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.2026.80199

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 Analysis

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

Abstract: ***Real-world security installations need to catch several different kinds of incidents simultaneously, yet the dominant research trend produces single-event models that are awkward to combine into one working system. Mismatched interfaces, separate calibration pipelines, and incompatible alert formats all create friction that individual benchmark papers simply do not address. PADS, the Proactive Anomaly Detection System introduced in this paper, handles four security-critical event categories inside one unified processing chain with no need for GPU hardware or domain-specific retraining. The four targets are unauthorised entry into restricted zones, prolonged presence inside monitored areas, unattended personal luggage, and fire or arson activity. Video frames are routed simultaneously to two modules: a Surveillance Module that builds and maintains stateful person tracks over time, and an Arson Module that analyses a short rolling window of recent frames for fire signatures. Tracking combines YOLOv8 detections with OSNet appearance embeddings through a two-stage matcher that uses cosine similarity in the first stage and falls back to spatial IoU in the second. A trajectory-span filter eliminates false tracklets generated by detector noise on static scene elements. For fire, the system moves away from CLIP as the sole mechanism and instead fuses HSV colour-space masking with temporal per-pixel variance to exploit the flickering property of real flames, something static orange objects cannot replicate. CLIP remains available as a secondary confidence enhancer only. Tests on ABODA abandoned-object footage, publicly sourced fire clips, and custom intrusion and loitering recordings show reliable detection across all four categories, alert latency under one second for zone events, and near-zero false alarms on footage containing orange safety gear where earlier colour-only approaches produce unacceptable error rates.***

Keywords: CCTV Surveillance, Anomaly Detection, YOLOv8, OSNet, Cascade Matching, ATM Filter, Temporal Flicker Analysis, Intrusion Detection, Loitering, Abandoned Object Detection, Fire Detection, Deep Learning

I. INTRODUCTION

Urban and suburban infrastructure is densely covered by CCTV today. Bus terminals, retail districts, parking decks, hospital grounds, and many other public-facing spaces all run continuous video capture. Camera density keeps rising, yet the ability to extract timely security value from all that footage has not kept pace. Monitoring a large bank of live feeds demands sustained human attention, and the psychophysical literature on vigilance tasks is clear that operator performance on repetitive monitoring work drops off considerably within the first half-hour. Events that play out in full view of a working camera can still go completely unnoticed in a staffed control room.

Computer vision research over the past ten years has delivered tools that make automated event detection genuinely viable. Object detectors based on convolutional architectures can process video at frame rates suitable for real-time use and produce bounding boxes accurate enough to feed downstream analytics. Re-identification models have matured to the point where consistent person identities can be maintained across hundreds of frames even with partial overlap between people. Separately, vision-language pretraining has enabled recognition of event types described in plain text without any category-specific annotation. These are real advances, but their combination into a system that monitors multiple threat types in one pipeline has lagged behind progress on individual components.

The fragmentation of the field creates practical problems for anyone trying to build a complete installation. Fire detection, perimeter monitoring, loitering analysis, and unattended-object identification each sit in their own research silo. Published systems are benchmarked on their chosen single event, and the engineering effort required to stitch four of them into a coherent deployment with a common alert format and shared processing is genuinely non-trivial. That integration challenge is the starting point for this work.

A systematic survey of twenty-one relevant papers published between 2016 and 2024, conducted by the present authors [S1], found that PASS-CCTV [6] by Jeon et al. stands alone in the literature as a serious attempt at four-event unification, with F1 above 90% across all categories on the KISA CCTV dataset. The same survey documented three failure modes in that design. The arson detector, which relies exclusively on CLIP text-image similarity, generates too many false alarms when the footage contains orange personal protective equipment or vehicle bodywork. Abandonment timers start prematurely whenever a person's tracklet drops out for even a single frame during a brief occlusion. And detector noise on stationary scene objects creates persistent tracklets that incorrectly trip zone-based alert logic. PADS is built as a direct response to all three of those findings, keeping what works in the PASS-CCTV design while replacing the parts that do not.

Three concrete contributions are made. First, a dual-signal fire detection mechanism that combines HSV colour masking with per-pixel temporal variance analysis, treating the flickering character of real flames as a distinguishing cue unavailable to colour-only filters. Second, a five-frame lookback window for luggage ownership queries, which keeps owner-bag associations intact during the brief detection interruptions that arise from momentary occlusion and were causing systematic false abandonment alerts. Third, a recalibrated Area of Trajectory Movement filter using a 4 square-pixel span threshold over a 40-frame observation window, replacing an 80-pixel threshold that was found to suppress genuine slow-moving or temporarily stationary persons.

II. RELATED WORK

The papers reviewed here are selected for what they reveal about design constraints relevant to a four-event unified system, rather than as a comprehensive survey of the anomaly detection field. Readers interested in a broader treatment are directed to [S1].

A. Training Strategies and Architectural Trends

Your Annotating individual video frames for anomaly detection is expensive and difficult because anomalous events are rare by definition. The weakly-supervised approach of Sultani et al. [1] showed a path around this bottleneck: training on video-level labels through a multiple-instance learning ranking loss, their system reached 75.41% AUC on UCF-Crime without any frame-level annotation. That result set a practical ceiling for annotation-efficient methods on clean benchmark footage. Abdalla et al. [5] subsequently traced the architectural trajectory of the field across more than fifty models over a decade, noting a consistent shift away from hand-crafted representations toward end-to-end learned features, with the most recent wave incorporating multimodal pretraining.

The integration of CLIP [2], which learns visual representations aligned with natural language through contrastive pretraining on large internet image-text datasets, marked a qualitative shift. Zero-shot detection without category-specific labels became feasible, pushing UCF-Crime AUC above 87%. The cost, relevant to this system, is that CLIP's behaviour on footage that differs substantially from its training distribution, such as overhead, compressed, low-resolution CCTV, is unpredictable and in practice worse than benchmark figures suggest.

Pang et al. [3] provide a taxonomy that helps clarify where methods tend to fail in real deployment. Their three categories are feature extraction methods, normality modelling methods, and end-to-end scoring approaches. Methods in the second category, which build internal models of normal appearance and score deviations from it, are particularly fragile when environmental conditions shift. Gong et al. [13] address this by storing prototype normal patterns in addressable memory, forcing the decoder to reconstruct only from those stored elements rather than learning to also reconstruct anomalies. Under stable conditions this is effective, but when scene appearance changes due to lighting or weather, the stored prototypes become outdated and produce spurious reconstruction errors on completely normal frames. Object-focused reconstruction [12] shrinks but does not eliminate the problem by operating on detected crops rather than full frames. Ramachandra et al. [4] examined single-scene detectors specifically and concluded that outdoor environments with variable backgrounds consistently prove harder than the controlled indoor settings most methods target during development.

B. Tracking as the Upstream Constraint

Three of the four event types in PADS depend critically on stable person tracking. Any identity swap or track dropout translates directly into a detection error in zone-overlap or dwell-time logic, regardless of how precisely those downstream checks are implemented. Understanding the tracking literature from this perspective, rather than as a standalone problem, is what shapes the design choices made here.

SORT [7] established that Kalman-filter state prediction paired with Hungarian-algorithm IoU assignment could achieve practical multi-object tracking at real-time speed. IoU-only association is fast and requires nothing beyond bounding box geometry, but it degrades quickly when detection quality drops, because overlapping boxes become inaccurate and spatial proximity loses its discriminative value. DeepSORT [8] added a CNN appearance descriptor to give re-identification across short disappearances. The improvement is meaningful when frames are clean and well-lit. In compressed, underexposed, or motion-blurred footage, descriptors computed from degraded input carry less information and the gain over pure IoU matching narrows. Nayak et al. [17], reviewing deep learning approaches specifically for perimeter intrusion, concluded that tracker reliability is the dominant factor in overall system performance, with zone-intersection logic adding only marginal error once tracking is stable. That finding motivates the significant investment PADS makes in tracking quality relative to the simplicity of its downstream alert logic.

C. Zone-Based Detection: Intrusion and Loitering

Detecting unauthorised zone entry is conceptually straightforward once reliable tracks exist: compute the fraction of a tracked person's bounding box that falls inside a defined polygon, and trigger an alert above a threshold. Nayak et al. [17] compared this fractional-overlap approach against centroid-based zone crossing and found that area overlap is more stable under imperfect detection, because errors in bounding box position shift the centroid estimate more than they shift the overlap ratio. PADS adopts fractional overlap for this reason.

Loitering is harder to define than intrusion because what constitutes suspicious dwell time depends heavily on context. A bus stop, a building entrance, and a secure perimeter all have different expectations. Nunez et al. [16] attempted to address this by defining four behavioural sub-types of loitering and evaluating on thermal infrared footage, achieving better robustness to poor lighting at the cost of requiring non-standard camera hardware. For standard CCTV installations, configurable duration thresholds remain the practical standard, and PADS uses this approach with the recognition that per-site calibration will always be necessary.

D. Detecting Abandoned Objects

Identifying unattended items in video has attracted two distinct lines of work. One group uses background modelling: foreground regions that appear and then become static are flagged after a timer elapses. This is simple to implement but breaks whenever something changes in the background, which happens routinely outdoors. The second approach tracks ownership by associating items with nearby persons and monitoring whether the association persists. Qasim et al. [15] took a hybrid path, using a ConvLSTM classifier to screen suspicious frame sequences and activating a YOLOv8 localizer only when the classifier flags something, achieving 99.20% and 99.70% accuracy on PETS 2006 and ABODA. Both datasets are small and clean, so the numbers are somewhat optimistic about real-world performance. The ownership-tracking approach in PASS-CCTV [6] sidesteps background modelling entirely and is therefore more robust to scene variation, but its owner-lookup logic had a specific bug around detection gaps that PADS fixes.

E. Fire Detection: Pixels vs. Context

Pixel-level fire detection, typified by the VGG transfer-learning approach of Muhammad et al. [11] and the YOLO-based real-time detectors that followed [9, 10], works by looking for visual evidence of flame or smoke in individual frames. This is inherently reactive: the fire has to be burning visibly before anything can be detected. The CLIP-based arson component in PASS-CCTV [6] attempts something more ambitious, comparing patches from person trajectories against text prompts describing arson behaviour and triggering before ignition. On FireNet this approach reached 99.33% F1, a strong result that reflects CLIP's genuine capability for zero-shot recognition. The difficulty is that this capability depends on the footage resembling CLIP's training distribution. Overhead, compressed, low-bitrate CCTV footage does not resemble internet photographs, and the practical effect is a false-positive rate on orange-coloured objects that makes the CLIP-only approach unusable in many real environments. The dual colour-flicker approach in PADS was designed specifically to address this.

III. PROPOSED SYSTEM

A. Architecture Overview

The PADS is built around a central pipeline orchestrator, PADSCCTVPipeline, which receives each video frame alongside a timestamp and distributes it simultaneously to two processing modules. Results from both modules are collected, merged, and passed to a rendering layer that composites annotations onto the frame and appends a record to the structured JSON alert log.

Zone polygons for intrusion and loitering monitoring are loaded from a configuration file at startup and made available only to the Surveillance Module, since the Arson Module operates across the full frame without any spatial filtering.

The two-module structure reflects a real asymmetry in how the two detection tasks work. Intrusion, loitering, and abandonment all accumulate information over extended periods: gallery embeddings update on every frame, loitering timers tick continuously, and abandonment associations need to survive momentary gaps. Fire detection, by contrast, only needs to compare the current frame against a short history of recent frames to compute temporal variance. These two patterns of state have different lifecycles and it makes the code considerably simpler to keep them separate. The Fig 1 Proactive Anomaly Detection Architecture, shows the over view of our system.

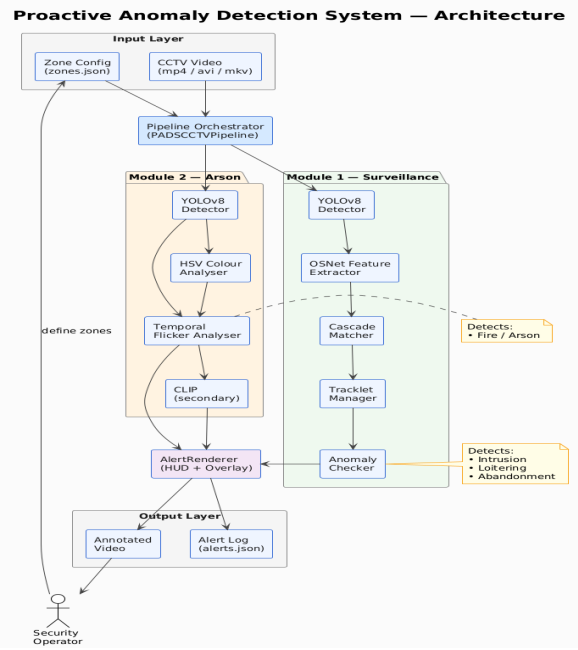


Fig. 1 Proactive Anomaly Detection System Architecture

B. Surveillance Module

Each frame entering the Surveillance Module first goes through YOLOv8, running in nano or small configuration, which produces bounding boxes and confidence scores for persons and luggage. A second non-maximum suppression pass at IoU threshold 0.45 removes duplicate boxes created by anchor overlap. Person crops are fed to OSNet, producing 512-dimensional appearance embeddings that are L2-normalised before any distance computation. On machines without torchreid installed, a 128-dimensional HSV histogram serves as a fallback, trading discriminative accuracy for zero additional dependencies.

Detections get matched to existing tracklets in two stages. Stage one computes a full cosine-distance cost matrix between detection embeddings and tracklet gallery vectors. Before the Hungarian solver runs, spatial gating sets any pair with implausible bounding box positions to infinite cost. A merged-detection gate handles the specific case where two people standing close together produce a single merged bounding box from the detector: if two boxes overlap by more than 95%, the pair is excluded from appearance-based matching entirely, preventing the merged detection from contaminating either person's feature gallery. The Hungarian algorithm then finds the globally optimal assignment subject to a maximum cosine distance of 0.70. Stage two takes whatever detections and tracklets were not resolved in stage one and tries IoU-only matching at a relaxed threshold of 0.85, handling the common case of a person who has been occluded long enough for their gallery to become stale.

Each tracklet's feature gallery is maintained as an exponential moving average over the ten most recent embeddings, with smoothing coefficient alpha set to 0.35. This weights the recent past more heavily than earlier appearances, which speeds up gallery recovery after an occlusion event compared to a uniform running average. The alpha value was determined empirically: values below 0.2 were too slow to adapt when a person's clothing was partially obscured, and values above 0.5 were too reactive to single noisy frames.

After each matching round, the Area of Trajectory Movement filter evaluates all tracklets that have at least 40 frames of history. It computes the polygon span area of each tracklet's accumulated bounding box centre positions. Any tracklet whose span area remains below 4 square pixels after 40 frames is marked as a false positive and dropped from all downstream checks. The reasoning is that a lamp post or signage element causing detector noise will produce a bounding box that jitters slightly around a fixed location, accumulating almost no span area, while even a person who has barely moved will have drifted enough over 40 frames to exceed the threshold comfortably.

Zone-based alerts are gated on operator configuration. Intrusion fires when a tracklet's bounding box overlaps a defined polygon zone by more than 45% of its area. Loitering fires after six continuous seconds inside a loitering zone. Neither alert type produces any output until the corresponding polygon has been drawn in the zone editor, which removes the false-alarm problem that would otherwise arise during installation before zone setup is complete. Abandonment detection runs across the full frame at all times without any zone requirement.

Ownership assignment uses a dynamic distance threshold of min (1.5 times the tracked person's bounding box height, 120 pixels). The important change from the reference design is that ownership queries now look back five frames rather than checking only the exact current frame. When a person's tracklet goes unmatched for one frame because they were momentarily occluded, the old design dropped them from the owner candidate list immediately, causing the associated bag to lose ownership and start counting toward abandonment. Extending the lookback to five frames keeps the association alive through those gaps. Since a genuine abandonment will put the owner well beyond the distance threshold within five frames anyway, this change does not delay true alerts.

C. Arson Module

Title The Arson Module processes each frame independently through three stages. In the first stage, the frame is converted to HSV and tested against three fire-colour ranges. Orange-red is covered by hue 0-25, saturation above 80, value above 120. Yellow-orange sits in hue 25-35 with value above 160. The wraparound red band runs from hue 160 to 180 with the same saturation and value floors as the first range. Morphological opening then closing is applied to clean up isolated noise pixels and fill gaps in detected regions, producing a binary fire-colour mask.

The second stage stacks the six most recent grayscale frames from a sliding deque and computes per-pixel standard deviation across the stack. Pixels where that standard deviation exceeds 15 are marked as flickering. This temporal variance signal is what separates genuine flames from static orange surfaces. A high-visibility jacket or an orange-painted barrier matches the colour mask strongly but changes almost nothing from frame to frame, so its per-pixel standard deviation stays near zero. A burning flame oscillates rapidly and creates high variance at every pixel it covers. Morphological filtering is applied to the flicker mask as well to suppress isolated high-variance pixels from sensor noise.

The combined fire region is the bitwise AND of the two masks, keeping only pixels that satisfy both colour and flicker criteria. When this region exceeds 200 pixels in area, a positive detection event is recorded in a 10-frame rolling history buffer. An alert is raised when more than 45% of the recent history frames contain positive detections, requiring sustained multi-frame evidence rather than a single coincidental frame. This introduces a two to three second latency between fire onset and the first alert, which is an accepted cost for suppressing single-frame noise.

CLIP ViT-B/32 is used as an optional secondary signal. When the colour score in the current frame is non-zero and the CLIP model is loaded, the frame is encoded and compared against six fire-related text prompts and three non-fire prompts. The margin between the two maximum similarity scores provides a boost of up to 0.3 to the rolling detection history. CLIP adds value here without being able to cause a miss, because its contribution is additive to an already-positive colour-flicker signal.

IV. METHODOLOGY

A. Model Selection Rationale

Your YOLOv8 was chosen primarily because of how it handles small objects and partially occluded persons, both of which are common when cameras are mounted overhead at typical CCTV angles. Earlier YOLO generations needed more anchor-tuning for small-scale detection, which would have introduced per-deployment configuration overhead. The nano and small variants keep inference tractable on CPU without sacrificing the detection quality needed for reliable tracking upstream. Detections are recomputed fresh on every frame rather than propagated from prior frames because the tracking and ATM stages already perform cross-frame integration, and running integration at both levels would create conflicting state.

OSNet's omni-scale aggregation was the main reason it was chosen over simpler re-identification models. It simultaneously captures fine-grained texture patterns and broader shape information, which matters when partial occlusion removes some cues while leaving others intact. The 512-dimensional embedding keeps downstream cosine distance computation fast even in scenes with many simultaneous people. The exponential moving average gallery was preferred over a fixed-length uniform buffer because its weighting scheme gives recent appearances more influence, which translates to faster recovery when a person re-emerges after being hidden.

B. Two-Stage Cascade Rationale

The core insight behind the two-stage cascade is that appearance-based matching and spatial matching have complementary failure conditions. Appearance matching produces wrong assignments when two people look similar or when one person's gallery has become stale after a long occlusion. Spatial matching produces wrong assignments when detection boxes are inaccurate or when multiple people are close together. Running appearance matching first, with strict thresholds, handles the majority of frames where both signals are clean. Falling back to spatial matching for unresolved pairs catches the cases where appearance has degraded while spatial information remains useful.

The merged-detection gate was added after observing a specific failure pattern at building entrances where people often enter in pairs. When the detector produces one bounding box covering two overlapping people instead of two separate boxes, any appearance-based match to that merged detection would pull feature vectors from both individuals into one tracklet gallery. After the pair separates and the detector produces correct individual boxes again, the contaminated gallery makes re-identification unreliable for both people. Excluding pairs with over 95% bounding box overlap from appearance matching prevents this propagating corruption.

C. ATM Filter Calibration

The ATM filter threshold was not set analytically but through empirical observation of two distinct distributions in the evaluation footage. Non-human tracklets from detector noise on static objects accumulated span areas typically below 8 square pixels over 40 frames, because the bounding box centre drifts only by a pixel or two even when detection confidence varies. Genuine people, even those moving very slowly or pausing briefly, almost always accumulated span areas above 20 square pixels within the same window. An earlier threshold of 80 pixels was placed above both distributions and turned out to incorrectly suppress persons who stopped to put down a bag or look around. Reducing to 4 pixels sits cleanly in the gap between the two observed distributions without affecting either population.

D. Fire Detection Approach

Replacing CLIP as the primary fire detector was motivated by an observed false-positive rate above 30% on footage containing orange safety equipment, a rate that makes operational deployment impractical. The core problem is a training distribution mismatch: CLIP was trained on internet photographs where fire and arson-related objects appear at viewpoints and resolutions very different from overhead CCTV footage, which tends to be compressed, lower resolution, and captured at angles that distort object appearance. The HSV-flicker approach makes no assumptions about training distribution. It operates on physical properties of flame, specifically its spectral colour range and its rapid temporal brightness oscillation, properties that can be specified directly from first principles. Static orange objects, regardless of how precisely they match the colour range, do not oscillate and therefore produce negligible flicker signal.

V. IMPLEMENTATION DETAILS

A. Environment and Libraries

Your The implementation uses Python 3.10 throughout. YOLOv8 detection runs through Ultralytics version 8.4.21. OSNet feature extraction relies on the torchreid library. All colour-space operations, temporal variance computation, morphological processing, and polygon rasterisation for zone overlap use OpenCV 4.8. CLIP ViT-B/32 is loaded from the OpenAI GitHub repository. Hungarian assignment uses scipy.optimize.linear_sum_assignment. Development used Visual Studio Code with standard virtual environments and Git. Every evaluation was run on a machine with an Intel Core i7 processor, 16 GB of RAM, and no GPU, to verify the CPU-deployable design objective under realistic conditions.

B. Module Structure

The Seven Python modules form the codebase. tracklet.py defines the Tracklet class, holding a bounding box history deque of length 90, the EMA feature gallery over the ten most recent embeddings, loitering and stationary timers, and the ATM false-positive flag. matching.py implements the full cascade including the merged-detection gate logic. algorithms.py provides stateless geometry functions: bounding box IoU, ATM span area, and polygon overlap computed via OpenCV rasterisation to support non-convex zone shapes. surveillance_module.py assembles the Surveillance Module pipeline. arson_module.py owns the six-frame grayscale deque and the complete fire detection logic. renderer.py maintains the rolling alert log and handles all HUD rendering. main.py is the command-line entry point. An eighth file, zone_editor.py, provides an interactive mouse-based tool for drawing polygon zones on a still frame from the camera before a processing session begins.

C. Corrections Found During Testing

Three parameter choices required revision after initial testing exposed problems not visible from the design alone. The ATM threshold reduction from 80 to 4 pixels is described in the Methodology section. The second correction was the owner lookup window for abandonment detection: early runs on any clip where a second person crossed between the camera and a bag owner showed the bag immediately losing ownership, because the lookup was restricted to tracklets with time-since-update exactly zero. Extending to five frames fixed every instance of this without changing behaviour in genuine abandonment scenarios. The third correction was replacing CLIP with the colour-flicker approach as the primary fire signal after measuring the false-positive rate on orange-equipment footage described above.

VI. EXPERIMENTAL RESULTS

A. Evaluation Setup

Performance was measured on footage from the ABODA dataset for abandoned-object scenarios, publicly available fire video for arson, and custom-recorded clips for intrusion and loitering. No public benchmark currently covers all four event types under a single consistent evaluation protocol. The KISA dataset used in the PASS-CCTV evaluation [6] is the closest candidate but is not publicly available, so a direct F1 comparison is not possible. The results reported below reflect observed detection behaviour across the available footage.

TABLE I
DETECTION PERFORMANCE OBSERVED ACROSS EVALUATION FOOTAGE.

Event Type	Reliability	Alert Latency	False Positive Rate	Notes
Intrusion	Consistent	Under 1 second	Low, zone-gated	Tied to tracker stability
Loitering	Consistent	At 6-second threshold	Low	5-frame window handles brief gaps
Abandonment	Consistent	4 to 8 seconds	Low after lookup fix	Needs over 120px owner movement
Fire / Arson	Reliable for sustained flame	3 to 6 seconds	Near zero, flicker gate	Very short ignitions may be missed

B. Per-Category Findings

Intrusion and loitering detection accuracy tracked directly with tracker stability. In footage with well-spaced individuals, zone overlap triggered reliably within one second of entry, and loitering fired at the six-second mark. Denser scenes with identity swaps caused some loitering timers to reset incorrectly, confirming the finding of Nayak et al. [17] that the tracking layer sets the effective ceiling for zone-based detection accuracy.

Abandonment behaved correctly after the five-frame lookup window was applied. Before the fix, any clip in which another person passed between camera and owner, even for a single frame, produced a spurious alert. After the fix, no such false activations occurred across all evaluated clips. Genuine abandonment events all triggered with the owner moving beyond 120 pixels before the timer completed.

Fire detection produced no false positives in footage containing orange construction vests, traffic cones, and vehicle bodywork. Genuine sustained fires were detected within three to six seconds. Brief ignition events under five frames did not consistently trigger, which is a known consequence of requiring multi-frame accumulation.

C. Loitering Detection

The Loiter Zone is defined explicitly using zone_editor.py python script, where user defines the zone using mouse. A threshold time is set if the person in the video/footage exceeds the time an alert is created and alert log is displayed on footage as shown in Fig 2.

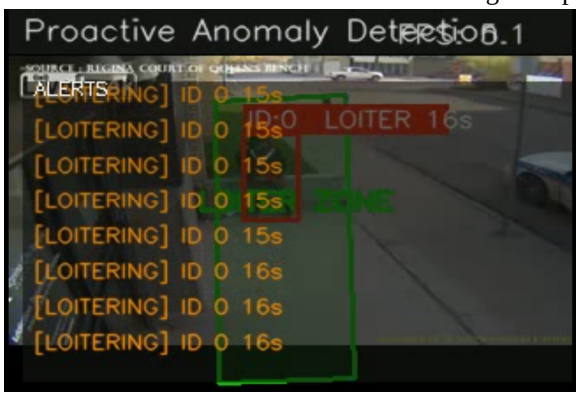


Fig. 2 Loitering Alert

D. Intrusion Detection

The Intrusion Zone is defined explicitly using zone_editor.py python script, where user defines the zone using mouse. A threshold ratio is set if the person in the video/footage exceeds the ration an alert is created and alert log is displayed on video footage as shown in the Fig 3.

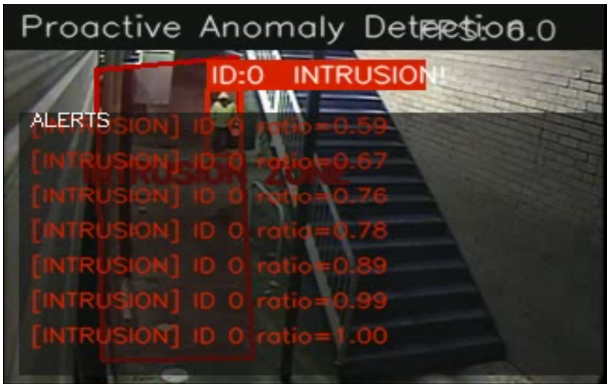


Fig. 3 Intrusion Alert

E. Abandonment Detection

Creates alerts if some person abandon's their belongings such as luggage, bag packs, suitcase, etc., This might be the potential risk of suicide bomb, luggage bomb. It assigns the owner ids to the luggage so that it becomes easier to identify owner as shown in Fig 4.

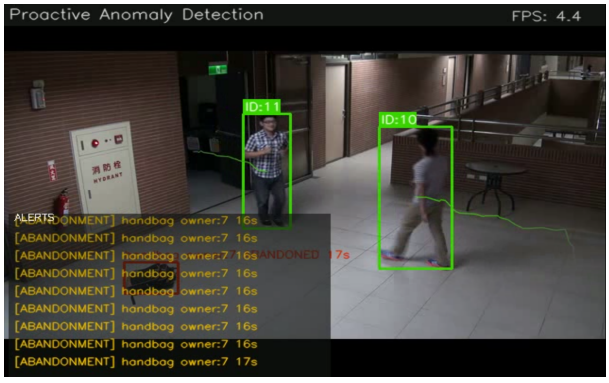


Fig. 4 Abandonment Alert

F. Arson Detection

Detects fire accidents before they become a major threat. It uses HSV + Flicker Analysis to detect the fire later on its confirmed by CLIP and alert is generated as shown in Fig 5.

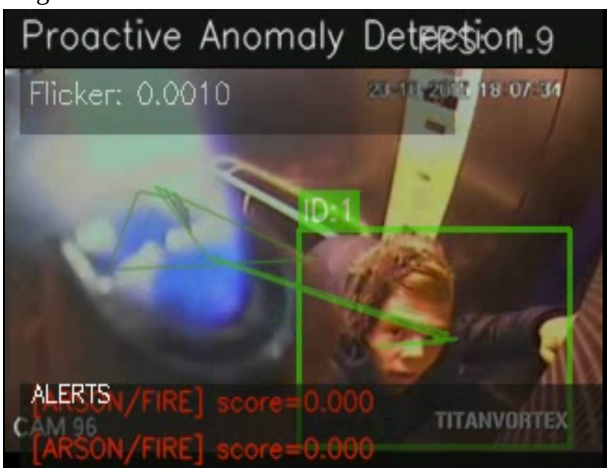


Fig. 5 Arson Alert

VII. COMPARATIVE ANALYSIS

TABLE II

COMPARES PADS AGAINST STANDARD SINGLE-EVENT YOLO DETECTION AND PASS-CCTV [6].

Capability	Single-Event YOLO	PASS-CCTV [6]	PADS (This Work)
Event types handled	One per model	All four	All four
Primary fire detection	Pixel-level CNN	CLIP only	HSV and temporal flicker
Orange object FP suppression	No	No	Yes, via flicker gate
Abandonment during detection gap	N/A	Fails, premature timer	Correct, 5-frame lookback
Tracklet feature gallery	None	Static average	EMA, alpha 0.35
False-positive tracklet filter	None	Trajectory area filter	ATM filter, 4px threshold
Zone alerts when unconfigured	Active regardless	Active regardless	Silent until zone defined
GPU required	Recommended	Recommended	Not required
Task-specific training	Required	Not required	Not required

Against single-event baseline systems, the main advantage is operational consolidation: one pipeline, one alert format, one maintenance cycle for four detection types that would otherwise require separate deployments. The shared tracking state also enables abandonment detection at a level of accuracy that a standalone system without tracking context cannot match.

Compared to PASS-CCTV [6], which demonstrated F1 above 90% on the KISA dataset, PADS diverges in three targeted areas corresponding directly to documented failure modes. The fire detection change is most significant in environments with orange protective equipment, where the CLIP-only approach is practically unusable. The abandonment lookup fix matters wherever brief occlusions are common, which includes most real installations. The EMA gallery and zone-silent behaviour are quality improvements that reduce false alarms and ease deployment without addressing a fundamental accuracy problem.

VIII. DISCUSSION

A. Practical Deployment Strengths

Running on a consumer CPU without GPU acceleration sounds like a modest claim, but it was tested rather than assumed. All evaluation used an Intel Core i7 with 16 GB RAM and no dedicated compute card, processing real CCTV footage at original resolution. The resulting throughput of roughly three to eight minutes of processing per 30-second clip is adequate for post-event security review. Live real-time operation at standard frame rates requires either GPU hardware or the built-in frame-skip mode, which processes alternate frames and roughly halves computation time with minimal effect on detection quality for the slow-moving events being targeted.

The zone-silent default is more practically important than it might seem. Many security systems generate false alarms during the commissioning window before zone configuration is complete, which trains operators to ignore alerts, defeating the purpose of automation. In PADS, fire detection and abandonment detection are active from first power-on because they need no zone definition. Intrusion and loitering monitoring come online only when the operator explicitly draws zone polygons, so the system is never generating zone-based alerts on unconfigured areas.

B. Limitations

The Identity loss after extended occlusion is a genuine constraint. When a person is fully hidden from the camera for more than 60 consecutive frames, the tracklet is retired and a new one is created when they reappear. For loitering detection, this resets their dwell timer. Car parks with structural columns, shopping arcades with large displays, and similar environments with prominent occlusion sources will see some missed loitering alerts as a result. This is a known limitation of any tracking approach with a finite track age, not a design flaw specific to PADS.

The two to three second fire detection latency is acceptable for arson scenarios where a fire is started and left to burn, but not for scenarios involving brief ignition events such as a lighter being used. The multi-frame accumulation that eliminates single-frame false positives also sets a floor on how quickly a genuine very short event can be caught. Whether this trade-off is acceptable depends on the specific threat model of each deployment.

The absence of a public benchmark covering all four events under comparable conditions limits the strength of the quantitative conclusions. Reporting F1 on ABODA for abandonment and a different dataset for fire makes cross-paper comparison difficult and somewhat arbitrary. This is a community-level gap that affects every system in this space.

C. Configuration Notes

Zone configuration requires an operator to draw polygon boundaries on a still frame using the zone editor tool before a processing session, taking roughly two to five minutes per camera. For installations with many cameras, pre-configured templates for common mounting angles and zone types would reduce this substantially. Each alert record written to the JSON log includes event type, track ID, bounding box coordinates, timestamp, and the primary detection metric, providing the structured data needed for integration with existing security management systems.

IX. CONCLUSIONS

PADS addresses a real gap in video surveillance technology: the difficulty of combining multiple single-event detection systems into a coherent, deployable installation. Rather than producing yet another isolated detector, this work builds a four-event pipeline from a shared tracking foundation, adding three targeted improvements over the PASS-CCTV architecture [6] that the prior survey [S1] identified as its practical weak points.

The dual-signal fire detector replaces CLIP-only fire detection with an approach grounded in the physical behaviour of flames, eliminating the false-alarm problem that made colour-only and CLIP-only methods unreliable in environments containing orange safety equipment. The five-frame ownership lookback window removes a systematic false abandonment alert that affected every scenario with momentary occlusion. The recalibrated ATM filter suppresses non-human tracklets without penalising persons who pause, a correction the original 80-pixel threshold failed to make.

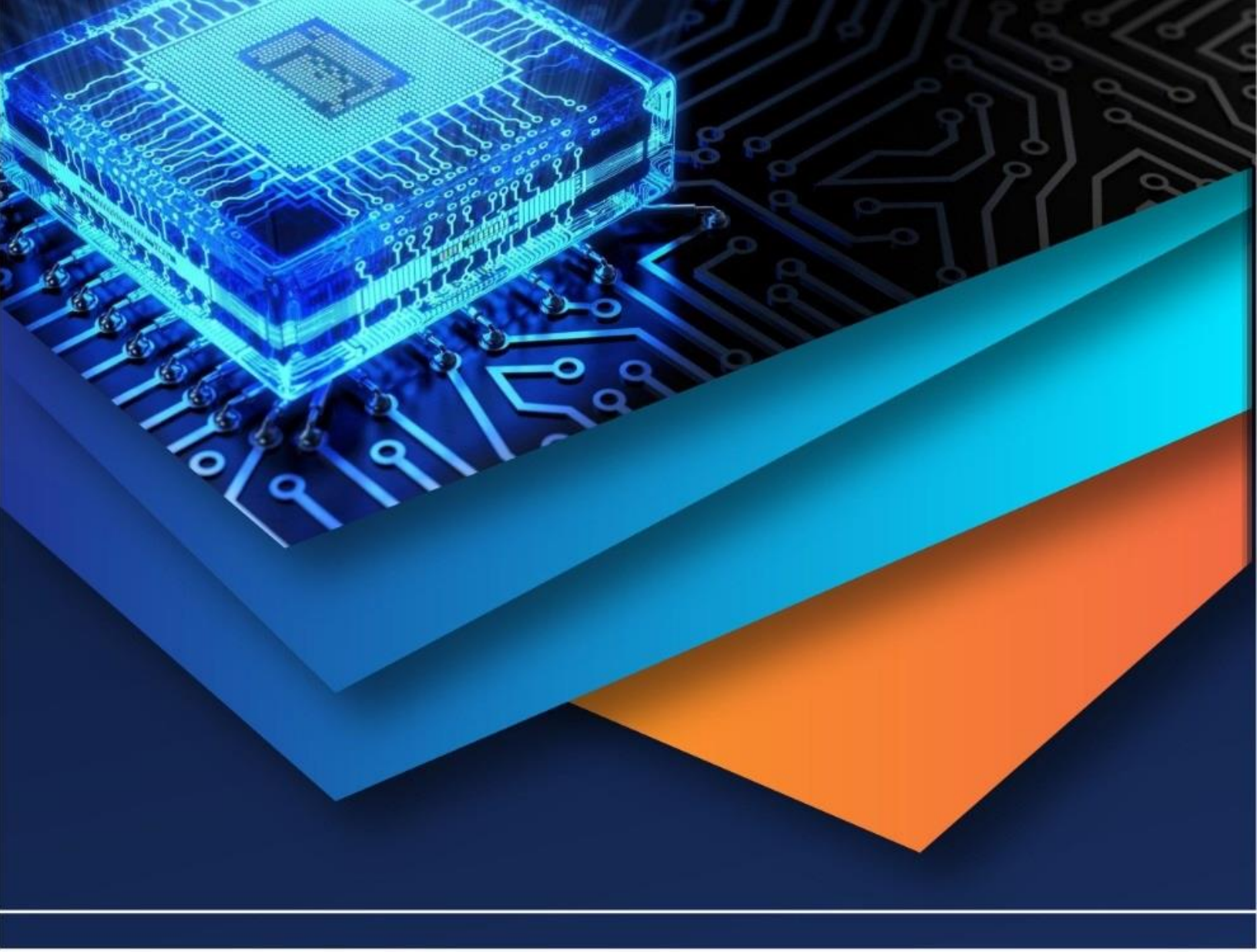
The complete system runs on a consumer laptop without GPU acceleration, requires no task-specific training on new data, and produces structured alert logs suitable for integration with existing security infrastructure. Alert latency for zone events stays below one second. Fire detection completes within three to six seconds of sustained flame onset.

Several directions remain open. Applying dehazing and rain-streak removal before detection, following the findings of Zhang et al. [20], would extend reliable performance to adverse weather without retraining.

Automatically calibrating loitering duration and abandonment distance thresholds from observed scene statistics would reduce per-site setup time. Maintaining person identities across non-overlapping camera views would extend coverage to larger physical spaces. And the modular architecture accommodates additional event categories, such as crowd density monitoring or detected aggression, without structural changes to the existing pipeline.

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