

A Survey on Intelligent Surveillance Systems for Classification of Usual and Unusual Events Using Deep Learning Models

Rathan J¹, Harisha Naik T²

¹Department of Computer Science, Dr. Manmohan Singh Bengaluru City University

²Department of Computer Applications, Presidency College (Autonomous)

Peer Review Information	Abstract
Type: Article Received: 09 July 2026 Revised: 13 Aug 2026 Accepted: 11 Sept 2026 Published: 15 Oct 2026	The proliferation of closed-circuit television and networked camera infrastructure has outpaced the human capacity to monitor it, motivating sustained research into intelligent surveillance systems capable of distinguishing usual from unusual events automatically. Over the past decade the field has migrated decisively from handcrafted motion and appearance descriptors towards deep representation learning, and more recently towards weakly supervised, self-supervised, and vision–language paradigms. This survey provides a critical and comparative examination of that trajectory, drawing on more than ninety studies published principally between 2019 and 2026 together with a smaller set of earlier works that remain foundational. The classification problem is first formalised, and it is established that the field employs at least four distinct and only partially compatible problem definitions, a circumstance that materially complicates cross-study comparison. A two-axis taxonomy organised by supervision regime and by learning objective is then proposed and used to structure a detailed analysis of reconstruction-based, prediction-based, memory-augmented, adversarial, graph-based, transformer-based, and vision–language architectures. Particular attention is given to the weakly supervised multiple-instance paradigm, which now dominates work on real-world untrimmed footage. The principal benchmarks and evaluation protocols are consolidated, and it is argued that the widespread reliance on frame-level area under the receiver operating characteristic curve has produced an appearance of saturation that conceals persistent weaknesses in localisation, generalisation, and false-alarm behaviour. Comparative performance tables are presented alongside an analysis of the experimental conditions that govern their comparability. The survey closes by identifying research gaps relating to scene dependence, dataset bias, domain shift, explainability, adversarial robustness, and deployment under edge and fog computing constraints, and by outlining the directions that appear most promising for the next phase of the field. Keywords: Video Anomaly Detection; Intelligent Video Surveillance; Deep Learning; Weakly Supervised Learning; Unusual Event Classification; Spatiotemporal Representation Learning; Vision–Language Models

How to Cite This Article

Rathan, J., & Naik, H. T. (2026). A survey on intelligent surveillance systems for classification of usual and unusual events using deep learning models. *International Journal on Advanced Computer Engineering and Communication Technology*, 15(2), 292–315.

Introduction

Motivation And Context

Video surveillance has become one of the most pervasive sensing modalities in the built environment. Municipal authorities, transport operators, industrial facilities, retail chains, and private householders collectively operate an installed base of cameras that generates a volume of footage far exceeding what trained personnel can meaningfully review. The operational consequence is well known: human attention degrades rapidly during sustained monitoring tasks, and the events that surveillance systems are nominally deployed to capture—assaults, thefts, road accidents, unattended baggage, crowd crushes, equipment failures—are rare by nature. A monitoring regime where an operator has to sit and watch for hours on end to spot an event that might happen once in several weeks is not one that can be relied on to work reliably.

Automated analysis is thus an attractive proposition, the most common task being binary or multi-class discrimination of usual vs unusual events. The framing is intuitive and appears tractable at first glance. Closer examination reveals considerable subtlety. What constitutes an unusual event is not an intrinsic property of the visual signal but a relational property defined against a context that is spatial, temporal, cultural, and operational. A person running is unremarkable on a jogging path and alarming in a bank lobby. A vehicle travelling against the flow of traffic is anomalous on a motorway and routine in a car park. The same visual pattern may be normal at one hour and anomalous at another. Any system that seeks to classify events as usual or unusual must therefore encode, explicitly or implicitly, a model of context — and it is precisely the difficulty of doing so that has sustained the research problem for more than two decades.

The Definitional Problem

A recurring source of confusion in the literature is that several distinct problems are conducted under a common vocabulary. It is worth separating them at the outset.

- The first is *semi-supervised* or *one-class* anomaly detection, in which the training set contains only normal footage and the system must flag departures from the learned distribution at test time. This is the formulation underlying the classical benchmarks such as UCSD Pedestrian and CUHK Avenue.
- The second is *weakly supervised* anomaly detection, in which training videos carry video-level labels indicating whether an anomaly occurs somewhere within them, but no frame-level annotation. This formulation, popularised by the introduction of the UCF-Crime dataset, has become dominant for realistic untrimmed surveillance footage.
- The third is *fully supervised* event classification, in which the system is trained to recognise a closed set of named event categories such as fighting, shoplifting, or arson. This is essentially an action recognition problem applied to surveillance data, and it inherits both the strengths and the limitations of that field.
- The fourth is *unsupervised* anomaly detection, in which the training data are unlabelled and may themselves contain anomalies. This is the most realistic setting for deployment at scale and the least studied.

These four problems differ in their assumptions, their evaluation protocols, and their failure modes. Results obtained under one formulation are not directly comparable to results obtained under another, yet such comparisons appear routinely. A significant contribution of the present survey is to keep these formulations carefully distinguished throughout.

Scope And Contributions

This survey concerns the automated classification of usual and unusual events in video acquired from fixed or quasi-fixed surveillance cameras, with an emphasis on deep learning approaches. We consider single-scene and multi-scene settings, frame-level and segment-level decisions, and both detection and localisation objectives. We exclude from scope: anomaly detection in non-visual modalities except where multimodal fusion is directly relevant; person re-identification and tracking except where they serve as components of an event classification pipeline; and general-purpose action recognition except where architectures developed there have been adopted as surveillance backbones.

The contributions of the survey are as follows.

- First, we provide a structured taxonomy organised along two orthogonal axes — supervision regime and learning objective — and demonstrate that this two-axis view accounts for the architectural diversity of the field more economically than the single-axis taxonomies commonly employed.
- Second, we conduct a critical rather than descriptive comparison. For each architectural family we identify the inductive bias it encodes, the assumption on which that bias depends, and the circumstances under which the assumption fails. We regard this analysis of failure conditions as more informative than the recitation of reported accuracies.
- Third, we consolidate the benchmark landscape and subject the prevailing evaluation methodology to scrutiny. We argue that frame-level area under the receiver operating characteristic curve, the field's default metric, is poorly matched to the operational requirements of surveillance and has encouraged a form of progress that does not transfer to deployment.

- Fourth, we build comparative performance tables on the main benchmarks, indicating the supervision regime, backbone and evaluation protocol used by each method, so that the reader can judge which comparisons are fair.
- Fifth, we highlight deployment issues such as latency, computational cost, placement at the edge and fog, energy consumption, and privacy, which are often neglected in surveys, but decide if a method is practically usable.
- Sixth, we identify a set of research gaps and outline future directions, distinguishing those that seem technically tractable in the near term, from those that require conceptual, rather than incremental, advance.

Review Methodology

The candidate studies were found by structured searches in IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and arXiv, using combinations of the keywords video anomaly detection, abnormal event detection, unusual event classification, intelligent video surveillance, weakly supervised anomaly detection, and violence detection, primarily restricted to the period 2019–2026. We employed backward and forward citation chaining on the most highly cited results to capture foundational work outside the primary window and very recent work not yet well indexed.

The study had to (i) tackle the detection or classification of anomalous, abnormal or unusual events in video; (ii) use a learned representation, be it deep or classical; and (iii) evaluate quantitatively on at least one recognized benchmark or propose a benchmark, taxonomy or evaluation protocol of general relevance. We excluded studies where the evaluation was based on a proprietary dataset only, with no adequate description, where the contribution was a minor architectural tweak with no analysis, and where the protocol reported could not be inferred from the text.

Approximately 95 studies met these requirements and are covered in the survey's body. For context, a smaller number of works are cited- foundational architectures, systems and networking research relevant to deployment, and prior surveys- bringing the total bibliography to over one hundred entries.

Organisation Of the Survey

Section 2 formalises the problem and presents the proposed taxonomy. Section 3 traces the transition from handcrafted representations to learned features. Section 4 examines the principal deep architectural families in detail. Section 5 is devoted to the weakly supervised paradigm. Section 6 surveys datasets and benchmarks; Section 7 examines evaluation protocols and metrics. Section 8 presents the comparative performance analysis. Section 9 addresses deployment, systems, and ethical considerations. Section 10 identifies research gaps and open challenges, Section 11 outlines future directions, and Section 12 concludes.

Problem Formulation and Taxonomy

Formal Statement

Let $V = \{f_1, f_2, \dots, f_T\}$ denote a video sequence of T frames, optionally partitioned into segments $S = \{s_1, \dots, s_N\}$ where each s_i comprises a fixed number of consecutive frames. The objective is to learn a scoring function $\phi: \mathcal{X} \rightarrow \mathbb{R}$ that assigns to each frame or segment an anomaly score, such that scores associated with unusual events exceed those associated with usual events. A decision is obtained by thresholding, $\hat{y} = \mathbb{1}[\phi(x) > \tau]$, where the threshold τ is either selected on a validation set or swept to produce a performance curve.

The formulations enumerated in Section 1.2 differ in the supervision available for learning ϕ . In the one-class setting the training set $\mathcal{D}_{train}$ contains only normal instances, and ϕ is derived from a model of normality. In the weakly supervised setting $\mathcal{D}_{train}$ contains bags of segments with bag-level labels, and ϕ must be induced from the bag labels alone. In the fully supervised setting instance-level labels are available. In the unsupervised setting no labels are available and the training distribution is contaminated.

An important observation is that the one-class and weakly supervised formulations impose different implicit definitions of “unusual”. Under the one-class formulation, an event is unusual if it is statistically improbable under the learned model of normality; the definition is distributional. Under the weakly supervised formulation, an event is unusual if it resembles the events that human annotators chose to label as such; the definition is semantic. These two definitions coincide only when the semantic categories of interest are also the statistically rare ones — an assumption that holds more often in laboratory benchmarks than in operational deployment.

Terminological Distinctions

The literature uses *anomaly*, *abnormality*, *novelty*, *outlier*, and *irregularity* with substantial overlap. We adopt the following working distinctions, which are consistent with the broader anomaly detection literature.

- An outlier is an instance that lies far from the bulk of the data in some representation space; the notion is purely geometric.
- A novelty is an instance drawn from a class not represented in training; the notion is distributional and presumes a closed training set.

- An anomaly in the surveillance context is an event whose occurrence is both statistically rare and operationally significant. The conjunction matters: a rare but insignificant event, such as an unusual lighting condition, is an outlier that a well-designed system should not report.
- An unusual event we treat as synonymous with anomaly, following the convention of the surveillance literature, while noting that the term carries a stronger connotation of contextual rather than absolute rarity.
- Violence detection is a related but narrower problem in which the target class is defined semantically rather than by rarity, and is more properly regarded as a specialised action recognition task.

Supervision-Based Taxonomy

The first axis of our taxonomy is the supervision regime, summarised in Table 1.

Table 1. *Supervision regimes in intelligent surveillance event classification.*

Regime	Training data	Label granularity	Principal advantage	Principal limitation
Fully supervised	Normal and anomalous	Frame or segment	Highest achievable precision on known classes	Annotation cost; closed-set assumption
Weakly supervised	Normal and anomalous	Video level	Realistic annotation burden; open to diverse real events	Label noise propagation; coarse localisation
One-class (semi-supervised)	Normal only	Not required	No anomalous data needed; open-set by construction	High false-alarm rate; sensitive to scene change
Self-supervised	Normal only	Derived from pretext tasks	Rich representations without annotation	Pretext–downstream misalignment
Unsupervised	Unlabelled, contaminated	None	Most realistic for deployment	Weak performance; unstable training
Open-set / few-shot	Mixed, limited	Partial	Adaptation to unseen scenes and classes	Immature; benchmarks scarce

Objective-Based Taxonomy

The second axis concerns the learning objective through which normality or abnormality is encoded. Six principal objectives recur.

- **Reconstruction.** A model is trained to reconstruct normal input; high reconstruction error at test time is interpreted as evidence of abnormality. The underlying assumption is that a model fitted to normal data will generalise poorly to abnormal data. As Section 4.1 discusses, this assumption is frequently violated because sufficiently expressive networks reconstruct unseen input adequately well.
- **Prediction.** A model forecasts a future frame or feature from a history of past ones, and prediction error serves as the anomaly score. Prediction imposes a stronger temporal constraint than reconstruction and has generally proved more discriminative.
- **Distance or compactness.** Normal instances are mapped into a compact region of an embedding space, and the anomaly score is a distance from the centre or from a learned prototype set.
- **Ranking under weak supervision.** A ranking loss encourages the highest-scoring segment of an anomalous video to exceed the highest-scoring segment of a normal video, permitting learning from video-level labels alone.
- **Probabilistic density estimation.** The density of normal data is modelled explicitly, whether by Gaussian mixtures, normalising flows, or score-based methods, and low density indicates abnormality.
- **Semantic alignment.** Recent vision–language approaches score abnormality by the similarity between a video representation and textual descriptions of normal or anomalous behaviour, thereby importing external semantic knowledge.

The two axes are largely independent: a reconstruction objective may be pursued under one-class or unsupervised supervision, and a ranking objective presupposes weak supervision but may be combined with distance-based or probabilistic components. Table 2 maps prominent methods onto both axes.

Table 2. *Two-axis classification of principal method families.*

Method family	Typical supervision	Typical objective	Representative works
Convolutional autoencoder	One-class	Reconstruction	Hasan et al. (2016); Chong & Tay (2017)
Spatiotemporal /	One-class	Reconstruction	Luo, Liu & Gao (2017a); Zhao et al. (2017)

ConvLSTM AE			
Frame prediction	One-class	Prediction	Liu et al. (2018); Ye et al. (2019)
Memory-augmented	One-class	Reconstruction + prototype distance	Gong et al. (2019); Park et al. (2020)
Adversarial	One-class	Reconstruction + adversarial	Ravanbakhsh et al. (2017); Sabokrou et al. (2018)
Object-centric	One-class / self-supervised	Multi-task	Ionescu et al. (2019); Georgescu et al. (2021a)
MIL ranking	Weak	Ranking	Sultani et al. (2018); Tian et al. (2021)
Graph-based	Weak	Ranking + relational smoothing	Zhong et al. (2019)
Transformer-based	Weak / one-class	Ranking or prediction	Li et al. (2022); Ristea et al. (2022)
Vision–language	Weak / zero-shot	Semantic alignment	Wu et al. (2024); Zanella et al. (2024)
Diffusion / score-based	One-class	Density estimation	Yan et al. (2023); Micorek et al. (2024)

Figure 1 presents the taxonomy in schematic form.

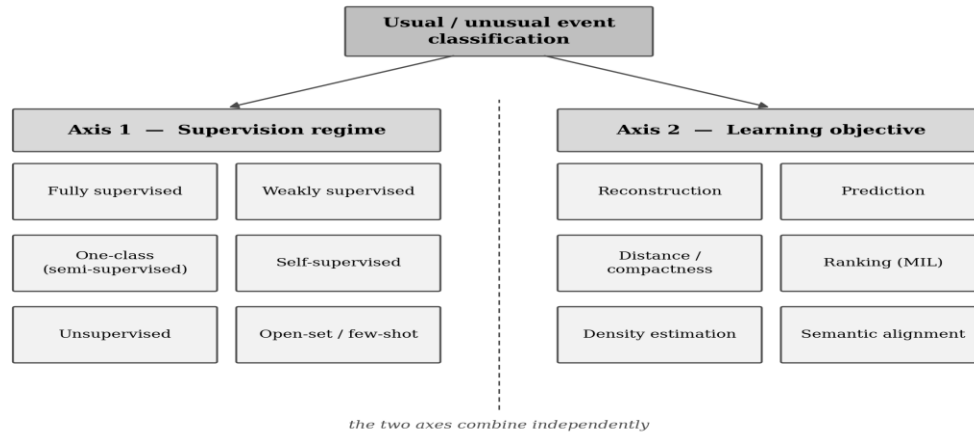


Fig. 1. Two-axis taxonomy of approaches to usual and unusual event classification. The supervision regime and the learning objective vary independently, so that a given objective may be pursued under several regimes.

Scene Dependence as A Third Consideration

A dimension that the two axes do not capture, but which strongly conditions performance, is the degree of scene dependence. Single-scene methods learn a model of normality specific to one camera view and are unusable elsewhere without retraining. Multi-scene methods learn a model intended to transfer, typically by relying on object-level or semantic representations that abstract away from background. Scene-agnostic and open-vocabulary methods aim to require no scene-specific training at all.

The distinction matters because the classical benchmarks are overwhelmingly single-scene, and methods optimised against them have absorbed a strong dependence on background statistics that is invisible under those benchmarks and catastrophic outside them. Section 10 returns to this point at length.

From Handcrafted Representations to Learned Features

Trajectory And Motion Descriptors

The earliest generation of intelligent surveillance systems reduced video to object trajectories. Detected objects were tracked across frames and the resulting paths clustered; trajectories that fell outside the dominant clusters were flagged as anomalous. Piciarelli et al. (2008) formalised this approach with a single-class support vector machine over trajectory representations, and the paradigm retains appeal for traffic monitoring, where the semantics of interest — wrong-way driving, illegal turns, sudden stops — are naturally expressed in trajectory terms (Santhosh et al., 2020).

The limitation is structural rather than incidental. Trajectory extraction presupposes reliable detection and tracking, which degrade sharply as scene density rises. In the crowded settings where surveillance is most valuable, occlusion renders per-object tracking unreliable, and the representation collapses.

Motion-field descriptors were developed in response. Rather than tracking individuals, these methods characterise the aggregate motion of a region. Adam et al. (2008) used fixed location monitors with accumulated local optical flow statistics, resulting in a real time system that is robust to crowding. The social force model was proposed by Mehran et al. (2009) where pedestrians were modeled as particles under interaction forces and anomalies were detected as deviations from normal force patterns. The method was effective for crowd panic but it assumes a specific physical model of the scene. Kim and Grauman (2009) modelled local optical flow with a mixture of probabilistic principal component analysers coupled by a Markov random field, imposing spatiotemporal consistency on local decisions.

Mahadevan et al. (2010) introduced mixtures of dynamic textures, jointly modelling appearance and dynamics, and in the same work released the UCSD Pedestrian dataset that would anchor the field's evaluation for the following decade.

Sparse Coding and Dictionary Learning

A parallel line of work modelled normality as a dictionary over which normal patterns admit sparse reconstruction. Cong et al. (2011) proposed sparse reconstruction cost as an anomaly measure, and Zhao et al. (2011) extended the framework to online operation with dynamic dictionary updating. Lu et al. (2013) achieved a substantial efficiency gain by learning multiple small dictionaries over spatiotemporal cuboids, reporting detection at 150 frames per second, and released the CUHK Avenue benchmark.

Sparse coding remains conceptually influential. Luo et al. (2017b) subsequently reformulated it within a recurrent neural architecture, demonstrating that the sparse coding objective could be embedded in a learned end-to-end system rather than abandoned.

Texture, Morphology, And Low-Level Image Analysis Foundations

Underlying all of the above is a body of image analysis research concerned with the extraction of discriminative low-level structure. Suresha and Harisha Naik (2017) surveyed texture-based image analysis and catalogued the statistical, structural, and transform-domain families of descriptor that supplied the feature vocabulary for early surveillance systems. Amalorpavam et al. (2013) examined morphological operations as a means of extracting and refining shape information from digital images, a class of operation still used for foreground mask cleanup in contemporary background-subtraction pipelines.

Related work has addressed acquisition conditions that still affect surveillance imagery. Suresha et al. (2019) studied the problem of improving faces reflected on semi-reflecting surfaces. This is a common case in retail and transport environments with many glass partitions, where the reflections create phantom motion that is easily misinterpreted by anomaly detectors.

These contributions are of interest to the present survey in two ways. First, they record the representational vocabulary that deep networks subsequently learn to induce, rather than to prescribe. Second, they pinpoint acquisition-level problems such as reflection, illumination variation, color inconsistency that have not been eradicated by deep architectures and still remain as one of the main sources of false alarm in deployed systems.

Why Handcrafted Approaches Reached Their Limit

There are three reasons for the plateau. The first is representational. Handcrafted descriptors encode a fixed hypothesis about which visual properties are informative. No fixed hypothesis has proved adequate across the diversity of surveillance scenes. The second is compositional: anomalies often arise from the combination of otherwise normal elements—a normal person in a normal place at an inappropriate time—and hand-designed features rarely capture these interactions. Third, empirically, as the benchmark scenes became more complex, the gap between handcrafted and learned representations consistently increased, and by the late 2010s, the gap on multi-scene data had become decisive (Ramachandra et al., 2020a).

This transition was accelerated by the general maturity of deep visual representation learning (Krizhevsky et al., 2012; He et al., 2016), and by the availability of video backbones pretrained on large action recognition corpora, providing transferable spatiotemporal features without scene-specific engineering.

Deep Architectural Families

This section, in turn, looks at the main architectural families. For each one, we provide the inductive bias, state the assumption that it relies upon, review representative work, and analyze the failure modes of the assumption.

Reconstruction-Based Autoencoders

- Inductive bias and assumption: It shows that a convolutional autoencoder trained on normal footage learns a compact representation that is sufficient for normal patterns. The assumption is that such a representation would be too poor to represent abnormal patterns and thus will have high reconstruction error.
- Representative work: This paradigm was set by Hasan et al. (2016) who trained a fully convolutional autoencoder on stacked frames and learned what the authors called temporal regularity. Chong and Tay (2017) inserted convolutional long short-term memory layers between encoder and decoder to allow the bottleneck to carry temporal and spatial structure. Zhao et al. (2017)

extended this with a three-dimensional convolutional autoencoder with a reconstruction branch and a prediction branch, an early recognition that reconstruction alone is insufficient. Luo et al. (2017a) proposed a temporally coherent variant with convolutional LSTM units and explicitly regularized successive latent codes to be similar.

- Failure conditions: The central assumption is unreliable. Deep autoencoders generalise well by design, and a network with sufficient capacity reconstructs unseen input — including anomalous input — with low error. Gong et al. (2019) demonstrated this directly. Compounding the problem, reconstruction error correlates with visual complexity rather than with semantic abnormality, so that cluttered but normal regions produce high error while simple but anomalous events produce low error. Astrid et al. (2021) addressed the issue by explicitly training the model *not* to reconstruct synthetically generated pseudo-anomalies, an approach that improves discrimination but introduces a dependence on the realism of the synthetic anomalies.

Prediction-Based Models

- Inductive bias and assumption: Predicting a future frame from its predecessors requires a model of scene dynamics. Abnormal events, being dynamically unlike the training distribution, should be poorly predicted.
- Representative work: Liu et al. (2018) introduced future frame prediction as an anomaly detection baseline, combining intensity, gradient, and optical flow constraints with an adversarial term, and reported substantial improvements over reconstruction baselines across UCSD Ped2, CUHK Avenue, and ShanghaiTech. This method has been adopted as the standard benchmark for the one-class setting. Ye et al. (2019) reformulated prediction in a predictive coding framework by alternating prediction and error refinement. Nguyen and Meunier (2019) trained a shared encoder with two decoders, one to reconstruct appearance and one to predict motion, so that abnormality was indicated by inconsistency between appearance and motion rather than error in either alone. Yu et al. (2020) cast the problem as a cloze test where the model should infer a purposely erased frame from its temporal context, which proved to be remarkably effective.
- Failure conditions: Prediction is biased to motion magnitude. Fast motion is inherently more difficult to predict and so high scores are associated with fast but normal activity. Optical flow is widely used in many prediction pipelines, bringing with it the failure modes of flow estimation, such as degradation under illumination change and low texture. Prediction-based methods also assume that the future is determined by the past, which does not hold well in scenes with high stochasticity such as busy intersections.

Memory-Augmented Networks

- Inductive bias and assumption: For autoencoders, the problem is over-generalisation and the cure is to restrict reconstruction to a finite set of learned normal prototypes.
- Representative work: Gong et al. (2019) proposed a memory-augmented autoencoder where the encoder output is replaced by a sparse convex combination of items fetched from an external memory that is learned during the training process. The reconstruction cannot be faithful for anomalous input since it has to be built from normal prototypes. Park et al. (2020) enhanced the method by using a compact memory updated at test time, and separate feature compactness and separateness losses to reduce memory size and improve discrimination. Cai et al. (2021) proposed consistency between appearance and motion memories, and Liu et al. (2021) combined memory-augmented flow reconstruction and flow-guided frame prediction into a hybrid framework, which is still among the best one-class results reported.
- Failure conditions: Memory size becomes a sensitive hyperparameter: too small and normal variation is suppressed, producing false alarms; too large and the original over-generalisation returns. Memory contents are learned per scene, so the approach is strongly scene-dependent. Test-time memory updating, while beneficial for gradual drift, risks absorbing a persistent anomaly into the normal model.

Adversarial And Generative Approaches

- Inductive bias and assumption: A generator trained on normal data, coupled with a discriminator, yields a model of the normal manifold; abnormality is signalled by generation failure or by discriminator response.
- Representative work: Ravanbakhsh et al. (2017) trained cross-domain generators mapping frames to optical flow and flow to frames, using generation failure as the anomaly signal. Sabokrou et al. (2018) proposed an adversarially learned one-class classifier in which a denoising autoencoder and a discriminator are trained jointly, the discriminator serving directly as the novelty score. The same group extended the approach to pixel-level irregularity localisation. Vu et al. (2019) employed multilevel representations with adversarial training at several abstraction levels, arguing that anomalies manifest at different scales.
- Failure conditions: Adversarial training is notoriously unstable, and reported results are sensitive to initialisation and stopping criteria — a reproducibility concern examined further in Section 7.4. Mode collapse narrows the learned normal manifold and inflates false alarms. Discriminator scores are not calibrated probabilities and transfer poorly across scenes.

Object-Centric and Multi-Task Frameworks

- Inductive bias and assumption: Anomalies in surveillance are typically properties of objects and their interactions rather than of whole frames. Restricting analysis to detected objects removes background confounds and yields a more transferable representation.
- Representative work: Ionescu et al. (2019) detected objects, encoded each with an autoencoder, and clustered the resulting representations, training one-versus-rest classifiers with cluster assignments as pseudo-labels and dummy anomalies drawn from unrelated data. Georgescu et al. (2021a) trained an object-level network on several self-supervised proxy tasks simultaneously — arrow of time discrimination, motion irregularity, middle box prediction, and knowledge distillation from a pretrained detector — arguing that a representation useful for all four is unlikely to encode scene-specific shortcuts. The same authors developed a background-agnostic framework with adversarial training (Georgescu et al., 2021b), explicitly targeting the background dependence discussed in Section 2.5. Sun et al. (2020) incorporated scene-level context reasoning, and Sun and Gong (2023) proposed hierarchical semantic contrast to structure the representation by semantic level.
- Failure conditions: Performance is bounded by the object detector: events involving objects outside the detector’s vocabulary, or objects too small or occluded for detection, are invisible to the system. Removing background also removes contextual information that is sometimes essential, since the same object behaviour may be normal or anomalous depending on where it occurs. The pipeline is computationally heavier than frame-level alternatives, which bears on deployment.

Recurrent And Convolutional–Recurrent Temporal Models

- Inductive bias and assumption: Event abnormality is a temporal property, and recurrent state provides a natural mechanism for accumulating evidence across time.
- Representative work: Long short-term memory units (Hochreiter & Schmidhuber, 1997) and their convolutional extension (Shi et al., 2015) underpin a large body of surveillance work. Luo et al. (2017b) embedded sparse coding within a stacked recurrent framework, and composite recurrent architectures have since been applied widely to normality modelling. Recurrent components appear throughout the reconstruction and prediction families discussed above rather than as a wholly separate lineage.
- Failure conditions: Recurrent models are difficult to parallelise, constraining throughput in real-time deployment. They exhibit limited effective memory over long horizons, which matters for anomalies defined by extended temporal context — loitering, for instance, is defined precisely by duration. Transformer architectures have largely displaced recurrence for long-range modelling, though at higher memory cost.

Graph-Based and Relational Models

- Inductive bias and assumption: A relational structure is defined by the video segments or the objects in them; information propagated over that structure regularizes predictions and reveals interaction-level anomalies.
- Representative work: Zhong et al. (2019) rephrased weakly supervised detection as supervised learning with label noise, and adopted a graph convolutional network (Kipf & Welling, 2017) to denoise the segment labels by propagating information over graphs constructed from feature similarity and temporal closeness. Markovitz et al. used graph representations on skeletal data for pose-based anomaly detection, which is inherently privacy preserving because it discards appearance. Previous work has explicitly modeled relations between objects, allowing for the detection of anomalies that are in the interaction rather than in the individual behavior [32].
- Failure conditions: Graph construction requires a similarity metric and results are sensitive to its choice. Iterative label cleaning is expensive in terms of computation during training. Skeleton based variants inherit the limitations of pose estimation and are unable to represent anomalies involving objects or environmental change.

Transformers And Attention Mechanisms

- Inductive bias and assumption: Self-attention (Vaswani et al., 2017) is well suited for anomalies whose evidence is distributed across a video, capturing dependencies between arbitrary positions without the locality constraint of convolution.
- Representative work: Later, representations provided by video transformer backbones, such as ViViT (Arnab et al., 2021), TimeSformer (Bertasius et al., 2021), Video Swin (Liu et al., 2022), and the masked autoencoding pretraining of VideoMAE (Tong et al., 2022), were used for anomaly detection. In the field, Li et al. (2022) proposed multi-sequence learning with a transformer for weakly supervised detection, which gradually refined segment-level pseudo-labels by self-training. Ristea et al. (2022) proposed a self-supervised predictive convolutional attentive block that can be embedded into existing architectures. Such a block combines masked convolution and channel attention to reconstruct a masked region from its surroundings. Multi-task self-supervised learning with transformer components was revisited by Barbalau et al. (2023), who reported consistent gains over the convolutional predecessor.

The wider utility of attention in classification pipelines beyond the video domain has also been shown. KS et al. (2026) applied inverse transform sampling with attentional deep learning to robust respiratory sound classification, addressing the class imbalance in rare-event detection in general. The parallel is instructive: surveillance anomaly detection is also an extreme-imbalance problem, and sampling strategies that reshape the effective training distribution deserve closer attention in the video setting, than they have so far received.

- Failure conditions: Attention scales quadratically with sequence length, imposing a practical limit on temporal context. Transformers crave data, and surveillance datasets are small compared to the corpora on which the architecture's advantages were demonstrated; most surveillance transformers consequently depend on features extracted by frozen pretrained backbones, inheriting whatever biases those backbones encode.

Probabilistic, Score-Based, And Diffusion Models

- Inductive bias and assumption: Explicit density estimation over normal data provides a principled anomaly score, avoiding the proxy character of reconstruction error.
- Representative work: Yan et al. (2023) proposed a feature prediction diffusion model, using the denoising process to model the distribution of normal features and treating prediction error under diffusion as the anomaly signal. Micorek et al. (2024) estimated multiscale log-density through denoising score matching, obtaining a score with a clearer probabilistic interpretation than reconstruction error affords. Normalising-flow approaches have similarly been applied to feature-level density estimation.
- Failure conditions: Density in a learned feature space is not the same as density in the data space, and the correspondence depends on the feature extractor. Diffusion models are computationally demanding at inference, which is a serious obstacle for real-time surveillance. Likelihood-based anomaly detection is known to exhibit counter-intuitive behaviour on out-of-distribution data in other modalities, and the extent to which this affects video remains inadequately characterised.

Vision–Language and Foundation-Model Approaches

- Inductive bias and assumption: Large-scale image–text pretraining (Radford et al., 2021) yields representations aligned with natural language semantics. If abnormality can be described in language, it can be detected by semantic alignment, without scene-specific training data.
- Representative work: Joo et al. (2023) employed CLIP features with temporal self-attention for weakly supervised detection, showing that language-aligned features outperform conventional action recognition features on the same architecture. Wu et al. (2024) proposed VadCLIP, adapting vision–language alignment to weakly supervised detection through dual-branch coarse- and fine-grained alignment and enabling open-vocabulary operation. Yang et al. (2024) introduced textual prompts encoding normality as explicit guidance. Pu et al. (2024) learned prompt-enhanced context features to improve discrimination under weak supervision.

A more radical departure is training-free detection. Zanella et al. (2024) used large language models over captioned frames to score abnormality without any training on the target domain, and multimodal large language models have been applied to produce not only a score but a textual explanation of the detected anomaly (Zhang et al., 2024; Lv & Sun, 2024). The explanatory capability is significant: an operator confronted with an alert requires justification, and no earlier architectural family provided one.

- Failure conditions: Language-aligned representations encode the biases of web-scale training corpora, with implications for fairness that Section 9.4 examines. Captioning-based pipelines are slow and unsuitable for real-time operation. Performance depends on prompt formulation to a degree that complicates reproducibility, and there is a genuine risk of benchmark contamination where pretraining corpora overlap with evaluation data.

Comparative Synthesis of Architectural Families

Table 3 summarises the analysis.

Table 3. *Comparative synthesis of deep architectural families.*

Family	Core assumption	Chief strength	Chief weakness	Real-time viability
Reconstruction AE	Poor generalisation to anomalies	Simplicity; no anomalous data	Assumption often violated	High
Prediction	Anomalies are unpredictable	Stronger temporal constraint	Motion-magnitude bias; flow dependence	Moderate
Memory-augmented	Normality is prototype-composable	Controls over-generalisation	Memory sizing; scene dependence	Moderate

Adversarial	Generation failure signals abnormality	Sharp outputs; pixel localisation	Training instability	Moderate
Object-centric	Anomalies are object properties	Background invariance; transfer	Detector-bounded; loses context	Low to moderate
Recurrent	Temporal state accumulates evidence	Natural sequence handling	Limited horizon; poor parallelism	Moderate
Graph-based	Relational structure is informative	Interaction-level anomalies	Graph design sensitivity; cost	Low
Transformer	Long-range dependency matters	Global context; strong features	Data-hungry; quadratic cost	Low to moderate
Diffusion / density	Explicit density is principled	Interpretable score	Inference cost; feature-space caveat	Low
Vision–language	Abnormality is describable	Open-vocabulary; explainable	Pretraining bias; latency	Low

Two observations follow. First, there is a systematic inverse relationship between representational sophistication and real-time viability, and the field’s recent progress has been concentrated at the sophisticated end. Second, the assumptions underlying the earlier families are the ones most often violated in practice, yet those families remain the ones most likely to be deployable — a tension that Section 10 identifies as a central research gap.

The Weakly Supervised Paradigm

Motivation And Formulation

The one-class formulation assumes that a curated corpus of purely normal footage is available for the target scene and that anything departing from it is worth reporting. Neither assumption survives contact with operational surveillance. Real archives contain anomalies; and departures from statistical normality include a great many events — weather, lighting, maintenance activity, unusual but innocuous behaviour — that no operator wishes to see flagged.

Weak supervision offers a pragmatic alternative. Sultani et al. (2018) framed the problem as multiple-instance learning: a video is a bag, its segments are instances, and the video-level label indicates whether the bag contains at least one anomalous instance. The training objective is a ranking loss enforcing that the maximum segment score within an anomalous bag exceed the maximum within a normal bag, augmented with smoothness and sparsity regularisers reflecting the priors that anomaly scores vary gradually over time and that anomalies are temporally sparse. The accompanying UCF-Crime dataset, comprising 128 hours of untrimmed real-world footage across thirteen anomaly categories, reoriented the field.

Limitations Of the Original Formulation and Successive Refinements

The maximum-score selection at the heart of the original formulation is fragile. If the highest-scoring segment of an anomalous video is in fact normal — which occurs frequently in early training when scores are near-random — the gradient reinforces an incorrect association, and the error can persist. Much subsequent work can be read as an attempt to make instance selection more reliable.

- **Feature magnitude:** Tian et al. (2021) observed that the magnitude of features extracted by a pretrained backbone is systematically larger for anomalous segments, and proposed robust temporal feature magnitude learning, selecting the top- k segments by feature magnitude rather than by predicted score. Combining this with dilated convolutions and self-attention for temporal context, the method achieved a marked improvement and became a standard baseline.
- **Self-training and pseudo-label refinement:** Feng et al. (2021) introduced a two-stage framework in which a multiple-instance pseudo-label generator produces segment-level labels that then fine-tune a task-specific encoder, so that features are adapted to anomaly detection rather than inherited unchanged from action recognition pretraining. Li et al. (2022) generalised the selection to sequences of varying length, progressively shortening the selected sequence as training proceeds and pseudo-labels become more reliable.
- **Label noise cleaning:** Zhong et al. (2019) treated the problem explicitly as supervised learning under noisy labels and applied graph convolution to correct segment labels through feature-similarity and temporal-consistency propagation.
- **Clustering and normalcy suppression:** Zaheer et al. (2020) proposed a framework combining a normalcy suppression mechanism, which suppresses scores in segments confidently identified as normal, with clustering-based supervision, reducing the reliance on the maximum-score heuristic.

- Memory and uncertainty: Zhou et al. (2023) maintained separate memory units for normal and abnormal patterns with explicit uncertainty regulation, addressing the ambiguity of segments that are neither clearly normal nor clearly anomalous.
- Contrastive and magnitude-contrastive objectives: Chen et al. (2023) introduced a glance-and-focus network with a magnitude-contrastive loss, addressing the observation that feature magnitude is scene-dependent and therefore an unreliable cue across scenes. Huang et al. (2021) applied deep contrastive learning to the surveillance setting more generally.
- Debiasing: Lv et al. (2023) analysed context bias in the weakly supervised setting, showing that models frequently learn to associate anomaly with scene context rather than with the event itself, and proposed an unbiased formulation using causal reasoning to sever the spurious association.

Multimodal Weak Supervision

Wu et al. (2020) argued that audio provides complementary evidence that vision alone cannot supply — screams, breaking glass, gunshots — and introduced XD-Violence, a large-scale audio-visual dataset with weak labels, together with a network modelling holistic, localised, and score-level dependencies. Subsequent work refined audio-visual fusion under weak supervision (Wu et al., 2022), and Wu and Liu (2021) developed causal temporal relation modelling to improve feature discrimination.

Audio is not universally available in surveillance deployments, and where available raises privacy questions more acute than those attaching to video. Nevertheless, the multimodal results establish that the visual channel alone is not saturated for information but for exploitation.

Unsupervised Extension

Zaheer et al. (2022) removed the video-level labels entirely, proposing generative cooperative learning in which a generator and a discriminator co-operate rather than compete, with cross-supervision derived from a negative learning objective. Performance is below the weakly supervised state of the art, but the setting is substantially more realistic, and the gap has narrowed.

Critical Assessment

The weakly supervised paradigm has produced the most operationally relevant results in the field, but three concerns warrant attention.

First, most methods depend on features extracted by a frozen backbone pretrained on action recognition data, typically I3D (Carreira & Zisserman, 2017) or, more recently, CLIP (Radford et al., 2021). Reported improvements frequently reflect changes in the feature extractor rather than in the anomaly detection method, and the literature is not consistently careful in isolating the two.

Second, the thirteen anomaly categories of UCF-Crime constitute a closed set defined by the annotators, and models optimised against it learn a category-specific rather than a general notion of abnormality. This is a semantic classifier presented as an anomaly detector.

Third, evaluation on UCF-Crime is dominated by videos in which anomalies occupy a substantial fraction of the duration, which favours methods that produce smooth, high-recall score profiles. Sparse, brief anomalies — arguably the harder and more operationally significant case — are under-represented.

Datasets and Benchmarks

Overview

Progress in the field has been closely coupled to dataset availability, and the characteristics of the dominant benchmarks have shaped the methods developed against them. Table 4 consolidates the principal resources.

Table 4. Principal benchmarks for usual and unusual event classification.

Dataset	Year	Scenes	Total frames / duration	Supervision	Anomaly types	Annotation
UCSD Ped1 / Ped2	2010	2	~18,560 / ~4,560 frames	One-class	Non-pedestrian entities, anomalous paths	Frame and pixel
Subway Entrance / Exit	2008	2	~144,000 frames	One-class	Wrong direction, no payment, loitering	Frame
CUHK Avenue	2013	1	~30,652 frames	One-class	Throwing, running, wrong direction	Frame and pixel
ShanghaiTech	2017	13	~317,000 frames	One-class	Cycling, brawling, chasing, jumping	Frame and pixel
Street Scene	2020	1	~203,000 frames	One-class	Jaywalking, illegal turns,	Track-level boxes

					loitering	
IITB-Corridor	2020	1	~483,000 frames	One-class	Protest, fighting, chasing, hiding	Frame
UCF-Crime	2018	Many	128 hours	Weak	13 real-world crime categories	Video level (test: frame)
XD-Violence	2020	Many	217 hours	Weak, audio-visual	6 violence categories	Video level (test: frame)
UBnormal	2022	29 (synthetic)	~236,000 frames	Open-set supervised	22 event types	Frame and pixel
NWPU Campus	2023	43	~1,466,000 frames	One-class	28 event types, scene-dependent	Frame
Hockey Fight	2011	Varied	1,000 clips	Supervised	Fighting	Clip level
RWF-2000	2021	Varied	2,000 clips	Supervised	Violence	Clip level

The Classical Single-Scene Benchmarks

UCSD Pedestrian (Mahadevan et al., 2010) and CUHK Avenue (Lu et al., 2013) defined the evaluation regime for a decade. Both are small, single scene, low resolution and contain visually conspicuous anomalies – a cyclist among pedestrians, a thrown object. The area under the curve on Ped2 now routinely exceeds 0.98. This is best interpreted as evidence that the benchmark is exhausted rather than that the problem is solved.

The case was improved by ShanghaiTech (Luo et al., 2017b) with thirteen camera views with varying angle and lighting, and it remains a meaningful test of one-class methods. However, its anomalies are staged rather than natural, and the actors’ awareness of the camera lends subtle cues to behavior that models might exploit.

Real-World And Large-Scale Benchmarks

The UCF-Crime dataset (Sultani et al., 2018) marked a departure to untrimmed, uncured, real-world videos from heterogeneous sources. Its main virtues are scale and realism; its weaknesses are variable video quality, occasional annotation inconsistency, and the closed-set category structure mentioned in Section 5.5.

XD-Violence (Wu et al., 2020) is even larger and it is the only one which provides synchronized audio. It is taken from films and online video however, leading to a distribution shift from fixed-camera surveillance.

Street Scene (Ramachandra & Jones, 2020) was proposed with a track-based evaluation protocol specifically to address the localization deficiencies of earlier benchmarks and remains under-used given its methodological importance. IITB-Corridor (Rodrigues et al., 2020) provides multi-timescale activity in a single high-traffic scene.

Recent Benchmarks Addressing Structural Weaknesses

Two recent datasets deserve particular attention as they focus on recognized deficiencies rather than just increasing scale.

UBNormal (Acsintoae et al., 2022) is synthetically generated allowing pixel-perfect annotation and, importantly, disjoint anomaly types in the training and test sets. This, in turn, enables true open-set evaluation, where the ability of a method to detect unseen anomaly categories can be measured, not assumed. The price is a domain gap between synthetic and real images.

NWPU Campus (Cao et al., 2023) is the largest one-class benchmark to date, consisting of forty-three scenes, and explicitly shows scene-dependent anomalies, i.e., behaviors that are normal in one place but anomalous in another. It directly tests the ability to reason about context, identified as central in Section 1.1 and not assessable by previous benchmarks. The dataset also contains an anomaly anticipation task which involves prediction before an event occurs.

Structural Limitations of The Benchmark Landscape

Four limitations recur across the benchmark set.

- Scale relative to method capacity: The classical datasets contain fewer frames than a single hour of one camera’s output. Training data-hungry architectures on such corpora invites overfitting that evaluation on the same distribution cannot detect.
- Annotation subjectivity: Frame-level anomaly boundaries are inherently imprecise. Where an anomalous event begins is a judgement, and inter-annotator agreement is rarely reported.
- Scene bias: Because most benchmarks provide few scenes, models can achieve strong scores by learning background statistics. Georgescu et al. (2021b) and Lv et al. (2023) both provide evidence that this occurs.

- Demographic and geographic narrowness: The benchmarks are drawn from a limited range of locations and populations. The fairness implications are examined in Section 9.4.

Evaluation Protocols and Metrics

Frame-Level AUC and Its Variants

The dominant metric is the area under the receiver operating characteristic curve computed over frame-level anomaly scores. Two protocols are in use and are frequently conflated. Under the micro protocol, scores from all test videos are concatenated and a single curve computed. Under the macro protocol, a curve is computed per video and the results averaged. Micro AUC is generally higher, and the difference can exceed two percentage points. Comparisons that do not specify the protocol are uninterpretable, and a non-trivial fraction of published comparisons do not specify it.

A further complication is score normalisation. Many one-class methods normalise scores to the unit interval per video using the minimum and maximum observed within that video. This is a form of test-time information leakage: it presumes knowledge of the score range of the complete video, which is unavailable in streaming operation, and it guarantees that every video contains a maximally normal and a maximally anomalous frame even when the video is entirely normal.

Localisation Metrics

Frame-level AUC rewards a system that identifies *when* an anomaly occurs but says nothing about *where*, and a system may achieve high frame-level scores while localising incorrectly. Pixel-level AUC was introduced as a corrective but is satisfied by a detection that overlaps the ground-truth region by a small margin.

Ramachandra and Jones (2020) proposed region-based and track-based detection criteria, which require that a detection overlap a ground-truth region by a specified threshold and that a ground-truth track be detected across a specified fraction of its duration. These metrics are considerably more demanding, and scores under them are far lower than frame-level figures suggest — a discrepancy that constitutes one of the more important, and less frequently discussed, findings in recent evaluation research.

Metrics For the Weakly Supervised Setting

For UCF-Crime the convention is frame-level AUC on the test partition; for XD-Violence it is average precision. Average precision is the more informative choice given the extreme class imbalance, since ROC-based measures are insensitive to the false-positive rate when negatives vastly outnumber positives. Some authors additionally report AUC computed over anomalous test videos only, which isolates localisation quality from the trivial component of the task.

False alarm rate at a fixed detection rate is reported inconsistently despite being the metric that most directly determines operational viability. A system with an AUC of 0.85 may generate hundreds of daily false alarms per camera, rendering it unusable regardless of the headline figure.

Reproducibility

Several reproducibility concerns affect the interpretation of reported results.

Model selection is frequently performed on the test set, either explicitly through epoch selection by test performance or implicitly through hyperparameter search reported only at its optimum. Few benchmarks define a validation partition, so it makes sense that a validation partition is correspondingly common and hard to avoid.

Post-processing choices, such as temporal smoothing window, score aggregation, normalization scheme, have material effects on results and are inconsistently documented. Reported gains of one to two AUC points are often within the range attributable to post-processing alone.

Variance across random seeds is rarely reported. Given that adversarially trained methods in particular exhibit substantial seed sensitivity, single-run comparisons at fine margins carry limited evidential weight.

Towards More Informative Evaluation

We suggest that the field would benefit from a protocol comprising: region- and track-based detection criteria alongside frame-level AUC; average precision rather than AUC as the primary aggregate; explicit reporting of false alarm rate at a fixed operating point; mandatory declaration of micro versus macro protocol and of normalisation scheme; cross-dataset evaluation as a standard component; and reporting of mean and standard deviation across at least three seeds. None of these requirements is onerous, and their collective adoption would substantially improve the interpretability of the literature.

Comparative Performance Analysis

Basis Of Comparison

The tables in this section collate results across the principal benchmarks together with the experimental conditions under which they were obtained. Differences in feature extractor, temporal segment length, evaluation protocol, post-processing, and score normalisation are frequently larger in effect than the differences between the detection mechanisms themselves, and the comparison is therefore presented with these conditions made explicit so that the sources of variation can be separated.

One-Class Methods on the Classical Benchmarks

Table 5. Frame-level AUC reported for representative one-class methods across the three principal single-scene and multi-scene benchmarks.

Method	Year	Objective	Ped2 AUC	Avenue AUC	SHTech AUC
Hasan et al.	2016	Reconstruction	0.850	0.702	—
Luo et al. (ConvLSTM-AE)	2017	Reconstruction	0.881	0.770	—
Luo et al. (sRNN)	2017	Sparse coding + RNN	0.922	0.817	0.680
Liu et al. (FFP)	2018	Prediction	0.954	0.851	0.728
Nguyen & Meunier	2019	Appearance–motion	0.962	0.869	—
Gong et al. (MemAE)	2019	Memory + reconstruction	0.941	0.833	0.712
Ionescu et al. (object-centric)	2019	Multi-stage object	0.978	0.906	0.847
Park et al. (MNAD)	2020	Memory + prediction	0.970	0.885	0.705
Yu et al. (cloze test)	2020	Video completion	0.973	0.899	0.736
Georgescu et al. (SSMTL)	2021	Self-supervised multi-task	0.992	0.918	0.826
Liu et al. (HF2VAD)	2021	Hybrid memory + prediction	0.993	0.911	0.767
Ristea et al. (SSPCAB)	2022	Masked prediction block	0.985	0.923	0.836
Micorek et al. (MULDE)	2024	Score-based density	0.993	0.918	0.862

Three patterns are evident. Ped2 performance has saturated, with the leading methods separated by margins smaller than plausible run-to-run variance. Avenue has largely followed. ShanghaiTech retains discriminative power, and the ordering of methods on ShanghaiTech differs materially from that on Ped2 — the object-centric and multi-task families gain most, consistent with the argument that multi-scene evaluation penalises background-dependent shortcuts.

It is also notable that the improvement from Hasan et al. (2016) to the current leaders is approximately fourteen AUC points on Ped2 but eighteen on ShanghaiTech, despite ShanghaiTech being the harder benchmark. This suggests that the gains are real rather than artefacts of benchmark exhaustion, while remaining consistent with the view that absolute figures overstate operational capability.

Weakly Supervised Methods

Table 6. Performance reported for weakly supervised methods, with the feature extractor stated because it accounts for a substantial share of the variation.

Method	Year	Features	UCF-Crime AUC	XD-Violence AP
Sultani et al.	2018	C3D	0.751	0.730
Zhong et al. (GCN)	2019	TSN	0.821	0.784
Zaheer et al. (CLAWS)	2020	C3D	0.831	—
Wu et al. (HL-Net)	2020	I3D + audio	0.822	0.786
Feng et al. (MIST)	2021	I3D	0.821	—
Tian et al. (RTFM)	2021	I3D	0.842	0.778
Li et al. (MSL)	2022	VideoSwin	0.859	0.786
Chen et al. (MGFN)	2023	I3D	0.869	0.800

Zhou et al. (UR-DMU)	2023	I3D	0.868	0.815
Lv et al. (UMIL)	2023	CLIP	0.863	—
Joo et al. (CLIP-TSA)	2023	CLIP	0.874	0.822
Wu et al. (VadCLIP)	2024	CLIP	0.884	0.849
Zanella et al. (LAVAD)	2024	Training-free LLM	0.807	0.621

The increase from 0.751 to 0.884 over six years on UCF-Crime is large, but the feature column shows that a large part of this is due to backbone improvement rather than the anomaly detection mechanism. Irrespective of the detection head, the switch from C3D to I3D and then to CLIP features adds several points.

The result of LAVAD deserves a separate comment. It is below the supervised state of the art at 0.807 but it is obtained without any training at all on the target dataset. Interpreted correctly, this figure indicates both that general-purpose semantic knowledge captures much of what the benchmark measures, and that the benchmark rewards semantic category recognition more than genuine anomaly detection. Figure 2 represents the performsnce over time.

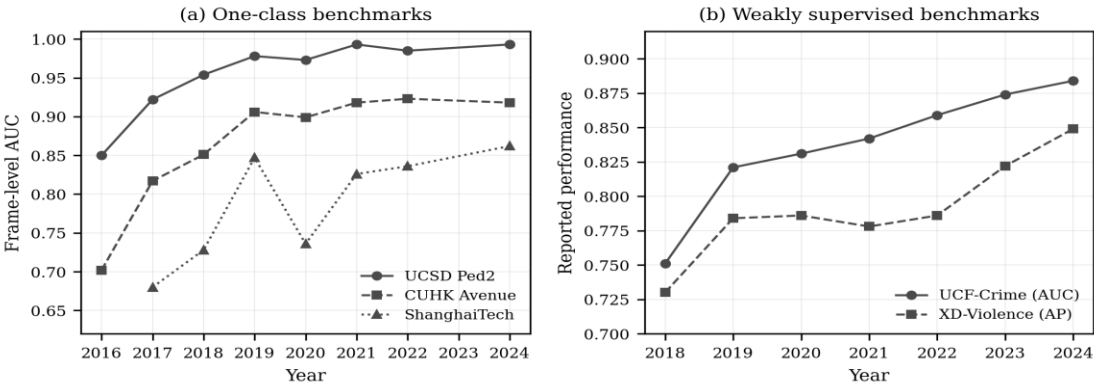


Fig. 2. Progression of reported performance over time. Panel (a) shows frame-level AUC on the one-class benchmarks; panel (b) shows AUC on UCF-Crime and average precision on XD-Violence for weakly supervised methods.

Computational Characteristics

Table 7. Computational characteristics by architectural family. Throughput excludes feature extraction where a frozen backbone is used; its inclusion typically halves the observed rate.

Method family	Typical parameters	Reported throughput	Hardware class assumed
Convolutional AE	1–5 M	100–200 fps	Single mid-range GPU
Frame prediction	10–20 M	20–30 fps	Single high-end GPU
Memory-augmented	5–15 M	20–40 fps	Single high-end GPU
Object-centric	30–60 M (incl. detector)	5–20 fps	Single high-end GPU
Transformer (feature-level)	5–15 M head + frozen backbone	10–30 fps	Single high-end GPU
Diffusion / score-based	20–50 M	1–10 fps	Single high-end GPU
Captioning + LLM	Billions	Well below real time	Multi-GPU or hosted

Table 7 is the counterpart to Table 5 and should be read alongside it. A method achieving 0.99 AUC at three frames per second is not a solution to the surveillance problem, since a single operational deployment may involve hundreds of concurrent streams. The frequent omission of throughput from published comparisons has permitted a divergence between research progress and deployment capability that Section 10.6 treats as a first-order concern.

Cross-Dataset Generalisation

Cross-dataset evaluation, in which a model trained on one benchmark is tested on another without adaptation, is reported infrequently but is highly informative where available. The consistent finding is that performance degrades severely — often by fifteen to thirty AUC

points — under scene change. Object-centric and semantically aligned methods degrade least, consistent with the argument that they encode less background-specific information.

Few-shot scene adaptation (Lu et al., 2020) and continual learning formulations (Doshi & Yilmaz, 2022) address the problem directly, learning to adapt to a new scene from a small number of normal examples. This line of work is under-developed relative to its practical importance.

Deployment, Systems, and Ethical Considerations

Architectural Placement: Edge, Fog, And Cloud

The computational profile documented in Table 7 makes placement a design decision of first importance. Transmitting all video to a central facility for analysis is bandwidth-prohibitive at scale and introduces latency incompatible with intervention. Performing all analysis at the camera is constrained by the thermal and power envelope of embedded hardware.

The intermediate fog tier has consequently attracted attention. Safinaz et al. (2025) examined quality-of-service-driven service placement and data analytics for industrial internet-of-things applications in fog–cloud systems, addressing the allocation problem of which analytic components should execute at which tier subject to latency and reliability constraints. The formulation transfers directly to surveillance: lightweight motion filtering and normality screening are naturally placed at the edge, with computationally heavy semantic analysis invoked selectively at the fog or cloud tier only for segments the edge stage flags as candidates. Such cascaded designs are common in industrial practice but rarely evaluated in the academic literature, where end-to-end accuracy on pre-segmented clips remains the norm. Figure 3 presents the tiered processing architecture for deployed surveillance analytics.

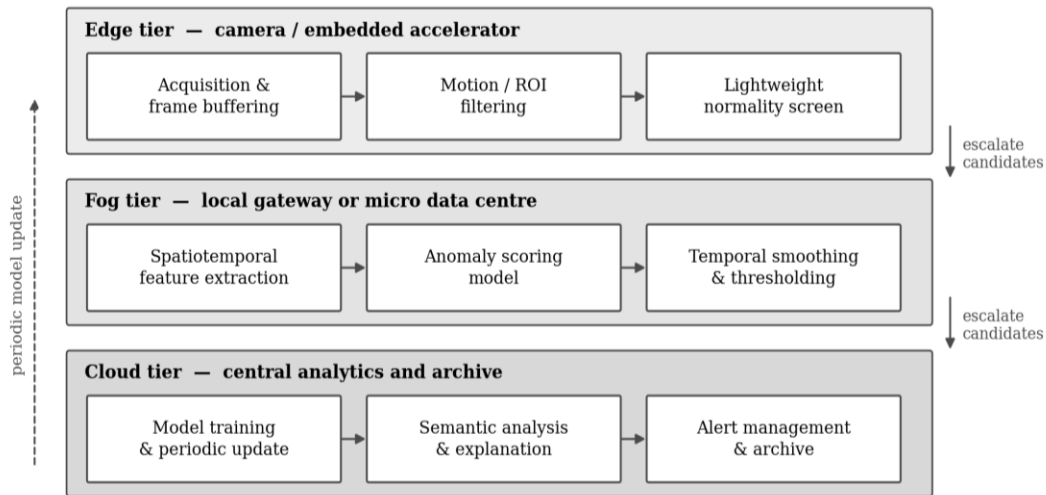


Fig. 3. Tiered processing architecture for deployed surveillance analytics. Lightweight screening at the edge suppresses the majority of normal frames; only candidate segments are escalated to the computationally heavier stages at the fog and cloud tiers.

Network characteristics condition what is achievable. Goswami et al. (2025) surveyed latency minimisation in dense urban fifth-generation mobile networks, a setting directly relevant to municipal camera deployments that rely on cellular backhaul. The end-to-end latency budget of an alerting system comprises acquisition, transmission, inference, and notification, and inference-time optimisation alone addresses only one component.

Efficiency, Parallelism, And Energy

Model compression through pruning, quantisation, and knowledge distillation is the standard route to edge deployment, and the accuracy cost is generally modest for the convolutional families though more significant for transformers.

Parallel execution strategy is an orthogonal consideration. NT et al. (2025) studied meta-heuristic methods for parallel processing in signal representation and real-time analysis. The authors scheduled computation on parallel resources under real-time constraints. Multi-camera surveillance is a natural application: the workload is embarrassingly parallel across streams but subject to shared memory bandwidth and accelerator contention, and naive allocation leaves substantial throughput unrealised.

Energy consumption is a growing concern for large installations and for camera nodes operating on harvested or battery power. The governing problem is the allocation of a constrained and variable energy budget across competing computational demands, which arises directly in solar-powered remote deployments where inference frequency must be modulated against available power. Energy-aware anomaly detection, in which model complexity adapts to the current power budget, is essentially unexplored.

Relationship To Anomaly Detection in Other Domains

Surveillance anomaly detection shares its formal structure with anomaly detection in network traffic, industrial process monitoring, and clinical signal analysis, and cross-domain transfer of methodology has been productive. Naik et al. (2025) proposed a fused generative adversarial network intrusion detection system employing randomised tree classification, addressing the same extreme class imbalance and the same requirement for low false-alarm operation that characterise video surveillance, and demonstrating the value of hybridising generative modelling with ensemble classification.

The parallel with network intrusion detection is especially close. That field confronted the false-alarm problem earlier and more directly than video surveillance has, and its accumulated experience — with alert correlation, with adaptive thresholding, with the operational cost of analyst fatigue — is directly applicable and largely unexploited in the video literature.

Privacy, Fairness, And Governance

The continuous video analysis involves privacy interests that a purely technical treatment risks underestimating. Regulatory systems in a number of jurisdictions impose requirements such as proportionality, purpose limitation and data minimization, and automated behavioral analysis may be processing requiring specific justification.

There are technical mitigations. Pose based methods discard appearance altogether, keeping skeletal geometry sufficient for many behavioral anomalies, but removing identity. Processing at the edge keeps raw imagery from being transmitted. Selective obfuscation obscures regions of identification, but preserves the motion structure on which detection relies.

Fairness concerns are substantive. A system trained predominantly on footage from a narrow demographic range may exhibit differential error rates across populations, and in a surveillance context differential false-positive rates translate into differential exposure to police attention. The benchmark datasets do not support disaggregated evaluation, since demographic attributes are neither annotated nor, for good reason, easily annotatable. This leaves the field in the uncomfortable position of being unable to measure a harm it has grounds to suspect. Vision–language methods introduce an additional pathway, since the pretraining corpora of foundation models are known to encode social biases that may propagate into abnormality judgements.

Adversarial robustness compounds these concerns. Mumcu et al. (2022) demonstrated that video anomaly detection systems are susceptible to adversarial perturbation, and a surveillance system is by construction deployed in an environment containing adversaries with an interest in evading it.

Finally, explainability is not merely desirable but in many jurisdictions a governance requirement. Defensible action cannot be taken in response to an alert that is not justified. The recent emergence of multimodal language models that generate textual explanations along with scores (Zhang et al., 2024; Lv & Sun, 2024; Doshi & Yilmaz, 2023) is thus important beyond its technical interest.

Research Gaps and Open Challenges

Context Dependence Remains Unsolved

The definition of abnormality depends on the context but most systems obtain a context independent notion of statistical rareness. The NWPU Campus (Cao et al., 2023) is the first large-scale benchmark that explicitly measures scene-dependent abnormality, and the reported performance on it is significantly lower than that on previous benchmarks. The most basic limitation of the current generation is that it lacks a system to represent and reason about context – location semantics, time of day, day of week, expected activity patterns. There is no convincing way.

Dataset Bias and Background Dependence

There is a wealth of evidence that models learn spurious associations between background and label (Georgescu et al., 2021b; Lv et al., 2023). Not only is this dependence unobserved, it is actively rewarded in evaluation, which is typically within-distribution. Debiasing methods exist but are evaluated on the same biased benchmarks. Progress requires benchmarks constructed to break the association, for instance by presenting identical anomalies across multiple backgrounds and identical backgrounds across normal and anomalous conditions.

The Long Tail and Open-Set Operation

Real surveillance encounters an unbounded set of possible anomalies, most of which are individually vanishingly rare. Weakly supervised methods trained on a closed category set do not generalise to unseen categories; one-class methods are open-set in principle but achieve this by treating all deviation as anomalous, producing unacceptable false-alarm rates. UBnormal (Acsintoae et al., 2022) enables measurement of open-set capability, and the results indicate that current methods possess it to a limited degree. Open-vocabulary approaches (Wu et al., 2024) offer a route forward but presuppose that the anomaly can be described in advance.

Domain Shift and Adaptation

Cross-scene degradation of fifteen to thirty AUC points implies that deployment at each new camera requires scene-specific data collection and training — a cost that scales linearly with installation size and that few operators will accept. Few-shot adaptation (Lu et al., 2020) and continual learning (Doshi & Yilmaz, 2022) address this but remain immature. The problem is compounded by temporal drift: a scene’s normality changes with season, construction, and shifting usage, so adaptation must be ongoing rather than one-off, while remaining resistant to absorbing persistent anomalies into the normal model.

The False-Alarm Problem

This is the gap with the greatest distance between literature and practice. A system operating at 0.85 AUC may still generate an alert volume that renders it unusable. The literature reports AUC; operators require precision at a workable recall. Doshi and Yilmaz (2021) addressed the question directly by developing an online framework with an asymptotic bound on false-alarm rate, and this line of work deserves considerably more attention than it has received. Systematic reporting of false-alarm rate at fixed detection rate would be a low-cost, high-value change in reporting practice.

The Accuracy–Efficiency Divergence

Tables 5 and 7 read together show that the field’s accuracy gains have been purchased with computation that deployment cannot supply. There is no established benchmark that jointly evaluates accuracy and efficiency, and consequently no incentive to optimise the trade-off. A leaderboard reporting accuracy at fixed computational budget would redirect effort productively.

Explainability and operator trust

Anomaly scores are uninterpretable. An operator cannot determine why a segment was flagged, cannot calibrate trust, and cannot defend a resulting intervention. Language-model-based explanation is a promising development but currently trades away the throughput required for real-time operation. Lightweight explanation — attention visualisation, prototype retrieval, counterfactual highlighting — merits systematic investigation.

Temporal Granularity and Long-Horizon Events

Most methods operate on segments of sixteen to sixty-four frames, a horizon adequate for sudden events but inadequate for anomalies constituted by extended behaviour: loitering, repeated approach and withdrawal, systematic surveillance of a premises by an intruder. Detecting such events requires reasoning over minutes or hours, and no current architecture addresses this scale.

Multimodality Beyond Audio

XD-Violence showed audio’s value, but surveillance settings often include more modalities — thermal imaging, depth, access-control logs, radio-frequency identification, environmental sensors. The fusion of video with structured non-visual context is largely un-explored and looks a promising route to the contextual reasoning identified in Section 10.1.

Evaluation Reform

The methodological concerns of 7 represent a research gap in themselves. Misalignment of the field’s primary metric with operational needs, non-uniform reporting of protocols, and infrequent quantification of variance. These are central questions which decide the meaning of what is reported as progress in the literature.

Emerging Models and Future Research Directions*Foundation Models as Scene Priors*

The most consequential recent development is the availability of vision–language models encoding broad semantic knowledge. Their present use is largely as improved feature extractors, which under-exploits them. A more ambitious use would treat the foundation model as a prior over what constitutes normal activity in a described setting, so that a natural-language scene description — “a hospital corridor during visiting hours” — would condition the model of normality without any scene-specific training data. Early open-vocabulary work (Wu et al., 2024; Yang et al., 2024) gestures in this direction.

Causal And Counterfactual Reasoning

Lv et al. (2023) demonstrated that causal analysis can identify and remove context bias. The framework extends naturally: counterfactual reasoning would allow a system to ask whether an observed behaviour would be anomalous had it occurred elsewhere or at another time, which is precisely the contextual judgement human operators make.

Continual And Federated Learning

Deployed systems must adapt without catastrophic forgetting and, where multiple installations are involved, without centralising footage. Federated learning (McMahan et al., 2017) offers a framework in which models improve across installations while raw video remains local, addressing both the data-scarcity and the privacy constraints simultaneously. Application to surveillance anomaly detection is nascent.

Synthetic Data and Simulation

UBnormal demonstrated that synthetic generation permits controlled, exhaustively annotated data with disjoint train and test anomaly sets. Extending this to systematic variation of scene, viewpoint, illumination, and demographic composition would enable the bias-controlled evaluation that Section 10.2 identifies as necessary, and would allow rare anomaly categories to be represented adequately in training.

Anomaly Anticipation

Anticipation was defined by Cao et al. (2023) as the task of anticipating an anomaly before it happens. This links the technical objective to the operational one, because surveillance is essentially for intervention and detecting something after the fact is of little use for prevention. The task is much harder than detection and has not seriously been addressed.

Hierarchical And Cascaded Architectures

The accuracy-efficiency gap is bridgeable in architecture. A cascade where a light-weight edge model filters most of the normal frames and calls heavier and heavier models only on candidates would achieve the accuracy of the heavy model at a fraction of the average cost. This design is common in other detection domains and is rarely used here, in part because prevailing benchmarks evaluate on pre-segmented clips where cascading yields no benefit.

Human-In-The-Loop Operation

Full autonomy in surveillance is not only not currently achievable accuracy-wise but arguably not desirable. Systems intended for triage, in which segments are ranked for human review rather than producing independent alerts, have weaker precision requirements and permit operator feedback as a continuing training signal. Annotation burden constrains supervised paradigms which can be alleviated by active learning formulations that select maximally informative segments to be annotated.

Robustness And Security

Certified robustness and detection of evasion attempts are areas worthy of sustained attention given documented adversarial vulnerability (Mumcu et al., 2022). The realistic threat model is physical-world attacks -- clothing, lighting, and motion patterns designed to evade detection -- and these are hardly studied.

Conclusion

This survey has discussed the classification of normal and abnormal events in intelligent surveillance systems using deep learning and reviewed over ninety studies published mainly during 2019-2026. Four conclusions follow.

First, the field has made real and substantial technical advances. The transition from handcrafted descriptors through reconstruction and prediction objectives to weakly supervised ranking and vision–language alignment has produced measurable improvement on every established benchmark, and recent semantic approaches have introduced capabilities — open-vocabulary operation, textual explanation, training-free detection — that had no counterpart a few years ago.

Second, this progress is systematically overstated by prevailing evaluation practice. Frame-level area under the curve on small, single-scene, within-distribution benchmarks measures a task materially easier than the operational one. Localisation metrics, cross-dataset evaluation, and false-alarm reporting all reveal a considerably less favourable picture, and their relative neglect has allowed a divergence between reported and practical capability to widen.

Third, the central intellectual difficulty is unchanged since the field’s inception. Abnormality is contextual, and no current method represents context adequately. The recent benchmarks that test contextual reasoning directly show performance well below that on benchmarks that do not, which suggests that much of the apparent progress has been made on the aspects of the problem that do not require it.

Fourth, the gap between research and deployment is widening rather than narrowing. The most accurate contemporary methods are the least deployable, and the field’s incentive structure — leaderboards reporting accuracy alone — provides no counterweight. Addressing this requires changes to evaluation practice as much as to architecture.

The directions we regard as most promising are those that address these structural issues rather than the incremental improvement of established architectures: foundation models employed as contextual priors rather than as feature extractors; causal and counterfactual formulations of contextual abnormality; bias-controlled synthetic benchmarks; cascaded architectures reconciling accuracy with throughput; anticipation rather than detection as the target task; and evaluation protocols aligned with operational requirements. Progress on these fronts would advance the field more substantially than a further decade of refinement within the current paradigm.

References

1. Acsintoae, A., Florescu, A., Georgescu, M. I., Mare, T., Sumedrea, P., Ionescu, R. T., Khan, F. S., & Shah, M. (2022). UBnormal: New benchmark for supervised open-set video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20143–20153.
2. Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3), 555–560.
3. Amalorpavam, G., Naik, H. T., Kumari, J., & Suresha, M. (2013). Analysis of digital images using morphological operations. *International Journal of Computer Science & Information Technology*, 5(1).
4. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836–6846.
5. Astrid, M., Zaheer, M. Z., Lee, J. Y., & Lee, S. I. (2021). Learning not to reconstruct anomalies. *Proceedings of the British Machine Vision Conference*.
6. Barbalau, A., Ionescu, R. T., Georgescu, M. I., Dueholm, J., Ramachandra, B., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M. (2023). SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229, 103656.
7. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *Proceedings of the 38th International Conference on Machine Learning*, 813–824.
8. Cai, R., Zhang, H., Liu, W., Gao, S., & Hao, Z. (2021). Appearance-motion memory consistency network for video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2), 938–946.
9. Cao, C., Lu, Y., Wang, P., & Zhang, Y. (2023). A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20392–20401.
10. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.
11. Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey* (arXiv:1901.03407). arXiv.
12. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58.
13. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107(8), 1655–1674.
14. Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., & Wu, Y. C. (2023). MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1), 387–395.
15. Cheng, M., Cai, K., & Li, M. (2021). RWF-2000: An open large scale video database for violence detection. *2020 25th International Conference on Pattern Recognition (ICPR)*, 4183–4190. IEEE.
16. Chong, Y. S., & Tay, Y. H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. *International Symposium on Neural Networks*, 189–196. Springer.
17. Climent-Pérez, P., & Florez-Revuelta, F. (2021). Protection of visual privacy in videos acquired with RGB cameras for active and assisted living applications. *Multimedia Tools and Applications*, 80(15), 23649–23664.
18. Cong, Y., Yuan, J., & Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3449–3456.
19. Doshi, K., & Yilmaz, Y. (2020). Any-shot sequential anomaly detection in surveillance videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 934–935.
20. Doshi, K., & Yilmaz, Y. (2021). Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114, 107865.
21. Doshi, K., & Yilmaz, Y. (2022). Rethinking video anomaly detection: A continual learning approach. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3961–3970.
22. Doshi, K., & Yilmaz, Y. (2023). Towards interpretable video anomaly detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2655–2664.
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
24. Feichtenhofer, C. (2020). X3D: Expanding architectures for efficient video recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 203–213.
25. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6202–6211.
26. Feng, J. C., Hong, F. T., & Zheng, W. S. (2021). MIST: Multiple instance self-training framework for video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14009–14018.

27. Georgescu, M. I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021a). Anomaly detection in video via self-supervised and multi-task learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12742–12752.
28. Georgescu, M. I., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021b). A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 4505–4523.
29. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & van den Hengel, A. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1705–1714.
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.
31. Goswami, S., Raj, K. S., Durdi, V. B., Naik, T. H., Sunil, M. P., Harish, K. K., & Dash, S. (2025). Latency minimization techniques in dense urban 5G mobile networks. *Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications*, 16(2), 154–167.
32. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 733–742.
33. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
34. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
35. Huang, C., Wu, Z., Wen, J., Xu, Y., Jiang, Q., & Wang, Y. (2021). Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics*, 18(8), 5171–5179.
36. Ionescu, R. T., Khan, F. S., Georgescu, M. I., & Shao, L. (2019). Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7842–7851.
37. Joo, H. K., Vo, K., Yamazaki, K., & Le, N. (2023). CLIP-TSA: CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection. *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234. IEEE.
38. Kim, J., & Grauman, K. (2009). Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2928.
39. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.
40. Kiran, B. R., Thomas, D. M., & Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2), 36.
41. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
42. KS, H., Naik T, H., Kartik, N., & N Nanda, K. (2026). Attentional deep learning with inverse transform sampling for robust respiratory sound classification. *Journal of Information Technology Management*, 18(1), 123–140.
43. Li, S., Liu, F., & Jiao, L. (2022). Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2), 1395–1403.
44. Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection: A new baseline. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6536–6545.
45. Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., Sun, P., & Song, L. (2024). Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 56(7), 1–38.
46. Liu, Z., Nie, Y., Long, C., Zhang, Q., & Li, G. (2021). A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13588–13597.
47. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3202–3211.
48. Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 fps in MATLAB. *Proceedings of the IEEE International Conference on Computer Vision*, 2720–2727.
49. Lu, Y., Yu, F., Reddy, M. K. K., & Wang, Y. (2020). Few-shot scene-adaptive anomaly detection. *European Conference on Computer Vision*, 125–141. Springer.
50. Luo, W., Liu, W., & Gao, S. (2017a). Remembering history with convolutional LSTM for anomaly detection. *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 439–444. IEEE.

51. Luo, W., Liu, W., & Gao, S. (2017b). A revisit of sparse coding based anomaly detection in stacked RNN framework. *Proceedings of the IEEE International Conference on Computer Vision*, 341–349.
52. Lv, H., & Sun, Q. (2024). *Video anomaly detection and explanation via large language models* (arXiv:2401.05702). arXiv.
53. Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., & Zhang, H. (2023). Unbiased multiple instance learning for weakly supervised video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8022–8031.
54. Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1975–1981.
55. Markovitz, A., Sharir, G., Friedman, I., Zelnik-Manor, L., & Avidan, S. (2020). Graph embedded pose clustering for anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10539–10547.
56. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.
57. Mehran, R., Oyama, A., & Shah, M. (2009). Abnormal crowd behavior detection using social force model. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 935–942.
58. Micorek, J., Possegger, H., Narnhofer, D., Bischof, H., & Kozinski, M. (2024). MULDE: Multiscale log-density estimation via denoising score matching for video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18868–18877.
59. Mumcu, F., Doshi, K., & Yilmaz, Y. (2022). Adversarial machine learning attacks against video anomaly detection systems. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 206–213.
60. Naik, N. H., Kumar, K. S., Suri, S., Parida, S., Reddy, B., & Sevukamoorthy, L. (2025, February). Fused IGAN-IDS system leveraging randomized tree classification for extreme performance improvement. *2025 IEEE International Conference on Emerging Technologies and Applications (MPSec ICETA)*, 1–6. IEEE.
61. Nayak, R., Pati, U. C., & Das, S. K. (2021). A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106, 104078.
62. Nguyen, T. N., & Meunier, J. (2019). Anomaly detection in video sequence with appearance-motion correspondence. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1273–1283.
63. Nievas, E. B., Suarez, O. D., García, G. B., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. *International Conference on Computer Analysis of Images and Patterns*, 332–339. Springer.
64. NT, H. N., Dev, S., Shivam, A., Singh, N., HS, S., & Mishra, A. (2025, March). Meta-heuristic approaches to parallel processing in signal representation and real-time analysis. *2025 International Conference on Automation and Computation (AUTOCOM)*, 1093–1097. IEEE.
65. Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2), 1–38.
66. Park, H., Noh, J., & Ham, B. (2020). Learning memory-guided normality for anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14372–14381.
67. Piciarelli, C., Micheloni, C., & Foresti, G. L. (2008). Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11), 1544–1554.
68. Pu, Y., Wu, X., & Wang, S. (2024). Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 33, 4923–4936.
69. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.
70. Ramachandra, B., & Jones, M. (2020). Street Scene: A new dataset and evaluation protocol for video anomaly detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2569–2578.
71. Ramachandra, B., Jones, M., & Vatsavai, R. R. (2020a). A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5), 2293–2312.
72. Ramachandra, B., Jones, M. J., & Vatsavai, R. (2020b). Learning a distance function with a Siamese network to localize anomalies in videos. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2598–2607.
73. Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., & Sebe, N. (2017). Abnormal event detection in videos using generative adversarial nets. *2017 IEEE International Conference on Image Processing (ICIP)*, 1577–1581. IEEE.
74. Ristea, N. C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M. (2022). Self-supervised predictive convolutional attentive block for anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13576–13586.

75. Rodrigues, R., Bhargava, N., Velmurugan, R., & Chaudhuri, S. (2020). Multi-timescale trajectory prediction for abnormal human activity detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2626–2634.
76. Sabokrou, M., Khalooei, M., Fathy, M., & Adeli, E. (2018a). Adversarially learned one-class classifier for novelty detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3379–3388.
77. Sabokrou, M., Pourreza, M., Fayyaz, M., Entezari, R., Fathy, M., Gall, J., & Adeli, E. (2018b). AVID: Adversarial visual irregularity detection. *Asian Conference on Computer Vision*, 488–505. Springer.
78. Saffinaz, S., Mittal, S., Subramoniam, M., NT, H. N., Asha, S., & Dutta, A. K. (2025, July). QoS-driven IIoT service placement and data analytics for smart applications in fog-cloud systems. *2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, 382–386. IEEE.
79. Santhosh, K. K., Dogra, D. P., & Roy, P. P. (2020). Anomaly detection in road traffic using visual surveillance: A survey. *ACM Computing Surveys*, 53(6), 1–26.
80. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 802–810.
81. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27, 568–576.
82. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6479–6488.
83. Sun, C., Jia, Y., Hu, Y., & Wu, Y. (2020). Scene-aware context reasoning for unsupervised abnormal event detection in videos. *Proceedings of the 28th ACM International Conference on Multimedia*, 184–192.
84. Sun, S., & Gong, X. (2023). Hierarchical semantic contrast for scene-aware video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22846–22856.
85. Suresha, M., & Harisha Naik, T. (2017). A survey on image analysis based on texture. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7, 686–695.
86. Suresha, M., Madhusudhan, S., & Harish Naik, T. (2019, November). Enhancement of reflected faces on semi-reflecting surfaces. *2019 Fifth International Conference on Image Information Processing (ICIIP)*, 205–209. IEEE.
87. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., & Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4975–4986.
88. Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35, 10078–10093.
89. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 4489–4497.
90. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
91. Vu, H., Nguyen, T. D., Le, T., Luo, W., & Phung, D. (2019). Robust anomaly detection in videos using multilevel representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 5216–5223.
92. Wan, B., Fang, Y., Xia, X., & Mei, J. (2020). Weakly supervised video anomaly detection via center-guided discriminative learning. *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.
93. Wang, X., Ma, X., & Grimson, W. E. L. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3), 539–555.
94. Wu, P., & Liu, J. (2021). Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30, 3513–3527.
95. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. *European Conference on Computer Vision*, 322–339. Springer.
96. Wu, P., Liu, X., & Liu, J. (2022). Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25, 1674–1685.
97. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., & Zhang, Y. (2024). VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6), 6074–6082.
98. Yan, C., Zhang, S., Liu, Y., Pang, G., & Wang, W. (2023). Feature prediction diffusion model for video anomaly detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5527–5537.
99. Yang, Z., Liu, J., Wu, Z., Wu, P., & Liu, X. (2024). Text prompt with normality guidance for weakly supervised video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18899–18908.

- 100.Ye, M., Peng, X., Gan, W., Wu, W., & Qiao, Y. (2019). AnoPCN: Video anomaly detection via deep predictive coding network. *Proceedings of the 27th ACM International Conference on Multimedia*, 1805–1813.
- 101.Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., & Kloft, M. (2020). Cloze test helps: Effective video anomaly detection via learning to complete video events. *Proceedings of the 28th ACM International Conference on Multimedia*, 583–591.
- 102.Zaheer, M. Z., Mahmood, A., Astrid, M., & Lee, S. I. (2020). CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. *European Conference on Computer Vision*, 358–376. Springer.
- 103.Zaheer, M. Z., Mahmood, A., Khan, M. H., Segu, M., Yu, F., & Lee, S. I. (2022). Generative cooperative learning for unsupervised video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14744–14754.
- 104.Zanella, L., Menapace, W., Mancini, M., Wang, Y., & Ricci, E. (2024). Harnessing large language models for training-free video anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.
- 105.Zhang, D., Gatica-Perez, D., Bengio, S., & McCowan, I. (2005). Semi-supervised adapted HMMs for unusual event detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 611–618.
- 106.Zhang, H., Xu, X., Wang, X., Zuo, J., Han, C., Huang, X., Gao, C., Wang, Y., & Sang, N. (2024). *Holmes-VAD: Towards unbiased and explainable video anomaly detection via multi-modal LLM* (arXiv:2406.12235). arXiv.
- 107.Zhao, B., Fei-Fei, L., & Xing, E. P. (2011). Online detection of unusual events in videos via dynamic sparse coding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3313–3320.
- 108.Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., & Hua, X. S. (2017). Spatio-temporal autoencoder for video anomaly detection. *Proceedings of the 25th ACM International Conference on Multimedia*, 1933–1941.
- 109.Zhong, J. X., Li, N., Kong, W., Liu, S., Li, T. H., & Li, G. (2019). Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1237–1246.
- 110.Zhou, H., Yu, J., & Yang, W. (2023). Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3), 3769–3777.
- 111.Zhu, Y., & Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. *Proceedings of the British Machine Vision Conference*.