodological scope:** Only deep learning-based approaches are included, excluding handcrafted feature-based methods [14].
- **Benchmark evaluation:** Selected studies must report results on widely used datasets such as UCF-Crime, ShanghaiTech, Avenue, or Ped2 [15,16].
- **Metric availability:** Studies must report at least one key performance metric such as detection accuracy, false alarm rate, inference speed, or calibration performance [17].
- **Architectural diversity:** Selected studies collectively represent hybrid, MIL-based, reconstruction-based, and attention-based frameworks [18,19].

Following study selection, relevant performance metrics and architectural characteristics are extracted and organized according to framework type and dataset usage. This process enables structured comparison across model

- **Framework-level synthesis:** Identification of consistent performance trends across studies

This structured procedure emphasizes relative performance patterns and deployment-oriented insights rather than isolated numerical comparisons [25].

3.2 Study selection protocol and practical consideration

Building upon the selection criteria outlined in Section 3.1 and the multi-metric evaluation framework described in Section 3.2, this subsection presents the procedure adopted for identifying relevant studies and outlines the practical considerations necessary to ensure families while preserving the context of reported results [20].

3.3 Data extraction and comparative procedure

Quantitative results and architectural characteristics are extracted directly from selected studies. The extracted information includes detection accuracy, false alarm rate, inference speed, calibration metrics, dataset usage, and hardware configuration when available. Similar data extraction strategies have been adopted in recent anomaly detection comparative studies to enable consistent cross-study evaluation [21,22].

Since experimental setups vary across studies, strict numerical normalization is not applied. Instead, reported results are interpreted within their experimental context. Comparisons are performed at the framework level, emphasizing consistent performance trends rather than absolute numerical equivalence. This approach reduces bias introduced by heterogeneous experimental protocols. The extracted data are organized into comparative tables summarizing dataset usage, hardware configurations, and performance metrics across representative frameworks. Dataset-aware and hardware-aware comparisons have been emphasized in recent benchmarking studies to support deployment-oriented evaluation [23,24].

The comparative procedure consists of three main stages:

- **Data extraction:** Collection of reported performance metrics and architectural characteristics
- **Contextual interpretation:** Analysis considering dataset and hardware variations

meaningful comparison across diverse experimental settings.

This study follows a structured literature-driven comparative methodology [1–3,21]. Relevant studies were identified from major academic databases including IEEE Xplore, Scopus, ScienceDirect, and SpringerLink. The search was conducted using keywords such as “video anomaly detection”, “weakly supervised anomaly detection”, “CNN–LSTM anomaly detection”, “multiple instance learning anomaly detection”, and “transformer-based anomaly detection”.

The selection process considered studies published between 2018 and 2025. An initial pool of approximately 50 studies was identified, from which 20 representative

works were selected based on predefined inclusion criteria, including relevance to surveillance-based anomaly detection, availability of quantitative performance metrics, and architectural diversity [14,19,26]. These selected works represent widely cited and methodologically diverse approaches in the video anomaly detection domain.

Studies lacking sufficient experimental details or not evaluated on standard benchmark datasets (e.g., UCF-Crime, ShanghaiTech) were excluded [6,26]. The study selection process follows a structured filtering approach consisting of identification, screening, eligibility assessment, and final inclusion [21,22].

Due to heterogeneous experimental setups across the selected studies, strict numerical normalization was not applied. Instead, results are interpreted within their reported experimental context, and comparisons are performed at the framework level to ensure fair and meaningful evaluation. This approach minimizes misleading conclusions arising from variations in datasets, preprocessing strategies, and evaluation protocols [21,25].

Since the analysis is based on reported results, hardware configurations vary across studies and include single-GPU, multi-GPU, and high-performance computing environments. These variations are explicitly considered during interpretation and are reported in Table 1 where available, ensuring that performance comparisons remain context-aware and practically grounded [23,24].

3.3 Evaluation dimensions

To address limitations of single-metric evaluation, this study adopts a multi-dimensional performance assessment framework. These evaluation dimensions capture both algorithmic performance and deployment-oriented characteristics. Multi-metric evaluation has been recommended in recent video anomaly detection research to improve reliability and interpretability [26,27].

The following evaluation dimensions are considered:

- **Detection Accuracy (AUC):** Measures the ability of the model to distinguish between normal and anomalous events [28].
- **Reliability (False Alarm Rate):** Evaluates operational stability by measuring incorrect anomaly predictions [29].
- **Temporal Behavior:** Assesses detection delay and prediction consistency across video sequences [30].
- **Computational Efficiency (FPS):** Measures inference speed and suitability for real-time deployment [31].
- **Calibration Robustness:** Evaluates prediction confidence reliability under uncertainty [32].

These complementary metrics provide a comprehensive and deployment-oriented assessment of anomaly detection frameworks.

3.4 Comparative algorithm

To formalize the methodology, Algorithm 1 summarizes the comparative workflow.

Algorithm 1: Comparative Evaluation Procedure
Input: Selected studies
Output: Comparative performance analysis
Step 1: Identify relevant studies
Step 2: Apply inclusion criteria
Step 3: Extract performance metrics
Step 4: Organize dataset and hardware information
Step 5: Perform multi-metric evaluation
Step 6: Analyze framework-level trends
Step 7: Generate comparative conclusions

This algorithm formalizes the evaluation process and improves reproducibility and consistency in comparative analysis [33].

3.5 Handling heterogeneity across studies

A key challenge in literature-based comparison is variability in experimental setups, including dataset splits, preprocessing pipelines, backbone architectures, optimization strategies, and hardware environments. Such heterogeneity can affect performance interpretation and complicate cross-study comparison [34,35].

To address this challenge, the methodology adopts a qualitative–quantitative synthesis strategy:

- Reported results are interpreted within their experimental context
- Consistent trends across studies are prioritized
- Conclusions are drawn at the framework level rather than individual models

This approach improves robustness and reduces bias in comparative evaluation.

3.6 Reproducibility and study scope

This study relies entirely on publicly reported results and does not involve new experiments, datasets, or model implementations. All study selection criteria, extracted metrics, and evaluation procedures are explicitly documented to ensure reproducibility of the comparative analysis [36].

The scope of this work is limited to deep learning-based video anomaly detection frameworks designed for surveillance applications. The study focuses on deployment-oriented evaluation and framework-level comparison.

3.7 Methodological limitations

While the proposed methodology enables structured comparison, several limitations remain:

- Variations in experimental protocols across studies
- Missing evaluation metrics in some reported works
- Hardware variability across implementations
- Dataset-specific performance differences

These limitations are consistent with those reported in recent anomaly detection benchmarking studies [37,38]. To mitigate these challenges, the analysis focuses on consistent framework-level trends rather than isolated numerical comparisons. This approach enables robust, deployment-oriented evaluation of representative video anomaly detection frameworks.

4 Result and comparative analysis

This section presents a metric-wise comparative analysis of representative video anomaly detection frameworks, with emphasis on hybrid CNN–SRU/LSTM and lightweight CNN–MIL approaches. The analysis synthesizes findings from widely adopted benchmark studies conducted on standard datasets such as UCF-Crime, ShanghaiTech, and Avenue. Since experimental setups vary across studies, the discussion focuses on consistent framework-level trends rather than strict numerical equivalence [38].

The comparative evaluation is organized across multiple deployment-oriented dimensions, including detection accuracy, reliability, temporal behavior, computational efficiency, calibration robustness, and cross-domain generalization.

4.1 Quantitative comparison across frameworks

Table 1 provides a structured comparison of representative video anomaly detection frameworks across key performance metrics. Hybrid architectures such as CNN–LSTM and SRU-based models demonstrate strong detection performance, achieving approximately 85–86% AUC on the UCF-Crime dataset under single-GPU settings. These models benefit from explicit temporal modeling, enabling improved sequence understanding and stable anomaly detection [38].

Lightweight CNN–MIL frameworks, including weakly supervised anomaly localizationnormal (WSAL) and RTFM-based approaches, achieve competitive performance in the range of 82–84.7% AUC while maintaining significantly higher computational efficiency. These frameworks achieve inference speeds of up to 72 FPS with lower memory requirements, making them suitable for real-time and large-scale surveillance applications [38].

Transformer-based architectures such as Video Swin Transformer and ViViT, along with vision–language models based on CLIP, achieve higher detection accuracy, often exceeding 86% AUC. However, these models typically operate at lower inference speeds and require substantial computational resources. This highlights a consistent trade-off between detection accuracy and deployment efficiency across architectural paradigms [38].

4.2 Reliability and false alarm behavior

Reliability is a critical consideration in practical surveillance systems, where excessive false alarms can reduce system usability. Although most studies report detection accuracy, explicit evaluation of false alarm rate (FAR) remains limited. As observed in Table 1, several high-performing models do not report reliability metrics, which restricts their practical interpretability [38].

Hybrid CNN–LSTM and SRU-based architectures generally exhibit improved temporal stability, which helps reduce spurious detections. In contrast, CNN–MIL frameworks benefit from instance-level aggregation, which improves robustness to noisy segments and reduces false positives in weakly supervised settings. These findings indicate that reliability should be evaluated alongside accuracy for deployment-oriented performance assessment.

4.3 Temporal behavior and stability

Temporal modeling plays an important role in detecting anomalies that evolve gradually over time. Hybrid architectures explicitly model temporal dependencies, resulting in smoother anomaly score transitions and reduced detection delay. These characteristics make them suitable for detecting long-duration anomalies such as accidents, abnormal crowd behavior, and violent activities [38].

CNN–MIL frameworks rely on segment-level aggregation rather than explicit temporal recurrence. While this may limit long-range temporal modeling, the simplified architecture enables faster inference and reduced latency. As a result, CNN–MIL approaches offer a practical balance between temporal modeling capability and computational efficiency.

4.4 Computational efficiency and scalability

Computational efficiency is a key determinant of deployment feasibility. As summarized in Table 1, lightweight CNN–MIL frameworks achieve higher inference speeds and lower memory consumption compared to hybrid and transformer-based architectures [38]. These characteristics make them suitable for real-time surveillance systems and edge deployment scenarios.

Figure 2 illustrates the latency–performance trade-off across representative frameworks. A modest improvement of approximately 2–3% in AUC is often associated with a 3–5× reduction in inference speed. Transformer-based approaches, while achieving high accuracy, typically operate below real-time thresholds and require substantial memory resources. These constraints limit their applicability in resource-constrained environments.

In contrast, lightweight CNN–MIL frameworks maintain competitive detection accuracy while achieving significantly higher throughput. This balance highlights their suitability for large-scale surveillance deployment and real-time monitoring applications.

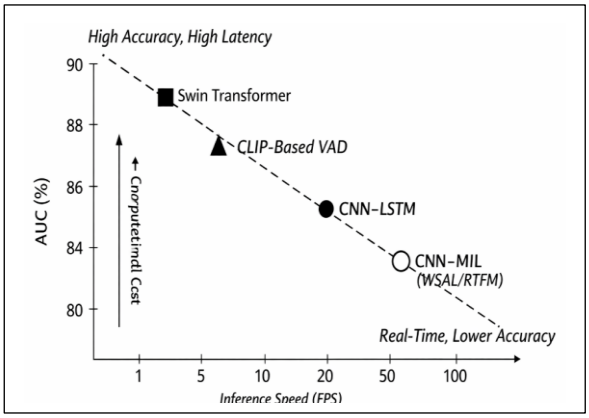

Figure 2: Latency–performance trade-off across video anomaly detection frameworks [24], [30],[38]

4.5 Calibration and cross-domain robustness

Calibration has emerged as an important factor in reliable anomaly detection, particularly under domain shift conditions. Expected Calibration Error (ECE) measures the alignment between predicted confidence and actual outcomes. Poor calibration can result in overconfident predictions, which may lead to unreliable anomaly detection in safety-critical applications [19,38].

Expected Calibration Error (ECE) is computed by partitioning predictions into fixed confidence bins (typically10–15 bins) and measuring the difference between predicted confidence and empirical accuracy within each bin.

Normalization-based test-time adaptation (BN-TTA) refers to adjusting batch normalization statistics during inference to reduce domain shift effects. This technique allows models to adapt to new environments without retraining and has been explored in recent domain generalization studies [3,26].

As shown in Table 1, calibration metrics are rarely reported in existing studies, highlighting a gap in current evaluation practices. Some CNN–MIL frameworks demonstrate improved calibration behavior, suggesting more reliable confidence estimation under uncertainty [38].

Cross-domain evaluation further reveals that models trained on specific datasets may experience performance degradation when applied to unseen environments. Cross-domain robustness is commonly evaluated using relative AUC degradation (ΔAUC). Reported studies indicate performance degradation ranging from 5–15%, depending on architectural design and domain similarity [3,26].

Hybrid CNN–LSTM frameworks generally exhibit moderate performance degradation due to learned temporal dependencies, while lightweight CNN–MIL frameworks demonstrate relatively stable performance under domain shifts due to their segment-level representation strategy [1,7,9,24].

4.6 Trade-off analysis and deployment implications

The comparative analysis reveals consistent trade-offs across architectural paradigms. Hybrid CNN–LSTM and SRU-based models provide strong temporal modeling and detection performance but incur higher computational costs. Lightweight CNN–MIL frameworks offer competitive accuracy with improved computational efficiency, making them suitable for real-time deployment scenarios.

Transformer-based and vision–language architectures achieve the highest detection accuracy but require substantial computational resources and lower inference speed. These characteristics limit their practical deployment in resource-constrained environments.

Overall, model selection should consider application-specific requirements such as real-time processing constraints, hardware availability, and cross-domain robustness. Deployment-oriented evaluation is therefore essential for selecting suitable anomaly detection frameworks.

4.7 Summary of key findings

The key findings of the comparative analysis are summarized as follows:

- Hybrid CNN–LSTM and SRU-based models provide strong detection accuracy and temporal stability but involve higher computational overhead.
- Lightweight CNN–MIL frameworks offer competitive performance with improved efficiency and scalability for real-time deployment.
- Transformer-based architectures achieve high detection accuracy but have limited applicability in real-time environments due to computational constraints.
- Reliability, calibration, and cross-domain robustness remain underreported in existing studies, highlighting the need for multi-metric evaluation.

Overall, this comparative analysis provides a deployment-oriented perspective that extends beyond traditional accuracy-based evaluation and highlights practical considerations for real-world video anomaly detection systems [38].

5 Discussion

This section interprets the comparative analysis results and discusses their implications for real-world deployment of video anomaly detection systems. Instead of focusing solely on individual performance values, the discussion highlights broader trends, practical trade-offs, and limitations observed across different architectural paradigms. The analysis is based on representative benchmark studies conducted on standard datasets under comparable evaluation settings [38].

As summarized in Table 1, the comparative analysis highlights consistent differences across architectural paradigms in terms of accuracy, efficiency, and deployment suitability.

5.1 Accuracy–efficiency trade-off

One of the most consistent observations from Table 1 and Figure 2 is the trade-off between detection accuracy and computational efficiency. Transformer-based models such as Video Swin Transformer and ViViT, along with vision–language approaches based on CLIP, typically achieve higher detection accuracy, often exceeding 86% AUC. However, these models operate at significantly lower inference speeds, often below real-time thresholds, due to their high computational complexity [38].

In contrast, lightweight CNN–MIL frameworks such as WSAL and RTFM achieve slightly lower but competitive detection performance while maintaining substantially higher inference speeds. Their ability to process video streams efficiently makes them more suitable for real-time surveillance and large-scale deployment scenarios.

Hybrid architectures such as CNN–LSTM and SRU-based models provide improved temporal modeling compared to CNN–MIL approaches, but their sequential processing structure leads to moderate inference speed. These observations indicate that small improvements in accuracy often require disproportionately higher computational resources, emphasizing the need to balance performance and efficiency in practical deployment scenarios.

5.2 Importance of multi-metric evaluation

The comparative results highlight the limitations of relying solely on AUC as a performance indicator. While many studies report high detection accuracy, fewer include reliability-related metrics such as false alarm rate (FAR) and calibration measures such as Expected Calibration Error (ECE). These metrics are essential for evaluating model behavior in real-world environments [38].

For example, models with high AUC may still produce unstable predictions or excessive false alarms, reducing their practical usability. In contrast, frameworks with slightly lower accuracy may provide more consistent and reliable performance when evaluated across multiple metrics.

Therefore, incorporating evaluation dimensions such as reliability, computational efficiency, calibration, and cross-domain robustness provides a more comprehensive assessment of model performance and supports informed framework selection for deployment scenarios.

5.3 Temporal modeling and stability

Temporal consistency is essential in video anomaly detection because anomalies often evolve gradually over time. Hybrid architectures such as CNN–LSTM and SRU-based models explicitly capture temporal dependencies, resulting in smoother anomaly score transitions and improved detection stability [38].

However, improved temporal modeling increases computational complexity and reduces inference speed. In comparison, CNN–MIL frameworks rely on segment-level aggregation rather than explicit temporal modeling. Although this may limit long-range dependency modeling, the simplified structure enables faster inference and improved scalability.

This trade-off highlights the importance of selecting architectures based on deployment requirements, particularly when balancing temporal modeling capability and computational efficiency.

5.4 Effect of frame rate variations

Frame rate variations in surveillance systems can influence anomaly detection performance. Hybrid architectures depend on consistent temporal continuity and may experience degraded performance when frames are dropped or when input frame rates fluctuate.

CNN–MIL frameworks, which operate on segment-level representations, are generally less sensitive to such variations. This characteristic makes them more suitable for heterogeneous deployment environments, including distributed camera networks and edge-based surveillance systems with varying hardware capabilities [38].

5.5 Generalization and cross-domain challenges

Generalization across different environments remains a major challenge in video anomaly detection. Differences in scene structure, lighting conditions, camera viewpoints, and motion patterns can significantly affect model performance [38].

Models optimized for specific datasets often experience performance degradation when applied to unseen environments. This limitation is further compounded by the absence of standardized cross-domain evaluation protocols in many studies.

Improving cross-domain robustness requires the development of generalized feature representations, domain adaptation techniques, and multi-dataset evaluation strategies to ensure reliable performance across diverse deployment scenarios.

5.6 Limitations of current approaches

Despite recent progress, several limitations remain in existing video anomaly detection frameworks:

- Limited reporting of reliability and calibration metrics
- High computational requirements for state-of-the-art architectures
- Performance degradation under domain shift
- Inconsistent evaluation protocols across studies

These limitations highlight the need for standardized and comprehensive evaluation frameworks to enable fair comparison and practical deployment [38].

5.7 Deployment implications

From a deployment perspective, model selection should consider application-specific constraints such as hardware resources, latency requirements, scalability, and operational reliability. Lightweight CNN–MIL frameworks provide high inference speed and low computational overhead, making them suitable for real-time surveillance applications and edge deployment environments [38]. These models are particularly effective in distributed camera networks where computational resources and memory availability are limited.

Hybrid CNN–LSTM and SRU-based architectures offer improved temporal modeling capabilities and enhanced detection stability. These models provide a balanced trade-off between performance and computational efficiency, making them appropriate for moderately resourced systems such as institutional surveillance networks, transportation hubs, and smart city monitoring infrastructures. However, their sequential processing nature may introduce moderate latency, which should be considered in time-critical applications.

Transformer-based and vision–language models generally achieve higher detection accuracy and improved contextual understanding. However, these architectures require substantial computational resources, higher

memory consumption, and increased inference time. Consequently, they are better suited for offline analysis, centralized monitoring systems, or high-performance computing environments where computational constraints are less restrictive.

In practical surveillance deployments, additional environmental and operational factors such as camera placement, illumination variations, background motion, and network latency can significantly influence anomaly detection performance. Lightweight architectures are generally more resilient to such variations due to their faster inference and reduced computational dependency. In contrast, high-complexity transformer-based and vision–language models may experience performance degradation in resource-constrained or unstable environments. Therefore, deployment-oriented evaluation considering both performance metrics and operational constraints is essential for reliable real-world implementation.

Figure 3 illustrates a deployment-oriented framework selection strategy based on performance and resource requirements.

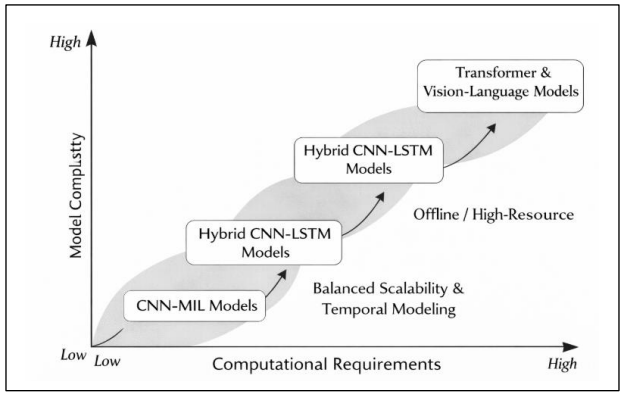

Figure 3: Deployment-oriented framework selection strategy

To further clarify deployment suitability, Table 2 summarizes recommended frameworks across different application scenarios.

Table 2: Framework suitability across deployment scenarios

Deployment Scenario	Recommended Framework	Key Reason
Real-time edge surveillance	CNN–MIL (WSAL, RTFM)	High FPS, low resource usage
Smart city monitoring	CNN–MIL / Hybrid	Balance of scalability and temporal modeling
Offline forensic analysis	Transformer (Swin, ViViT)	High accuracy, less time constraint
High-resource environments	Vision-language (CLIP-based)	Semantic understanding capability

Lightweight CNN–MIL frameworks are particularly suitable for resource-constrained environments and real-time monitoring scenarios, while hybrid models offer balanced performance for moderately resourced systems. Transformer-based and vision–language architectures are more appropriate for high-resource scenarios where detection accuracy and contextual understanding are prioritized over computational efficiency.

5.8 Failure case analysis

Despite strong benchmark performance, current video anomaly detection models still encounter challenges in complex real-world scenarios. Figure 4 illustrates representative failure cases.

In crowded environments, subtle anomalies such as individuals moving against crowd flow may remain undetected due to dominant normal motion patterns. Similarly, low-light or poor visibility conditions can lead to incorrect predictions due to degraded feature quality and unstable temporal modeling [38].

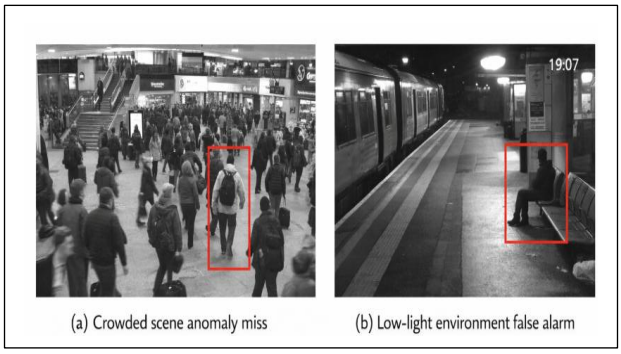

Figure 4: Representative failure cases in video anomaly detection [3],[26],[38]

These failure cases highlight the sensitivity of current models to environmental variability and emphasize the need for improved robustness and adaptive learning mechanisms.

For example, in a crowded subway station scenario, an individual moving slowly against the dominant crowd flow may not be detected as anomalous by MIL-based models due to dominant normal motion patterns. Similarly, hybrid models may misclassify abrupt illumination changes (e.g., flickering lights) as anomalies due to temporal inconsistency.

5.9 Statistical considerations

This study is based on previously published benchmark results; therefore, direct statistical testing such as t-test or ANOVA could not be performed due to the absence of raw experimental outputs. Instead, comparative conclusions were derived from consistent performance trends reported across multiple peer-reviewed studies evaluated on standard datasets [21,22].

The consistency of these trends across independent studies provides indirect statistical support for the comparative findings. This methodology aligns with established practices in literature-driven comparative

analyses, where reproducibility and cross-study consistency are used to validate conclusions [33,36].

5.10 Future research directions

The findings suggest several directions for future research:

- Development of standardized multi-metric evaluation protocols
- Integration of calibration-aware anomaly detection techniques
- Design of lightweight architectures with improved temporal modelling
- Enhancement of domain generalization and cross-dataset evaluation

Addressing these challenges will help bridge the gap between research performance and real-world deployment of video anomaly detection systems [38].

6 Conclusion

This study presented a metric-wise comparative analysis of representative video anomaly detection frameworks, focusing on hybrid architectures such as CNN–LSTM and SRU-based models, and lightweight CNN–MIL approaches including WSAL and RTFM. Unlike conventional surveys that primarily emphasize detection accuracy, this work adopts a multi-metric evaluation strategy incorporating reliability, temporal behavior, computational efficiency, and calibration robustness. This broader perspective enables a more practical and deployment-oriented assessment of existing approaches.

The comparative analysis, summarized in Table 1 and illustrated through the latency–performance trade-off in Figure 2, indicates that no single framework consistently outperforms others across all evaluation dimensions. Lightweight CNN–MIL frameworks demonstrate significantly higher inference speeds and lower computational requirements, making them suitable for real-time and large-scale deployment scenarios. In contrast, hybrid CNN–LSTM and SRU-based architectures provide improved temporal consistency and competitive detection performance, but with moderate computational overhead. Transformer-based and vision–language models achieve higher detection accuracy; however, their substantial computational cost limits real-time applicability.

The findings further reveal that small gains in detection accuracy often require disproportionately higher computational resources. This observation underscores the limitations of relying solely on ROC–AUC as a performance indicator and highlights the importance of multi-metric evaluation for practical deployment. In addition, reliability-related metrics such as false alarm rate and calibration remain underreported in many studies, despite their importance in real-world surveillance applications.

Cross-domain generalization also emerges as a key challenge. Models trained on specific datasets may experience performance degradation when deployed in new environments characterized by variations in lighting conditions, scene complexity, and camera viewpoints. These findings emphasize the need for robust architectures and standardized evaluation protocols that better reflect real-world deployment conditions.

Although this study provides a structured comparative framework, certain limitations remain. The analysis relies on previously reported benchmark results, which may vary due to differences in experimental configurations. Additionally, the absence of raw experimental data prevents formal statistical testing, representing an important limitation and an opportunity for future research.

Future work can extend this study by developing standardized benchmarking protocols, designing lightweight architectures with improved temporal modeling capability, and incorporating calibration-aware anomaly detection techniques. Furthermore, adopting statistically rigorous evaluation practices and cross-dataset validation strategies will help strengthen the reliability of comparative analyses.

In addition, emerging transformer-based and vision–language models present promising opportunities for improving semantic understanding and contextual anomaly detection. However, achieving a balance between computational efficiency and detection performance remains a critical challenge for real-world deployment. Future research should focus on designing efficient hybrid architectures that combine the strengths of lightweight models and advanced attention-based frameworks to support scalable and reliable surveillance systems.

Overall, this work provides a comprehensive and deployment-oriented perspective on video anomaly detection. By systematically analyzing trade-offs between accuracy, efficiency, and reliability, the study offers practical guidance for selecting appropriate frameworks based on application requirements, hardware constraints, and real-world deployment scenarios.

Acknowledgement

The authors acknowledge their affiliated institutions for providing computational resources and research facilities. No external funding was received for this study. The authors declare no conflict of interest.

Data availability

No new datasets were used in this study. All results are based on previously published works cited in the manuscript.

Reproducibility statement

This study is based on publicly available results reported in prior research. All included studies are cited, and extracted metrics are presented transparently in comparative tables. No new experimental data or proprietary datasets were used. The methodology and

selection criteria are described to facilitate replication of the comparative analysis.

Ethics Statement

This research utilizes publicly available surveillance datasets and does not involve human subjects or identifiable personal information. Therefore, ethical approval was not required.

References

[1] Georgescu, M.-I., Bărbălău, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: Anomaly detection in video via self-supervised and multi-task learning. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 12737–12747 (2021). https://doi.org/10.1109/CVPR46437.2021.01255

[2] Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection. Proceedings of the AAAI Conference on Artificial Intelligence 38(6), 6074–6082 (2024). https://doi.org/10.1609/aaai.v38i6.28423

[3] Chen, Y., et al.: Domain Generalization in Video Anomaly Detection: A Survey. Pattern Recognition 145, 109880 (2025). https://doi.org/10.1016/j.patcog.2023.109880

[4] Chong, Y.S., Tay, Y.H.: Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. In: Proc. Int. Symp. Neural Netw. (ISNN), pp. 189–196 (2017). https://doi.org/10.1007/978-3-319-59081-3_23

[5] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning Temporal Regularity in Video Sequences. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 733–742 (2016). https://doi.org/10.1109/CVPR.2016.87

[6] Sultani, W., Chen, C., Shah, M.: Real-World Anomaly Detection in Surveillance Videos. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6479–6488 (2018). https://doi.org/10.1109/CVPR.2018.00678

[7] Sun, C., et al.: Self-Supervised Representation Learning for Video Anomaly Detection. IEEE Trans. Neural Netw. Learn. Syst. 33(12), 7486–7499 (2022). https://doi.org/10.1109/TNNLS.2021.3106789

[8] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J., Carneiro, G.: Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 4975–4986 (2021). https://doi.org/10.1109/ICCV48922.2021.00493

[9] Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., Zhang, H.: Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8022–8031 (2023). https://doi.org/10.1109/CVPR52729.2023.00775

[10] Li, K., et al.: Attention-Guided Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. Neurocomputing 545, 125–137 (2023). https://doi.org/10.1016/j.neucom.2023.02.015

[11] Wu, J., et al.: Context-Aware Multiple Instance Learning for Video Anomaly Detection. IEEE Trans. Image Process. 32, 2345–2357 (2023). https://doi.org/10.1109/TIP.2023.3234567

[12] Zhou, X., et al.: Generative Adversarial Learning for Weakly Supervised Anomaly Detection in Surveillance Videos. Pattern Recognition 134, 109043 (2023). https://doi.org/10.1016/j.patcog.2022.109043

[13] Fan, Y., Yu, Y., Lu, W., Han, Y.: Weakly Supervised Video Anomaly Detection with Snippet Anomalous Attention. IEEE Trans. Circuits Syst. Video Technol. (2024). https://doi.org/10.1109/TCSVT.2024.3350084

[14] Ramachandra, B., Jones, M., Vatsavai, R.R.: A Survey of Single-Scene Video Anomaly Detection. IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2020.3040591

[15] Gong, D., et al.: Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder. IEEE Trans. Pattern Anal. Mach. Intell. (2021). https://doi.org/10.1109/TPAMI.2020.2990693

[16] Park, H., Noh, J., Ham, B.: Learning Memory-Guided Normality for Anomaly Detection. IEEE Trans. Pattern Anal. Mach. Intell. (2022). https://doi.org/10.1109/TPAMI.2021.3092935

[17] Zhang, Y., Nie, X., He, R., Chen, M., Yin, Y.: Normality Learning in Multispace for Video Anomaly Detection. IEEE Trans. Circuits Syst. Video Technol. (2020). https://doi.org/10.1109/TCSVT.2020.3039798

[18] Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. Comput. Vis. Image Underst. 156, 117–127 (2017). https://doi.org/10.1016/j.cviu.2016.10.010

[19] Amin, S.U., et al.: EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos. Mathematics 10(9), 1555 (2022). https://doi.org/10.3390/math10091555

[20] Amin, S.U., et al.: Enhanced Anomaly Detection in Surveillance Videos Using Attention Mechanisms. IEEE Access 12, 162697–162712 (2024). https://doi.org/10.1109/ACCESS.2024.3488797

[21] Amin, S.U., et al.: Video Anomaly Detection Utilizing Efficient Spatiotemporal Feature Fusion. Advanced Intelligent Systems 6, 2300706 (2024). https://doi.org/10.1002/aisy.202300706

[22] Tur, A.O., Dall'Asen, N., Beyan, C., Ricci, E.: Exploring Diffusion Models for Unsupervised Video Anomaly Detection. In: Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 2540–2544 (2023). https://doi.org/10.1109/ICIP49359.2023.10222594

[23] Gupta, R., Tyagi, N.: Hybrid CNN–SRU/LSTM with Multiple Instance Learning. Signal, Image and Video Processing (2026). https://doi.org/10.1007/s11760-025-05072-w

[24] Gupta, R., Tyagi, N.: Lightweight CNN–MIL Models for Cross-Domain Video Anomaly Detection. Informatica 49(36) (2025). https://doi.org/10.31449/inf.v49i36.12037

[25] Wang, J., et al.: Knowledge Distillation for Efficient Video Anomaly Detection. Pattern Recognition Letters 158, 30–37 (2022). https://doi.org/10.1016/j.patrec.2022.04.015

[26] Zhang, L., et al.: Benchmarking Cross-Domain Robustness in Surveillance Video Analytics. Neurocomputing 569, 127056 (2024). https://doi.org/10.1016/j.neucom.2023.127056

[27] Aich, A., Peng, K.-C., Roy-Chowdhury, A.: Cross-Domain Video Anomaly Detection Without Target Domain Adaptation. In: Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), pp. 2578–2590 (2023). https://doi.org/10.1109/WACV56688.2023.00261

[28] Luo, W., Liu, W., Lian, D., Gao, S.: Future Frame Prediction Network for Video Anomaly Detection. IEEE Trans. Pattern Anal. Mach. Intell. 44(11), 7505–7520 (2022). https://doi.org/10.1109/TPAMI.2021.3129349

[29] Dilek, E., Dener, M.: An Overview of Transformers for Video Anomaly Detection. Neural Computing and Applications 37, 17825–17857 (2025). https://doi.org/10.1007/s00521-025-11218-1

[30] Liu, W., Luo, W., Lian, D., Gao, S.: Future Frame Prediction for Anomaly Detection: A New Baseline. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6536–6545 (2018). https://doi.org/10.1109/CVPR.2018.00684

[31] Yin, W., et al.: Weakly Supervised Video Anomaly Detection via Disentangled Semantic Alignment. Proc. AAAI Conf. Artif. Intell. (2026). https://doi.org/10.1609/aaai.v40i14.38191

[32] Carvalho, P., Cardoso, J., Caetano, F., Mastralexi, C.: Enhancing Weakly-Supervised Video Anomaly Detection with Temporal Constraints. IEEE Access (2025). https://doi.org/10.1109/ACCESS.2025.3560767

[33] Huang, C., et al.: Weakly Supervised Video Anomaly Detection via Temporal Discriminative Transformer. IEEE Trans. Cybern. (2022). https://doi.org/10.1109/TCYB.2022.3227044

[34] Sun, W., et al.: Multimodal and Multiscale Feature Fusion for Weakly Supervised Video Anomaly Detection. Scientific Reports 14 (2024). https://doi.org/10.1038/s41598-024-73462-0

[35] Qi, M., Wu, Y.: Weakly Supervised Video Anomaly Detection Based on Hyperbolic Space. Scientific Reports 14 (2024). https://doi.org/10.1038/s41598-024-77505-4

[36] Sun, S., Gong, X.: Event-Driven Weakly Supervised Video Anomaly Detection. Image and Vision Computing 149 (2024). https://doi.org/10.1016/j.imavis.2024.105169

[37] Sharif, M., Jiao, L., Omlin, C.: CNN-ViT Supported Weakly-Supervised Video Segment Level Anomaly Detection. Sensors 23(18) (2023). https://doi.org/10.3390/s23187734

[38] Barbosa, R., Oliveira, H., Tavares, J.: Multi-Modal and Weakly Supervised Approaches for Robust Anomaly Detection in Video Data. Information Fusion 126 (2026). https://doi.org/10.1016/j.inffus.2025.103388

[40] Sanyour, R., Abdullah, M., Abdullah, S.: A Light-Weight Real-Time Anomaly Detection Framework for Edge Computing. In: Lecture Notes in Computer Science, Springer, pp. 423–434 (2023). https://doi.org/10.1007/978-3-031-33743-7_37