In: 2025 30th International Conference on Automation and Computing (ICAC), 27-29Aug 2025, Loughborough, UK
DOI:10.1109/icac65379.2025.11196302

Generalised Robotic Anomaly Detection in Dynamic Public Environments

Kelvin Benny
School of Aerospace, Transport, and Manufacturing
Cranfield University
Cranfield, UK
kelvinbenny@cranfield.ac.uk

Sharaf Mohanna
School of Aerospace, Transport, and Manufacturing
Cranfield University
Cranfield, UK
s.mohanna@cranfield.ac.uk

Emily K. Rose
School of Aerospace, Transport, and Manufacturing
Cranfield University
Cranfield, UK
emily.rose@cranfield.ac.uk

Dauood Tariq
School of Aerospace, Transport, and Manufacturing
Cranfield University
Cranfield, UK
dauood.tariq@cranfield.ac.uk

Gilbert Tang
School of Aerospace, Transport, and Manufacturing
Cranfield University
Cranfield, UK
g.tang@cranfield.ac.uk

Rebecca Raper
School of Aerospace, Transport, and Manufacturing
Cranfield University
Cranfield, UK
rebecca.raper@cranfield.ac.uk

***Abstract*—This paper presents a real-time anomaly detection system integrated into a ROS-enabled mobile robot for public safety monitoring in dynamic environments such as shopping centers. The system targets three critical anomalies: fallen individuals, abandoned bags, and visible knives. Our final approach combines YOLO-World v2 for object detection, YOLO-Pose for posture estimation, and GPT-4V for contextual reasoning. In controlled and public scenarios, the system achieved 88.3% accuracy and high precision (fall: 95.6%, knife: 91.7%), improving on the state of the art in fall detection recall (79.6% vs. ~49.7%). We detail the architecture, deployment strategy, and performance evaluation, and discuss latency, lighting sensitivity, and privacy challenges.**

Keywords—Anomaly Detection, Mobile Robotics, YOLO, Pose Estimation, Vision-Language Models

I. INTRODUCTION

Recent advancements in mobile robots have led to their increasing deployment in public spaces for logistics and service tasks, such as delivering medication in hospitals or guiding visitors around shopping centers. Expanding their role to include anomaly detection could further enhance their utility. This capability could be integrated into their existing workflows—either while performing their primary functions, during dedicated surveillance tasks, or opportunistically when no active jobs are queued. For example, in care home environments, the ability to detect anomalies such as patient falls could enable faster emergency responses without relying solely on human witnesses. Similarly, applying these techniques to existing CCTV infrastructure could enhance maintenance and safety monitoring in public spaces.

Anomaly detection, or unusual event detection, is key in video analytics for spotting abnormal activities in videos. This process faces challenges like object identification issues, occlusion, clutter, changes in lighting, and shape variations. To overcome these challenges, researchers have turned to machine learning and deep learning, with the latter showing notable effectiveness. Machine learning has significantly improved the ability to detect anomalies, particularly in identifying specific frames with unusual events. The primary objective of anomaly detection is to highlight moments in video that show unusual behaviour, which is extremely useful in various environments such as shopping centers, banks, airports, and hospitals, helping monitor and detect suspicious activities.

This project focuses on the integration of advanced anomaly detection into a ROS-enabled mobile robotic platform, enabling the detection of three key public safety threats: individuals who have fallen, abandoned bags, and visible knives. These anomalies were selected for their real-world relevance in public venues where crowd density, line-of-sight interruptions, and changing lighting conditions present unique challenges. Our approach combines YOLO-World v2 for object detection, YOLO-Pose for posture estimation, and GPT-4V for contextual reasoning, allowing the system to generate location-tagged, human-readable alerts in real time.

The final goal is to enhance the autonomous capabilities of mobile robots by enabling them to detect and respond to anomalies, either independently or in collaboration with human operators. This will support broader applications in public safety, logistics, and service robotics, particularly in environments where human supervision is limited or inconsistent.

II. RELATED WORK

Anomaly detection is utilized in areas varying from structured industrial environments to dynamic public spaces, with increasing applications for mobile robots. This section reviews key approaches across various domains and their relevance to our work.

Industrial anomaly detection (IAD) typically employs convolutional neural networks (CNNs) for feature estimation or mapping [1], with recent approaches integrating large language models to describe anomalies [2]. While effective in controlled settings, these methods require adaptation for dynamic environments. Crowd anomaly detection presents different challenges, with recent reviews [3] finding CNNs particularly effective for identifying unusual behaviours in

Published by IEEE. This is the Author Accepted Manuscript issued with: Creative Commons Attribution License (CC:BY 4.0).
The final published version (version of record) is available online at DOI:10.1109/icac65379.2025.11196302. Please refer to any applicable publisher terms of use.

public spaces. Vision-language models have shown promise for detecting suspicious activities [4], with approaches ranging from specific anomaly detection to identifying deviations from expected pedestrian behaviour.

Fall detection systems have advanced beyond traditional posture recognition through integrated approaches combining depth sensing with deep learning [5] and standalone camera-based solutions using YOLOv8 [6]. Unmanned aerial vehicle (UAV)-based surveillance has leveraged various techniques including one-class support vector machines with Haralick textural features, frameworks incorporating temporal correlations, two-branch GANs, and unsupervised generative learning [7]. For item detection, AI-driven security screening combines deep learning with blockchain technology, utilizing models like Faster R-CNN and YOLO [8]. Shoplifting detection has benefited from AI and computer vision to track movement patterns and identify suspicious behaviours before crimes occur [9].

Image processing techniques address challenges like occlusion through methods such as weighted special transforms. CNNs remain fundamental to anomaly detection, with transfer learning approaches adapting pre-trained models like VGGNet and ResNet through various methods. Some approaches use denoising autoencoders for efficiency, while adaptive frameworks based on teacher-student knowledge distillation show promise for mobile applications. Unsupervised and self-supervised models [10] offer alternatives to labor-intensive labelling.

YOLO-based object detection provides real-time capabilities crucial for dynamic environments, with recent optimizations incorporating EfficientRep backbones [11] and integration with temporal action classification through MobileLSTM. Large language models (LLMs) offer novel approaches by transforming visual data into textual descriptions for anomaly analysis [2,4], with systems like AnomalyGPT demonstrating effectiveness in industrial settings.

VisionGPT [12] represents an integrated approach combining object detection, image processing, and LLM-based interpretation. This system processes camera feeds through YOLO, filters and structures the data, and leverages LLMs to deliver context-aware analyses of potential anomalies. The integration of Open-Vocabulary Detection with LLMs enables flexible, interpretable, and dynamic anomaly detection suitable for complex environments like shopping centres.

TABLE 1. BENCHMARK VALIDATION CRITERIA FROM LITERATURE

Anomaly	Fall Detection [13]	Bag Detection [14]	Knife Detection [15]	Benchmark
Precision	~ 80.1%	~ 85.7%	~ 78.7%	>80%
Recall	~ 49.7%	~ 75%	~ 71.8%	>60%
Accuracy	~ 93.0%	~ 91.1%	~ 85.4%	>85%
Latency	~2-3s	~2-3s	~2s	< 3s

The system will focus on detecting specific anomalies that are most relevant to the application, with a particular emphasis on falls, abandoned bags, and knives, as these present differing chances of success but will have significant impact in a range of environments. Benchmark validation criteria were produced by finding the performance of the current best-in-class algorithms for detection of each anomaly [13,14,15] (Table 1).

III. SYSTEM OVERVIEW

Initially, the system relied on a text-based reasoning framework. The robot's camera captured visual input, which was processed by object detection and captioning models to generate textual descriptions of the scene. These descriptions were then passed to large language models, such as GPT-3.5, for reasoning and anomaly classification. One early implementation, known as VisionGPT, used YOLO-World v2 as an Open Vocabulary Detector (OVD). This allowed the system to detect objects using natural language prompts—for instance, identifying terms like "abandoned bag" or "knife" without retraining the model. The detected object labels were sent to GPT-3.5 to evaluate whether an anomaly was present.

However, a key limitation of this approach was the lack of contextual awareness—particularly the inability to distinguish between a regular bag and an abandoned one. The model could not assess human behaviours or spatial relationships. To address this, the next iteration integrated YOLO-Pose and BLIP. YOLO-Pose provided posture estimation to infer human activity, while BLIP, a vision transformer, generated image captions that added contextual understanding. This integration improved behavioural interpretation but introduced new issues: BLIP was computationally intensive and frequently inaccurate in its descriptions, reducing the system's reliability.

At this point, the architecture still depended on converting visual data into text for the LLM to reason over. A paradigm shift was introduced by directly passing entire image frames to a Vision-Language Model (VLM), enabling it to visually analyse the environment rather than rely solely on pre-labelled text (Table 2). Various VLMs were evaluated, and GPT-4V was ultimately selected due to its superior description quality and faster response times.

The final pipeline combines YOLO-World v2 for detecting objects beyond traditional training sets, YOLO-Pose for inferring human posture and behaviour, and GPT-4V to interpret the entire visual frame and classify anomalies with contextual reasoning (Fig. 1).

TABLE 2. COMPARISON OF DEVELOPED ALGORITHMS

Methods	Strength	Weakness
OVD + GPT-3.5	Fast, simple reasoning	No spatial context
OVD + BLIP-2 + GPT-3.5	Adds visual context	Slow due to captioning, also not very accurate
OVD + YOLO Pose + GPT-3.5	Detects posture too	Still label-based, not visual
OVD + Vicuna/ LLaVA	Fully visual, image input	Slower, LLM latency
OVD + YOLO Pose+ GPT4V	Image input and pose	Less accurate than Vicuna and LLava

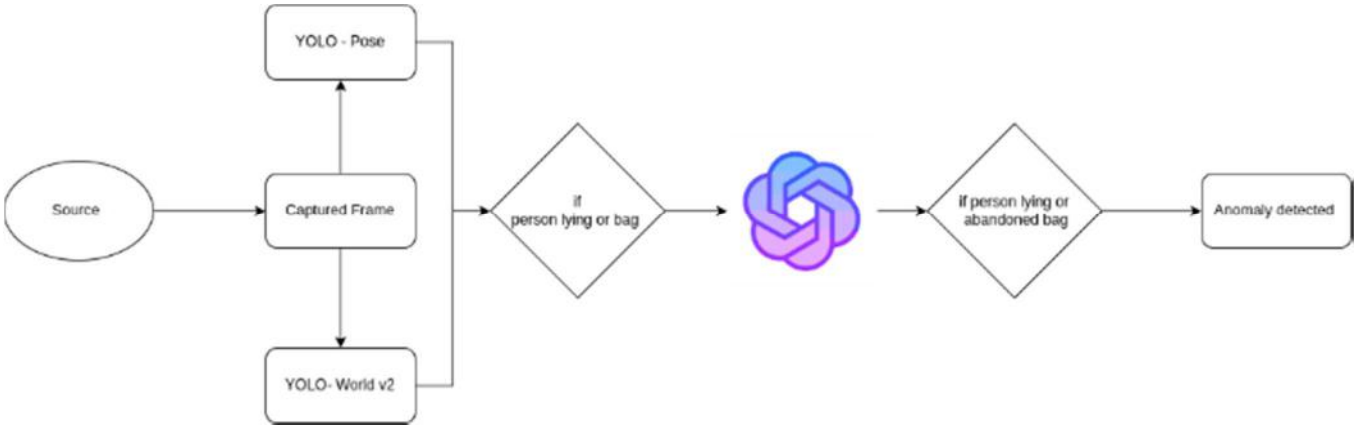


Fig. 1. Final algorithm architecture

Importantly, the pipeline is highly scalable. Since it does not rely on task-specific training, it can be extended to detect new types of anomalies by modifying prompt structures and updating the logic that flags which frames or objects are sent to the vision-language model for classification. This modularity makes it well-suited for deployment in varied and evolving environments.

IV. Methodology

The system testing was conducted in two stages, in a controlled laboratory environment and in a dynamic public environment.

A. Controlled Testing

Tests used simulated anomalies in a 4 m x 5 m laboratory environment. The robot followed a pre-defined route around the laboratory while the anomaly detection system ran on an external PC. The robot completed more than 20 circuits, each time being exposed to one or more anomaly: at least three of each anomaly type were tested. In addition, the robot ran three times with no anomalies present. Finally, the effect of lighting was tested by varying the brightness in the room while both the robot and the anomalies remained stationary.

B. Dynamic Pubic Environment

The first floor of a university building was used to demonstrate a dynamic public environment with open spaces, obstacles, and people going about their business. Again, the robot moved around the space while being exposed to simulated anomalies. This time, all three anomaly types were present while the robot patrolled the area. This was tested three times with anomalies in different places, and three times with no anomalies present.

V. Results

Every frame captured from each scenario was analysed and labelled as true positive, true negative, false positive, or false negative, depending on whether the frame contained an anomaly, and if the algorithm detected it.

A. Negative Control

The algorithm performed well, with only 2.4% of frames indicating a false positive (Table 3). Furthermore, the program was able to distinguish attended and abandoned bags, ensuring alerts are not raised when not required (Fig. 2).

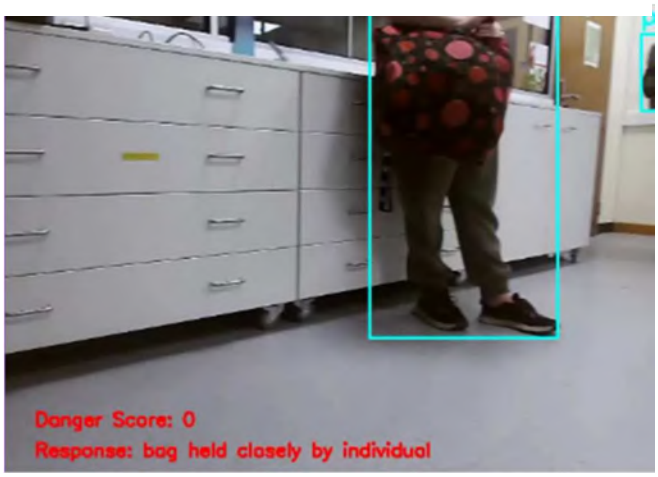


Fig. 2. A correctly identified attended bag

Table 3. Confusion matrix for negative controls

		Detected			
		Abandoned bag	***Lying person***	***Knife***	***No anomaly***
Real	***No anomaly***	3	1	0	165

B. Abandoned bags

Each of four different abandoned bags were detected at least once on the robot's patrol, meaning an alarm was raised. However, not every shot including an abandoned bag was recognised as such. The program had particular trouble with a white tote bag (Fig. 3), achieving only 50% accuracy. The remaining three bags were more successful (Table 4). The VLM was logical in regards to proximity of a person to a bag, as shown in Fig. 4, where both a bag and a person are identified but sufficiently far apart to be suspicious.



Fig. 3. The white bag was often not detected

TABLE 4. ABILITY TO DETECT ABANDONED BAGS

	Accuracy	Recall	False Positive Rate	Precision
Abandoned bag	0.796	0.263	0.047	0.625
Abandoned bags minus problematic white bag	0.896	0.500	0.037	0.692

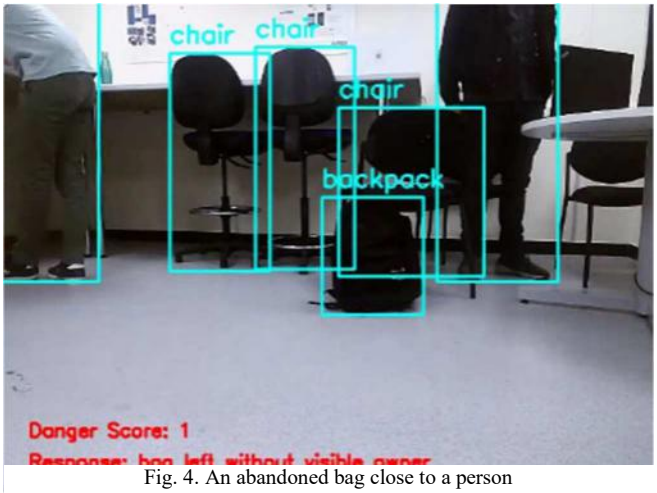


Fig. 4. An abandoned bag close to a person

C. *Fallen person*

Fallen persons were detected by the algorithm at least once per robot patrol, but not consistently in every frame. Recall dropped when feet faced the camera due to occlusion and body foreshortening. Dark clothing also reduced detection accuracy. Including these disadvantages, the algorithm achieved 79.6% recall (Table 6).

D. *Knife*

Knife detection showed high precision, meaning detections were almost always correct (Table 6). This reduces false alarms, avoiding unnecessary panic or wasted emergency resources, and makes the system reliable for real-world threat detection.

E. *Multiple Anomalies*

The system struggled to detect multiple anomalies at once. In tests, it often detected only one person lying down when two were present and prioritized fallen people over unattended bags. GPT sometimes ignored bags in this situation, possibly assuming they belonged to the fallen person. Although the system is designed to be human-in-the-loop, it should ideally alert operators to all dangers.

F. *Lighting*

The system sometimes generated false positives in low-light conditions by hallucinating anomalies. Interestingly, recall for abandoned bags and knives was higher in these tests than under consistent lighting (Table 5). This may be because both the robot and anomalies stayed still during lighting changes, suggesting that scene stability improves recall and accuracy.

G. *Controlled Enviroment*

Table 6 and Fig. 5 show the system has a low false positive rate, making its alerts trustworthy. However, recall is generally low (except for falls), meaning many anomalies are missed. Still, detecting an anomaly once is enough to trigger an alert, so consistent single detections remain crucial.

H. *Dynamic Public Environment*

The system showed high accuracy and a low false positive rate in dynamic environments, making its alerts reliable (Table 7). However, the low recall means many real anomalies were missed, especially all 67 frames including knives. This highlights a major issue with false negatives, risking critical situations being overlooked.

I. *Comparison Controlled environment with busy environment*

The anomaly detection system maintains high accuracy in both controlled (88.3%) and dynamic (86.8%) environments, with consistently low false positive rates (2.3%). However, recall drops sharply in dynamic settings (28.4%), especially for knife detection, which had no successes. Controlled environments show better recall and precision, highlighting the impact of environmental complexity. Fall detection is strong in controlled tests but weaker in public spaces.

TABLE 5. RESULTS IN VARYING LIGHTING

Accuracy	Recall	False positive rate	Precision
0.742	0.574	0.061	0.917

TABLE 6. RESULTS OF DETECTION ALGORITHM IN A CONTROLLED ENVIRONMENT

	Accuracy	Recall	False Positive Rate	Precision
Abandoned bag	0.796	0.263	0.047	0.625
Fall	0.917	0.796	0.019	0.956
Knife	0.846	0.282	0.007	0.917
All	0.883	0.489	0.024	0.831

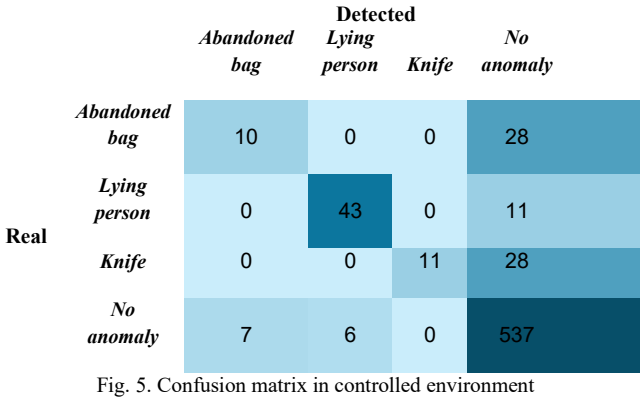


Fig. 5. Confusion matrix in controlled environment

TABLE 7. RESULTS IN A DYNAMIC PUBLIC ENVIRONMENT

Accuracy	Recall	False positive rate	Precision
0.868	0.284	0.023	0.694

Despite low recall, the system's reliable alerts and ability to trigger on a single detection make it practically useful, though improvements are needed for better performance in dynamic environments.

VI. DISCUSSION

The system demonstrates precision in identifying safety-critical anomalies, particularly for fall detection where it outperformed current benchmarks with 95.6% precision and 79.6% recall—surpassing typical recall values of ~49.7% (Table 1). These improvements validate the effectiveness of combining pose estimation and vision-language reasoning.

However, recall remains a challenge in bag and knife detection, especially in scenarios involving partial occlusion, poor contrast, or atypical object presentation. For example, a white tote bag yielded only 50% detection accuracy, illustrating the limitations of current models when visual distinctiveness is low. Small or partially visible knives were similarly difficult to detect, with recall hovering around 28.2%.

Due to the rarity of anomalies, the dataset was mostly comprised of negative examples. In lab conditions, only 77 out of 681 frames contained an anomaly. This class imbalance inflates accuracy and highlights the importance of high precision metrics, which the system achieved. Future testing should focus on collecting more positive cases to improve recall evaluation.

Latency (~4 seconds per detection) limits the system's responsiveness to fast-moving threats. Integrating the algorithm onboard (e.g., Jetson Orin Nano) or using lighter LLMs could mitigate this. Additionally, implementing multi-frame aggregation and confidence-based filtering would help improve recall without sacrificing reliability.

From a systems perspective, the reliance on Wi-Fi connectivity between robot and external computer highlights the need for robust network infrastructure or autonomous edge processing. Real-world deployment would benefit from better camera positioning and hardware integration (e.g., depth sensors) to address detection failures due to angle or object height.

Multi-frame aggregation, where an anomaly must persist over consecutive frames before triggering an alert, could reduce false positives and improve detection robustness. Adaptive confidence thresholds—potentially modified based on time of day or observed crowd density—could help refine system sensitivity in real time.

Ethical and privacy considerations are fundamental for public deployment. Though the robot is designed to enhance safety, it captures video of individuals, invoking GDPR and related legislation. Clear signage, privacy policies, and strict access controls must accompany deployment. Raw footage should be anonymized or deleted within a limited retention period unless flagged for review. Operator transparency and accountability must be maintained through detailed system logs and access tracking.

Scalability also presents challenges. The current anomalies—falls, bags, and knives—are diverse but not exhaustive. The VLM prompt assumes a general public setting, but future iterations could benefit from specifying environments more precisely, such as a shopping center. This would help tailor detection to context-specific risks and constraints. Future iterations should expand detection to include examples like liquid spills, smoke, or dynamic human behaviours like fighting. YOLO-World's open vocabulary detection, combined with GPT-4V's reasoning, allows for extensibility. However, more abstract behaviours may require temporal pattern analysis over multiple frames.

Human-Robot Interaction (HRI) is another area for refinement. Currently, the system sends multiple alerts for repeated detections of the same anomaly. This redundancy could cause alert fatigue. De-duplication logic, combined with a unified dashboard for operators, would streamline response and improve usability.

In summary, while the proposed system demonstrates strong potential—particularly in modularity, fall detection, and contextual awareness—it still requires development in areas such as recall for less visible anomalies, latency, infrastructure independence, and ethical readiness to be fully suited for real-world deployment in public safety applications.

VII. CONCLUSION

This work demonstrates a practical, modular anomaly detection system for mobile robots using YOLO, pose estimation, and vision-language reasoning. The system improves fall detection recall over previous methods and shows strong potential for real-world deployment. Ongoing work will enhance recall, reduce latency, and expand the types of detectable anomalies while addressing ethical concerns related to privacy and consent.

ACKNOWLEDGMENT

We would like to thank Smart City Robotics and Swift Robotics for their collaboration, advice, and equipment.

REFERENCES

[1] Wan Q, Gao L, Li X, Wen L. Unsupervised Image Anomaly Detection and Segmentation Based on Pretrained Feature Mapping. IEEE Trans Industr Inform. 2023 Mar 1;19(3):2330–9.

[2] Gu Z, Zhu B, Zhu G, Chen Y, Tang M, Wang J. AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models [Internet]. 2024. Available from: www.aaai.org

[3] Sharif MH, Jiao L, Omlin CW. Deep crowd anomaly detection: state-of-the-art, challenges, and future research directions. Artif Intell Rev. 2025 May 1;58(5)

[4] Zanella L, Menapace W, Mancini M, Wang Y, Ricci E. Harnessing Large Language Models for Training-free Video Anomaly Detection. 2024 Apr 1; Available from: http://arxiv.org/abs/2404.01014

[5] Li Y, Zhang W, Ma H, Chen F, Sun D, Cheng X, et al. Posture Detection Based on Kinect and YOLO Technologies. In: 2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI). IEEE; 2023. p. 205–9

[6] Nguyen DL, Vo XT, Priadana A, Nguyen DV, Nguyen TLH, Jo KH. Simple Human Fall Surveillance System Based on Person Detection. In: 2024 International Workshop on Intelligent Systems (IWIS). IEEE; 2024. p. 1–6.

[7] Hamdi S, Bouindour S, Snoussi H, Wang T, Abid M. End-to-end deep one-class learning for anomaly detection in UAV video stream. J Imaging. 2021 May 1;7(5).

[8] Chen Y, Mao G, Yang X, Du M, Song H. Research on airport baggage anomaly retention detection technology based on machine vision, edge computing, and blockchain. IET Blockchain. 2024 Dec 15;4(4):393–406.

[9] Kilic S, Tuceryan M. Crime Detection from Pre-crime Video Analysis with Augmented Pose Information. In: 2023 IEEE International Conference on Electro Information Technology (eIT). IEEE; 2023. p. 1–6.

[10] Jain Y, Dabouei A, Xu M. Cross-Domain Learning for Video Anomaly Detection with Limited Supervision. 2024 Aug 9; Available from: http://arxiv.org/abs/2408.05191

[11] Shaik A, Basha SM. Leveraging YOLOv5s with optimizationbased effective anomaly detection in pedestrian walkways. Expert Syst. 2024 Feb 1;

[12] Wang H, Qin J, Bastola A, Chen X, Suchanek J, Gong Z, et al. VisionGPT: LLM-Assisted Real-Time Anomaly Detection for Safe Visual Navigation. 2024 Mar 18; Available from: http://arxiv.org/abs/2403.12415

[13] Broadley RW, Klenk J, Thies SB, Kenney LPJ, Granat MH. Methods for the Real-World Evaluation of Fall Detection Technology: A Scoping Review. Sensors(Basel). 2018 Jun 27;18(7):2060. doi: 10.3390/s18072060. PMID: 29954155; PMCID: PMC6068511.

[14] Zhou L, Xu J. Enhanced Abandoned Object Detection through Adaptive Dual-Background Modeling and SAO-YOLO Integration. Sensors (Basel). 2024 Oct12;24(20):6572. doi: 10.3390/s24206572. PMID: 39460053; PMCID: PMC11510867.

[15] Moran, Maira & Conci, Aura & Sánchez, Ángel. (2022). Automatic Detection of Knives in Complex Scenes. 10.1007/978-3-031-06307-7_4.



https://dspace.org/

Generalised robotic anomaly detection in dynamic public environments

Benny, Kelvin; Mohanna, Sharaf; Rose, Emily K.; Tariq, Dauood; Tang, Gilbert ...

2025-08-27
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/icac65379.2025.11196302
Benny K, Mohanna S, Rose EK, et al., (2025) Generalised robotic anomaly detection in dynamic public environments. In: 2025 30th International Conference on Automation and Computing (ICAC), 27-29 August 2025, Loughborough, UK

Downloaded from a DSpace repository. https://dspace.org/