iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79308

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Proactive Anomaly Detection for CCTV Footage: A Systematic Literature Review

Unnati Khanapurkar[1], Nagari Roshan[2], Mohd Abdul Baseer[3], Syed Saif Hashmi[4]
[1]*Assistant Professor, Department of Computer Science and Engineering, Methodist, College of Engineering and Technology, Abids, Hyderabad, Telangana, 500001, India*
[2, 3, 4]*Studenent, Department of Computer Science and Engineering, Methodist College of Engineering and Technology, Abids, Hyderabad, Telangana, 500001, India*

Abstract: ***Automated anomaly detection in surveillance video has received considerable research attention over the past decade, yet a recurring limitation across the literature is that most systems are designed and evaluated under clean, controlled conditions. In practice, outdoor cameras routinely operate in fog, rain, low light, and snow, and very few published methods have been tested under such conditions. This survey examines twenty-one papers published between 2016 and 2024 that are relevant to detecting four specific event types: arson, physical intrusion, loitering, and abandoned objects. The papers cover a range of techniques including multiple instance learning [1], YOLO-based detection [9, 10], multi-object tracking via SORT [7] and DeepSORT [8], memory-augmented autoencoders [13], and vision-language approaches built on CLIP [2]. Our main finding is that while individual event detectors perform well on standard benchmarks, no existing system handles all four event types reliably under adverse weather. We document the specific failure modes and identify the research gaps that need to be addressed.***

Keywords: Anomaly Detection, CCTV Surveillance, Arson, Intrusion, Loitering, Abandoned Object, Adverse Weather, Deep Learning, YOLO, CLIP, ConvLSTM, DeepSORT, SORT.

I. INTRODUCTION

The number of surveillance cameras deployed in public and private spaces has grown substantially over the past two decades. In most cities, cameras now cover transit stations, commercial areas, building perimeters, and roadways. However, simply having cameras installed does not guarantee security events will be caught. Monitoring large numbers of live feeds continuously requires human operators, and sustained attention on static video is known to degrade within roughly twenty minutes. This means that even in well-equipped control rooms, incidents can and do go unnoticed. Automating the detection of specific dangerous events is therefore not just a matter of convenience but a practical requirement for any large-scale deployment.

Deep learning has made meaningful progress on this problem. Sultani et al. [1] showed in 2018 that surveillance anomaly detectors can be trained using only video-level weak labels, avoiding the expensive and time-consuming process of annotating individual frames. Their multiple instance learning approach achieved an AUC of 75.41% on UCF-Crime, a dataset covering thirteen criminal event types including arson, and it established a baseline that later work continued to build on. Separately, the YOLO architecture introduced by Redmon et al. [9] made real-time object detection fast enough for surveillance applications, and subsequent versions like YOLOv5 [10] made it widely accessible. More recently, Radford et al. [2] introduced CLIP, a vision-language model pre-trained on large amounts of image-text data that can recognise visual concepts from text descriptions without requiring task-specific training. This has opened up new possibilities for detecting rare or visually ambiguous events in surveillance settings.

Despite this progress, a consistent issue appears when examining how these systems are evaluated. The overwhelming majority of published methods are tested under stable, well-lit, clear-weather conditions. Real outdoor cameras do not operate this way. Rain introduces noise into every frame. Fog reduces contrast to the point where person detectors fail to produce any output. Nighttime conditions remove the colour information that many classifiers depend on. Snow disrupts background modelling by continuously changing what a normal scene looks like. The gap between benchmark performance and what a system actually delivers when deployed outdoors is something we found in nearly every paper reviewed here.

A second issue concerns scope. Most published detectors focus on a single event type. Fire detectors, intrusion detectors, loitering detectors, and abandoned-object detectors each have their own research communities, benchmark datasets, and preferred architectures. Combining these into a single operational system is not straightforward, as each component has different resource requirements and produces alerts in different formats.

©IJRASET: All Rights are Reserved

Among all the papers reviewed, only Jeon et al. [6] make a genuine attempt to handle all four event types within one pipeline while also evaluating under weather-degraded footage.

This survey reviews twenty-one papers with both of these issues in mind. Section 2 covers the literature organized by theme. Section 3 presents a comparison table across all reviewed papers. Section 4 identifies the main gaps that remain open. Section 5 concludes with how these findings connect to the research proposed in the accompanying project.

II. LITERATURE SURVEY

A. The Trajectory of Video Anomaly Detection Research

To understand where the field currently stands, it is useful to look at how it has developed. Abdalla et al. [5] reviewed more than fifty anomaly detection models over a ten-year span, documenting a clear shift from hand-crafted feature methods toward end-to-end deep learning. On UCF-Crime, early CNN-based methods achieved AUC scores in the mid-seventies, while more recent approaches using vision-language features from CLIP [2] have reached above 87%. That improvement reflects genuine architectural progress rather than just additional data or tuning.

Pang et al. [3] provide a useful taxonomy for understanding why different methods succeed or fail in different conditions. They group deep anomaly detection approaches into feature extraction methods, normality modelling methods, and end-to-end scoring methods. Normality modelling approaches, which learn what a normal scene looks like and flag deviations from it, are particularly relevant to the weather problem. When fog or rain alters the appearance of a scene, the background looks different from what the model learned as normal, generating false alarms even when nothing suspicious is happening. Ramachandra et al. [4] make a similar observation specifically for single-scene detectors, noting that outdoor environments with variable backgrounds are considerably harder than the controlled indoor scenes most methods are developed on.

Khan et al. [21] surveyed video anomaly detection from a broader perspective, covering CNN, LSTM, GAN, and two-stream architectures. Their coverage is helpful for situating the range of approaches that have been tried, though weather robustness is not a focus. Among the more specific architectural contributions, memory-augmented autoencoders proposed by Gong et al. [13] stand out. Their MemAE model stores prototype representations of normal scenes in an addressable memory and forces the decoder to reconstruct only from those stored prototypes. This reduces the risk of the model also learning to reconstruct anomalies well. However, when weather changes the appearance of a normal scene, the stored prototypes no longer match incoming frames, which produces elevated reconstruction error even for normal footage. This is a concrete example of how a method that works on benchmarks can fail in outdoor deployment.

Ionescu et al. [12] addressed a different aspect of the reconstruction approach by focusing on detected objects rather than full frames. Their object-centric autoencoders reduce the influence of background variation on anomaly scoring, which is a sensible direction. The limitation is that the method still depends on object detection working correctly upstream. In poor visibility, when the detector misses people or produces inaccurate bounding boxes, the reconstruction pipeline receives bad inputs and produces unreliable scores. Doshi and Yilmaz [14] took a different angle by addressing gradual scene appearance drift through continual learning, which is relevant to long-term deployment but does not address the acute performance drops caused by weather events.

B. The Unique Position of PASS-CCTV Among Multi-Event Systems

Among all the systems reviewed in this work, PASS-CCTV by Jeon et al. [6] occupies a distinct position. It is the only published system that handles intrusion, loitering, abandoned-object detection, and arson within a single pipeline, and it is the only one evaluated on footage that explicitly includes adverse weather conditions. This makes it the central reference point for this survey. It has real limitations, but it also shows in concrete terms what the integration challenges look like when they are taken seriously.

The first problem PASS-CCTV addresses is tracking reliability under degraded conditions. In fog, low light, or overhead camera angles, individuals appear visually similar and standard trackers tend to lose or swap identities. SORT [7] uses bounding box overlap alone for frame-to-frame association, which degrades when detector confidence drops in poor visibility. DeepSORT [8] adds a learned appearance descriptor, which helps in normal conditions but provides less benefit when frames are blurry or low-contrast. Jeon et al. [6] combine a re-identification model with mid-level features from the object detector to create a richer tracking representation that degrades more gradually. A trajectory-level filter complements this by checking whether a tracked object has actually traversed the scene over time. Jittering non-human objects, like a lamp post bounding box that oscillates slightly due to detection noise, are removed because they do not sweep out meaningful ground area, even when they superficially resemble pedestrian tracks.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

With stable tracking in place, each of the four event types is handled through a dedicated mechanism. Intrusion and loitering both use fractional bounding box overlap with operator-defined alert zones. Loitering adds a ten-second persistence requirement before triggering an alert. Abandoned-object detection uses a localizer centred on each tracked person to identify associated items, then monitors the distance between owner and item over time. Arson detection uses CLIP [2] to compare image patches from trajectory segments against text descriptions of arson-related behaviour, enabling pre-ignition detection based on human actions rather than visible flame. On the KISA dataset, where roughly half the footage includes fog, rain, snow,
or low-light conditions, the system achieved F1 scores above 90% across all four event types on both validation and certification subsets [6]. Performance dropped to the 89-93% range on an abroad subset representing environments not seen during development, which points to some degree of environment-specific tuning in the current design.
One practical limitation worth noting is that PASS-CCTV does not include any mechanism for automatically adjusting its thresholds when deployed in a new location. Zone boundaries, duration thresholds, and detection confidence cutoffs all require manual configuration per site. The system also slows noticeably when many individuals are tracked simultaneously, which is a concern for busy public spaces.

C. Arson and Fire: Pixel Detection versus Behaviour Analysis

Two distinct approaches appear in the fire and arson detection literature. Pixel-level methods look directly for flame or smoke regions in video frames. Behaviour-level methods instead analyse the actions of people who may be starting a fire, using appearance or trajectory information. Each approach has a different trade-off between early detection and reliability.
Muhammad et al. [11] were among the earlier groups to apply deep CNNs to fire detection in surveillance video, using transfer learning from a pre-trained VGG network followed by a classifier. This produced reasonable results on controlled fire video datasets and represented a clear improvement over colour-based and texture-based methods that had difficulty distinguishing fire from similarly-coloured backgrounds. The YOLO architecture [9] and its successors like YOLOv5 [10] later became common choices for fire detection because their speed makes real-time processing feasible. The shared limitation of all pixel-level approaches is that they require visible fire to produce a detection. By that point in an arson scenario, the perpetrator may already have left.
The CLIP-based arson detector in PASS-CCTV [6] sidesteps this limitation by focusing on human behaviour rather than fire pixels. Text prompts describing arson-related actions are compared against image patches extracted from each person's trajectory, and a high similarity score raises an alert. On the FireNet benchmark this approach achieved 99.33% F1 without any fire-specific training, which demonstrates the genuine strength of the pre-trained CLIP model [2] for zero-shot event recognition. The cost is computational: CLIP inference adds processing overhead on top of the detection and tracking pipeline. Prompt design also matters; poorly chosen prompts can produce false positives from scenes that only superficially resemble the target descriptions.

D. Intrusion Detection: Tracking Quality as the Bottleneck

Physical intrusion detection has a clear logical definition: a person enters a zone they are not authorized to be in. The implementation challenge is less about the detection logic and more about having a reliable tracker feeding into it. Nayak et al. [17] reviewed deep learning-based perimeter intrusion detection systems and found that tracker quality is the primary factor determining overall system performance. A tracker that regularly loses or swaps identities will produce both missed detections and spurious alerts regardless of how precisely the alert zone is configured.
The progression from SORT [7] to DeepSORT [8] illustrates how this problem has been approached. SORT relies on bounding box IoU for frame-to-frame association, which is fast but fails easily when detection quality drops due to weather or crowding. DeepSORT adds a learned appearance descriptor, providing more robust re-identification across occlusions and brief disappearances. Under fog or heavy rain, however, appearance descriptors computed from low-contrast frames carry less discriminative information, reducing their benefit.
The feature coupling used in PASS-CCTV [6] can be seen as a pragmatic extension of this progression, drawing on multiple information sources to compensate for degradation in any one of them.
For zone-based alert logic, using fractional bounding box overlap rather than centroid-based crossing is more robust under weather degradation. When bounding boxes become less accurate due to reduced detection confidence, centroid-based crossing detection produces premature or delayed alerts because the estimated centroid position is noisier. Fractional overlap distributes the uncertainty across the whole box area, which gives more stable and calibrated alert timing [6, 17].

E. Loitering: The Challenge of Context-Dependent Thresholds

Loitering is arguably the most context-dependent of the four event types. Intrusion has a clear spatial definition. Abandoned objects have a clear ownership definition. But what counts as suspicious loitering depends heavily on location, time of day, and surrounding activity. Two minutes of stationary behaviour outside a locked building at 2am raises concern; the same behaviour at a bus stop at noon does not. Most published systems handle this using a fixed zone occupancy duration threshold, which is simple to implement but cannot capture contextual variation without manual recalibration per deployment site.

Nunez et al. [16] take a more structured approach by defining four distinct loitering categories: seated loitering, random walking, abnormal trajectory, and stationary. Their trajectory analysis method was evaluated on thermal infrared footage. Thermal cameras are less sensitive to poor lighting and some types of fog, which makes this dataset somewhat more relevant to adverse-weather conditions than most loitering work. The per-category accuracy results are informative, but the hardware dependency limits how widely the method can be applied. Most existing CCTV installations use visible-spectrum cameras, and replacing them with thermal units is expensive.

The PASS-CCTV loitering results [6] are worth examining closely. F1 on the validation subset reached 99.3%, but dropped to 92.9% on the abroad subset covering novel environments. This gap suggests that the ten-second duration threshold, while effective for the KISA training environments, did not transfer perfectly to new deployment sites. There is no adaptive mechanism described in [6] that would automatically adjust this threshold for a new location, which is a practical limitation for any real-world rollout.

F. Abandoned Object Detection: Background Modelling versus Owner Tracking

The two main approaches to abandoned-object detection differ in how they identify that an item has been left behind. Background modelling methods identify foreground objects that become stationary and persist beyond a time threshold. Owner-tracking methods follow people and their associated belongings, alerting when the owner leaves without an item. The two approaches have complementary strengths and weaknesses.

Qasim et al. [15] propose a two-stage pipeline that combines temporal and spatial reasoning. A ConvLSTM network first classifies short video sequences as suspicious or non-suspicious, and only when a sequence is flagged does a YOLOv8l localizer activate to identify the specific item. The ConvLSTM captures temporal dynamics across multiple frames, which single-frame detectors cannot, and the two-stage design avoids running an expensive localizer on every frame. On PETS 2006 and ABODA, the method achieved 99.20% classifier accuracy and 99.70% localizer accuracy. These are strong results, though both datasets are small and recorded under controlled conditions. ABODA in particular contains only eleven videos, which limits how much can be concluded about generalization.

The owner-tracking approach in PASS-CCTV [6] avoids background modelling entirely, which makes it inherently more robust to weather-induced background variation. By associating detected items with their nearest tracked owner and monitoring the spatial separation over time, the system detects abandonment as the moment a person moves beyond a threshold distance from a previously associated object. This works well when the owner is visible during the abandonment event, which is the common case. The method misses situations where an object is placed before the person enters the camera's field of view, since no owner association can be established

G. Adverse Weather and Low-Light: Available Mitigation Approaches

Weather robustness is the area where the reviewed literature is weakest relative to what practical deployment requires. Most papers acknowledge weather as a challenge but then proceed to evaluate only on clean footage. There are a few exceptions that offer concrete technical direction.

Liang et al. [18] address low-light conditions using an event-camera guided enhancement approach. Event cameras record brightness changes at very high temporal resolution rather than capturing full frames at fixed intervals, so they avoid the exposure time trade-off that forces standard cameras to choose between motion blur and noise in dim scenes. By using event-stream data to guide noise reduction in conventional frame video, the method achieved PSNR of 22.81 dB and SSIM of 0.818 on a synthetic low-light benchmark, outperforming frame-only enhancement. The practical barrier is hardware: event cameras are expensive and absent from most existing CCTV infrastructure.

Li et al. [19] cover deep learning methods for low-light image and video enhancement using standard camera setups. Their survey covers paired training approaches that learn to map underexposed images to well-lit references. Results on standard low-light datasets are solid. The open question is how well these methods generalise to genuine surveillance footage rather than the relatively clean dataset images they are trained and evaluated on.

Zhang et al. [20] provide one of the more systematic evaluations of how deep learning detection models perform under fog, rain, and snow. Their key finding is that models trained on clear-weather data lose a substantial portion of their accuracy under adverse weather, but that training with synthetic weather augmentation recovers most of that loss without any architectural modification. They also observe that fog and rain require different mitigation strategies: fog primarily degrades spatial resolution and responds to dehazing preprocessing, while rain introduces temporal noise that responds better to cross-frame smoothing. This distinction is largely absent from the surveillance-specific literature, where adverse weather tends to be treated as a single undifferentiated problem.

III.COMPARATIVE SUMMARY OF REVIEWED WORKS

Table 1 summarizes all 21 papers covered in this survey, ordered by reference number. Only papers with identifiable authors, a verifiable publication venue, and reproducible results are included.

Table I
Comparison of Reviewed Papers
Note: Dataset names, metric abbreviation, and technique labels are standard field terminology and are used without modification.

Authors (Ref)	Description	Techniques Used	Evaluation Parameters	Dataset	Limitations
Sultani et al. [1] (2018)	Weakly supervised anomaly detection on real surveillance footage using MIL ranking loss on video-level labels	C3D features, Multiple Instance Learning (MIL), Ranking loss, Temporal segment pooling	AUC 75.41% on UCF-Crime	UCF-Crime (13 crime types incl. arson)	Single event type per model; no adverse weather evaluation; video-level labels only
Radford et al. [2] (2021)	CLIP: vision-language model pre-trained contrastively enabling zero-shot scene classification from text	Contrastive Language-Image Pre-training, ViT and ResNet backbones, Natural language supervision	Zero-shot accuracy on ImageNet; linear probe on 27 datasets	400M internet image-text pairs; 27 downstream datasets	Not surveillance-specific; requires careful prompt design; high inference cost
Pang et al. [3] (2021)	Deep learning anomaly detection review proposing three-category taxonomy for AD approaches	Autoencoder, GAN, One-class NN, Deep SVDD, End-to-end anomaly scoring	AUC, AP, F1 across multiple benchmark datasets	KDD-Cup, CIFAR-10, MVTec and others	Survey only; surveillance-specific limitations underemphasized; no weather analysis
Ramachandra et al. [4] (2020)	Survey of single-scene VAD covering distance-based, probabilistic and reconstruction-based methods	Background subtraction, GMM, Sparse coding, Deep autoencoders, Optical flow	EER, AUC, Frame-level accuracy	UCSD Pedestrian, CUHK Avenue, Subway datasets	Single-scene focus; outdoor and weather degradation not discussed
Abdalla et al. [5] (2024)	10-year VAD survey covering 50+ models across supervised, weakly supervised and vision-language paradigms	CNN, GAN, ViT, CLIP-based VLM, MIL, Unsupervised AE	AUC on UCF-Crime, ShanghaiTech, XD-Violence; EER; AP	UCF-Crime, ShanghaiTech, XD-Violence, CUHK Avenue	Survey only; weather robustness not covered; arXiv preprint

Jeon et al. [6] (2024)	PASS-CCTV: unified pipeline for intrusion, loitering, abandonment and arson with adverse weather testing	YOLOv8, OSNet re-ID, Feature coupling, Zone intersection ratio, CLIP arson scorer, Top-down luggage tracker	F1 greater than 90% on all event types, Precision, Recall, Alert delay (sec)	KISA CCTV (fog, rain, snow, low-light), ABODA, FireNet	Slows with many simultaneous tracklets; per-site threshold tuning needed; no adaptive recalibration
Bewley et al. [7] (2016)	SORT: lightweight real-time multi-object tracking using Kalman filter and Hungarian IoU assignment	Kalman filter, Hungarian algorithm, IoU bounding box association	MOTA, MOTP on MOTChallenge 2015	MOTChallenge 2015	No appearance modelling; identity switches common in occlusion or poor visibility
Wojke et al. [8] (2017)	DeepSORT: extends SORT with CNN appearance descriptor for more robust re-identification across occlusions	Kalman filter, Hungarian algorithm, CNN appearance descriptor, Cosine distance metric	MOTA, MOTP, ID switch count on MOT16	MOT16	Appearance descriptor degrades in low contrast or fog; higher compute cost than SORT
Redmon et al. [9] (2016)	YOLO: unified single-pass object detection treating bounding box prediction as regression on grid cells	Single CNN, Grid-based regression, Bounding box prediction, Class probability output	mAP on PASCAL VOC 2007 and 2012, FPS	PASCAL VOC 2007, VOC 2012, COCO	Misses small and clustered objects; later versions needed for most surveillance use cases
Jocher et al. [10] (2020)	YOLOv5: improved YOLO with CSPDarknet backbone and multi-scale neck; widely used for fire and object detection	CSPDarknet53, PANet + FPN neck, Anchor-based multi-scale head, Mosaic augmentation	mAP on MS-COCO, FPS, Model size	MS-COCO	No peer-reviewed paper; GitHub and Zenodo repository only; anchor tuning needed per dataset
Muhammad et al. [11] (2018)	CNN-based fire detection in surveillance video using VGG transfer learning and SVM classification	VGG pre-trained CNN, Transfer learning, Feature extraction, SVM classifier	Accuracy, Precision, Recall, FPS on fire video datasets	VisiFire, Bilkent fire video dataset	Requires visible flame or smoke to trigger; no adverse weather or nighttime evaluation
Ionescu et al. [12] (2019)	Object-centric autoencoders for anomaly detection reconstructing detected objects rather than full frames	Faster R-CNN object proposals, Object-centric AE, Per-object reconstruction error scoring	AUC on ShanghaiTech, UCSD Ped2, CUHK Avenue	ShanghaiTech, UCSD Ped2, CUHK Avenue	Depends on upstream detection quality; weather-induced detection errors propagate through pipeline

Authors (Ref)	Description	Techniques Used	Evaluation Parameters	Dataset	Limitations
Gong et al. [13] (2019)	MemAE: memory-augmented autoencoder storing prototype normal patterns to prevent anomaly reconstruction	Encoder-decoder AE, Memory module with attention-based retrieval, Sparse memory addressing	AUC on UCSD, CUHK Avenue, ShanghaiTech	UCSD Ped1 and Ped2, CUHK Avenue, ShanghaiTech	Stored normal prototypes confused by weather-induced appearance changes; no outdoor testing

Doshi and Yilmaz [14] (2020)	Continual learning framework for anomaly detection to handle gradual scene appearance drift over time	Online learning, Experience replay, Elastic Weight Consolidation (EWC), AE base model	AUC, Forgetting rate across sequential evaluations	UCSD, CUHK Avenue, custom long-term sequences	Targets gradual drift not acute weather events; single event type focus
Qasim et al. [15] (2024)	Two-stage pipeline: ConvLSTM classifier activates YOLOv8l localizer only on suspicious sequences	ConvLSTM (5 layers), YOLOv8l, CSPDarknet53, Transfer learning, FPN+PAN, Binary cross-entropy	Classifier accuracy 99.20%, Localizer 99.70%, mAP50, Precision, Recall	PETS 2006, ABODA (11 videos)	No adverse weather testing; small datasets limit generalization conclusions; high memory footprint
Nunez et al. [16] (2024)	Trajectory-based loitering detection on thermal infrared imagery with four sub-type classification	Thermal infrared camera, Trajectory feature analysis, LSTM temporal modelling, Sub-type classifier	Precision, Recall, Accuracy per loitering sub-type	Long-term Thermal Drift Dataset	Requires thermal camera hardware; not evaluated on standard visible-spectrum CCTV
Nayak et al. [17] (2022)	Survey of deep learning-based perimeter intrusion detection comparing zone-based and behaviour-based methods	CNN detectors, SORT, DeepSORT, Kalman filter, Zone overlap and trajectory methods	Precision, Recall, F1, False alarm rate per reviewed system	Multiple PIDS benchmark datasets	Survey only; weather impact on tracking and zone detection accuracy not analysed
Liang et al. [18] (2023)	Event-camera guided low-light video enhancement using event-stream data to guide frame-based processing	Event camera, Multimodal coherence modelling, Temporal coherence propagation, CNN optical flow	PSNR 22.81 dB, SSIM 0.818, LPIPS 0.2747	Synthetic low-light video benchmark	Requires hybrid event plus frame camera hardware; no downstream anomaly detection evaluation
Li et al. [19] (2021)	Survey of deep learning methods for low-light image and video enhancement covering standard camera setups	CNN-based enhancement networks, Paired training data, Perceptual and reconstruction loss functions	PSNR, SSIM on standard low-light datasets	LOL, VE-LOL low-light image datasets	Evaluation uses dataset images; generalization to real surveillance footage not demonstrated
Zhang et al. [20] (2021)	Benchmarking deep learning detection models under fog, rain and snow with analysis of augmentation strategies	Multiple CNN detection models, Domain adaptation, Synthetic weather augmentation per weather type	mAP drop vs. clear weather; recovery rate after augmentation; per-weather analysis	Foggy Cityscapes, ACDC, DAWN adverse weather datasets	Driving scene domain; transfer to fixed-camera surveillance requires additional adaptation
Khan et al. [21] (2021)	Survey of video anomaly detection covering CNN, LSTM, GAN, and two-stream network architectures	CNN, LSTM, GAN, Two-stream networks, Spatiotemporal feature extraction methods	AUC, Accuracy, F1 across standard VAD benchmarks	UCF-Crime, ShanghaiTech, CUHK Avenue, UCSD datasets	arXiv preprint; covers methods up to 2021 only; weather robustness not analyzed

IV. RESEARCH GAPS AND OPEN CHALLENGES

After going through all the reviewed papers, five gaps stand out as particularly significant and consistently unaddressed across the literature.

A. The Trajectory of Video Anomaly Detection Research

This is the most important gap we identified. PASS-CCTV [6] is the only system covering all four event types, and it does test on some weather-degraded footage from the KISA dataset. However, its weather robustness is not by design: there is no weather-specific adaptation mechanism, no weather-augmented training, and no analysis of which event type degrades most under which condition. Zhang et al. [20] showed that fog and rain each need different mitigation strategies, and that augmentation-based adaptation can recover much of the lost performance. Applying those findings within a multi-event surveillance pipeline is the most direct way to address this gap, and it is the main technical direction of our proposed project.

B. Benchmark Datasets Do Not Reflect Deployment Conditions

UCF-Crime [1] covers arson but not under weather-varied conditions. ABODA has only eleven videos. MOTChallenge datasets [7, 8] are useful for tracking evaluation but contain no anomaly labels and no weather conditions. The KISA dataset used in [6] is probably the best match for this problem but is not publicly available, so other researchers cannot use it for comparison. Without a public benchmark that combines all four event types with systematic weather variation, it is difficult to fairly compare approaches or to measure real-world readiness. This is a gap that the community would benefit from addressing collectively.

C. Most Approaches Are Reactive Rather Than Proactive

The term proactive appears in the title of this survey and in the name of the most relevant system [6], but in practice most methods alert only after an anomaly is fully visible. Fire detectors trigger when flame is present. Intrusion detectors trigger after zone entry. The CLIP-based arson component in [6] is a meaningful exception: by analysing human behaviour before ignition, it can alert earlier than pixel-level fire detectors. But this pre-event reasoning is not extended to intrusion, loitering, or abandonment. Detecting suspicious approach trajectories before someone crosses a zone boundary, or identifying that a person has surveyed an area before returning to linger, would require behavioural prediction that current methods do not support.

D. Computational Cost Limits Edge Deployment

Adding capabilities to a detection system generally means adding compute. PASS-CCTV [6] slows considerably when tracking many individuals simultaneously. CLIP [2] inference adds overhead beyond the detection and tracking pipeline. By contrast, single-event systems like the fire detector of Muhammad et al. [11] can be fast because they do relatively little. Achieving real-time performance on a four-event system under adverse weather, on the kind of edge hardware that is realistically available in surveillance installations, is an engineering problem that has not yet been solved.

E. Explainability Has Not Been Addressed

When an automated system generates a security alert, the operator who receives it needs to assess it quickly. The text-prompt similarity scores used for arson in PASS-CCTV [6] provide some interpretability for that specific event type. For intrusion, loitering, and abandonment, the output is a flag and a confidence score with no supporting explanation. As automated surveillance is integrated into consequential security workflows, the ability to understand and audit why a specific alert was raised will increasingly matter from both an operational and regulatory standpoint. We did not find any paper in the reviewed set that treats this as a primary design concern.

V. CONCLUSION

This survey reviewed twenty-one papers covering deep learning-based anomaly detection for surveillance video, focusing on arson, intrusion, loitering, and abandoned objects under adverse weather conditions.

Across the reviewed work, genuine progress has been made on individual components. The MIL framework of Sultani et al. [1] established a viable path for weakly supervised training on real surveillance footage. SORT [7] and DeepSORT [8] provide practical multi-object trackers that support zone-based anomaly analysis. YOLO [9, 10] made detection fast enough for real-time use. MemAE [13] and object-centric autoencoders [12] improved the selectivity of reconstruction-based anomaly scoring. CLIP [2] enabled zero-shot event recognition that outperformed task-specific detectors on arson in [6]. The ConvLSTM pipeline of Qasim et al. [15] achieved strong accuracy on available abandoned-object benchmarks. These are solid individual contributions.

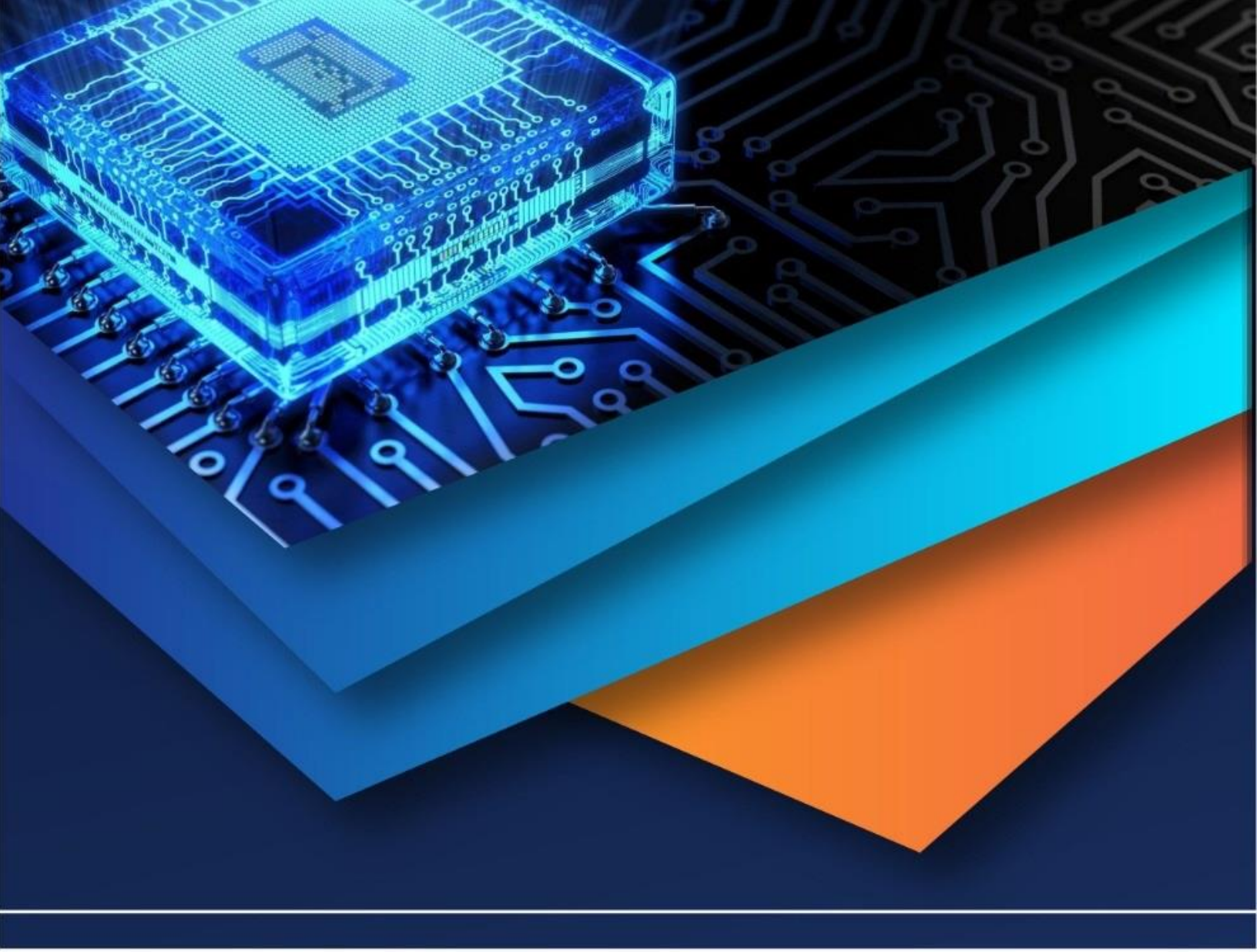
©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The more difficult question is how these contributions hold up when integrated and tested under conditions that reflect actual outdoor deployment. PASS-CCTV [6] is the closest thing to an answer that the current literature provides. It shows that multi-event detection is technically feasible and that reasonable performance under some adverse conditions is achievable. The gap between its validation results and its abroad-subset results also shows that there is still work to do on generalization and weather-specific robustness.

For the accompanying project, we plan to build on the PASS-CCTV architecture [6] and incorporate weather-aware preprocessing strategies informed by the findings of Zhang et al. [20] and the enhancement approaches of Liang et al. [18] and Li et al. [19]. We acknowledge that the available public datasets do not fully represent the target deployment conditions, which will constrain the strength of our experimental conclusions. Even so, demonstrating measurable improvement under controlled weather-variation scenarios, and clearly documenting what does and does not transfer across conditions, would represent a useful step toward a practically deployable system.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," Proc. IEEE/CVF CVPR, Salt Lake City, UT, USA, Jun. 2018, pp. 6479-6488. DOI: 10.1109/CVPR.2018.00657.
[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," Proc. ICML, 2021, pp. 8748-8763. arXiv:2103.00020.
[3] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection: A Review," ACM Computing Surveys, vol. 54, no. 2, pp. 1-38, Mar. 2021. DOI: 10.1145/3439950.
[4] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A Survey of Single-Scene Video Anomaly Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, pp. 2293-2312, 2020. DOI: 10.1109/TPAMI.2020.2994089.
[5] M. Abdalla, S. Javed, M. Al Radi, A. Ulhaq, and N. Werghi, "Video Anomaly Detection in 10 Years: A Survey and Outlook," Neural Computing and Applications, 2024. arXiv:2405.19387. DOI: 10.1007/s00521-025-11659-8.
[6] H. Jeon, H. Kim, D. Kim, and J. Kim, "PASS-CCTV: Proactive Anomaly Surveillance System for CCTV Footage Analysis in Adverse Environmental Conditions," Expert Systems with Applications, vol. 254, p. 124391, Nov. 2024. DOI: 10.1016/j.eswa.2024.124391.
[7] A.Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," Proc. IEEE ICIP, Phoenix, Z, USA, Sep. 2016, pp. 3464-3468. arXiv:1602.00763.
[8] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," Proc. IEEE ICIP, Beijing, China, Sep. 2017, pp. 3645-3649. arXiv:1703.07402.
[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Proc. IEEE/CVF CVPR, Las Vegas, NV, USA, Jun. 2016, pp. 779-788. DOI: 10.1109/CVPR.2016.91.
[10] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, L. Alabdulmohsin et al., "ultralytics/yolov5: v6.0," Zenodo, Nov. 2021. DOI: 10.5281/zenodo.3908559.
[11] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, "Convolutional Neural Networks Based Fire Detection in Surveillance Videos," IEEE Access, vol. 6, pp. 18174-18183, Mar. 2018. DOI: 10.1109/ACCESS.2018.2812835.
[12] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video," Proc. IEEE/CVF CVPR, Long Beach, CA, USA, Jun. 2019, pp. 7842-7851. DOI: 10.1109/CVPR.2019.00796.
[13] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," Proc. IEEE/CVF ICCV, Seoul, South Korea, Oct. 2019, pp. 1705-1714. DOI: 10.1109/ICCV.2019.00180.
[14] K. Doshi and Y. Yilmaz, "Continual Learning for Anomaly Detection in Surveillance Videos," Proc. IEEE/CVF CVPR Workshops, Seattle, WA, Jun. 2020, pp. 254-255. arXiv:2008.02787.
[15] A. M. Qasim, N. Abbas, A. Ali, and B. A. A. Al-Ghamdi, "Abandoned Object Detection and Classification Using Deep Embedded Vision," IEEE Access, vol. 12, pp. 30786-30798, Feb. 2024. DOI: 10.1109/ACCESS.2024.3369233.
[16] J. C. Nunez, M. Berge, and T. Moeslund, "Identifying Loitering Behavior with Trajectory Analysis," Proc. IEEE/CVF WACVW, Waikoloa, HI, Jan. 2024. DOI: 10.1109/WACVW60836.2024.00035.
[17] R. Nayak, U. C. Pati, and S. K. Das, "A Survey on Deep Learning-Based Methods for Perimeter Intrusion Detection," MDPI Sensors vol. 22, no. 9, p. 3601, 2022. DOI: 10.3390/s22093601.
[18] J. Liang, Y. Yang, B. Li, P. Duan, Y. Xu, and B. Shi, "Coherent Event Guided Low-Light Video Enhancement," Proc. IEEE/CVF ICCV, Paris, France, Oct. 2023, pp. 10615-10625.
[19] C. Li, C. Guo, W. Han, J. Gu, M.-M. Cheng, J. Cheng, and C. C. Loy, "Low-Light Image and Video Enhancement Using Deep Learning: A Survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 9396-9416, Dec. 2021. arXiv:1805.10536.
[20] Y. Zhang, H. Chen, X. Li, K. Wang, and Y. Yu, "Benchmarking Deep Learning for Adverse Weather Object Detection," arXiv:2103.15114, 2021.
[21] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A Survey on Video Anomaly Detection," arXiv:2105.03858, May 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)