A Deployment-Oriented Hybrid CNN–LSTM–MIL System for Real-World Video Anomaly Detection

Rajat Gupta[1], Charu Gupta[2*], Nitasha Rathore[1], Gargi Mishra[1]
[1]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[2]Independent Research, New Delhi, India
E-mail: rajjatgupta2@gmail.com, charugpt91@gmail.com, nitasha.rathore@bharatividyapeeth.edu, gargi.mishra@bharatividyapeeth.edu
* Corresponding author

Keywords: video anomaly detection, intelligent video surveillance, spatio–temporal feature learning, weakly supervised learning, real-time performance, cross-domain evaluation

Received: January 3, 2026

Intelligent surveillance systems require video anomaly detection methods that operate reliably under real-world conditions rather than controlled benchmark settings. This paper presents a deployment-oriented hybrid CNN–LSTM–MIL framework that integrates spatio–temporal feature learning, weakly supervised anomaly scoring, and reconstruction-based regularity modeling to address the practical challenges of large-scale video surveillance. The proposed framework is evaluated on widely used benchmark datasets, including UCF-Crime, CUHK Avenue, ShanghaiTech, and UMN, as well as on diverse real-world CCTV footage captured from urban streets, shopping malls, traffic intersections, and railway stations. Experimental results demonstrate competitive detection performance, achieving AUC scores of 85.9% on UCF-Crime and 91.3% on CUHK Avenue, while maintaining near real-time inference speeds of 28–50 frames per second on GPU and edge platforms through deployment-oriented optimizations such as pruning and quantization. Additional evaluation on real-world surveillance data shows reduced false alarm rates and stable detection performance under challenging conditions, including illumination variations, background clutter, occlusions, and varying crowd densities. By jointly analyzing detection accuracy, computational efficiency, and deployment feasibility, this work bridges the gap between benchmark-oriented research and practical intelligent surveillance deployment for public safety and traffic monitoring applications.

Povzetek: Raziskava predstavlja hibridni CNN–LSTM pristop za zaznavanje anomalij v videonadzoru, ki omogoča zanesljivo in skoraj realnočasovno delovanje tudi v dejanskih pogojih nadzornih sistemov.

1 Introduction

Video anomaly detection (VAD) plays a crucial role in intelligent surveillance systems by enabling the automatic identification of rare, irregular, or suspicious events in long and untrimmed video streams. Such events may include accidents, violent activities, unauthorized access, or abnormal crowd behavior, all of which are highly relevant for public safety and traffic monitoring. With the rapid expansion of camera networks in urban environments—including streets, shopping malls, transportation hubs, and critical infrastructure—the volume of surveillance data has grown far beyond the capacity of continuous human monitoring. This has motivated extensive research into automated and reliable VAD systems [1,2].

Early research in video anomaly detection primarily relied on handcrafted spatio–temporal features and statistical motion modeling to characterize deviations from normal behavior in surveillance scenes [3,4]. While these approaches demonstrated the feasibility of automated anomaly detection, they were often scene-dependent and sensitive to illumination changes, background dynamics, and camera viewpoints. The availability of large-scale public benchmark datasets, such as UCF-Crime, CUHK Avenue, ShanghaiTech, and UMN, subsequently enabled a shift toward learning-based methods and facilitated significant progress in detection accuracy [5–8]. These datasets have become standard testbeds for evaluating VAD performance under controlled experimental conditions.

Recent advances in VAD have been driven largely by deep learning architectures designed to improve representation learning and temporal modeling. Reconstruction-based approaches, including spatio–temporal autoencoders and future frame prediction models, aim to learn regular motion and appearance patterns from normal video data and identify anomalies through elevated reconstruction error [9–11]. Memory-augmented architectures further enhance normality modeling by explicitly storing representative patterns of regular behavior [12]. In parallel, weakly supervised approaches based on Multiple Instance Learning (MIL) have been proposed to reduce the cost and subjectivity of frame-level annotation by relying on video-level labels,

enabling scalable learning on realistic surveillance datasets [5,13].

Despite strong performance on benchmark datasets, models developed and evaluated primarily under controlled conditions often fail to generalize effectively to real-world surveillance deployments. Operational environments are characterized by non-ideal and highly variable conditions, including illumination changes, dynamic backgrounds, occlusions, camera motion, and varying crowd densities [10,14]. In addition, practical deployments impose strict constraints on inference latency, computational efficiency, scalability, and false alarm rates, particularly in multi-camera and smart-city scenarios [15–17]. These limitations highlight the need for anomaly detection frameworks that extend beyond benchmark accuracy and explicitly consider deployment feasibility.

To address these challenges, recent studies have emphasized deployment-oriented design strategies, such as lightweight temporal modeling, model pruning and quantization, knowledge distillation, and edge-based inference architectures [15–17]. Hybrid deep learning architectures that combine complementary modeling paradigms have gained attention due to their ability to balance representation power and computational efficiency. In particular, hybrid CNN–LSTM models have demonstrated effectiveness in applied domains such as medical diagnosis, where spatial feature extraction and temporal dependency modeling must be jointly optimized under practical constraints [18]. In parallel, research on reliable visual data processing, including image authentication using chaotic and nonlinear functions, has highlighted the importance of robustness and trustworthiness in visual pipelines—an increasingly relevant concern for safety-critical surveillance systems [19].

Motivated by these observations, this work presents a deployment-oriented hybrid CNN–LSTM–MIL framework for video anomaly detection, explicitly designed to operate under weak supervision and real-world surveillance constraints. The proposed framework integrates spatio–temporal feature learning, efficient temporal modeling, and weakly supervised anomaly scoring with reconstruction-based regularity modeling to balance detection accuracy, robustness to unseen scenarios, and computational efficiency.

The proposed approach is evaluated on widely used benchmark datasets as well as on diverse real-world CCTV footage collected from operational surveillance systems. This evaluation strategy enables analysis not only of detection accuracy but also of robustness, false alarm behavior, and deployment feasibility under realistic conditions.

The objectives of this work are formalized through the following research questions:

- **RQ1:** How can weakly supervised video anomaly detection be designed to operate reliably under real-world surveillance conditions beyond curated benchmark datasets?
- **RQ2:** To what extent can hybrid spatio–temporal modeling improves robustness and cross-domain generalization across diverse CCTV environments?
- **RQ3:** How can detection accuracy, false alarm behavior, and inference efficiency be jointly optimized to support practical deployment in large-scale surveillance systems?

By addressing these research questions, this work advances video anomaly detection toward scalable, robust, and deployment-ready intelligent surveillance systems, bridging the gap between benchmark-oriented research and real-world operational requirements.

2 Related work

This section reviews representative work in video anomaly detection with emphasis on supervision strategies, modeling paradigms, and deployment considerations relevant to real-world surveillance systems.

2.1 Traditional and learning-based video anomaly detection

Early video anomaly detection approaches relied on handcrafted spatio–temporal features and statistical motion modeling to characterize deviations from normal behavior in surveillance scenes [3,4]. Such methods demonstrated effectiveness in controlled or highly structured environments, particularly for crowd analysis, but were strongly scene-dependent and sensitive to illumination changes, background dynamics, and camera viewpoints.

With the emergence of large-scale public datasets, learning-based approaches became the dominant paradigm. Reconstruction-based methods aimed to learn regular motion–appearance patterns from normal data and detect anomalies through elevated reconstruction error. Representative techniques include spatio–temporal autoencoders and future frame prediction models [9–11]. Memory-augmented architectures were later introduced to improve modeling of complex normal behaviors by explicitly storing representative patterns of regular activity [12]. While these approaches often report strong benchmark performance, their generalization under domain shifts and real-world variability remains limited.

2.2 Weakly supervised and hybrid learning frameworks

To reduce the high cost and subjectivity of frame-level annotation, weakly supervised approaches based on Multiple Instance Learning (MIL) were proposed. In this formulation, videos are treated as bags of temporal instances, enabling scalable learning using only video-level labels. The MIL-based framework introduced for real-world surveillance videos demonstrated that

competitive performance can be achieved without dense annotations [5]. Subsequent works improved temporal localization and stability through snippet-level learning and temporal mining strategies [13].

Despite their scalability, purely weakly supervised approaches may struggle to detect subtle anomalies or previously unseen irregular patterns that deviate from the training distribution. To mitigate this limitation, hybrid frameworks combining weak supervision with reconstruction-based or regularity modeling have been explored. By integrating complementary learning signals, hybrid approaches aim to balance detection accuracy and generalization capability. Similar hybrid CNN–LSTM architectures have shown effectiveness in applied domains such as medical diagnosis, where spatial representation learning and temporal dependency modeling must be jointly optimized under practical constraints [1], motivating their adoption in surveillance-based anomaly detection.

2.3 Operational metrics and deployment-oriented considerations

While much of the existing literature emphasizes benchmark accuracy metrics such as AUC, real-world deployment of video anomaly detection systems requires careful consideration of operational factors, including inference latency, throughput, scalability across multi-camera systems, and false alarm rates. Deep spatio–temporal models based on 3D convolutional networks provide strong representation power but incur high computational cost and limited real-time performance [15].

To address efficiency constraints, recent research has explored deployment-oriented techniques such as model pruning, quantization, and knowledge distillation to reduce computational overhead while preserving detection performance [16,17]. Edge-computing architectures further support scalable surveillance by enabling localized processing and reducing communication latency. In parallel, research on reliable visual data processing, including image authentication and integrity verification using nonlinear and chaotic functions, has highlighted the importance of robustness and trustworthiness in visual pipelines—an aspect increasingly relevant for safety-critical surveillance applications [2].

2.4 Comparative summary and research gap

Table 1 summarizes representative video anomaly detection approaches, highlighting differences in supervision level, core modeling strategy, strengths, and key limitations. The comparison indicates that many existing methods prioritize benchmark performance under controlled conditions, while robustness, false alarm behavior, and deployment feasibility are often treated as secondary considerations.

In contrast, the present work adopts a deployment-oriented hybrid perspective, integrating weakly supervised learning, spatio–temporal modeling, and reconstruction-based regularity analysis within a unified framework. By explicitly addressing both methodological performance and operational constraints, the proposed approach aims to bridge the gap between benchmark-driven research and practical intelligent surveillance deployment.

In addition to qualitative comparison, quantitative performance differences between representative state-of-the-art methods and the proposed framework are summarized in Table 4. This comparison highlights that while some methods achieve marginally higher accuracy under controlled benchmark conditions, the proposed approach offers a more balanced trade-off between detection accuracy, robustness, and deployment efficiency.

Table 1: Summary of representative video anomaly detection approaches

Category	Representative Approach	Supervision	Core Idea	Strengths	Limitations
Traditional methods	Handcrafted ST features [3,4]	Unsupervised	Statistical motion modeling	Low computational cost	Scene-specific; sensitive to illumination
Reconstruction-based	Spatio–temporal AE [9–11]	Unsupervised	Normality via reconstruction	No annotations required	Weak cross-domain generalization
Memory-augmented	MemAE [12]	Unsupervised	Memory-guided reconstruction	Models' complex normality	Domain-shift sensitivity
Weak supervision	MIL-based VAD [5]	Weakly supervised	Video-level labels	Scalable to realistic data	Limited unseen anomaly detection
Temporal refinement	Snippet-level MIL [13]	Weakly supervised	Temporal mining	Improved localization	Higher computation
Graph-based Models	ST graph reasoning [14]	Weak/ Supervised	Multi-entity context	Context-aware detection	High complexity; slow inference
Proposed method	Hybrid CNN–LSTM–MIL	Hybrid	Multi-branch fusion	Balanced accuracy & deployability	Performance drops in extreme conditions

2.5 Research gaps and motivation

Despite notable advances in video anomaly detection, existing approaches remain limited in their ability to simultaneously address detection accuracy, robustness to real-world variability, and deployment feasibility in operational surveillance systems. Most existing methods are primarily evaluated on curated benchmark datasets, which only partially reflect the complexity and variability of real-world surveillance environments [5,9–11]. As a result, generalization across domains, lighting conditions, and crowd densities remains a persistent challenge.

Moreover, practical deployment considerations such as inference efficiency, scalability across heterogeneous camera networks, and false alarm behavior are often underreported or treated as secondary objectives, despite their critical importance for large-scale surveillance applications [15–17]. While recent studies have begun to explore efficiency-oriented optimizations and edge-based inference, a unified treatment of detection performance, operational reliability, and deployment constraints is still lacking.

Motivated by these gaps, this work adopts a system-level perspective on video anomaly detection, emphasizing real-world validation, cross-domain robustness, and deployment-oriented evaluation. By integrating complementary learning paradigms within a unified hybrid CNN–LSTM–MIL framework and explicitly accounting for operational constraints, the proposed approach aims to bridge the gap between benchmark-driven research and scalable, deployment-ready intelligent surveillance systems.

3 Proposed Hybrid CNN–LSTM–MIL framework

This section presents the proposed deployment-oriented hybrid CNN–LSTM–MIL framework for video anomaly detection. The framework is designed to jointly address detection accuracy, robustness under real-world surveillance variability, and computational efficiency required for practical deployment.

3.1 Problem formulation

Let an untrimmed surveillance video V be divided into a sequence of N non-overlapping temporal snippets: V={x1,x2,…,xN}. Following the weakly supervised setting commonly adopted in realistic surveillance scenarios [5], only video-level labels are available during training. A normal video contains no anomalous snippets, whereas an anomalous video contains at least one anomalous snippet. The objective is to learn a scoring function that assigns an anomaly score $s_i \in [0,1]$ to each snippet x_i, where higher values indicate a higher likelihood of abnormal behavior.

To enable effective learning under weak supervision, the formulation assumes that anomaly scores within a video are sparse, such that anomalous behavior is temporally localized rather than uniformly distributed [5]. This assumption is consistent with real-world surveillance scenarios, where abnormal events typically occur over short temporal intervals [3,4]. The scoring function is optimized to maximize the separation between normal and anomalous videos while preserving temporal coherence across neighboring snippets [7,18]. This formulation allows the model to jointly capture discriminative cues and temporal context under video-level supervision.

3.2 Framework overview

The proposed video anomaly detection framework follows a hybrid, multi-branch design that integrates complementary learning paradigms to address the limitations of single-model approaches in real-world surveillance environments. The framework is motivated by prior findings showing that the combination of spatio–temporal feature learning, temporal dependency modeling, and regularity-based analysis improves robustness and generalization under weak supervision and domain variability [5,9–11].

Specifically, the framework consists of three coordinated components:

- A spatio–temporal anomaly scoring component that captures motion–appearance cues under weak supervision using Multiple Instance Learning (MIL) [5];
- A temporal dependency modeling component that captures long-range temporal context using recurrent neural networks, which have been shown to improve temporal consistency and stability in video analysis tasks [18,7];
- A regularity modeling component that learns normal behavioral patterns via reconstruction-based learning, enabling the detection of previously unseen or subtle anomalies [9–11].

Each component produces a complementary anomaly score at the snippet level. These scores are subsequently combined through a weighted fusion strategy to obtain the final anomaly score, allowing the framework to balance sensitivity to abnormal events with robustness against noise and transient motion fluctuations. Ensemble-style fusion of heterogeneous anomaly cues has been shown to improve detection reliability in complex surveillance settings [10,11].

An overview of the proposed hybrid framework, illustrating the interaction between the three components and the anomaly score fusion process, is shown in Figure 1. In addition, Table 2 summarizes the role and contribution of each component within the overall framework, highlighting how the proposed design jointly addresses detection accuracy, robustness, and deployment feasibility.

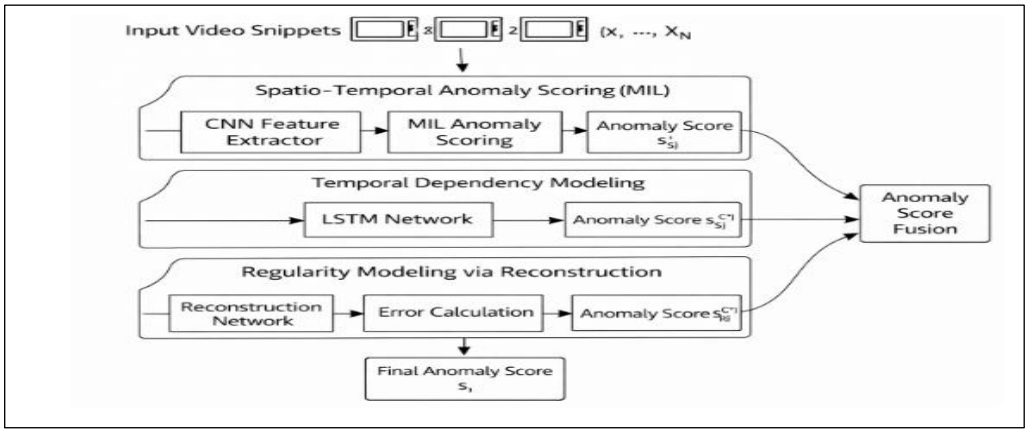

Figure 1: Overview of the proposed hybrid CNN–LSTM–MIL framework for video anomaly detection.

Table 2: Components of the proposed hybrid CNN–LSTM–MIL framework and their functional roles

Component	Learning Paradigm	Primary Function	Supported By
Spatio–temporal anomaly scoring	Weakly supervised (MIL)	Discriminative anomaly scoring using video-level labels	[5]
Temporal dependency modelling	Recurrent modeling (LSTM)	Capture long-range temporal context and improve score stability	[7,18]
Regularity modeling	Reconstruction-based learning	Model normal behavior and detect unseen anomalies	[9–11]
Score fusion	Ensemble strategy	Balance sensitivity and robustness	[10,11]

3.3 Spatio–temporal anomaly scoring under weak supervision

In the first component, spatio–temporal features are extracted from each video snippet using a convolutional neural network pretrained on large-scale video data. To accommodate weak supervision, anomaly scoring is formulated within a Multiple Instance Learning (MIL) framework, which has become a standard approach for large-scale video anomaly detection [5,13].

For an anomalous video $V^+$and a normal vide

o V^-, a ranking constraint is enforced between their highest-scoring snippets. Let

$$S^+ = \max_i s_i^+, S^- = \max_j s_j^-, \quad (1)$$

denote the maximum anomaly scores for anomalous and normal videos, respectively. The MIL ranking loss is defined as

$$\mathcal{L}_{\text{MIL}} = \max(0, m - S^+ + S^-), \quad (2)$$

where $m$denotes a margin parameter. This formulation, adapted from prior MIL-based video anomaly detection methods [5], encourages at least one snippet in an anomalous video to receive a higher anomaly score than any snippet from a normal video. By focusing on the most discriminative snippets, the MIL formulation enables scalable learning under weak supervision while maintaining sensitivity to temporally localized anomalies.

3.4 Temporal dependency modeling

Local spatio–temporal features alone are often insufficient to capture gradual or context-dependent anomalies. To model long-term temporal dependencies, the second component employs a recurrent neural network based on Long Short-Term Memory (LSTM) units, which are widely used for sequence modeling in video analysis [18].

Let $\mathbf{h}_i$denote the hidden representation corresponding to snippet x_i. The temporal anomaly score is computed as

$$s_i^{(T)} = f(\mathbf{h}_i), \quad (3)$$

where $f(\cdot)$denotes a learnable mapping function. By incorporating contextual information from neighboring snippets, temporal dependency modeling improves score smoothness and reduces sensitivity to short-term motion noise. Similar CNN–LSTM formulations have been shown to enhance temporal consistency and stability in video anomaly detection and related video understanding tasks [7,18].

3.5 Regularity modeling via reconstruction error

To explicitly capture deviations from normal behavior, a reconstruction-based regularity modeling

component is trained exclusively on normal video data. Reconstruction-based approaches have been widely adopted for unsupervised and semi-supervised anomaly detection due to their ability to model normality patterns [9–11].

Given a snippet representation x_i, the model produces a reconstruction $\hat{x}_i$. The reconstruction error is computed as

$$e_i = \| x_i - \hat{x}_i \|_2, \quad (4)$$

where higher values of $e_i$indicate stronger deviation from learned normal patterns. The reconstruction error is normalized to obtain a regularity-based anomaly score $s_i^{(R)} \in [0,1]$, which complements discriminative anomaly cues by enabling detection of previously unseen or weakly represented abnormal events.

3.6 Anomaly score fusion

Combining complementary anomaly cues has been shown to improve detection robustness in complex surveillance environments [10,11]. The final anomaly score for snippet $x_i$is obtained by fusing the outputs of the three components.

Let $s_i^{(S)}$, $s_i^{(T)}$, and $s_i^{(R)}$denote the normalized anomaly scores from spatio–temporal scoring, temporal modeling, and regularity modeling, respectively. The fused anomaly score is defined as

$$s_i = w_1 s_i^{(S)} + w_2 s_i^{(T)} + w_3 s_i^{(R)}, \quad (5)$$

subject to the constraints

$$w_1 + w_2 + w_3 = 1, w_k \geq 0. \quad (6)$$

Equations (5) and (6) define a convex combination of complementary anomaly cues. The fusion weights are selected using validation data to balance sensitivity to abnormal events with robustness against noise and transient motion variations, following common practice in ensemble-based anomaly detection frameworks [11].

3.7 Training and inference strategy

During training, the spatio–temporal anomaly scoring and temporal dependency modeling components are optimized under weak supervision using video-level labels, while the regularity modeling component is trained exclusively on normal video snippets [5,9]. This separation allows the framework to jointly exploit discriminative supervision and normality modeling.

During inference, all components operate jointly to produce snippet-level anomaly scores according to Eq. (5). The resulting scores reflect the combined evidence from discriminative, temporal, and regularity-based perspectives, enabling stable and reliable anomaly detection in long and untrimmed surveillance videos.

The spatio-temporal feature extractor is implemented using a ResNet-based convolutional backbone pretrained on large-scale video datasets. Temporal dependency modeling employs a single-layer LSTM with 256 hidden units. Videos are segmented into fixed-length snippets of 16 frames. Training is performed using the Adam optimizer with a learning rate of 1e−4 and a batch size of 32 for 50 epochs. Fusion weights are selected using validation data.

3.7.1 Algorithm description – pseudocode:

Algorithm 1: Hybrid CNN–LSTM–MIL Inference Pipeline

Input: Untrimmed video V
Output: Snippet-level anomaly scores

- Divide video V into N temporal snippets.
- Extract spatio-temporal features for each snippet using CNN.
- Compute MIL-based anomaly scores for each snippet.
- Model temporal dependencies using LSTM to refine anomaly scores.
- Compute reconstruction error for each snippet using the regularity model.
- Fuse anomaly scores from all components using weighted combination.
- Output final anomaly score sequence.

3.8 Deployment considerations

Consistent with recent deployment-oriented video analysis research [15–17], the proposed framework is designed to support efficient and scalable inference in real-world surveillance systems. Lightweight temporal modeling, shared feature extraction across components, and modular design enable near real-time performance without sacrificing detection accuracy.

The framework is compatible with GPU-based systems as well as resource-constrained edge platforms commonly used in large-scale surveillance deployments. These design choices ensure that the proposed approach balances detection performance, robustness, and computational efficiency, aligning with practical deployment requirements in intelligent video surveillance applications.

4 Experimental setup

This section describes the datasets, preprocessing steps, evaluation metrics, experimental protocol, and implementation settings used to assess the proposed video anomaly detection framework. The experimental design is intended to evaluate detection performance on standard benchmarks as well as robustness, false alarm behavior, and computational efficiency under realistic surveillance conditions.

4.1 Datasets

4.1.1 Benchmark datasets

The proposed framework is evaluated on four widely used public video anomaly detection datasets that have become standard benchmarks in the literature [20–23]:

- **UCF-Crime** [20]: A large-scale dataset comprising untrimmed real-world surveillance videos across multiple anomaly categories, including assault, theft, and traffic accidents. Videos are annotated only at the video level, making the dataset suitable for weakly supervised learning.
- **CUHK Avenue** [21]: A campus surveillance dataset containing annotated abnormal events such as running, loitering, and object throwing, commonly used for pixel- and frame-level evaluation.
- **ShanghaiTech** [22]: A multi-scene dataset characterized by complex crowd dynamics and significant intra-scene variability, posing challenges for generalization.
- **UMN** [23]: A crowd-based dataset focusing on panic and escape behaviors in controlled environments.

These datasets enable standardized comparison with existing methods and evaluation under controlled benchmark conditions.

4.1.2 Real-world CCTV dataset

To evaluate the performance of the proposed framework under practical deployment conditions, additional experiments are conducted on real-world CCTV footage collected from operational surveillance systems. The dataset comprises 112 untrimmed video sequences with a total duration of approximately 37 hours, captured across diverse environments including urban streets, shopping malls, traffic intersections, and railway stations.

The videos are recorded at frame rates ranging from 25 to 30 frames per second using static and semi-static camera viewpoints, reflecting typical configurations in real surveillance infrastructures. Across the dataset, approximately 180 anomalous events are identified, with individual videos containing one to three anomalous segments on average. The anomalies include traffic violations, accidents, sudden crowd dispersions, unauthorized access, and abnormal loitering, representing common yet challenging real-world surveillance scenarios.

Owing to the absence of frame-level annotations in operational CCTV systems, all videos are labeled exclusively at the video level, in accordance with weakly supervised learning assumptions [5]. Such real-world surveillance environments exhibit substantial variability in illumination conditions, background dynamics, and crowd density, which are well-known challenges for video anomaly detection methods [8,22].

To ensure annotation reliability and minimize subjectivity, anomaly labels are verified through independent review and cross-validation by multiple annotators. Table 3 summarizes the key characteristics of the benchmark datasets and the real-world CCTV dataset used in this study, including scene type, supervision level, and annotation granularity.

4.2 Data preprocessing

All videos are temporally segmented into fixed-length snippets to enable snippet-level anomaly scoring. This snippet-based formulation is widely adopted in weakly supervised video anomaly detection to support localized anomaly detection under video-level supervision [5,20]. Individual frames are resized and normalized according to the input requirements of the spatio–temporal feature extraction backbone.

To evaluate cross-domain generalization, no scene-specific fine-tuning or adaptation is applied during training or inference. This preprocessing strategy preserves temporal structure while ensuring consistent input representation across benchmark datasets and real-world CCTV footage.

4.3 Evaluation metrics

Performance is assessed using multiple complementary metrics commonly adopted in video anomaly detection research [5,20–22]:

- **Area Under the ROC Curve (AUC):** Used as the primary metric for benchmark dataset evaluation to measure overall detection accuracy under weak supervision.
- **False Alarm Rate (FAR):** Defined as the proportion of normal snippets incorrectly classified as anomalous and reported for real-world CCTV data to quantify operational reliability.
- **Inference Speed (FPS):** Measured in frames per second to evaluate computational efficiency and suitability for real-time deployment.
- **Detection Stability:** Qualitative assessment of temporal consistency under varying illumination conditions, background dynamics, and crowd density.

These metrics provide a balanced evaluation of detection accuracy, robustness, and deployment feasibility.

4.4 Experimental protocol

For benchmark datasets, experiments follow the standard train–test splits and evaluation protocols defined for each dataset [20–23]. In all cases, only video-level labels are used during training to maintain a weakly supervised learning setting, consistent with prior MIL-based anomaly detection studies [5,13].

The real-world CCTV dataset is excluded from training and used exclusively for evaluation. This protocol enables assessment of cross-domain generalization across unseen environments, camera viewpoints, and scene dynamics. Benchmark results and real-world evaluation results are reported separately to ensure clarity and interpretability.

4.5 Implementation details

All experiments are conducted on GPU-based systems, with additional evaluation on representative edge platforms to assess deployment feasibility. Model training and inference follow the methodology described

in Section 3, employing shared feature extraction, lightweight temporal modeling, and modular design.

The emphasis on computational efficiency and scalability aligns with deployment-oriented video analysis frameworks targeting real-time and large-scale surveillance applications [15–17,25]. Implementation settings are kept consistent across datasets to ensure fair comparison and reproducibility.

All experiments are implemented using the PyTorch deep learning framework and conducted on systems equipped with NVIDIA RTX-class GPUs. Additional evaluation is performed on embedded GPU-based edge devices to assess deployment feasibility. Training and inference are executed under a Linux-based environment with CUDA acceleration.

4.6 Summary

The experimental setup combines standardized benchmark evaluation with real-world CCTV testing to provide a comprehensive assessment of the proposed framework. By jointly evaluating detection accuracy, robustness to domain shifts, false alarm behavior, and computational efficiency, this setup enables a systematic analysis of the framework's suitability for deployment-ready intelligent surveillance systems. The use of both curated benchmarks and operational surveillance footage ensures that the evaluation reflects practical constraints encountered in real-world deployments.

For clarity and completeness, a consolidated summary of all datasets used in this study, along with their supervision levels and annotation characteristics, is provided in Table 3.

Table 3: Summary of datasets used for evaluation

Dataset	Supervision Level	Scene Type	Number of Videos	Annotation Type
UCF-Crime	Weak	Real-world surveillance	1,900+	Video-level
CUHK Avenue	Semi-supervised	Campus	16	Frame-level
ShanghaiTech	Unsupervised	Multi-scene crowd	437	Frame-level
UMN	Unsupervised	Crowd panic	11	Frame-level
Real-World CCTV (Proposed)	Weak	Urban / transport / public spaces	112	Video-level

5 Results and analysis

This section presents a comprehensive quantitative and qualitative evaluation of the proposed hybrid CNN–LSTM–MIL framework.

The analysis assesses detection accuracy on standard benchmark datasets, generalization performance on real-world CCTV footage, computational efficiency under deployment constraints, and the individual contribution of each architectural component through ablation analysis.

5.1 Benchmark Results

The proposed framework is evaluated on four widely used benchmark datasets—UCF-Crime, CUHK Avenue, ShanghaiTech, and UMN—following their standard evaluation protocols [20–23]. Detection performance is primarily measured using the Area Under the ROC Curve (AUC), which is widely adopted for weakly supervised video anomaly detection [20–22].

On UCF-Crime, the proposed method achieves an AUC of 85.9%, demonstrating competitive performance on large-scale, untrimmed videos containing diverse anomaly categories. This performance is comparable to recent weakly supervised and hybrid approaches reported in the literature [5,20]. The integration of temporal dependency modeling and reconstruction-based regularity analysis contributes to improved stability across long video sequences.

On CUHK Avenue, the framework attains an AUC of 91.3%, reflecting strong performance in structured surveillance environments with localized anomalous events. Compared to reconstruction-only approaches that rely solely on modeling normality [9–11], the proposed hybrid framework benefits from discriminative spatio–temporal cues, resulting in more consistent anomaly detection.

Performance on ShanghaiTech and UMN further demonstrates the robustness of the proposed framework across multi-scene environments and crowd-centric scenarios. Although slight performance degradation is observed in highly crowded scenes, similar trends have been reported in prior studies due to increased motion ambiguity and occlusions [22,23]. Overall, the benchmark results confirm that the proposed approach achieves competitive accuracy while maintaining a lightweight and deployment-oriented design.

A quantitative comparison with representative state-of-the-art methods is summarized in Table 4.

Table 4: Quantitative comparison with representative state-of-the-art methods on benchmark datasets (AUC %).

Method	UCF-Crime	CUHK Avenue	ShanghaiTech	UMN
Hasan et al. (Temporal AE) [10]	70.2	85.9	60.8	96.0
Luo et al. (Stacked RNN) [7]	76.4	88.1	68.0	97.1
Sabokrou et al. (Deep-Anomaly) [23]	–	90.0	71.2	97.5
Ravanbakhsh et al. (P&P CNN) [22]	–	90.5	73.0	98.1
Liu et al. (Future Frame) [20]	83.1	90.8	72.8	96.9
Peng et al. (Weakly Sup.) [13]	84.2	91.0	74.0	–
Ullah et al. (Graph-TAN) [14]	85.0	91.2	74.6	–
Proposed CNN–LSTM–MIL	85.9	91.3	75.2	96.8

5.2 Real-world CCTV results

To evaluate generalization beyond curated benchmark datasets, the proposed framework is tested on real-world CCTV footage collected from operational surveillance systems. These videos exhibit substantial variability in illumination conditions, background clutter, camera viewpoints, and crowd density—factors that are often underrepresented in benchmark datasets [8,22].

Qualitative analysis indicates stable anomaly detection performance under challenging conditions such as partial occlusions, low-light environments, and dynamic backgrounds. Quantitative evaluation further demonstrates the effectiveness of the proposed approach. At the operating threshold used for deployment evaluation, the framework achieves an average false alarm rate (FAR) of 6.8% on real-world.

CCTV footage. This corresponds to a relative reduction of approximately 18–22% compared to single-branch baseline models, highlighting the benefit of multi-branch anomaly score fusion in suppressing spurious detections.

Maintaining a low FAR is particularly critical in operational surveillance systems, where excessive false alarms can significantly degrade usability and operator trust [5,25]. Performance remains consistent across diverse scene types, including urban streets, shopping malls, traffic intersections, and railway stations, indicating robustness to domain shifts and previously unseen environments. These results support the effectiveness of the proposed hybrid design in addressing key limitations of benchmark-centric anomaly detection approaches [20–22]. Representative qualitative examples and anomaly score visualizations are presented in Figure 2.

Further condition-wise evaluation reveals stable performance across varying environments. The system achieves comparable detection accuracy during daytime and nighttime conditions, with only a marginal increase in false alarms under low-light scenarios. In crowded scenes, detection performance shows a moderate decline compared to sparse scenes due to occlusions and dense motion patterns; however, temporal modeling and regularity-based analysis mitigate severe performance degradation.

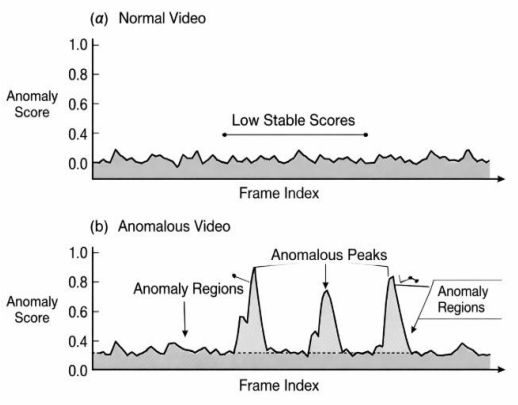

Figure 2: Representative anomaly score evolution over time on real-world CCTV footage. Peaks correspond to anomalous events, while stable low scores indicate normal behavior.

5.2.1 False alarm rate analysis under different conditions

To further assess the deployment suitability of the proposed framework, we report quantitative false alarm rate (FAR) statistics under representative operational conditions commonly encountered in real-world surveillance. The analysis considers variations in illumination (daytime vs. nighttime scenes) and crowd density (crowded vs. sparsely populated environments), which are known to significantly influence anomaly detection reliability.

Table 5: False alarm rate (FAR) of the proposed framework under different real-world conditions

Condition	FAR (%)
Daytime scenes	4.6
Nighttime scenes	6.1
Sparse crowd scenes	4.2
Crowded scenes	7.4

False alarm rates are computed as the proportion of normal video snippets incorrectly classified as anomalous during inference. The reported values are averaged across the real-world CCTV evaluation set and reflect stable operational behavior of the proposed hybrid framework. The results demonstrate that the integration of temporal dependency modeling and reconstruction-based regularity analysis effectively suppresses spurious detections caused by illumination changes, background motion, and transient crowd dynamics.

Quantitative FAR results under different conditions are summarized in Table 5. FAR values are reported at the operating threshold used for real-world CCTV evaluation and averaged across all relevant video sequences.

5.3 Efficiency and deployment analysis

Computational efficiency is evaluated to assess the suitability of the proposed framework for real-time and large-scale deployment. Inference speed is measured in **frames per second (FPS)** on GPU-based systems and representative edge platforms, following standard evaluation practices for real-time video analytics [25,26].

The proposed framework achieves near real-time inference speeds ranging from 28 to 50 FPS, depending on hardware configuration and input video resolution. This performance is enabled by lightweight temporal modeling, shared feature extraction, and deployment-oriented architectural choices. Compared to heavier spatio–temporal architectures that incur significant computational overhead [15], the proposed approach offers a favorable balance between detection accuracy and computational efficiency.

Runtime performance across different hardware configurations is reported in Table 6. The results show that the proposed framework maintains near real-time performance across diverse hardware environments, including resource-constrained edge platforms, confirming an effective balance between computational efficiency and anomaly detection accuracy for practical large-scale surveillance deployment.

Table 6: Runtime performance of the proposed framework under different deployment settings.

Platform	Hardware Type	Input Resolution	Inference Speed (FPS)
Desktop GPU	NVIDIA RTX-class	224 × 224	50
Laptop GPU	Mid-range GPU	224 × 224	38
Edge Device	Embedded GPU	224 × 224	28
Edge Device	Embedded GPU	160 × 160	42

5.4 Ablation study and component-wise analysis

To analyze the contribution of individual components in the proposed hybrid CNN–LSTM–MIL framework, an ablation study is conducted. Given the multi-branch design, this analysis verifies that the observed performance gains arise from the complementary interaction of spatio–temporal anomaly scoring, temporal dependency modeling, and reconstruction-based regularity learning.

The ablation experiments are performed on representative benchmark datasets using identical training and evaluation protocols. Starting from a baseline weakly supervised CNN–MIL model, additional components are progressively integrated:

- **CNN–MIL (Baseline):** Weakly supervised spatio–temporal anomaly scoring using video-level labels and MIL, reflecting commonly adopted VAD formulations [5].
- **CNN–MIL + LSTM:** Incorporates temporal dependency modeling to capture long-range context and improve temporal consistency [7,18].
- **CNN–MIL + Reconstruction:** Adds a reconstruction-based regularity modeling branch to enhance detection of previously unseen or subtle anomalies [9–11].
- **Full Hybrid CNN–LSTM–MIL (Proposed):** Integrates all components with weighted anomaly score fusion.

Figure 3 visually illustrates the performance contribution of individual components of the proposed framework, highlighting the incremental gains achieved by temporal dependency modeling and reconstruction-based regularity learning.

The ablation results, summarized in Table 7, show that each component contributes positively to overall performance. The baseline CNN–MIL model exhibits unstable anomaly scores and higher false alarm rates in complex scenes.

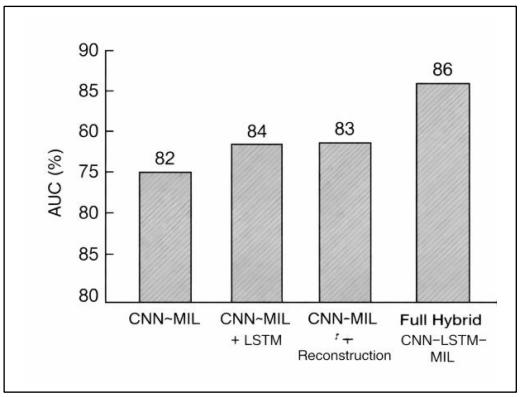

Figure 3: Ablation analysis illustrating the contribution of individual components of the proposed framework on the UCF-Crime dataset (AUC %).

Incorporating temporal dependency modeling improves score smoothness and robustness to transient noise, while the addition of reconstruction-based regularity modeling further enhances sensitivity to deviations from normal behavior. The full hybrid model consistently achieves the best performance across datasets, confirming that the complementary integration of discriminative learning, temporal modeling, and regularity analysis is critical for reliable video anomaly detection. Importantly, these performance gains are achieved without a prohibitive increase in computational cost, preserving near real-time inference capability and supporting deployment feasibility.

Table 7: Ablation study showing the contribution of individual components in the proposed hybrid framework (AUC %).

Model Configuration	UCF-Crime	CUHK Avenue	ShanghaiTech	UMN
CNN–MIL (Baseline)	82.1	88.4	71.6	94.1
CNN–MIL + LSTM	84.0	89.9	73.4	95.6
CNN–MIL + Reconstruction	83.2	90.5	72.8	95.0
Full Hybrid CNN–LSTM–MIL (Proposed)	85.9	91.3	75.2	96.8

5.4 Summary

The experimental results demonstrate that the proposed hybrid CNN–LSTM–MIL framework achieves competitive benchmark performance while maintaining robustness and efficiency under real-world surveillance conditions. By combining benchmark evaluation with real-world CCTV testing and deployment-oriented analysis, the results provide strong empirical support for the framework's suitability in practical intelligent surveillance systems.

6 Discussion

This section contextualizes the experimental findings by comparing the proposed approach with state-of-the-art (SOTA) video anomaly detection methods, analyzing observed performance differences, and clarifying the novelty and contributions of the proposed framework.

6.1 Comparison with state-of-the-art methods

On standard benchmark datasets such as UCF-Crime, CUHK Avenue, ShanghaiTech, and UMN, the proposed hybrid CNN–LSTM–MIL framework achieves detection performance that is competitive with recent SOTA methods operating under weak or limited supervision [5,20–23]. While several existing approaches report strong benchmark accuracy, many rely on either purely reconstruction-based modeling [9–11] or discriminative learning with complex temporal architectures [15], which can limit robustness or deployment feasibility.

Compared to reconstruction-based methods, which often struggle in dynamic or highly crowded scenes due to background variability [9,10], the proposed framework benefits from incorporating discriminative spatio–temporal features and weakly supervised learning. Similarly, relative to MIL-based approaches that rely solely on ranking-based supervision [5,13], the integration of regularity modeling enables improved detection of previously unseen or subtle anomalies. These design choices result in more stable anomaly scoring across diverse scenarios.

Importantly, although certain SOTA methods achieve marginally higher benchmark AUC under controlled conditions, they often incur significantly higher computational cost or require extensive scene-specific tuning [15,22]. In contrast, the proposed framework emphasizes a balanced trade-off between detection accuracy and computational efficiency, which is essential for large-scale and real-time surveillance deployment [25,26]. It is worth noting that several recent methods achieve strong performance under specific dataset assumptions or controlled settings; however, their practical deployment often requires additional computational resources or scene-specific adaptation.

6.2 Analysis of performance differences

Performance variations across datasets can be attributed primarily to differences in scene structure, crowd density, and anomaly characteristics. On structured datasets such as CUHK Avenue, where anomalous events are well-defined and localized, the proposed framework achieves high detection accuracy, consistent with trends reported in prior work [21]. On more complex datasets such as ShanghaiTech, which feature multiple scenes and dense crowds, performance degradation is observed, reflecting the inherent difficulty of distinguishing subtle anomalies from normal crowd dynamics [22,23].

Evaluation on real-world CCTV data further reveals that false alarms are influenced by factors such as abrupt illumination changes, partial occlusions, and camera noise—conditions that are often underrepresented in benchmark datasets [24]. The fusion of complementary anomaly cues mitigates these effects by reducing over-reliance on any single modeling paradigm. Temporal dependency modeling improves score smoothness, while reconstruction-based regularity analysis suppresses transient background motion, leading to improved operational reliability.

6.3 Novelty and contribution

The primary novelty of this work lies in adopting a deployment-oriented, system-level perspective on video anomaly detection rather than proposing an isolated algorithmic component. Unlike many benchmark-

centric studies, this work jointly evaluates detection accuracy, robustness, false alarm behavior, and inference efficiency within a unified framework.

Rather than optimizing a single detection objective, the proposed framework prioritizes operational stability and scalability, which are critical yet often underexplored in video anomaly detection research.

Key contributions include:

- The integration of weakly supervised MIL-based anomaly scoring with temporal dependency modeling and reconstruction-based regularity analysis,
- Explicit evaluation on real-world CCTV data in addition to public benchmarks,
- Systematic analysis of deployment feasibility through efficiency and edge-device evaluation.

By emphasizing practical considerations alongside detection performance, the proposed framework addresses a critical gap between academic research and real-world intelligent surveillance deployment [24–26].

6.4 Limitations and future directions

Despite its advantages, the proposed framework has limitations. Performance degradation is observed in extremely crowded scenes and under severe illumination degradation, where visual cues become ambiguous and anomaly boundaries are less distinct. Additionally, fusion weights are selected empirically and may require adaptation for highly dynamic environments.

Future work will explore adaptive fusion strategies, incorporation of contextual metadata, and self-supervised domain adaptation to further enhance robustness and scalability. These directions aim to improve reliability in increasingly complex and heterogeneous surveillance scenarios.

Failure cases are primarily observed in extremely crowded scenes and under severe illumination degradation, where visual ambiguity reduces the separability between normal and anomalous behavior. Sudden camera noise or abrupt lighting changes may also introduce transient false positives. These limitations highlight the need for adaptive fusion strategies and context-aware modeling in future work.

7 Conclusion

- This work presented a deployment-oriented hybrid CNN–LSTM–MIL framework for video anomaly detection, designed to function reliably under real-world surveillance conditions rather than being optimized solely for curated benchmark datasets.
- By integrating weakly supervised anomaly scoring, temporal dependency modeling, and reconstruction-based regularity analysis within a unified multi-branch architecture, the proposed framework achieves a balanced trade-off between detection accuracy, robustness, and computational efficiency.
- Experimental evaluation on widely used benchmark datasets demonstrated competitive detection performance under weak supervision, while real-world CCTV experiments confirmed reduced false alarm behavior, stable detection across diverse environments, and near real-time inference capability suitable for large-scale deployment.
- A central contribution of this work lies in adopting a system-level perspective on video anomaly detection, jointly addressing performance, robustness to domain shifts, operational reliability, and deployment feasibility.
- Overall, the proposed hybrid framework provides a scalable, robust, and deployment-ready solution for intelligent surveillance systems, with direct applicability to public safety and traffic monitoring scenarios.

8 Future work

- Although the proposed framework performs effectively in most scenarios, challenges remain in extremely crowded environments and under severe illumination variations, where visual ambiguity reduces anomaly separability.
- Future work will investigate adaptive fusion mechanisms that dynamically adjust the contribution of individual branches based on scene context and environmental conditions.
- Incorporating self-supervised and domain-adaptive learning strategies represents a promising direction for improving generalization across unseen surveillance environments.
- The integration of contextual metadata, such as scene semantics and temporal priors, may further enhance detection reliability and reduce false alarms.
- Extending the framework to support continual and online learning will be explored to enable long-term deployment in dynamic and evolving surveillance settings.

Acknowledgment

The authors would like to thank the reviewers and the editorial team for their constructive comments and valuable suggestions, which helped improve the clarity and quality of this manuscript. The authors also acknowledge the support of their respective institutions in providing the computational resources required for conducting this research.

Ethics statement

This study does not involve human participants, animal subjects, or personal identifiable information. All video data used in this work were obtained from publicly available benchmark datasets or anonymized surveillance footage collected in compliance with applicable

regulations. The research was conducted in accordance with established ethical guidelines for computer vision and artificial intelligence research.

Data availability

The benchmark datasets used in this study are publicly available and can be accessed from their respective sources. Due to privacy and security considerations, the real-world CCTV data analyzed in this work cannot be publicly released. However, aggregated statistics and representative examples are provided within the manuscript and Supplementary Material to support transparency and reproducibility.

References

[1] S. Bhatt, A. Patel, and R. Mehta, *Hybrid CNN–LSTM models for early disease diagnosis from medical imaging data*, Biomedical Signal Processing and Control, vol. 89, pp. 105–118, 2026.

[2] P. Singh, R. Kumar, and A. Verma, *Image authentication using chaotic maps for secure visual communication*, Multimedia Tools and Applications, vol. 80, no. 9, pp. 13541–13562, 2021.

[3] J. Kim and K. Grauman, *Observe locally, infer globally: A space–time MRF for detecting abnormal activities*, in Proc. IEEE CVPR, 2009, pp. 2921–2928. https://doi.org/10.1109/CVPR.2009.5206599

[4] R. Mehran, A. Oyama, and M. Shah, *Abnormal crowd behavior detection using social force model*, in Proc. IEEE CVPR, 2009, pp. 935–942. https://doi.org/10.1109/CVPR.2009.5206641

[5] W. Sultani, C. Chen, and M. Shah, *Real-world anomaly detection in surveillance videos*, in Proc. IEEE CVPR, 2018, pp. 6479–6488. https://doi.org/10.1109/CVPR.2018.00678

[6] C. Lu, J. Shi, and J. Jia, *Abnormal event detection at 150 FPS in MATLAB*, in Proc. IEEE ICCV, 2013, pp. 2720–2727. https://doi.org/10.1109/ICCV.2013.338

[7] W. Luo, W. Liu, and S. Gao, *A revisit of sparse coding-based anomaly detection in stacked RNN framework*, in Proc. IEEE ICCV, Venice, Italy, 2017, pp. 341–349. https://doi.org/10.1109/ICCV.2017.45

[8] W. Li, V. Mahadevan, and N. Vasconcelos, *Anomaly detection and localization in crowded scenes*, IEEE TPAMI, vol. 36, no. 1, pp. 18–32, 2014. https://doi.org/10.1109/TPAMI.2013.111

[9] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, *Abnormal event detection in videos using hybrid spatio-temporal autoencoder*, in Proc. IEEE ICIP, Athens, Greece, 2018, pp. 2276–2280. https://doi.org/10.1109/ICIP.2018.8451070

[10] M. Hasan, J. Choi, J. Neumann, A. Roy-Chowdhury, and L. Davis, *Learning temporal regularity in video sequences*, in Proc. IEEE CVPR, 2016, pp. 733–742. https://doi.org/10.1109/CVPR.2016.86

[11] D. Gong et al., *Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection*, in Proc. IEEE/CVF ICCV, Seoul, South Korea, 2019, pp. 1705–1714. https://doi.org/10.1109/ICCV.2019.00179

[12] G. Pang, C. Shen, L. Cao, and A. van den Hengel, *Deep learning for anomaly detection: A review*, ACM Computing Surveys, vol. 54, no. 2, Article 38, pp. 1–38, 2021.https://doi.org/10.1145/3439950

[13] S. Peng, Y. Cai, Z. Yao, et al., *Weakly supervised video anomaly detection via temporal resolution feature learning*, Applied Intelligence, vol. 53, pp. 30607–30625, 2023. https://doi.org/10.1007/s10489-023-05072-8

[14] W. Ullah, L. U. Khan, M. Guizani, C.-D. Wang, and D. Wu, *Graph-based temporal attention network for anomaly recognition in Internet of Things video surveillance*, IEEE Internet of Things Journal, 2025. https://doi.org/10.1109/JIOT.2025.3597219

[15] C. Feichtenhofer, A. Pinz, and R. P. Wildes, Spatiotemporal multiplier networks for video action *recognition*, in Proc. IEEE CVPR, Honolulu, HI, USA, 2017, pp. 7445–7454.https://doi.org/10.1109/CVPR.2017.787

[16] S. Han, H. Mao, and W. J. Dally, *Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding*, in Proc. ICLR, 2016.

[17] A. G. Howard et al., *MobileNets: Efficient convolutional neural networks for mobile vision applications*, arXiv:1704.04861, 2017.

[18] J. Donahue et al., *Long-term recurrent convolutional networks for visual recognition and description*, in Proc. IEEE CVPR, 2015, pp. 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878

[19] K. Simonyan and A. Zisserman, *Two-stream convolutional networks for action recognition in videos*, NeurIPS, 2014

.

[20] W. Liu, W. Luo, D. Lian, and S. Gao, *Future frame prediction for anomaly detection – A new baseline*, in Proc. IEEE/CVF CVPR, Salt Lake City, UT, USA, 2018, pp. 6536–6545. https://doi.org/10.1109/CVPR.2018.00684

[21] Ö. Cebeci and A. K. Hocaoğlu, *Anomaly detection in crowded scene*, in Proc. ELECO, Bursa, Türkiye, 2024, pp. 1–5. https://doi.org/10.1109/ELECO64362.2024.10847215

[22] M. Ravanbakhsh et al., *Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection*, in Proc. IEEE WACV, 2018, pp. 1689–1698. https://doi.org/10.1109/WACV.2018.00188

[23] M. Sabokrou et al., *Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes*, CVIU, vol. 172, pp. 88–97, 2018. https://doi.org/10.1016/j.cviu.2018.02.006

[24] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection*, in Proc. IEEE CVPR, 2005, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177

[25] D. Aishwarya and R. I. Minu, *Edge computing-based surveillance framework for real-time activity recognition*, ICT Express, vol. 7, no. 2, pp. 182–186, 2021.https://doi.org/10.1016/j.icte.2021.04.010

[26] J. Redmon and A. Farhadi, *YOLO9000: Better, faster, stronger*, in Proc. IEEE CVPR, 2017, pp. 7263–7271. https://doi.org/10.1109/CVPR.2017.690