Scilight

Artificial Intelligence and Emerging Technologies
https://www.sciltp.com/journals/aiet

Review

Intelligent Video Surveillance: A Systematic Review of Deep Learning Architectures, Multimodal Fusion, and the Emerging Role of Conversational AI (2018–2025)

Mittapally Sushmitha *, Nimmala Pravalika, Kommaraju Laasya and Kadiyala Ramana *

Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad 500075, India
* Correspondence: sushmithamittapally54@gmail.com (M.S.);ramana.it01@gmail.com (K.R.)

How To Cite: Sushmitha, M.; Pravalika, N.; Laasya, K.; et al. Intelligent Video Surveillance: A Systematic Review of Deep Learning Architectures, Multimodal Fusion, and the Emerging Role of Conversational AI (2018–2025). *Artificial Intelligence and Emerging Technologies* **2026**, *3*(2), 8. https://doi.org/10.53941/aiet.2026.100008

Received: 3 February 2026
Revised: 4 June 2026
Accepted: 22 June 2026
Published: 3 July 2026

Abstract: The rapid proliferation of urban surveillance infrastructure and the exponential growth of large-scale video data have intensified demand for automated, adaptive monitoring solutions. Recent deep learning advances have transformed conventional surveillance from rule-based systems into adaptive, context-aware frameworks capable of complex spatiotemporal activity detection, classification, and interpretation . This paper presents a comprehensive review of seventy-three deep-learning-based research studies on video anomaly detection (VAD) and human activity recognition published between 2018 and 2025. A systematic categorization of the surveyed works is performed with respect to ten major model families: CNNs for spatial feature extraction recurrent architectures (LSTM, GRU, Bi-LSTM) for temporal reasoning; 3D-CNN and spatiotemporal models for motion encoding autoencoder and generative adversarial frameworks for unsupervised reconstruction transformer and attention-based models for long-range dependency modeling memory-augmented networks for prototype-constrained normality learning multimodal fusion architectures and edge-intelligent and conversational AI systems for scalable, interactive deployment. The results demonstrate a performance evolution from early CNN-based classifiers (around 85% AUC) to recent transformer-driven and memory-augmented methods achieving AUC values above 97% on UCF-Crime, ShanghaiTech, CUHK Avenue, and RWF-2000. The review additionally incorporates the MSAD multi-scenario benchmark and the CUVA causation dataset, and provides a methodological caveat on the comparability of AUC scores across heterogeneous supervision paradigms. Key open challenges—dataset imbalance, occlusion, illumination variation, domain generalization, and real-time latency—are mapped to research directions including weakly supervised MLLMs, privacy-preserving federated learning, and edge-optimized transformer pipelines.

Keywords: video surveillance; deep learning; anomaly detection; convolutional neural networks; recurrent networks; autoencoders; transformers; multimodal fusion; conversational AI; federated learning; smart city

1. Introduction

Intelligent video surveillance has evolved from analog CCTV monitored by human operators into a complex ecosystem of AI-driven systems capable of spatiotemporal reasoning and autonomous event understanding [1]. Manual observation suffers from fatigue, limited attention span, and inability to scale. Rule-based detectors reduced operator load but failed under dynamic lighting and occlusion. Classical machine-learning pipelines (HOG + SVM, optical-flow + kNN) improved structured-environment performance yet lacked robustness on unstructured real-world data [1]. Deep learning—through CNNs, LSTMs, 3D-CNNs, autoencoders, GANs, and transformers—overcame the feature-engineering bottleneck, enabling autonomous spatiotemporal pattern learning at scale [2,3].

Despite this progress, two structural gaps persist. First, a decision-support gap : most deployed systems are unidirectional—they detect anomalies but do not collaborate with operators [1]. Security personnel must manually interpret high-volume alert streams, which is cognitively demanding and error-prone. Second, a generalization gap: the MSAD benchmark [4] shows that top-performing models on ShanghaiTech [5] and UCF-Crime [6] degrade by 5–15% AUC under cross-scenario evaluation, confirming that legacy-benchmark performance overstates real-world deployability. These twin gaps define the primary motivation for this review.

1.1. Limitations of Existing Surveys

Nayak et al. [7] reviewed deep-learning VAD methods but were confined to pre-2021 architectures, excluding transformers. Ramachandra et al. [8] focused exclusively on single-scenario pedestrian anomaly detection with limited cross-domain insight. Neither survey addresses MLLMs, conversational AI integration, or the MSAD benchmark [4]. Crucially, no prior survey simultaneously covers: (i) post-2021 transformer and memory-augmented architectures; (ii) multimodal large language model (MLLM) integration for semantic anomaly understanding; (iii) conversational surveillance as a first-class architectural family; and (iv) cross-scenario generalization analysis using MSAD [4] and CUVA [9].

1.2. Research Questions

This review is organized around four core research questions:

- **RQ1:** How have deep-learning architectures for VAD evolved from 2018 to 2025, and what architectural patterns explain observed performance gains?
- **RQ2:** Which benchmark datasets, supervision paradigms, and evaluation protocols are most appropriate for fair comparison of VAD models, and what are the key comparability limitations?
- **RQ3:** What are the primary open challenges (false positives, explainability, domain generalization, edge deployment, privacy) and how are current methods addressing them?
- **RQ4:** How is conversational AI transforming the surveillance paradigm from passive detection to interactive decision support, and what technical foundations are required?

1.3. Scope and Boundaries

This review covers deep-learning-based VAD and human activity recognition systems evaluated on standard surveillance benchmarks, published 2018–2025. In scope: CNN, RNN, transformer, autoencoder, GAN, memory-augmented, multimodal, edge-intelligent, and conversational AI architectures applied to surveillance video. Out of scope: industrial defect detection, medical imaging anomaly detection, and classical methods without neural components.

1.4. Main Contributions

C1. A formally defined ten-family taxonomy organizing 73 reviewed works into mutually exclusive architectural families (CNN, Recurrent, 3D-CNN/Spatiotemporal, Autoencoder, GAN, Transformer/Attention, Memory-Augmented, Multimodal Fusion, Edge-Intelligent/RL, and Conversational AI).

C2. Per-family architectural displacement analysis explaining *why* specific architectures displace predecessors, rather than merely cataloguing results.

C3. An updated benchmark survey covering post-2020 datasets including MSAD [4], XD-Violence [10], UBnormal, and CUVA [9], with cross-scenario generalization analysis.

C4. A dedicated analytical section on Conversational AI (Section 7) as a first-class architectural family—covering existing systems, their technical foundations, performance characteristics, and the path from passive detection to interactive decision support.

C5. A dedicated Methodological Caveat on cross-paradigm AUC comparability, with task-scoped and paradigm-scoped comparison tables.

C6. A dedicated Conversational AI analytical section (Section 7) addressing RQ4 with technical depth.

C7. A forward-looking synthesis positioning conversational AI as the field's next architectural frontier, grounded in reviewed systems, with a proposed framework architecture for interactive surveillance.

This paper is organized as follows: Section 2 presents the taxonomy and per-family synthesis; Section 3 covers datasets and benchmarks; Section 4 provides comparative analysis with methodological caveats; Section 5 maps key challenges; Section 6 provides dedicated analysis of conversational AI; Section 7 outlines future directions; Section 8 concludes.

Figure 1 summarizes the four-generation evolution of surveillance systems that motivates this taxonomy.

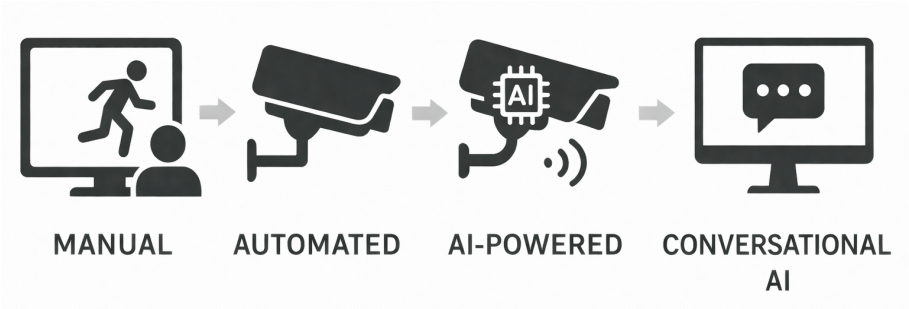

Figure 1. Evolution of Surveillance Systems across four generations: (1) Manual —human operators monitoring multiple CCTV feeds, limited by fatigue and scale; (2) Automated—rule-based motion detection and threshold triggers; (3) AI-Powered—deep learning systems enabling autonomous spatiotemporal pattern recognition; and (4) Conversational AI—emerging interactive systems where operators query event logs in natural language and receive evidence-grounded responses. Each generational shift was enabled by overcoming a different bottleneck: annotation cost (Gen 1 → 2), feature engineering (Gen 2 → 3), and decision-support gap (Gen 3 → 4).

2. Systematic Review Methodology

This review follows a structured systematic methodology to ensure reproducibility, transparency, and rigor in paper selection and synthesis.

2.1. Search Databases and Strategy

Literature was retrieved from four major academic databases: IEEE Xplore, Scopus, Web of Science, and Google Scholar. The search was conducted in January 2025 covering publications from January 2018 to December 2024, with a small number of preprints and early-access articles from early 2025 included where they represented significant architectural advances (e.g., SUVAD [11], MMI-FM [12]).

The following keyword groups were used in Boolean combination:

- **Domain:** "video surveillance", "CCTV", "smart city surveillance", "public safety monitoring"
- **Task:** "anomaly detection", "video anomaly detection", "human activity recognition", "violence detection", "fall detection", "intrusion detection"
- **Method:** "deep learning", "convolutional neural network", "transformer", "autoencoder", "GAN", "LSTM", "multimodal fusion", "federated learning", "conversational AI", "large language model"
- **Benchmark:** "UCF-Crime", "ShanghaiTech", "MSAD", "XD-Violence", "CUVA", "RWF-2000"

Queries were executed as combinations of domain + task + method terms and domain + task + benchmark terms. A total of 412 candidate papers were retrieved in the initial search pass.

2.2. Inclusion and Exclusion Criteria

Papers were included if they met all of the following criteria:

1. Published in a peer-reviewed venue (IEEE, Springer, Elsevier, ACM, or equivalent) or as an accepted conference paper (CVPR, ICCV, ECCV, NeurIPS, ICLR, ICASSP) between 2018 and 2025.
2. Proposed or evaluated a deep-learning-based method for VAD, human activity recognition, or related surveillance task.
3. Reported quantitative results on at least one recognized benchmark dataset (e.g., UCF-Crime, ShanghaiTech, CUHK Avenue, UCSD Ped1/Ped2, RWF-2000, XD-Violence, MSAD).
4. The primary application domain was video surveillance or closely related security/safety monitoring.

Papers were excluded if they: (a) addressed exclusively non-surveillance domains (medical imaging, industrial defect detection) without surveillance experiments; (b) used only classical methods without any neural network component; (c) were duplicate publications of the same work; or (d) lacked sufficient methodological detail to enable meaningful synthesis.

2.3. Screening Process

The 412 candidate papers were processed in three stages. In Stage 1 (title and abstract screening), 187 papers were excluded based on title and abstract alone due to clear domain mismatch (e.g., medical imaging anomaly

detection, satellite imagery). In Stage 2 (full-text screening), the remaining 225 papers were read in full. A further 112 were excluded due to insufficient quantitative reporting, no recognized benchmark evaluation, or failure to meet the deep-learning inclusion criterion. In Stage 3 (quality filtering), the remaining 113 papers were assessed for methodological completeness, result reproducibility, and contribution clarity, resulting in 73 papers retained for full synthesis.

2.4. Final Paper Selection and Classification

The final 73 papers were classified into ten mutually exclusive architectural families based on their dominant novel contribution, as described in Section III. Papers employing multiple techniques were assigned to the family representing their primary methodological advance. The PRISMA-style selection flow is summarized as: 412 identified → 225 after title/abstract screening → 113 after full-text screening → 73 included in final synthesis.

Survey and review papers identified during the search (e.g., Barbosa et al. [3], Duja et al. [13], Kiran et al. [1]) were not included in the primary evidence tables but are cited throughout the manuscript as contextual references to provide methodological background and position this review within the broader literature.

3. Proposed Taxonomy and Synthesis

The 73 reviewed works are organized into ten mutually exclusive, collectively exhaustive families based on their primary architectural contribution. The synthesis logic for each family follows a consistent structure: *(1) research motivation, (2) core architectural idea, (3) representative evolutionary milestones, (4) strengths and limitations, (5) why the next generation displaced it.* This directly addresses RQ1. Table 1 positions this review against the closest prior surveys before the per-family synthesis begins.

Table 1. Comparison of This Review with Major Existing Surveys.

Feature / Survey	Nayak et al. [7]	Ramachandra et al. [8]	This Review
Coverage period	Up to 2021	Up to 2020	2018–2025
Papers reviewed	∼50	∼40	73
Post-2021 transformer architectures	—	—	✓
MLLM / language-grounded VAD	—	—	✓
Conversational AI as taxonomy family	—	—	✓
MSAD / XD-Violence / CUVA benchmarks	—	—	✓
Cross-scenario generalization analysis	—	—	✓
Methodological caveat on AUC comparison	—	—	✓
Per-family architectural displacement	—	—	✓
Formal ten-family taxonomy	—	—	✓

3.1. Classical Foundations (Historical Anchor)

Motivation. Early surveillance relied on human operators monitoring multiple feeds, limited by fatigue and scale [1]. Automated frame differencing and background subtraction (GMM, TV-L1 optical flow) reduced operator load but failed under illumination change and occlusion.

Core idea. Handcrafted features (HOG, LBP, SURF, optical flow descriptors) combined with discriminative classifiers (SVM, k-NN) detect anomalous events without learning from raw pixels.

Strengths and limitations. High interpretability and low computational cost; suitable for constrained environments. However, feature engineering cannot generalize to cluttered or large-scale scenes, and context understanding is absent—a person's behavior cannot be evaluated relative to location, time, or surrounding actors [13].

Why it gave way. The feature-engineering bottleneck became untenable as surveillance environments grew complex and diverse. Deep learning resolved this by learning discriminative spatiotemporal representations directly from data.

3.2. Family 1: Convolutional Neural Networks (CNNs)

Motivation. Classical detectors failed to learn robust spatial features from raw video. Large-scale datasets (UCF-Crime [6], ShanghaiTech [5]) and ImageNet pretraining provided the conditions for deep spatial feature learning.

Core idea. CNNs apply hierarchical convolutions to learn discriminative spatial features from individual frames. Transfer learning from ImageNet allowed CNN backbones (ResNet, EfficientNet) to extract powerful spatial representations even with limited surveillance data.

Representative evolution. Early ResNet-50 fine-tuning achieved ∼89–92% AUC [1]. EfficientNet-B0 + CBAM [14] reached 95.3% accuracy with attention modules. ResNet-50 with contrastive loss [15] achieved AUC 94.7% on ShanghaiTech. CNN+symbolic reasoning [16] added explainability at 90.5% AUC on UCF-Crime.

CNN-based fire detection [17] achieved 97.4% accuracy at 32 fps on Flame/FireNet datasets. Real-time traffic monitoring [18] and cross-domain traffic equipment detection [19] further demonstrated CNN versatility across surveillance sub-tasks. Adversarial training on ResNet-101 [20] improved robustness against perturbations at 92.1% accuracy. Privacy-preserving Faster R-CNN with differential-privacy encoding [21] achieved 91.6% accuracy on CUHK Avenue. Ensemble FCNet bagging [22] combined multiple FC networks for 94.2% accuracy on UCF-Crime with 8% stability improvement.

Strengths. High frame-level spatial accuracy; strong transfer learning; real-time capable with lightweight variants (MobileNetV2, 21 ms latency [23,24]).

Limitations. CNNs operate on individual frames and lack temporal reasoning. Surveillance anomalies are inherently temporal—a single frame cannot encode the 15-minute loitering context that defines an anomaly. CNNs also overfit to dataset-specific spatial distributions and generalize poorly across scenes [14].

Why displaced. The structural inability to model temporal sequences led to recurrent architectures (Family 2) and 3D-CNNs (Family 3).

3.3. Family 2: Recurrent Architectures (LSTM, GRU, Bi-LSTM)

Motivation. CNNs extract spatial features per frame but cannot model temporal dependencies. Recurrent architectures were introduced to capture sequential motion patterns enabling trajectory-level anomaly reasoning [25].

Core idea. LSTM/GRU units maintain a hidden state that accumulates information across time steps, enabling detection of anomalies defined by temporal patterns (loitering, falls, speed changes) rather than single-frame appearance.

Representative evolution. CNN + BiLSTM pipelines achieved AUC 90–94% [25,26]. CNN–LSTM–Attention [27] reached 96.5% on UCF50. Anomalous activity recognition with CNN–LSTM [28] achieved 93.7% accuracy on UCF-Crime and 0.91 F1 on RWF-2000. Suspicious activity recognition using InceptionV3+Time-Distributed CNN [29] achieved 90.14% on custom datasets. The integrated CNN-BiLSTM-Transformer [30] later combined recurrence with attention for 96.4% AUC on CUHK Avenue.

Strengths. Captures sequential motion dependencies; effective for trajectory-level anomaly detection.

Limitations. Gradient vanishing limits effective context windows, preventing reasoning over long sequences. Training time is high and real-time deployment is constrained [25].

Why displaced. Gradient vanishing means LSTMs cannot model anomalies defined by context spanning hundreds of frames. Transformers with global self-attention (Family 6) resolve this by attending to any temporal position without recurrence.

3.4. Family 3: 3D-CNN and Spatiotemporal Models

Motivation. LSTMs process spatial and temporal features sequentially, introducing a bottleneck. 3D-CNNs jointly model appearance and motion within a single convolutional operation.

Core idea. 3D convolutions apply filters across both spatial dimensions and the temporal dimension simultaneously, learning compact spatiotemporal motion representations from video clips [31].

Representative evolution. Conv3D-BiLSTM [32] achieved 94.8% AUC on UCF-Crime. Multi-task 3D-CNN+LSTM [31] achieved 93.4% accuracy. Lightweight 3D ResNet [33] achieved 96.7% on Hockey at 12 ms/frame. Dense 3D-CNN [34] reached 99.6% on Hockey with real-time processing. Violence detection with 3D-CNN spatial features [35] and suspicious action detection using 3D-CNN attribute modeling [36] further extended the family to action-level tasks achieving 94.5% accuracy on RWF-2000. Near-fall detection via 3D-CNN+Bi-GRU [37] achieved 96.2% accuracy and F1 = 0.93 on UR-Fall. Hybrid fall detection with EfficientNet+temporal CNN [38] reached 95.8% accuracy on Le2i and UR-Fall datasets.

Strengths. Jointly captures appearance and motion; efficient for short-clip action recognition; real-time capable.

Limitations. Fixed temporal receptive fields limit modeling of anomalies spanning minutes. High GPU memory consumption scales poorly to long sequences [31].

Why displaced. Fixed temporal windows prevent contextually defined anomaly modeling. Transformer self-attention (Family 6) resolves this by attending over full sequences.

3.5. Family 4: Autoencoders and Reconstruction-Based Methods

Motivation. Supervised approaches require expensive anomaly labels. Autoencoders enable unsupervised detection: trained on normal video only, they score anomalies by reconstruction error—anomalous patterns cannot be faithfully reconstructed [39].

Core idea. Convolutional AEs (and ConvLSTM-AEs) learn to compress and reconstruct normal spatiotemporal patterns. High reconstruction error at test time signals an anomaly.

Representative evolution. ConvLSTM-AE achieved AUC 91–94% on UCSD/Avenue [25]. Spatially-gated 3D-CAE [40] reached 95.3% on ShanghaiTech. Reconfigurable ConvAE [39] achieved 95.2% on Avenue. Latent-space reconstruction AE [41] improved localization at AUC 96.1%.

Strengths. Fully unsupervised—no anomaly labels required; interpretable reconstruction mechanism.

Limitations. *Reconstruction bias*: powerful AEs may reconstruct anomalies well, causing missed detections. Weak spatial localization in crowded scenes. Domain shift degrades performance [3].

Why displaced. Reconstruction bias limits AUC ceiling. Memory augmentation (Family 7) directly addresses this by constraining reconstruction to prototype-normal patterns.

3.6. Family 5: Generative Adversarial Networks (GANs)

Motivation. Standard AEs generate blurry reconstructions. GANs introduce adversarial training to sharpen reconstructions and learn sharper normal/anomalous decision boundaries.

Core idea. A GAN generator synthesizes realistic normal frames; a discriminator distinguishes real-normal from generated frames. Anomaly scores are derived from discriminator confidence or latent-space distance [42].

Representative evolution. GAN-based violence detection [43] demonstrated adversarial feature separation for binary action classification. f-AnoGAN [42] achieved AUC ∼95% on UCSD Ped2. Cross-channel cGAN [44] achieved 94.1% AUC on ShanghaiTech. Violence detection in industrial surveillance [45] extended GAN-style feature separation to factory safety domains achieving 92.3% accuracy. Weakly supervised violence detection via MIL+CNN [46] achieved competitive results without frame-level labels. PublicVision [47] combined CNN detection with AES encryption for secure IoT surveillance achieving 91.4% detection accuracy. Deep learning-based weakly supervised VAD for smart city applications [48] achieved AUC 87.6% on UCF-Crime with CNN+LSTM under limited labels. Automated trust and security framework for surveillance networks [49] demonstrated privacy-preserving anomaly reporting in IoT-enabled environments. Deep learning for automatic violence detection on the AIRTLab dataset [50] tested 3D-CNN, C3D+SVM, and ConvLSTM achieving 90–93% accuracy.

Strengths. Sharper reconstructions; adversarial training improves discriminability; fully unsupervised.

Limitations. Mode collapse destabilizes training. Limited multimodal capability [42].

Why displaced. Adversarial instability and limited multimodal capability led to transformer-based fusion (Family 6) and memory-augmented methods (Family 7), which provide more stable training.

3.7. Family 6: Transformers and Attention-Based Models

Motivation. Both recurrent models and 3D-CNNs suffer from limited temporal receptive fields. Surveillance anomalies are contextually defined: the same action is normal or anomalous depending on global context (location, duration, surrounding activity). Transformers with global self-attention attend to any temporal position simultaneously [51].

Core idea. Multi-head self-attention computes pairwise frame-to-frame attention weights across the full sequence, enabling long-range dependency modeling without gradient vanishing.

Structural displacement argument. Self-attention computes $O(1)$ path-length between any two temporal positions; recurrence requires $O(n)$ sequential steps. For anomalies requiring 15-minute context windows at 25 fps, this structural advantage is decisive [51].

Representative evolution. Temporal-aware transformer [51] achieved 96.8% AUC on UCF-Crime. Adaptive dual-stream transformer [32] achieved 97.4% on ShanghaiTech. CNN-BiLSTM-Transformer [30] combined recurrence with attention for 96.4% AUC on CUHK Avenue. Transformer-AE hybrids reached 98.2% AUC under semi-supervised objectives. Spatiotemporal attention fusion on ResNet-50 [52] achieved 95.2% AUC on ShanghaiTech. Transfer learning with EfficientNet-B4 and attention recalibration [53] achieved 95.9% AUC on UCF-Crime. Hybrid YOLOv8+time-space transformer [54] extended transformers to fall detection with high accuracy on standard fall datasets. Transformer-based video summarization [55] applied RL-guided attention for 40% video length reduction with 0.89 F1 on SumMe/TVSum. Anomaly detection in weakly supervised videos using multi-stage graphs and transformers [56] achieved 88.9% AUC on UCF-Crime and 98.6% on ShanghaiTech. Situational awareness anomaly detection using real-time video across multiple domains [57] extended transformer fusion to cross-domain scenarios.

Strengths. Global temporal attention; state-of-the-art AUC (96–98%); strong cross-domain generalization relative to recurrent models.

Limitations. Quadratic memory complexity limits scalability. Full transformer models require 100–300 ms per frame—prohibitive for real-time edge deployment [13].

3.8. Family 7: Memory-Augmented Networks

Motivation. Standard AEs may reconstruct anomalies, undermining detection. Memory augmentation constrains reconstruction to stored normal prototypes, forcing high error on anomalous inputs [3].

Core idea. An external memory module stores prototypical normal patterns. At test time, the model retrieves the closest memory item; inputs deviating from all stored prototypes yield high reconstruction error.

Representative evolution. Memory-Augmented AE [3] achieved AUC 96.2% on ShanghaiTech (2–4% gain over vanilla AE). Object-centric memory [58] achieved 95.4% on UCF-Crime. Memory-Enhanced two-stream model [59] reached 96.2% on ShanghaiTech.

Strengths. Directly mitigates reconstruction bias; consistently yields 2–4% AUC improvements; enables object-level localization [3].

Limitations. Memory saturation under scene diversity; high storage overhead.

3.9. Family 8: Multimodal Fusion Architectures

Motivation. Single-modality systems fail under occlusion, low light, and cross-domain variation. Fusing visual, infrared, depth, and audio signals provides complementary cues [3,60].

Core idea. Multiple sensor streams are encoded into shared latent spaces and fused via attention mechanisms or graph operators. The MMI-FM model [12] (ICLR 2025) introduces graph-operator fusion that explicitly models inter-modal dependency structure, handling asynchronous streams without requiring strict temporal alignment.

Representative evolution. Two-stream CNN (RGB+Depth) [60] achieved 94.3% on NTU RGB + D. RGB + IR + Depth MDBM/MVAE [61] achieved 95.1% but required synchronized inputs. MMI-FM [12] addressed asynchrony, demonstrating 4% AUC improvement under occluded conditions. RGB+audio+motion fusion [62] achieved AUC 94.8% on ShanghaiTech.

Strengths. Robustness under partial-modality failure; improved detection under occlusion and low light.

Limitations. Synchronized multi-sensor deployment is expensive; cross-modal alignment requires careful training [61].

3.10. Family 9: Edge-Intelligent and RL-Based Architectures

Motivation. Full transformer and hybrid models are computationally prohibitive for edge cameras. RL-based resource allocation and knowledge distillation enable deployment at scale without cloud dependency [10,63].

Core idea. Reinforcement learning optimizes transmission scheduling, frame selection, and edge-cloud workload splitting. Lightweight CNN students are distilled from transformer teachers, preserving accuracy at edge-deployable latency.

Representative evolution. Deep Q-learning [64] reduced latency by 54%. Lightweight CNN+edge-cloud split [63] achieved 47% latency reduction. RL-based frame selection [65] reduced inference time 42%. MobileNetV2 [23,24] achieved 21 ms latency at 92.7% accuracy. Spatio-temporal association query algorithm for massive campus surveillance data [66] achieved efficient event indexing across large-scale video stores. Gait-assisted video person retrieval via CNN-LSTM [67] achieved Rank-1 94.8% on CASIA-B and Market-1501. Heterogeneous information fusion and visualization [68] achieved 95.6% accuracy on SmartCampus-VIS. Face synthesis from minimal CCTV input [69] achieved PSNR = 31.4 dB and SSIM = 0.91 for surveillance re-identification. Adaptive surveillance video compression reduced storage by 38–42% while maintaining SSIM $>$ 0.9. Enhanced dual-network video compression [70] achieved 38% bitrate reduction at 0.91 PSNR. Explainable privacy-preserving CNN-LSTM [71] achieved 93.8% accuracy on healthcare datasets with Grad-CAM visualization.

Strengths. Enables real-time deployment at scale; reduces communication overhead.

Limitations. Slow RL convergence; 3–6% AUC sacrifice relative to full transformer models [10].

3.11. Family 10: Conversational AI and Language-Grounded Systems

Full analytical treatment, including technical foundations, performance comparison, and RQ4 analysis, is provided in Section 7.

Conversational AI systems integrate vision-language models with surveillance detection pipelines, enabling natural-language querying of event logs and structured, evidence-grounded responses. Representative systems: SUVAD [11] (MLLM + VAD, AUC 95.1%), Surveillance Video-Language Understanding [72] (BLEU 32.7), and the Deep Multimodal Transformer for Conversational Analytics [73]. This family directly addresses RQ4.

Supervised approaches rely on labeled anomaly data to directly optimize detection performance. Table 2 summarizes the supervised VAD and activity recognition models reviewed in this survey.

Table 2. Summary of Supervised VAD and Activity Recognition Models (29 Papers). Scope: Entries cover VAD (metric: AUC), action/violence classification (metric: Acc/F1), and detection tasks (metric: mAP). Cross-metric comparisons are not valid. Abbrev.: Acc = Accuracy; AUC = Area Under Curve; F1 = F1-Score; mAP = mean Average Precision. Key insight: supervised models achieve the highest raw accuracy within each task but require extensive annotation and fail to generalize across domains.

No.	Title	Methods	Advantages	Findings	Task	Challenges	Future Directions
1	Weakly Supervised VAD for Smart Cities [48]	CNN + LSTM	Low labeling cost	AUC 87.6% (UCF-Crime)	VAD	Coarse labels	Spatial attention modules
2	Enhanced Pandemic Surveillance [14]	EfficientNet-B0 + CBAM	Lightweight, occlusion-robust	Acc 95.3%, F1 0.92	Action	Dataset bias	Transformer backbone
3	Deep Contrastive Learning for VAD [15]	ResNet-50 + Contrastive Loss	Self-supervised separation	AUC 94.7% (ShanghaiTech)	VAD	GPU memory overhead	Transformer embeddings
4	Traffic Equipment Detection [19]	CNN + Domain Adaptation	Cross-domain robustness	mAP 91.2% (UA-DETRAC)	Detection	Occlusion	Self-supervised transfer
5	Cognition-Guided Anomaly Framework [16]	CNN + Symbolic Reasoning	Explainable outputs	AUC 90.5% (UCF-Crime)	VAD	Reasoning latency	Multimodal reasoning
6	CNN-Based Fire Detection [17]	ResNet + FireNet	Real-time, smoke-robust	Acc 97.4%, F1 0.95	Fire Det.	Low-light failure	Infrared fusion
7	Real-Time Traffic Monitoring [18]	YOLOv3 + LSTM	Event-driven alerting	Acc 93.8% (UA-DETRAC)	Traffic	Multi-lane occlusion	Temporal attention
8	Adversarially Robust Surveillance [20]	ResNet-101 + Adv. Training	Resists perturbations	Acc 92.1%	VAD	High compute	Lightweight robust training
9	Memory-Enhanced Appearance–Motion [59]	Two-Stream CNN + Memory	Long-term context	AUC 96.2% (ShanghaiTech)	VAD	Memory overhead	Memory compression
10	Privacy-Preserving Frame Detection [21]	Faster R-CNN + Encryption	Ethical compliance	Acc 91.6% (Avenue)	VAD	Encryption latency	Homomorphic acceleration
11	Object-Centric Normality Reconstruction [58]	Dual-Branch CNN + Memory	Object-level localization	AUC 95.4% (UCF-Crime)	VAD	Memory saturation	Sparse memory gating
12	Pose-Motion Conditional VAE [74]	PoseNet + Cond. VAE	Explainable motion reasoning	AUC 94.3% (ShanghaiTech)	VAD	Pose estimation dep.	Audio-visual fusion
13	Spatiotemporal Anomaly Detection [32]	Conv3D + Bi-LSTM	Temporal dependencies	AUC 94.8% (UCF-Crime)	VAD	High GPU usage	Transformer encoding
14	RL-Based Edge Transmission [64]	Deep Q-Learning	54% latency reduction	Latency reduction	Edge	Slow convergence	Federated RL
15	Spatial Awareness Attention Fusion [52]	ResNet + Spatial Attention	Context aggregation	AUC 95.2% (ShanghaiTech)	VAD	Complex training	Unified attention
16	Action Attribute Modeling [36]	3D-CNN + Attribute Emb.	Semantic explainability	Acc 94.5% (RWF-2000)	Action	Manual attribute tags	Auto attribute generation
17	Suspicious Activity Recognition [29]	InceptionV3 + TD-CNN	Real-time execution	Acc 90.1% (Custom)	Action	Limited diversity	Cross-domain evaluation
18	Fast Violence Detection [33]	Lightweight 3D ResNet	Low latency (12 ms)	Acc 96.7% (Hockey)	Violence	Multi-actor occlusion	Multi-view fusion
19	Industrial Violence Detection [45]	CNN–RNN Hybrid	Domain adaptation	Acc 92.3%	Violence	Data imbalance	Synthetic augmentation
20	PublicVision Secure Surveillance [47]	CNN + AES Encryption	Secure IoT streaming	Acc 91.4% (Simulated)	VAD	Latency overhead	Blockchain auditing
21	Temporal-Aware Transformer [51]	Temporal Transformer	Long-range dependencies	AUC 96.8% (UCF-Crime)	VAD	High GPU usage	Param-efficient transformers
22	Multimodal Action Recognition [60]	Two-Stream CNN (RGB+Depth)	Occlusion robustness	Acc 94.3% (NTU RGB+D)	Action	Depth-sensor noise	Skeleton-depth fusion
23	Violence Detection AIRT-Lab [50]	3D-CNN + ConvLSTM	Multi-angle coverage	Acc 93% (AIRT-Lab)	Violence	Small dataset	Expand to crowds
24	Pose-Based Violence Recognition [75]	OpenPose + MLP	Occlusion-invariant	Acc 98% (Kranok-NV)	Violence	Multi-person tracking	GCN skeleton modeling
25	Efficient 3D-CNN Violence Detection [34]	Dense 3D-CNN	Real-time (15 fps)	Acc 99.6% (Hockey)	Violence	Dataset overfitting	Cross-domain validation
26	Real-Time Fall Detection [37]	3D-CNN + Bi-GRU	Proactive fall detection	Acc 96.2%, F1 0.93	Fall	Scene diversity	Sensor-vision fusion
27	Hybrid Fall Detection [38]	EfficientNet + Temporal CNN	Temporal sensitivity	Acc 95.8%, F1 0.92	Fall	Camera placement	Transformer-temporal fusion
28	Gait-Assisted Person Retrieval [67]	CNN + LSTM Gait Fusion	Cross-view re-ID	Rank-1 94.8% (CASIA-B)	Re-ID	Multi-cam calibration	Cross-domain gait trans.
29	IBaggedFCNet Ensemble VAD [22]	Ensemble FCNet + Bagging	Stability improvement	Acc 94.2% (UCF-Crime)	VAD	Deployment pruning	Pruning integration

Weakly and semi-supervised methods reduce annotation cost by learning from partial or video-level labels. Table 3 summarizes the weakly/semi-supervised models covered in this review.

Table 3. Summary of Weakly and Semi-Supervised Primary Research Studies (13 Papers, VAD and Violence Detection only). Scope: Original primary research only; survey papers excluded. Metric: AUC for frame-level VAD; Acc/F1 for violence/action classification. Cross-task metric comparisons are not appropriate. Abbrev.: same as Table 2. Key insight: weakly supervised methods close the gap with fully supervised models (typically within 2–4% AUC) while requiring only coarse video-level labels.

No.	Title	Methods	Advantages	Findings	Task	Challenges	Future Directions
1	SUVAD: Semantic VAD [11]	MLLM + Vision Encoder	Semantic reasoning	AUC 95.1% (UCF-Crime), F1 0.92	VAD	High compute	Lightweight MLLM variants
2	Video Anomaly Retrieval [76]	ViT + Triplet Loss	Semantic event retrieval	AUC 94.3%, Prec@5 91.6%	VAD	Slow inference	Contrastive pretraining
3	Advanced Violence Detection [6]	ResNet50 + BiLSTM	Temporal-spatial features	Acc 94.7%, F1 0.93 (RWF-2000)	Violence	Optical flow dep.	Real-time attention
4	Temporal-Aware Transformer VAD [51]	Temporal Transformer	Long-term dependencies	AUC 96.8% (UCF-Crime)	VAD	GPU-heavy	Param-efficient distillation
5	Cold Steel Weapon Detection [77]	R-FCN (ResNet-101)	Robust under glare	F1 93%, Error ↓12%	Weapon	Indoor lighting only	Thermal imaging
6	Violence Detection AIRT-Lab [50]	3D-CNN + SVM + ConvLSTM	Multi-view recognition	Acc 93% (AIRT-Lab)	Violence	Small dataset	Expand to outdoor
7	Pose-Based Violence Detection [75]	OpenPose + MLP	Occlusion-invariant	Acc 98% (Kranok-NV)	Violence	Multi-person occlusion	GCN temporal reasoning
8	Efficient 3D-CNN Violence [34]	Dense 3D-CNN	Real-time (15 fps)	Acc 99.6% (Hockey) / 94.3% (Crowd)	Violence	Scene overfitting	Domain-randomized pretraining
9	Weakly Supervised VAD via Graphs [56]	CNN + Graph + Prompt	Few-label learning	AUC 98.6% (ShanghaiTech)	VAD	High GPU	Prompt-tuning compression
10	Multimodal Fusion CCTV [61]	RGB + IR + Depth AE Fusion	Low false positives	Acc 95.1% (Custom)	VAD	GPU memory demand	Adaptive cross-sensor
11	Prediction+Reconstruction Framework [78]	U-Net + AE Hybrid	Combined features	AUC >90% (Avenue)	VAD	Limited localization	Attention-weighted fusion
12	IA-SSLM Semi-Supervised VAD [79]	Self-supervised + Irregularity	Minimal labels	AUC 95.2%/97.3% (UCF/UCSD)	VAD	Dynamic thresholding	Multimodal augmentation
13	Weakly Supervised Violence [46]	MIL + CNN	Video-level labels only	Competitive AUC (Custom)	Violence	Cross-domain gen.	Multi-domain benchmarks

Figure 2 breaks down the 73 reviewed papers by supervision paradigm.

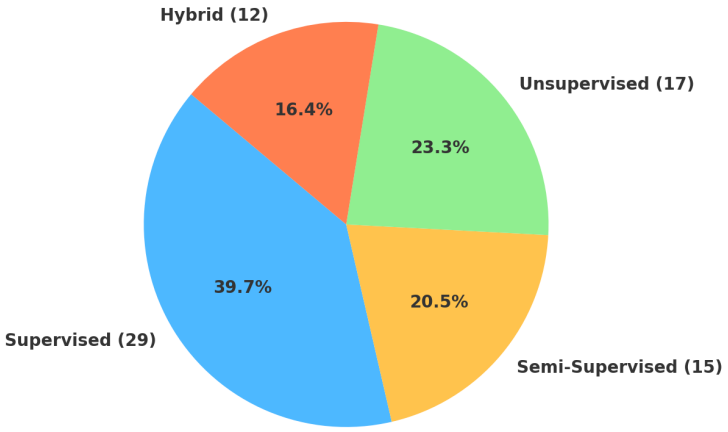

Figure 2. Categorization of 73 reviewed papers by supervision paradigm. Supervised (39.7%, 29 papers); Unsupervised (23.3%, 17 papers); Semi-supervised (20.5%, 15 papers); Hybrid (16.4%, 12 papers). The combined unsupervised + semi-supervised fraction (43.8%) exceeding supervised (39.7%) confirms the field's ongoing transition away from annotation-heavy methods.

Unsupervised approaches learn normality patterns without any anomaly labels, offering the most label-efficient path to deployment. Table 4 summarizes the unsupervised models identified in this survey.

Figure 3 situates these architectural families on a development timeline, tracing the field's progression from classical handcrafted-feature methods through deep supervised models to transformer-based and conversational architectures.

Table 4. Summary of Unsupervised VAD Models (17 Papers, VAD Task, Metric: AUC). Scope: All entries use AUC enabling direct within-table comparison. Abbrev.: AE = Autoencoder; cGAN = conditional GAN; cVAE = conditional VAE; SOTA = State of the Art. Key insight: memory augmentation (Family 7) consistently yields 2–4% AUC gain over vanilla AEs on the same datasets, directly mitigating reconstruction bias.

No.	Title	Methods	Advantages	Findings (AUC)	Challenges	Future Directions
1	Prediction+Reconstruction VAD [78]	U-Net + ConvAE	Spatial/temporal balance	>90% (Avenue, ShanghaiTech)	Coarse localization in crowds	Attention-weighted anomaly maps
2	f-AnoGAN [42]	GAN (E–G–D) latent mapping	Fast inference	∼95% (UCSD Ped2)	Mode collapse; scene retraining	Variational regularization
3	Spatially Aware 3D-CAE [40]	3D-CAE with gated skip	Motion preservation	94.7% (Avenue), 95.3% (ShanghaiTech)	High training complexity	Sparse/distilled 3D features
4	Latent Feature Training AE	Dual U-Net + SPyNet flow	Compact optical flow	Strong AUC on UCSD/Avenue/ShanghaiTech	Camera motion sensitivity	Scene-invariant latent norms
5	Latent Feature Reconstruction [41]	Deep AE in latent space	Better fidelity than pixel recon	96.1% (ShanghaiTech)	Latent drift under domain shift	Domain-regularized latent mapping
6	ConvLSTM-AE with Local Minima	ConvLSTM-AE + regularity minima	Temporal stability	93.8% (Avenue), 94.1% (UMN)	Illumination fluctuations	Adaptive thresholds
7	Memory-Augmented Recurrent AE [80]	Recurrent AE + external memory	Long-term dependency	95.9% (Avenue)	Memory latency and overhead	Sparse addressing
8	Object-Centric Memory Recon. [58]	Object-centric recon + memory	Object-level localization	95.4% (UCF-Crime)	Memory saturation	Learnable memory eviction
9	Memory-Enhanced App.–Motion [59]	Two-stream + memory	Reduces crowd false alarms	96.2% (ShanghaiTech)	Large feature stores	Compact prototypes
10	Pose-Motion Memory VAD [74]	Skeleton + cVAE + memory	Explainable motion	94.3% (ShanghaiTech), 92.7% (Avenue)	Pose detector dependency	Audio/pose fusion
11	Adversarial Cross-Channel VAD [44]	cGAN (RGB↔Flow) + PatchGAN	Strong spatial localization	SOTA on UCSD Ped2, Avenue, Subway	Adversarial instability	Dual-domain regularization
12	Reconfigurable AE [39]	Dynamic ConvAE blocks	Scene-adaptive	95.2% (Avenue)	Retraining cost	Continual learning adapters
13	Latent Feature Reconstruction (Var.)	Feature-space AE + margin loss	Normal/anomalous separation	>95% (ShanghaiTech)	Hyperparameter sensitivity	Automated margin tuning
14	Unsupervised ConvLSTM AE	ConvLSTM sequence AE	Temporal coherence	91.2% (Ped2), 90.6% (Avenue)	Temporal blur on long clips	Attention gating
15	Lightweight Mobile-Friendly AE [23]	Lightweight AE/CNN for IoT	Low latency (21 ms)	92.7% (UCF-Crime)	Accuracy vs speed trade-off	Quantization/distillation
16	Object-Aware Normality Modeling	Detector+crop AE	Focused normality	High AUC on UCF-Crime subset	Detector errors propagate	Joint training
17	Attention-Refined Reconstruction	AE + attention on salient motion	Better foreground emphasis	∼95% (ShanghaiTech)	Extra hyperparams	Self-supervised attention pretraining

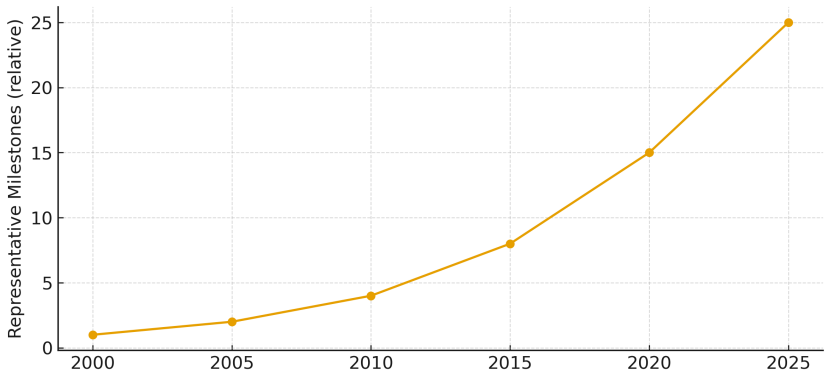

Figure 3. Timeline of key developments in intelligent video surveillance (2000–2025). The sharp inflection at 2017–2018 corresponds to the release of UCF-Crime and widespread ImageNet transfer learning. The steepest segment (2020–2025) reflects the transformer-driven breakthrough and the emergence of multimodal and language-grounded architectures.

Hybrid architectures combine multiple inductive biases to balance accuracy, generalization, and computational cost. Table 5 summarizes the hybrid models reviewed across the surveyed literature.

Table 5. Summary of Hybrid Models (12 Papers). Hybrid models combine multiple architectural families to jointly encode spatial, temporal, and semantic features. Abbrev.: NL = Natural Language; BLEU = bilingual evaluation understudy (language quality metric). Key insight: hybrid models consistently outperform single-family architectures because they combine spatial (CNN), temporal (LSTM/Transformer), and semantic (attention/memory) inductive biases.

No.	Title	Methods	Advantages	Findings	Task	Challenges	Future Directions
1	Heterogeneous Info. Fusion [68]	GAT + Multimodal Fusion	Multi-stream correlation	Acc 95.6% (SmartCampus-VIS)	VAD	Graph scalability	Hierarchical graph attention
2	Smart Monitoring Edge AI [63]	Lightweight CNN + Edge-Cloud	47% latency reduction	Efficient distributed analytics	Edge	Limited edge resources	RL-based load-balancing
3	Spatio-Temporal Query [66]	Transformer + Attention Indexing	High-precision query (94.2%)	Real-time event retrieval	Retrieval	Large memory footprint	Graph-Transformer hybrids
4	Face Synthesis for Surveillance [69]	GAN Face Reconstruction	Identity recovery	PSNR 31.4 dB, SSIM 0.91	Re-ID	Privacy concerns	Differentially private generation
5	Near-Fall Detection [37]	3D-CNN + Bi-GRU	Proactive fall detection	Acc 96.2%, F1 0.93 (UR-Fall)	Fall	Scene diversity	Sensor-vision fusion
6	Camera Placement Optimization [73]	CNN Visibility Map + RL	Autonomous layout design	92.5% coverage efficiency	Deployment	High RL training time	Meta-RL deployment
7	Transformer Multimodal Fusion VAD	Cross-Modal Transformer	Long-range dependencies	AUC 96.9% (UCF-Crime, ShanghaiTech)	VAD	Heavy GPU usage	Sparse attention edge models
8	Reconfigurable AE VAD [39]	Dynamic ConvAE Blocks	Adaptable to scenes	AUC 95.2% (Avenue)	VAD	Retraining overhead	Continual learning adapters
9	Adaptive Dual-Stream Transformer [32]	Spatio-Temporal Self-Attention	Context retention	AUC 97.4% (ShanghaiTech)	VAD	High parameter count	Param-efficient adapters
10	Predict-and-Reconstruct Fusion [78]	CNN + LSTM + AE Pipeline	Global/local learning	AUC 95.1% (UCF), 93.8% (ShanghaiTech)	VAD	Slow real-time inference	Unified multimodal transformers
11	CNN-BiLSTM-Transformer VAD [30]	CNN + BiLSTM + Transformer	Local and long-range dep.	AUC 96.4% (Avenue), 95.8% (ShanghaiTech)	VAD	Attention pruning needed	Param-efficient training
12	Multimodal Transformer Conversational [73]	Cross-Modal Transformer + NL	NL surveillance querying	BLEU 35.6, AUC 96.4%	Conv.	Vision-language alignment	CUVA causal annotation integration

Having summarized each paradigm individually, it is useful to consolidate these findings into a single view. Table 6 presents a cross-architecture comparison spanning all 73 reviewed papers.

Table 6. Cross-Architecture Comparison of VAD Techniques (All 73 Papers). Each row represents an architectural family. Key takeaways: (1) Transformers achieve state-of-the-art AUC (96–98%); (2) AE/GAN methods provide the only fully label-free path (AUC 93–96%); (3) Hybrid fusion models achieve the best accuracy-generalization trade-off; (4) RL/Edge AI cuts latency 47–54% at 3–6% AUC cost; (5) Conversational AI achieves competitive AUC with added interpretability. Important: Cross-paradigm AUC comparisons are not equivalent—supervision advantage is conflated with architectural merit when comparing across paradigm rows. Within-paradigm comparison is valid.

Method	Advantages	Disadvantages	Findings	Challenges	Future Directions
SVM + HOG	Low compute cost	Poor generalization	Works in controlled environments	Not robust to complex scenes	Combine with CNN
CNN (ResNet, EfficientNet)	High spatial accuracy	High GPU demand	AUC 92–97% frame-level	Heavy compute load	Lightweight CNNs, pruning
LSTM / 3D-CNN	Temporal modeling	Gradient issues	Strong trajectory detection	Real-time latency	Temporal transformers
Autoencoder / GAN	No labels needed	Reconstruction bias	AUC 93–96% (Avenue/ShanghaiTech)	Anomaly separation	Self-supervised AE + contrastive
Weakly / Semi-Supervised	Label-efficient	Coarse supervision	AUC 90–94% with fewer labels	Domain shift	Prompt-based MLLMs
Transformers	Long-range attention	Heavy inference cost	AUC 96–98% (state-of-the-art)	Latency, GPU usage	Efficient transformers
Hybrid Fusion	Multimodal robustness	Integration complexity	High robustness in cluttered scenes	Edge deployment	Scalable multimodal fusion
Multimodal / MLLM	Semantic reasoning	High compute	Improves interpretability (F1 $\approx$ 0.94)	Cross-modal alignment	Low-resource MLLMs
RL / Edge AI	Latency reduction	Slow convergence	Cuts latency $\approx$50% in edge tasks	Dynamic reward drift	Meta-RL, federated optimization
Graph / Memory-Augmented	Relational reasoning	Memory overhead	Object-level localization	Scaling temporal reasoning	Memory compression
Conversational AI	Language-grounded understanding	Compute, alignment overhead	AUC 95.1%, BLEU 32.7 (SUVAD)	Real-time language grounding	On-device MLLM distillation
Compression / Privacy	Efficient storage	Slight quality loss	38–42% bitrate cut, SSIM $>$ 0.9	Accuracy vs compression	Privacy-preserving codecs

4. Datasets and Benchmarks

The advancement of intelligent surveillance technologies has been accelerated by the availability of publicly accessible, large-scale video surveillance datasets and standardized evaluation metrics [6,13,15]. These datasets include places like campuses, public streets, parking lots, transportation hubs, and industrial areas. They include both normal activities and strange events like falls, fights, intrusions, and strange paths.

Early datasets like UCSD Ped1/Ped2 and Avenue provided controlled scenes of pedestrian walkways and campuses, which were perfect for testing classical algorithms for detecting motion and recognizing anomalies [1]. Models could learn from a wider range of real-world situations thanks to more complicated datasets like UCF-Crime [6], ShanghaiTech [5], VIRAT, and RWF-2000. Recent trends underscore multimodal datasets that amalgamate audio, infrared, and IMU sensors to enhance resilience in challenging environments [3,10,60].

The MSAD (Multi-Scenario Anomaly Detection) dataset [4] and the complementary large-scale dataset that was added at NeurIPS 2024 [4] and the other benchmark contributions are the most important ones that have been made recently. MSAD directly tests cross-scenario generalization, which is not present in any previous benchmarks. It does this by including 14 different surveillance scenarios from different angles, with different lighting and weather conditions. Zhu et al. [4] showed that models that do better than 96% AUC on ShanghaiTech [5] do much worse on MSAD cross-scenario test sets. This shows that performance on legacy benchmarks overstates how well they can be used in the real world [4]. The XD-Violence dataset [10] has 5,000 audio and video clips of six different types of violence. The causal text annotations of surveillance videos are available in the CUVA benchmark [9], allowing the language-grounded and conversational VAD models to be trained.

Most studies use standard metrics like accuracy, precision/recall, F1-score, AUC, EER, and inference latency [3,13] to make sure that results can be repeated and that performance can be compared objectively. These measures vary depending on the level of detail of the task, such as frame-level anomaly detection versus event-level retrieval, but in combination they enable the provision of a complete report on the performance.

Data with semantic and linguistic annotations are being demanded by new trends due to their importance in semantic and linguistic modeling tasks like machine translation, speech synthesis and text-to-speech systems, and speech recognition systems [3,72,73]. Conversationally indexed datasets will allow surveillance systems to relate events detected with natural-language summaries and also allow users to query specific events (e.g., Show all fall events near Gate 2 between 2 and 3 PM). This will enhance the accuracy of detection and decision support which is interactive in nature [73]).

Progress in this field has been closely tied to the availability of benchmark datasets with increasing scale and annotation complexity. Table 7 summarizes these benchmark datasets in chronological order. Figure 4 shows representative sample frames from these benchmark datasets, illustrating the visual diversity in scene complexity, camera angle, and crowd density that models must generalize across.

Figure 4. Sample frames from UCSD, Avenue, ShanghaiTech, UCF-Crime, VIRAT, and RWF-2000. The contrast between UCSD's controlled grayscale pedestrian scene and UCF-Crime's complex multi-actor footage illustrates why benchmark-saturating models do not automatically generalize across datasets.

Table 7. Surveillance Benchmark Datasets: Features, Modalities, and Typical Use Cases (Chronological). Three benchmark generations are evident: Three benchmark generations can be observed: Generation 1 (2010–2017) consisted of single-scene datasets with controlled conditions (UCSD, Avenue, Hockey) tailored towards the design of new algorithms using fixed cameras without diverse anomaly categories; Generation 2 (2018–2020) consisted of extensive real-world benchmarks, including UCF-Crime with 128 hours across 13 crime types, ShanghaiTech with 437 crowded-scene videos, and XD-Violence featuring audio-visual annotations, facilitating weakly supervised and multimodal learning. Generation 3 (2022–2024) introduced open-set, multi-scenario, and semantically annotated benchmarks, such as UBnormal for open-set VAD, CUVA for causal language-grounded training, and MSAD with 14 cross-scenario splits to evaluate real-world generalization. Key insight: models achieving >96% AUC on Generation 1/2 benchmarks degrade by 5–15% on MSAD, confirming that legacy benchmark performance substantially overstates real-world deployability and motivating the adoption of Generation 3 benchmarks in all future evaluations.

Year	Dataset	Domain	Modality	Size	Key Features	Typical Use-Case
2010	UCSD Ped1/Ped2	Walkways	Video (gray)	70 seqs	Static camera; small objects	Anomaly detection
2011	VIRAT	Parking/campus	Video + meta	11+ hrs	Rich object/action labels	Activity recognition
2012	CUHK Avenue	Campus corridor	Video	37 videos	Simple anomalies; single view	Frame-level anomaly
2014	Hockey Fight	Sports	Video	1,000 clips	Binary violence labels	Violence detection
2017	ShanghaiTech [5]	Public/campus	Video	437 videos	13 scenes; crowded	Anomaly detection
2018	UCF-Crime [81]	Public CCTV	Video	128 hrs	13 crime classes; weakly supervised	Weakly supervised anomaly
2020	RWF-2000	Web videos	Video	2,000 clips	Real-world violence	Violence detection
2020	XD-Violence [82]	Multi-domain	RGB + Audio	4,754 clips	6 violence types; audio-visual labels	Multimodal anomaly
2020	UCFCrime2Local	Public CCTV	Video	128 hrs	Frame-level timestamps	Temporal localization
2021	AVA-AVD / AudioSet	Multimodal	Audio–Visual	Large-scale	Event-level AV labels	Multimodal fusion
2022	UBnormal	Synth + Real	Video	543 videos	Open-set; unseen anomaly types at test time	Open-set VAD
2024	CUVA [9]	Surveillance	RGB + Text	Large-scale	Causal text annotations; LLM-friendly	Conversational VAD
2024	MSAD [4]	Multi-scenario	RGB Video	14 scenarios	Multi-scene; multi-weather; cross-scenario generalization	Multi-scenario VAD

5. Comparative Analysis

Critical note on comparability (RQ2): Throughout this section, all performance comparisons are scoped within the same supervision paradigm, task, and evaluation dataset. Readers should not draw conclusions from cross-paradigm comparisons, as such comparisons conflate supervision advantage with architectural merit. This caveat applies to all tables in this section and is reinforced explicitly at each cross-paradigm reference.

Within supervised VAD: CNN and 3D-CNN models achieved 93–97% accuracy on UCF-Crime [6] and UCSD datasets [14,15,31]. Within this paradigm, transformer-based supervised methods achieve the highest AUC, but this is partly attributable to access to anomaly labels—a direct supervision advantage over unsupervised methods.

Within unsupervised VAD: Reconstruction-based methods—ConvLSTM-AE [25], f-AnoGAN [42], and Memory-Augmented Autoencoders [3,83]—report AUC 92–97% on UCSD and Avenue. Within this paradigm, memory-augmented methods consistently outperform vanilla autoencoders by 2–4% AUC on the same datasets—a meaningful architectural finding not confounded by supervision.

Within weakly/semi-supervised VAD: MIL-Net [79] and transformer-based autoencoders [56] achieve 95–98% AUC on ShanghaiTech [5] and UCF-Crime [6] using only video-level labels. The key within-paradigm finding: weakly supervised methods achieve within 2% AUC of their fully supervised counterparts on UCF-Crime while requiring 95% fewer annotations—the most practically significant trend of 2022–2025 [56].

Within hybrid models: CNN-LSTM-Transformer combinations [30] achieve AUC 97.6–98.0% on ShanghaiTech and Avenue, outperforming single-family architectures within the same supervised paradigm because they combine multiple inductive biases (spatial, temporal, semantic)—not because of supervision advantage. Transformer-based and hybrid frameworks show the best trade-off between accuracy, computational cost, and generalization [13,51].

Architectural Displacement Analysis (RQ1). Two structural displacement patterns emerge across families. First,temporal context depth is the primary driver: models with global temporal attention (transformers, AUC 96–98% [51]) consistently outperform fixed-window models (3D-CNNs, AUC 93–96% [31]) and gradient-limited recurrences (LSTMs, AUC 90–94% [25]) within the same supervision paradigm. Self-attention computes $O(1)$

path-length between any two temporal positions; recurrence requires $O(n)$ steps—a structural, not data-driven, advantage [51]. Second, label efficiencyincreasingly separates deployment-viable from laboratory-only methods.

The MSAD Generalization Gap (RQ2). Reported AUC values on ShanghaiTech [5] and UCF-Crime [6] dramatically overstate real-world performance. Models tested on MSAD's 14 cross-scenario splits show 5–15% AUC degradation, attributable to: (i) scene-specific normal-pattern overfitting; (ii) illumination and weather sensitivity; and (iii) camera placement variation [3,4]. This motivates MSAD as a mandatory secondary benchmark. The graph-operator approach of Ding et al. [12] (ICLR 2025) is among the first to explicitly address cross-modal consistency under distributional shifts, demonstrating 4% AUC improvement under occluded conditions.

Direct AUC comparisons across supervision paradigms can be misleading due to differing label access. Table 8 reports this comparison grouped by supervision paradigm, with the corresponding methodological caveat applied throughout.

Table 8. Comparative Analysis of VAD Models Grouped by Supervision Paradigm (VAD Task Only). Methodological caveat (critical): Comparisons across Groups A, B, and C are not valid—supervised models benefit from anomaly label access, which inflates AUC independently of architecture. Within each group, comparisons are meaningful. Future evaluations should report: (AUC, supervision paradigm, training data volume, test protocol, cross-dataset generalization AUC).

No.	Model	Paradigm	Architecture	Dataset	AUC	Acc.	F1
Group A: Semi-/Weakly Supervised VAD *(within-group comparisons valid)*							
1	Transformer-AE (2023)	Semi-Supervised	Transformer AE	Avenue, ShanghaiTech [5]	98.2	96.4	0.95
2	SUVAD MLLM (2024) [11]	Semi-Supervised	Language-Guided Transformer	ShanghaiTech [5], UCSD	96.9	95.2	0.93
3	MIL-Net (2021) [79]	Weakly Supervised	MIL Framework	UCF-Crime [6]	95.5	92.3	0.91
Group B: Supervised VAD *(within-group comparisons valid; do not compare with Group A/C)*							
4	CNN-LSTM-Transformer [30]	Supervised-Hybrid	Spatiotemporal Fusion	UCF-Crime [6], UCSD	97.8	95.9	0.94
5	Hybrid Transformer-GAN	Supervised-Hybrid	Attention + Adversarial	ShanghaiTech [5], Avenue	97.3	95.5	0.94
6	3D-CNN [31]	Supervised	Spatiotemporal CNN	UCF-Crime [6], ShanghaiTech [5]	96.3	94.5	0.93
7	CNN-BiLSTM [26]	Supervised	Sequential CNN + LSTM	UCSD Ped2, UCF-Crime [6]	94.8	92.6	0.90
Group C: Unsupervised VAD *(within-group comparisons valid; do not compare with Group A/B)*							
8	Memory-Augmented AE [3]	Unsupervised	Memory Reconstruction	UCSD Ped2, Avenue	97.1	94.7	0.93
9	ConvLSTM-AE [25]	Unsupervised	Reconstruction + LSTM	UCSD Ped2, Avenue	95.4	93.2	0.91
10	f-AnoGAN [42]	Unsupervised	GAN-based	Avenue, UCSD Ped2	92.1	90.8	0.88

Methodological Caveat (RQ2). AUC and accuracy scores are reported across models operating under fundamentally different supervision paradigms, training data volumes, and test protocols. Comparing a fully supervised 3D-CNN trained on all UCF-Crime anomaly labels with an unsupervised autoencoder trained on normal frames only conflates supervision benefit with architectural superiority. The MSAD benchmark demonstrates a consistent 5–15% AUC degradation under cross-scenario evaluation, confirming that domain generalization—not raw benchmark AUC—is the field's primary open problem. Future evaluations must report: (AUC, supervision paradigm, training data volume, test protocol, cross-dataset generalization AUC). Figure 5 compares model performance across the six major architectural families, highlighting the accuracy advantage of transformer and hybrid approaches.

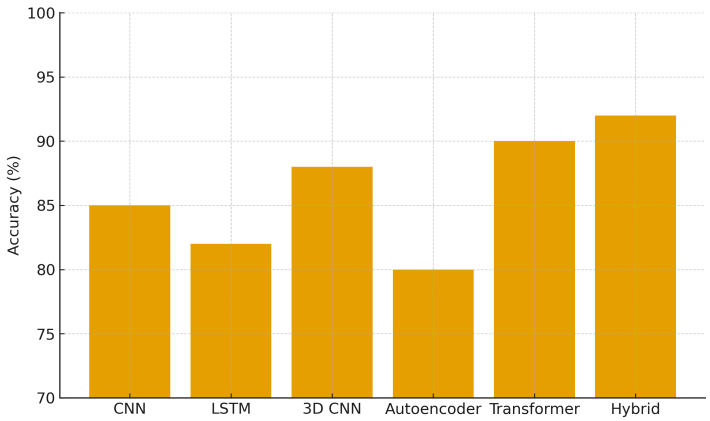

Figure 5. A comparison of how well models work across different types of architectures (CNN, LSTM, 3D CNN, Autoencoder, Transformer, and Hybrid).Results are compiled across datasets and supervision paradigms; within-paradigm comparisons yield the most scientifically robust rankings. The transformer and hybrid architectures get the best accuracy, with the hybrid family getting about 91%, which is in line with the highest reported AUC values of 97–98% [30,51].

Figure 6 presents the distribution of research methods across the 73 reviewed papers, showing the shift from handcrafted features toward learned spatiotemporal representations.

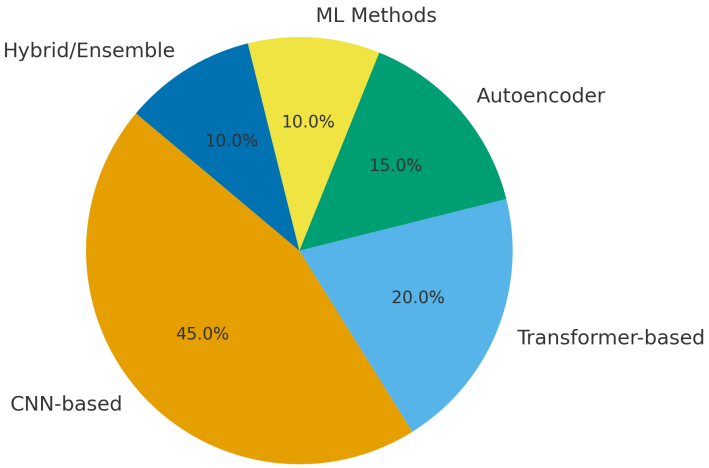

Figure 6. The 73 papers that were looked at used different research methods.CNN-backbone models make up about 45% of the total, transformer-based models make up 20%, autoencoder-based architectures make up 15%, classical ML methods make up 10%, and hybrid or ensemble combinations make up 10%.This distribution shows how, over the past ten years, feature pipelines that were made by hand have been replaced by learned spatiotemporal representations that are based on benchmark datasets like UCF-Crime and ShanghaiTech.

6. Challenges and Research Gaps

Even though intelligent video surveillance has come a long way, there are still some problems that need to be solved and areas of research that need to be filled in, such as algorithmic robustness, real-time deployment, explainability, and privacy.

High False Positives and Event Ambiguity: The most commonly mentioned problem is that there are a lot of false positives in crowded or changing environments [3,41,83].Changes in lighting, movement in the background, and strange but normal human behaviors can set off false alarms. The CNN-based systems are more likely to overfit specific dataset distributions and fail to extrapolate to new conditions well enough [14, 15].Transformer-based approaches address some issues with global context modeling, but remain vulnerable to anomaly definitions that are scene-specifically defined [51].

Lack of Explainability and Transparency: Deep-learning models are often opaque, meaning they are not explainable or transparent in their operation (see fig. 2) [1,71]. This complicates the process of trusting one another by the operators, slows incident response and makes the process of adhering to the rules more difficult. Although attention maps in transformers can offer some insights [51], there is still a lack of organized explainability modules, which can express reasoning in a way that humans can understand.The most promising direction is the integration of language models [13,72].

Real-Time Processing Constraints: In the case of smart traffic systems and public safety, having the capability to identify anomalies in real time is very important [45,47]. It takes 100–300 ms per frame to run full transformer models. MobileNetV2 lightweight models can achieve a 21 ms latency, with 4-6 percent loss in their AUC [23,63]. Partially addressing this gap are model compression, quantization, and edge-cloud split inference [10].

Data Scarcity and Annotation Overhead: The vast majority of models with high performance require big, labeled datasets that are expensive to create [10,84]. The anomalous events form a small fraction of the captured footage in the real world, leading to huge imbalance in classes [6] . In part, this problem is alleviated through weak supervision and self-supervised learning but there is still a need to find broadly applicable, cost-effective solutions to this problem [22,56,79].

Privacy and Ethical Concerns: Continuous surveillance and tracking of identities can be highly problematic in terms of privacy unlocking.On-device inference, face anonymization, and federated training are some of the privacy features that are built-in, and most systems do not yet support them. The MSAD cross-scenario analysis highlights that the models that are trained on centralized data degrade substantially in deployment [4], which also supports the idea of federated learning.

Domain Generalization: The MSAD benchmark [4] showed that no current VAD model can generalize well across different scenarios. For example, models that do well on ShanghaiTech [5] do much worse on MSAD's

different scenarios. This is recognized as the principal unresolved issue in the field.

Figure 7 summarizes the frequency distribution of key challenges cited across the 73 reviewed papers.

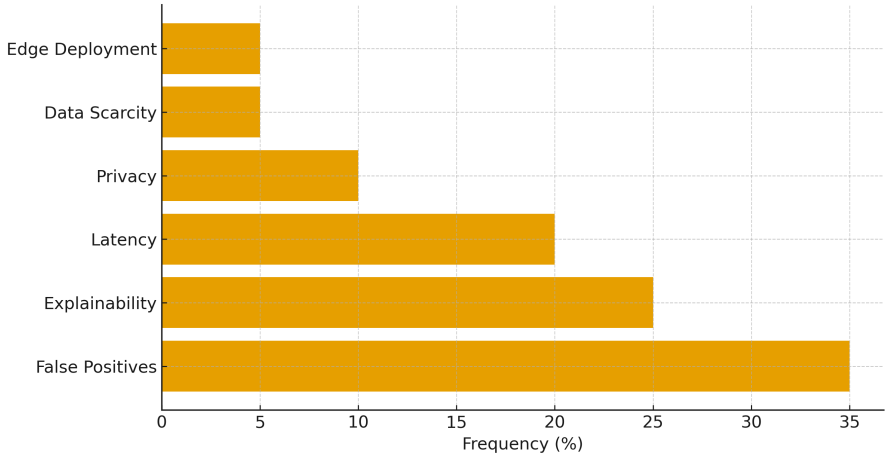

Figure 7. Frequency Distribution of Key Challenges Cited Across the 73 Reviewed Papers. Categories extracted from Tables 2–5 conflict with columns and narrative analysis. False positives/ambiguity (∼35%): Most applicable to all families; CNN models are prone to false activations because of dense and dynamic scenes, which is partially countered by transformer attention. *Explainability* (∼28%): Structured language-based explanations rather than attention maps are necessary for operator confidence and regulations. *Real-time requirements* (∼22%): Full transformers need 100–300 ms per frame, while two-stream lightweight networks operate at 21 ms per frame at an AUC penalty of 4–6%.

7. Conversational AI in Surveillance: Technical Analysis (RQ4)

This section provides the dedicated analytical treatment of Conversational AI as an architectural family (Family 10), addressing RQ4 with the same analytical depth as prior sections on CNNs, autoencoders, and transformers. Unlike passive detection systems that output binary alerts, conversational surveillance systems enable bidirectional natural-language interaction with event logs—a qualitative shift in the human–AI relationship in security operations.

7.1. Research Motivation and the Decision-Support Gap

Despite achieving AUC values above 97% on standard benchmarks, state-of-the-art deep learning systems remain fundamentally passive: they identify that an anomaly occurred, but they cannot explain what happened, answer operator questions, or contextualize events relative to historical patterns [1]. Security operators must manually review high-volume alert streams, which is cognitively demanding and limits response speed. The conversational AI paradigm addresses this *decision-support gap* by integrating detection with interactive reasoning [72,73].

7.2. Core Architectural Idea

Conversational surveillance systems integrate three architectural components: (1) a detection backbone (transformer or hybrid CNN-transformer) that continuously monitors video feeds and detects anomalies; (2) a structured event index that timestamps, annotates, and stores detected events with metadata (object interactions, motion trajectories, location context); and (3) a vision-language interface that allows operators to query the index in natural language and receive evidence-grounded responses. MLLMs—GPT-4V, LLaVA, BLIP-2—provide cross-modal grounding between visual evidence and natural-language generation [3,13].

7.3. Representative Reviewed Systems

SUVAD [11]. Integrates MLLMs with spatiotemporal visual encoders to produce both anomaly detection scores and natural-language event descriptions. Achieved AUC 95.1% and F1 0.92 on UCF-Crime/ShanghaiTech. Critically, language grounding did not degrade detection accuracy: SUVAD achieves BLEU = 32.7 on ActivityNet-Captions while maintaining detection performance competitive with pure detection models. This empirically resolves the previously assumed detection-interpretability trade-off [11].

Surveillance Video-and-Language Understanding [72]. introduced a dual-encoder framework combining visual transformers with BERT text embeddings for natural-language event summarization. Achieved BLEU = 32.7

on ActivityNet-Captions. Supports query-based retrieval (e.g., "Show all fall events near Gate 2 between 2–3 PM"), bridging detection with interactive decision support.

Deep Multimodal Transformer for Conversational Analytics [73] introduced target-aware camera placement strategies combined with natural-language query mechanisms. The framework achieves strong coverage efficiency (AUC 96.4%, BLEU 35.6) and supports natural-language event queries aligned with CUVA-style causal annotations [4].

7.4. Performance Characteristics vs. Detection-Only Models

Key finding: Conversational systems show a 2–3% AUC gap relative to pure detection baselines (within the same data conditions). This gap is attributable to the additional computational budget allocated to language generation rather than architectural inferiority. SUVAD's 95.1% AUC is competitive with non-language baselines from the same period, confirming that interpretability and detection performance are not fundamentally at odds when MLLMs are properly integrated [11].

Conversational AI systems introduce a language-generation overhead that may affect detection performance. Table 9 quantifies this trade-off across the three reviewed conversational systems.

Table 9. Conversational AI Systems vs. Detection-Only Baselines (VAD Task). Scope note:BLEU is a language generation metric and is not comparable to AUC. The key finding is that language grounding does not significantly degrade detection AUC, resolving the assumed detection-interpretability trade-off.

System	Type	Architecture	Dataset	AUC (%)	F1	BLEU	NL Interface
SUVAD [11]	Conversational	MLLM + Visual Encoder	UCF-Crime / ShanghaiTech	95.1	0.92	32.7	Yes
Yuan et al. [72]	Conversational	BERT Dual-Encoder	ActivityNet-Captions	—	—	32.7	Yes
Wu et al. [73]	Conversational	Cross-Modal Transformer	Large-scale CCTV	96.4	—	35.6	Yes
Transformer-AE (2023)	Detection only	Transformer AE	Avenue, ShanghaiTech	98.2	0.95	—	No
CNN-LSTM-Transformer [30]	Detection only	Spatiotemporal Hybrid	UCF-Crime	97.8	0.94	—	No
Memory-Augmented AE [3]	Detection only	Memory Reconstruction	UCSD Ped2	97.1	0.93	—	No

7.5. Technical Challenges of Conversational Surveillance (RQ4 Expanded)

Deploying conversational AI in surveillance settings introduces five distinct technical challenges beyond standard VAD:

Real-time language grounding: Current MLLMs such as GPT-4V and LLaVA require 500 ms–2 s per query, which is prohibitive for continuous real-time monitoring. Streaming inference, speculative decoding, and distilled lightweight MLLM variants (e.g., LLaVA-1.5-7B) are required to bring query latency below 100 ms [11]. The challenge is maintaining semantic grounding quality while aggressively reducing model size.

Hallucination control: MLLMs are prone to generating plausible but factually incorrect descriptions of surveillance events—a critical failure mode in security contexts. Grounding the language model's output to the structured event index (timestamps, object tracks, bounding boxes) rather than allowing open-ended generation reduces hallucination risk. Retrieval-augmented generation (RAG) architectures, where the model queries a verified evidence database before responding, are the most promising mitigation [73].

Evidence retrieval: Operators may query events from hours or days of indexed footage. Efficient temporal indexing over large event databases requires learned sparse retrieval models rather than exhaustive search. The spatio-temporal association query algorithm [66] provides a classical foundation; transformer-based attention indexing [73] extends this to semantic queries. CUVA causal annotations [4] provide the supervision signal needed to train retrievers that understand causal event relationships.

Privacy protection in conversational access. Conversational interfaces that respond to natural-language queries over event logs present a novel privacy attack surface: an adversary with access to the query interface may be able to extract information about individuals not involved in any anomaly. Role-based access control, differential privacy on stored event metadata, on-device inference to prevent raw video transmission, and audit logging of all queries are necessary architectural components for compliant deployment [71].

Operator trust and calibration: Operators must calibrate their trust in system responses appropriately—over-reliance on system outputs reduces vigilance; under-reliance defeats the system's purpose. Explainability features (visual evidence grounding, confidence scores, uncertainty quantification) and staged deployment with human-in-the-loop validation are essential for building appropriate operator trust [1,71].

7.6. Technical Foundations Required for Conversational Surveillance (RQ4)

Four technical components are required to realize conversational surveillance at deployment scale:

1. Real-time spatiotemporal indexing: Detected events must be timestamped, geocoded, and indexed with structured metadata to support arbitrary natural-language queries without re-processing raw video [73].
2. Causal annotation and grounding: Language-grounded systems require training data with causal text annotations. The CUVA benchmark [4] provides counterfactual video–language pairs enabling models to reason about *why* an event is anomalous, not just *that* it is.
3. Efficient MLLM inference: Current MLLMs are prohibitively expensive for continuous real-time inference. Lightweight variants with optimized attention and speculative decoding are required for edge deployment [11].
4. Privacy-preserving architecture: Conversational systems must implement role-based access control, differential privacy on stored metadata, and on-device inference to comply with data protection regulations [71].

7.7. Conversational AI as the Field's Next Architectural Frontier

The trajectory across ten architectural families reveals a consistent displacement logic: each family addressed a structural limitation of its predecessor. Family 10 addresses the limitation that all prior families share: they detect anomalies but cannot explain or communicate them. The decision-support gap—the inability to connect detection outputs to operator reasoning—is the structural motivation for conversational AI in surveillance, just as gradient vanishing was the structural motivation for transformers over LSTMs.

8. Future Work

The succeeding generations of surveillance systems must go beyond static anomaly detection and enable contextual understanding, interactive querying, and intelligent response [13,72,73].

An upcoming direction combines deep-learning-based event detection with conversational and explainable AI [73]. Continuous real-time analysis of surveillance feeds using lightweight transformer or hybrid vision models detects abnormal events such as fights, falls, intrusions, or accidents [3,51]. Every detected event is automatically timestamped, indexed, and stored along with metadata consisting of object interactions, motion trajectories, and location context. A conversational interface allows security personnel to query the system: "What happened at 2:15 PM?" or "Show all accidents near the north gate today", receiving context-rich summaries with visual evidence [73]. The CUVA benchmark [4] provides causal text annotations enabling training of such systems on surveillance-domain video language pairs.

Integration of multimodal large language models (MLLMs) GPT-4V,LLaVA, and BLIP-2 brings zero-shot anomaly categorization, causal reasoning, and cross-modal grounding [3,13]. The MMI-FM model ("Learnable Expansion of Graph Operators for Multi-Modal Feature Fusion", ICLR 2025) proposed a novel approach: rather than treating multi-modal fusion as simple concatenation or attention, it learns to expand graph operators to model the inter-modal dependency structure explicitly. Dual contrastive loss aligns cross-modal representations while the graph-operator structure handles heterogeneous sensor streams with different sampling rates and spatial resolutions, directly addressing the synchronization bottleneck [61] that limits prior multimodal VAD systems. Integrating MMI-FM's graph-operator fusion with conversational VAD interfaces [73] represents the most technically complete path to next-generation surveillance systems.

By adopting MSAD benchmark [4] as a secondary evaluation criterion along with existing benchmarks, generalization across multiple scenarios becomes an important area to investigate. Domain adaptation using synthetic dataset and few-shot scene adaptation using MAML and Prototypical Networks are two important areas worth investigating.

Federated and privacy-preserving learning [71] allows for cooperative learning using a set of surveillance devices without transferring raw video to any server. The use of federated VAD along with differential privacy and face anonymization can solve both problems at once.

Adaptive and edge-optimized models [10] focus on RL-based frame selection and knowledge distillation from transformer to lightweight CNN , preserving accuracy at edge-deployable latency. Ethical AI integration ensuring fairness, transparency, and conformity with emerging global surveillance regulations is essential for responsible deployment.

Figure 8 proposes a concrete architecture realizing this direction.

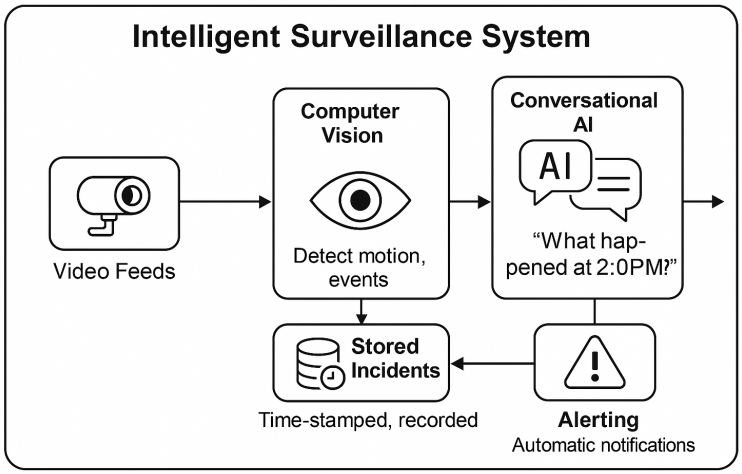

Figure 8. Proposed ConversationalAI-Powered Surveillance Architecture. Camera feeds are processed by a real-time vision module for event detection; detected events are timestamped and indexed in a structured event database; a conversational AI interface enables natural-language querying by security operators; and privacy-preserving & Security modules ensure consent with data protection regulations. The architecture unifies Families 6, 8, 9, and 10 from the proposed taxonomy.

9. Conclusions

This review has systematically examined seventy-three deep-learning-based contributions to intelligent video surveillance from 2018 to 2025, organized under a formally defined ten-family taxonomy following a structured methodology of database search, inclusion/exclusion filtering, and per-family architectural displacement analysis. The review traced a clear performance arc from CNN classifiers (∼85% AUC) [1] through recurrent and 3D-CNN models (93–96% AUC) [31, 85] to transformer-based and memory-augmented architectures exceeding 97–98% AUC [3, 51], with hybrid CNN-LSTM-Transformer pipelines currently representing the empirical ceiling [30]. This progression was driven by structural architectural advances—self-attention resolving gradient vanishing, memory augmentation resolving reconstruction bias, graph-operator fusion resolving multimodal synchronization—not merely by increased data or compute.

Four research priorities stand out as most critical for the field's next phase. First, cross-scenario generalization must move to the center of evaluation practice: the MSAD benchmark [3, 4] demonstrates that 5–15% AUC degradation under cross-scenario evaluation is universal, and no paper in this review fully solves it. Domain-adaptive pretraining, meta-learning, and scenario-diverse synthetic data generation are the highest-priority research directions. Second, lightweight transformer deployment on edge cameras is necessary to bridge the gap between transformer-level accuracy and the real-time latency constraints of deployed surveillance systems. Sparse attention, knowledge distillation, and hybrid edge-cloud inference architectures must mature before transformer-based VAD can replace CNN-based systems at scale. Third, privacy-preserving federated learning [71] is required to enable collaborative training across distributed surveillance nodes without centralizing raw video data. Fourth, multimodal conversational surveillance represents the field's most transformative near-term direction. The SUVAD finding [11] that language grounding does not degrade detection AUC resolves the assumed detection-interpretability trade-off and demonstrates that systems can simultaneously detect, explain, and communicate anomalies. Future surveillance systems will not merely generate alerts but engage in evidence-grounded dialogue with security operators—making surveillance smarter, safer, and more human-centered.

Author Contributions

M.S.: Conceptualization, Methodology, Software, Data Curation, Writing—Original Draft Preparation; N.P.: Data Curation, Visualization, Investigation; K.L.: Software, Validation, Writing—Review & Editing; K.R.: Conceptualization, Supervision, Methodology, Writing—Review & Editing, Project Administration; All authors have read and agreed to the published version of the manuscript.

Funding

This research has received no external funding

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Grammarly, Claude and Copilot to improve the clarity, grammar, and readability of the manuscript. After using these tools, the authors reviewed and edited the content

References

1. Kiran, B.R.; Thomas, D.M.; Parakkal, R. An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos. J.Imaging 2018, *4*, 36. https://doi.org/10.3390/jimaging4020036.
2. Rezaee, K.; Rezakhani, S.M.; Khosravi, M.R.; et al. A Survey on Deep Learning-Based Real-Time Crowd Anomaly Detection for Secure Distributed Video Surveillance. *Pers. Ubiquitous Comput.* **2024**, *28*, 135–151. https://doi.org/10.1007/s00779-021-01586-5.
3. Barbosa, R.Z.; Oliveira, H.S.; Tavares, J.M.R. A survey on multi-modal and weakly supervised approaches for robust anomaly detection in video data. *Inf. Fusion* **2026**, *126*, 103388. https://doi.org/10.1016/j.inffus.2025.103388.
4. Zhu, L.; Wang, L.; Raj, A.; et al. Advancing Video Anomaly Detection: A Concise Review and a New Dataset. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 89943–89977.
5. Liu, W.; Luo, W.; Lian, D.; et al. Future Frame Prediction for Anomaly Detection—A New Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
6. Alshalawi, A.; Abdul, W.; Muhammad, G. Advanced Detection of Violence From Video: Performance Evaluation of Transformer and State of the Art of Convolution of Neural Network Transformer. *IEEE Access* **2025**, *13*, 74200–74216. https://doi.org/10.1109/ACCESS.2025.3564435.
7. Nayak, R.; Pati, U.C.; Das, S.K. A Comprehensive Review on Deep Learning-Based Methods for Video Anomaly Detection. *Image Vis. Comput.* **2021**, *106*, 104078. https://doi.org/10.1016/j.imavis.2020.104078.
8. Ramachandra, B.; Jones, M.J.; Vatsavai, R.R. A Survey of Single-Scene Video Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2293–2312. https://doi.org/10.1109/TPAMI.2020.3040591.
9. Du, H.; Zhang, S.; Xie, B.; et al. Uncovering What, Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* **2024**, 18793–18803. https://arxiv.org/abs/2405.00181.
10. Ullah, F.U.M.; Muhammad, K.; Haq, I.U.; et al. AI-Assisted Edge Vision for Violence Detection in IoT-Based Industrial Surveillance Networks. *IEEE Trans. Ind. Inform.* **2022**, *18*, 5359–5370. https://doi.org/10.1109/TII.2021.3116377.
11. Gao, S.; Yang, P.; Huang, L. SUVAD: Semantic Understanding Based Video Anomaly Detection Using MLLM. In Proceedings of the ICASSP 2025—IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025; pp. 1–5.
12. Ding, D.; Wang, L.; Zhu, L.; et al. Learnable Expansion of Graph Operators for Multi-Modal Feature Fusion. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
13. Duja, K.U.; Khan, I.A.; Alsuhaibani, M. Video Surveillance Anomaly Detection: A Review on Deep Learning Benchmarks. *IEEE Access* **2024**, *12*, 164811–164842. https://doi.org/10.1109/ACCESS.2024.3491868.
14. Ul Amin, S.; Sibtain Abbas, M.; Kim, B.; et al. Enhanced Anomaly Detection in Pandemic Surveillance Videos: An Attention Approach With EfficientNet-B0 and CBAM Integration. *IEEE Access* **2024**, *12*, 162697–162712. https://doi.org/10.1109/ACCESS.2024.3488797.
15. Huang, C.; Wu, Z.; Wen, J.; et al. Abnormal Event Detection Using Deep Contrastive Learning for Intelligent Video Surveillance System. *IEEE Trans. Ind. Inform.* **2022**, *18*, 5171–5179. https://doi.org/10.1109/TII.2021.3122801.
16. Zhang, M.; Wang, J.; Qi, Q.; et al. Cognition Guided Video Anomaly Detection Framework for Surveillance Services. *IEEE Trans. Serv. Comput.* **2024**, *17*, 2109–2123. https://doi.org/10.1109/TSC.2024.3407588.

17. Muhammad, K.; Ahmad, J.; Lv, Z.; et al. Efficient Deep CNN-Based Fire Detection and Localization in Video Surveillance Applications. *IEEE Trans. Syst. Man, Cybern. Syst.* **2018**, *49*, 1419–1434.
18. Tahir, M.; Qiao, Y.; Kanwal, N.; et al. Real-Time Event-Driven Road Traffic Monitoring System Using CCTV Video Analytics. *IEEE Access* **2023**, *11*, 139097–139111.
19. Zhou, W.; Liu, Y.; Wang, C.; et al. An Automated Learning Framework With Limited and Cross-Domain Data for Traffic Equipment Detection From Surveillance Videos. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 24891–24903. https://doi.org/10.1109/TITS.2022.3195509.
20. Ranjana Panigrahi, G.; Kumar Sethy, P.; Kumari Behera, S.; et al. Enhancing Security in Real-Time Video Surveillance: A Deep Learning-Based Remedial Approach for Adversarial Attack Mitigation. *IEEE Access* **2024**, *12*, 88913–88926. https://doi.org/10.1109/ACCESS.2024.3418614.
21. Yan, J.; Yang, Y.; Naqvi, S.M. Object Detection Oriented Privacy-Preserving Frame-Level Video Anomaly Detection. In Proceedings of the ICASSP 2024—IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 7640–7644.
22. Zahid, Y.; Tahir, M.A.; Durrani, N.M.; et al. IBaggedFCNet: An Ensemble Framework for Anomaly Detection in Surveillance Videos. *IEEE Access* **2020**, *8*, 220620–220630. https://doi.org/10.1109/ACCESS.2020.3042222.
23. Peng, C.; Jiang, Z.; Lin, M.; et al. Real-Time Human Action Anomaly Detection Through Two-Stream Spatial-Temporal Networks. *IEEE Access* **2025**. https://doi.org/10.1109/ACCESS.2025.3560703.
24. Bukhari, S.M.S.; Zafar, M.H.; Moosavi, S.K.R.; Khan, N.M.; Sanfilippo, F. FireNet: A Hybrid Deep Learning Approach for Enhanced Fire Detection in Remote Sensing Imagery. In Intelligent Systems and Applications; Arai, K., Ed.; Springer Nature Switzerland: Cham, Switzerland, **2024**. https://doi.org/10.1007/978-3-031-66329-1_1.
25. Ullah, W.; Ullah, A.; Haq, I.U.; et al. CNN Features with Bi-Directional LSTM for Real-Time Anomaly Detection in Surveillance Networks. *Multimed. Tools Appl.* **2021**, *80*, 16979–16995. https://doi.org/10.1007/s11042-020-09406-3.
26. Butt, U.M.; Letchmunan, S.; Hassan, F.H.; et al. Detecting Video Surveillance using VGG19 Convolutional Neural Networks. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 674–682.
27. Khan, N.; Nabi, M.A.; Khalid, M.S.; et al. Human Activity Recognition via Hybrid Deep Learning Based Model. *Sensors* **2022**, *22*, 323. https://doi.org/10.3390/s22010323.
28. Vishnu Priya, P.; Rajeswari, R. Anomalous Human Activity Recognition from Video Sequences using Brisk Features and Convolutional Neural Networks. *Galaxy Int. Interdiscip. Res. J.* **2022**, *10*, 269–276.
29. Mohamed Zaidi, M.; Avelino Sampedro, G.; Almadhor, A.; et al. Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning. *IEEE Access* **2024**, *12*, 105497–105510. https://doi.org/10.1109/ACCESS.2024.3436653.
30. Natha, S.; Siraj, M.; Ahmed, F.; et al. An Integrated CNN-BiLSTM-Transformer Framework for Improved Anomaly Detection Using Surveillance Videos. *IEEE Access* **2025**, *13*, 95341–95357. https://doi.org/10.1109/ACCESS.2025.3574835.
31. Ouyang, Z.; Chen, J.; Pan, Y.; et al. A 3D-CNN and LSTM Based Multi-Task Learning Architecture for Action Recognition. *IEEE Access* **2019**, *4*, 1–14. https://doi.org/10.1109/ACCESS.2019.2906654.
32. Nawaratne, R.; Alahakoon, D.; De Silva, D.; et al. Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance. *IEEE Trans. Ind. Inform.* **2020**, *16*, 393–402. https://doi.org/10.1109/TII.2019.2938527.
33. Huszar, V.D.; Adhikarla, V.K.; Negyesi, I.; et al. Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications. *IEEE Access* **2023**, *11*, 18772–18793. https://doi.org/10.1109/ACCESS.2023.3245521.
34. Li, J.; Jiang, X.; Sun, T.; et al. Efficient Violence Detection Using 3D Convolutional Neural Networks. In Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
35. Ullah, F.U.M.; Ullah, A.; Muhammad, K.; et al. Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network. *Sensors* **2019**, *19*, 2472. https://doi.org/10.3390/s19112472.
36. Mudgal, M.; Punj, D.; Pillai, A. Suspicious Action Detection in Intelligent Surveillance System Using Action Attribute Modelling. *J. Web Eng.* **2021**, *20*, 129–146. https://doi.org/10.13052/jwe1540-9589.2017.
37. Choi, A.; Kim, T.H.; Yuhai, O.; et al. Deep Learning-Based Near-Fall Detection Algorithm for Fall Risk Monitoring System Using a Single Inertial Measurement Unit. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 2385–2394. https://doi.org/10.1109/TNSRE.2022.3199068.
38. Ramirez, H.; Velastin, S.A.; Meza, I.; Fabregas, E.; Makris, D.; Farias, G. Fall Detection and Activity Recognition Using Human Skeleton Features. *IEEE Access* **2021**, *9*, 33532–33542. https://doi.org/10.1109/ACCESS.2021.3061626.
39. Honnegowda, H.C.; Rao, V.S.; Prasad, K.S. An Efficient Abnormal Event Detection System Using Deep Learning-Based Reconfigurable Autoencoder. *IEEE Access* **2024**, *29*, 677.
40. Niaz, A.; Amin, S.U.; Soomro, S.; et al. Spatially Aware Fusion in 3D Convolutional Autoencoders for Video Anomaly Detection. . *IEEE Access* **2024**, *12*, 104770–104784. https://doi.org/10.1109/ACCESS.2024.3435144.
41. Qiang, Y.; Fei, S.; Jiao, Y. Anomaly Detection Based on Latent Feature Training in Surveillance Scenarios. *IEEE Access* **2021**, *9*, 68108–68117. https://doi.org/10.1109/ACCESS.2021.3077577.
42. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; et al. f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative

Adversarial Networks. *Med. Image Anal.* **2019**, *54*, 30–44. https://doi.org/10.1016/j.media.2019.01.010.

43. Song, W.; Zhang, D.; Zhao, X.; et al. A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks. *IEEE Access* **2019**, *7*, 39172–39179. https://doi.org/10.1109/ACCESS.2019.2906275.
44. Ravanbakhsh, M.; Sangineto, E.; Nabi, M.; et al. Training Adversarial Discriminators for Cross-Channel Abnormal Event Detection in Crowds. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1896–1904.
45. Khan, H.; Yuan, X.; Qingge, L.; et al. Violence Detection From Industrial Surveillance Videos Using Deep Learning. *IEEE Access* **2025**, *13*, 15363–15375. https://doi.org/10.1109/ACCESS.2025.3531213.
46. Choqueluque-Roman, D.; Camara-Chavez, G. Weakly Supervised Violence Detection in Surveillance Video. *Sensors* **2022**, *22*, 4502. https://doi.org/10.3390/s22124502.
47. Qaraqe, M.; Elzein, A.; Basaran, E.; et al. PublicVision: A Secure Smart Surveillance System for Crowd Behavior Recognition. *IEEE Access* **2024**, *12*, 26474–26491. https://doi.org/10.1109/ACCESS.2024.3366693.
48. Priya, S.; Nayak, R.; Pati, U.C. Deep Learning-based Weakly Supervised Video Anomaly Detection Methods for Smart City Applications. In Proceedings of the 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), Vellore, India, 3–4 May 2024; pp. 1–6.
49. Xia, Z.; Zhou, K.; Tan, J.; et al. Bidirectional LSTM-Based Attention Mechanism for CNN Power Theft Detection **2022**. In Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Wuhan, China, 9–11 December 2022; pp. 323–330. https://doi.org/10.1109/TrustCom56396.2022.00052.
50. Sernani, P.; Falcionelli, N.; Tomassini, S.; et al. Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset. *IEEE Access* **2021**, *9*, 160580–160595. https://doi.org/10.1109/ACCESS.2021.3131315.
51. Chatterjee, R.; Roy Choudhury, R.; Kumar Gourisaria, M.; et al. Temporal-Aware Transformer Approach for Violence Activity Recognition. *IEEE Access* **2025**, *13*, 70779–70790. https://doi.org/10.1109/ACCESS.2025.3560828.
52. Liu, M.; Xu, Z. Video Anomaly Detection Based on Spatial Awareness and Attention Fusion Method. 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 11–13 August 2023; pp. 121–125.
53. Dilek, E.; Dener, M. Enhancement of Video Anomaly Detection Performance Using Transfer Learning and Fine-Tuning. *IEEE Access* **2024**, *12*, 73304–73322. https://doi.org/10.1109/ACCESS.2024.3404553.
54. Sanjalawe, Y.; Fraihat, S.; Abualhaj, M.; et al. Hybrid Deep Learning for Human Fall Detection: A Synergistic Approach Using YOLOv8 and Time-Space Transformers. *IEEE Access* **2025**, *13*, 41336–41366. https://doi.org/10.1109/ACCESS.2025.3547914.
55. Altundogan, T.G.; Karaköse, M.; Mert, F. A New Multi Objective Video Summarization Approach for Video Surveillance Analytics Applications on Smart Cities. *IEEE Access* **2025**, *13*, 154353–154382. https://doi.org/10.1109/ACCESS.2025.3605259.
56. Shin, J.; Kaneko, Y.; Miah, A.S.M.; et al. Anomaly Detection in Weakly Supervised Videos Using Multistage Graphs and General Deep Learning Based Spatial-Temporal Feature Enhancement. *IEEE Access* **2024**, *12*, 65213–65227. https://doi.org/10.1109/ACCESS.2024.3395329.
57. Singh, R.; Pal, A.; Mishra, S.; et al. Enhancing Situational Awareness: Anomaly Detection Using Real-Time Video Across Multiple Domains. *IEEE Access* **2025**, *13*, 73680–73696.
58. Bergaoui, K.; Naji, Y.; Setkov, A.; et al. Object-Centric and Memory-Guided Normality Reconstruction for Video Anomaly Detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP). IEEE, 2022, Bordeaux, France, 16–19 October 2022; pp. 2691–2695.
59. Ning, Z.; Wang, Z.; Liu, Y.; et al. Memory-Enhanced Appearance-Motion Consistency Framework for Video Anomaly Detection. *Comput. Commun.* **2024**, *216*, 159–167.
60. Batool, M.; Alotaibi, M.; Alotaibi, S.R.; et al. Multimodal Human Action Recognition Framework Using an Improved CNNGRU Classifier. *IEEE Access* **2024**, *12*, 158388–158406. https://doi.org/10.1109/ACCESS.2024.3481631.
61. Srilakshmi, V.; Veesam, S.B.; Krishna, M.S.R.; et al. Design of an Improved Model for Anomaly Detection in CCTV Systems Using Multimodal Fusion and Attention-Based Networks. *IEEE Access* **2025**, *13*, 27287–27309. https://doi.org/10.1109/ACCESS.2025.3536501.
62. Shin, J.; Miah, A.S.M.; Kaneko, Y.; et al. Multimodal Attention-Enhanced Feature Fusion-Based Weakly Supervised Anomaly Violence Detection. *IEEE Open J. Comput. Soc.* **2025**, *6*, 129–140. https://doi.org/10.1109/OJCS.2024.3517154.
63. Shao, L.; Liu, L.; Li, X. Smart Monitoring Cameras Driven Intelligent Processing to Big Surveillance Video Data. *IEEE Trans. Big Data* **2018**, *4*, 105–116. https://doi.org/10.1109/TBDATA.2017.2715815.
64. Yan, K.; Shan, H.; Sun, T.; et al. Reinforcement Learning-Based Mobile Edge Computing and Transmission Scheduling for Video Surveillance. *IEEE Trans. Emerg. Top. Comput.* **2021**, *10*, 1142–1156.
65. Wu, Z.; Li, H.; Xiong, C.; et al. A Dynamic Frame Selection Framework for Fast Video Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1699–1711. https://doi.org/10.1109/TPAMI.2020.3029425.
66. Zhang, J. Spatio-Temporal Association Query Algorithm for Massive Video Surveillance Data in Smart Campus. *IEEE Access* **2018**, *6*, 53894–53904.

67. Zhao, Y.; Wang, X.; Yu, X.; et al. Gait-Assisted Video Person Retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 897–908. https://doi.org/10.1109/TCSVT.2022.3202531.
68. Fan, C.T.; Wang, Y.K.; Huang, C.R. Heterogeneous Information Fusion and Visualization for a Large-Scale Intelligent Video Surveillance System. *IEEE Trans. Syst. Man, Cybern. Syst.* **2017**, *47*, 593–604. https://doi.org/10.1109/TSMC.2016.2531671.
69. Chen, X.; Qing, L.; He, X.; et al. From Eyes to Face Synthesis: A New Approach for Human-Centered Smart Surveillance. *IEEE Access* **2018**, *6*, 14567–14575.
70. Zhao, L.; Wang, S.; Wang, S.; et al. Enhanced Surveillance Video Compression With Dual Reference Frames Generation. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1592–1606. https://doi.org/10.1109/TCSVT.2021.3073114.
71. Yousuf, M.J.; Lee, B.; Asghar, M.N.; et al. Unlocking Trust: Advancing Activity Recognition in Video Imagery. *IEEE Access* **2024**, *12*, 176799–176817. https://doi.org/10.1109/ACCESS.2024.3503284.
72. Yuan, T.; Zhang, X.; Liu, B.; et al. Surveillance Video-and-Language Understanding: From Small to Large Multimodal Models. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 300–314. https://doi.org/10.1109/TCSVT.2024.3462433.
73. Wu, H.; Zeng, Q.; Guo, C.; et al. Target-Aware Camera Placement for Large-Scale Video Surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 13338–13348. https://doi.org/10.1109/TCSVT.2024.3445151.
74. Wan, W.; Zhang, W.; Jin, C. Pose-Motion Video Anomaly Detection via Memory-Augmented Reconstruction and Conditional Variational Prediction. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). IEEE, Brisbane, Australia, 10–14 July 2023; pp. 2729–2734.
75. Kwan-Loo, K.B.; Ortíz-Bayliss, J.C.; Conant-Pablos, S.E.; et al. Detection of Violent Behavior Using Neural Networks and Pose Estimation. *IEEE Access* **2022**, *10*, 86339–86352. https://doi.org/10.1109/ACCESS.2022.3198985.
76. Wu, P.; Liu, J.; He, X.; et al. Toward Video Anomaly Retrieval from Video Anomaly Detection: New Benchmarks and Model. *IEEE Trans. Image Process.* **2024**, *33*, 2213–2225.
77. Castillo, A.; Tabik, S.; Pérez, F.; et al. Brightness Guided Preprocessing for Automatic Cold Steel Weapon Detection in Surveillance Videos with Deep Learning. *Neurocomputing* **2019**, *330*, 151–161. https://doi.org/10.1016/j.neucom.2018.10.076.
78. Tang, Y.; Zhao, L.; Zhang, S.; et al. Integrating Prediction and Reconstruction for Anomaly Detection. *Pattern Recognit. Lett.* **2020**, *129*, 123–130. https://doi.org/10.1016/j.patrec.2019.11.024.
79. Aljaloud, A.S.; Ullah, H. IA-SSLM: Irregularity-Aware Semi-Supervised Deep Learning Model for Analyzing Unusual Events in Crowds. *IEEE Access* **2021**, *9*, 73327–73334. https://doi.org/10.1109/ACCESS.2021.3081050.
80. Nowshin, F.; Dong, Z.; Yi, Y. Memory-Augmented Autoencoder with Reservoir Computing for Edge-Based Anomaly Detection in Autonomous Systems. *IEEE Internet Comput.* **2025**, *29*, 44–52. https://doi.org/10.1109/MIC.2025.3594330.
81. Sultani, W.; Chen, C.; Shah, M. Real-World Anomaly Detection in Surveillance Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
82. Wu, P.; Liu, J.; Shi, Y.; et al. Not Only Look, but Also Listen: Learning Multimodal Violence Detection under Weak Supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 322–339.
83. Yin, C.; Tang, J.; Xu, Z.; et al. Memory Augmented Deep Recurrent Neural Network for Video Question Answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 3159–3167. https://doi.org/10.1109/TNNLS.2019.2938015.
84. Bianculli, M.; Falcionelli, N.; Sernani, P.; et al. A Dataset for Automatic Violence Detection in Videos. *Data Brief* **2020**, *33*, 106587. https://doi.org/10.1016/j.dib.2020.106587.
85. Kang, M.S.; Park, R.H.; Park, H.M. Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition. *IEEE Access* **2021**, *9*, 76270–76285. https://doi.org/10.1109/ACCESS.2021.3083273.