Toward Plug-and-Play Asset Management: Spot-Based Visual Surveying in Untagged Industrial Spaces

Danial Rizvi, *Member, IEEE* Gavin Paul, *Member, IEEE*
Mariadas Capsran Roshan, *Member, IEEE* Amal Gunatilake, *Member, IEEE*

***Abstract*—The prevailing paradigm in industrial asset management couples radio-frequency identification (RFID) tags with dense Internet of Things (IoT) scanner grids to maintain real-time inventories. Such infrastructure demands substantial capital investment, which many small- and medium-sized manufacturers cannot justify. It is also unsuitable for assets whose geometry or material properties hinder reliable tagging, such as metal stock, sheet-metal stacks, or bulk items that reflect or shield RFID signals. To address this gap, we investigate a *device-free localisation* framework that dispenses with per-item tags by delegating the search and identification tasks to a mobile robot. The proposed pipeline integrates stereo and RGB-D vision sensors on a Boston Dynamics *Spot* quadruped, employs a vision–language model to identify assets, and synchronises detections with a Snipe-IT asset-management dashboard via a lightweight REST interface. Field trials achieved 64 % floor coverage with less than 13 % false identifications at an inference cost below USD 0.20 per inspection, demonstrating the practicality of the approach. The system generated structured and verifiable asset records without human intervention, highlighting the viability of robot-assisted, vision-based localisation as a plug-and-play alternative to infrastructure-heavy IoT or RFID solutions.**

I. INTRODUCTION

Routine misplacement of assets represents a significant operational cost in manufacturing. Saxena et al. [1] attribute discrepancies between recorded and on-hand inventory to aggressive supply chain strategies that surprisingly led to increases in holding costs for misplaced or unaccounted-for items. This highlighted the need for more agile asset-management practices. Recent research has explored modernising these processes through Digital Twin (DT) technology and other IoT-based approaches. In a systematic review [2], Alhadi et al. trace the historical development of DTs and conclude that they can integrate seamlessly into Industry 4.0 asset-management workflows by leveraging advancements in artificial intelligence (AI), the Internet of Things (IoT), augmented reality (AR), and geographic information systems (GIS). However, the associated infrastructure and implementation costs can be prohibitive for small- and medium-sized enterprises (SMEs). These organisations face many of the same challenges as larger manufacturers but lack the economies of scale and capital reserves to address them effectively, including the expense of installing and maintaining IoT tracking infrastructure, difficulty in tagging irregular or high-turnover items, and limited personnel capacity for regular audits. In [3], adoption costs and human resources were identified as significant barriers for SMEs pursuing Industry 4.0 implementation. Consequently, SMEs often operate with incomplete or outdated inventory records, leading to delays, misplaced tools or components, and unplanned downtime.

In most industrial contexts, asset-management software serves as the primary tool for mitigating inefficiencies, maintaining a centralised register of assets with associated metadata such as location, condition, and maintenance schedules. These platforms can integrate with barcode or RFID scanning to update records, automate replenishment through purchasing systems, and, at higher sophistication, host a DT for simulation, predictive maintenance, and spatial analytics [4]. This implementation is commonly termed Real-Time Location System (RTLS). However, their effectiveness depends on timely and accurate data entry—an area where SMEs often struggle due to limited automation in data capture [3]. The resulting lag between the physical and digital factory states compromises operational efficiency, underscoring the need for automated, real-time asset-localisation methods that minimise human intervention and reduce infrastructure costs.

Effective asset tracking in factories often depends on RTLS, which integrates asset tagging devices such as RFID readers, ultra-wideband (UWB) beacons, or Bluetooth Low Energy (BLE) nodes with software to monitor tagged items continuously. These systems provide constant visibility over tools, components, and work-in-progress items, enabling process optimisation, root-cause analysis, and bottleneck identification, while reducing search times, delays, and redundant purchases [5], [6]. However, where RTLS coverage is incomplete, asset visibility gaps are typically addressed through ad hoc searches and manual record updates by personnel [7]. This reliance on human intervention increases labour costs and diverts skilled personnel from higher-value activities, highlighting the need for more flexible, infrastructure-light asset-tracking solutions.

The costs of asset tagging and work-hours pose a unique challenge, which encourages exploration into whether it is possible to localise assets using "Device free localisation" (DFL) approaches. In [8], a comprehensive study to broaden the taxonomy of DFL approaches was conducted. This work highlights the necessity that asset localisation plays in industrial environments and delineates between active and passive localisation approaches based on whether or not the asset

Danial Rizvi, Amal Gunatilake and Gavin Paul are with the Robotics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Australia. Mariadas Capsran Roshan is with Swinburne University of Technology, Australia. Corresponding author: {Danial.Rizvi@student.uts.edu.au}

This research is supported by the UTS:RI. The authors would like to acknowledge the support received through the ARC Industrial Transformation Training Centre (ITTC) for Collaborative Robotics in Advanced Manufacturing under grant IC200100001.

itself "carries a tag or attached device." Passive localisation, corresponding to DFL, primarily relies on radio and radar-based sensing to detect assets, differentiating their presence using machine learning or statistical approaches [8], [9]. In manufacturing, DFL has been demonstrated using RF-based sensing networks to track the movement of high-value tools and mobile equipment within defined "areas of interest" (AoIs) on the shop floor[5]. These systems continuously monitor radio wave propagation characteristics to detect the presence and movement of objects without requiring them to carry tags.

For asset management, AoI-based RF sensing can provide coarse-grained localisation, sufficient for determining whether an asset is in a designated storage zone, in use on a particular production line, or misplaced in a restricted area. However, while this level of spatial resolution is valuable for high-level inventory control, it is often insufficient for tasks requiring precise asset location, such as rapid retrieval or automated handling. Vision-based methods, such as RGB-D or stereo imaging, differ from RF-based approaches by providing fine-grained localisation down to individual grid cells in an occupancy map, enabling reliable geotagging and visual confirmation of assets.

Vision-based DFL systems fit a unique niche by leveraging a host of public training sets and tools, and are also low-cost. In Naggar et al. [10], low-cost cameras and open source software such as OpenCV was demonstrated to achieve localisation accuracy within 1 cm as well as the coordination of multiple robots to in formation. They effectively benchmarked the capability of indoor positioning systems in a very controlled environment. This approach, however, lacks generalisability in large-scale dynamic environments because it focused on localisation in a fixed reference frame. Furthermore, Publicly available datasets, such as MS-COCO and ImageNet, combined with a highly competitive ecosystem of machine learning models, have made it feasible to train and deploy sophisticated visual recognition pipelines without prohibitive cost or bespoke data collection [11]. These advances in optical sensing and computer vision have dramatically expanded the capabilities of automated perception systems in industrial contexts. This creates a timely opportunity to re-evaluate how asset tracking can be implemented in factories, particularly in environments where physical tagging is impractical.

Bouman et al. [12] demonstrated that autonomous platforms can navigate and map extreme or hazardous environments without human intervention, providing a strong foundation for industrial applications. When integrated with advances such as scene and semantic recognition as presented in [13] and [14], these systems can construct detailed 3D models from RGB-D input, segment and identify 2D objects within them, and infer semantic context about their role or state in the environment. In a factory setting, these capabilities enable autonomous systems not only to operate robustly in the presence of noise, variable lighting, and cluttered backgrounds but also to exploit the environmental structure to improve localisation and classification. For example, consistent object arrangements, fixed workstations, and recurring asset layouts

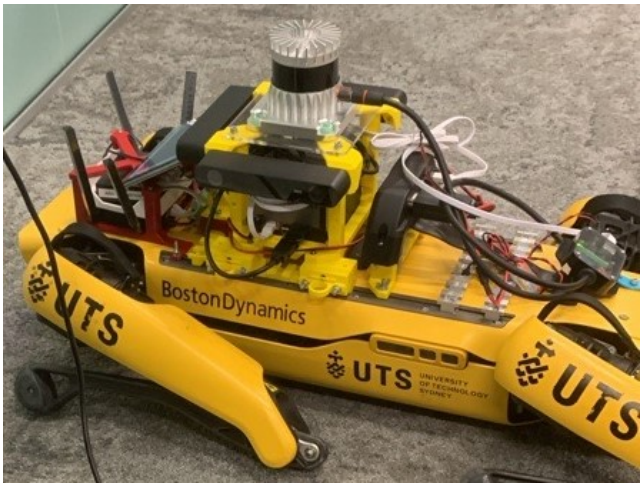

Fig. 1. Overview of the UTS:RI Boston Dynamics *Spot* [15] with an additional sensor array and compute unit.

can be leveraged as priors. Semantic segmentation of point-cloud datasets suck as KITTI [14] would constrain the set of plausible labels for an object, increasing classification confidence and reducing false positives in dynamic production environments.

The improved energy efficiency and increased payload capacity of modern autonomous mobile robots make them well-suited to carrying advanced sensing and processing systems for industrial tasks. Many platforms also provide turn-key solutions for autonomous navigation, including tools for planning and configuring repeatable routes [15], [16]. These capabilities enable the robot to follow predefined paths through a facility and perform data collection at key waypoints. In this study, we employ Boston Dynamics' *Spot* as the autonomous platform, selected for its payload capacity, mobility, and compatibility with custom sensing and software integration, as shown in Figure 1.

One of the most significant advances in processing raw visual data has been the introduction of transformer architectures, which enable highly parallelised computation and superior performance in tasks requiring long-range dependency modelling [17]. Vision transformers adapt this architecture for image analysis by dividing an image into patches, embedding them as tokens, and processing them through self-attention layers to capture both local and global context [18]. This approach has demonstrated competitive or superior accuracy to convolutional neural networks (CNNs) on a range of object detection, classification, and segmentation benchmarks.

Given the relative ease with which mobile autonomous systems can collect high-resolution camera data during navigation, transformers provide a compelling means of converting this unstructured input into actionable asset-management information. For example, raw RGB-D frames acquired by a robot such as *Spot* can be processed through a transformer-based model to segment the scene, classify identified assets, and associate each detection with a geospatial coordinate.

These structured outputs can then be ingested by an asset-management platform, where they update inventory records, flag anomalies, and inform retrieval or maintenance actions.

The primary contribution of this paper is the development of a transformer-based approach for camera-driven industrial asset tracking that overcomes key challenges in vision-based inventory systems. Specifically, it enhances classification robustness in cluttered and visually complex environments, mitigates false positives caused by similar object appearances, and enables seamless integration of vision-based detections with existing asset-management platforms, thereby providing a practical and low-infrastructure alternative to traditional RFID–IoT solutions.

This paper is organised into two main sections following the introduction. The methodology describes the approach used to implement the proposed asset-tracking system, beginning with an overview of Spot's kinematics and operational advantages in industrial environments, followed by the configuration of a fixed-path inspection routine. It then outlines the optimisation strategies employed for frame selection to balance computational efficiency and detection accuracy, and details the setup of the specific use case examined in this study. The results section presents the findings from field trials, including the generated map and the integration of detection outputs into the asset-management dashboard. It also discusses the actionable insights derived from the data, before reporting on inspection time, response time, and associated costs.

II. Methodology

The methodology for this study encompasses three principal components: the implementation of the asset-tracking system, the configuration of the asset-management software, and the implementation of the use-case study.

Spot has been deployed in a wide variety of industrial environments, demonstrating its versatility in navigating complex spaces due to its quadrupedal structure [12]. This mobility, combined with its ability to carry diverse payloads, enables the collection and transfer of sensor data to support the development of larger and more sophisticated systems. The Robotics Operating System (ROS) serves as the middleware that enhances *Spot*'s functionality and sensor integration. By maintaining a multi-threaded interface between sensors and processes, ROS allows onboard hardware to *publish* collected data. It enables other programs to *subscribe* to it in real time. This architecture facilitates efficient communication between independent hardware and software components, effectively allowing the integrated system to operate as more than the sum of its parts.

An outcome of this implementation is that *Spot* develops a situational awareness of its environment through the analysis of collected frames via the OpenAI API. Results from [14] indicate that, when provided with sufficient contextual information, *Spot* can generate observations that contribute valuable insights into the state and configuration of the factory floor as visualised in Figure 2. The following sub-sections describe the key elements of this perception pipeline: (i) the prompt-engineering strategy used to maximise the relevance and accuracy of the model outputs, and (ii) the rationale for selecting the specific transformer architecture employed in the analysis.

Fig. 2. Overview of the experimental workspace showing the robot's predefined navigation zones: Zone 1 (Storage), Zone 2 (Assembly), and Zone 3 (Tools). These zones correspond to AoIs used for contextual asset localisation.

The Vision–Language Model (VLM) layer processes the data collected by *Spot* using a detailed, context-rich prompt, as shown in Figure 3. This prompt is designed to prime the VLM to recognise specific asset classes within the environment and to structure its output in a format that can be directly synthesised by downstream components of the asset-management pipeline. This structured, context-aware output underpins the novelty of the proposed approach, enabling the seamless integration of natural-language model reasoning with automated inventory tracking.

The motivation behind this architecture was to evaluate the feasibility and effectiveness of using LLMs—or more specifically, vision-language models (VLMs)—for scene understanding and asset identification in mobile robotics applications [19]. VLMs, powered by transformer architectures, are pre-trained on vast and diverse datasets and are capable of interpreting multi-modal input. This characteristic allows them to generate semantic insights from visual data without requiring extensive fine-tuning for specific environments.

This contextual understanding is integrated with a topological map of the environment, identifying waypoints and traversal zones as illustrated in Figure 2 and Figure 4. The map defines Areas of Interest (AoIs) of 20 m^2 each, aligned

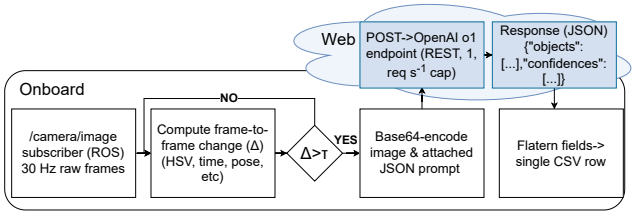

Fig. 3. Frame-selection and API bridge workflow for transmitting candidate images from onboard the Spot to a remote vision–language model for asset detection.

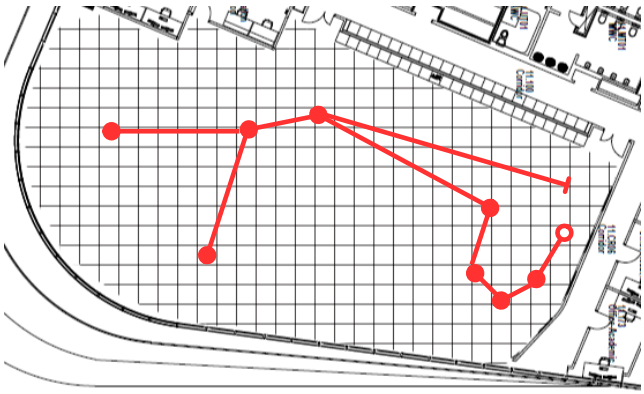

Fig. 4. Topological waypoint map used for the factory tour: Open cicrcle ○ indicates the start/home, filled circle ● represents the intermediate waypoints visited in sequence, red lines denote planned traversal paths, and **T** marks the terminus/docking point. The background grid corresponds to AoIs of 20 m^2 each, aligned with the 2D occupancy map generated from *Spot's* recorded trajectory.

precisely with the 2D occupancy map generated from *Spot*'s recorded trajectory.

The generated topological map enables operators to plan survey routes more effectively and provides spatial context that reduces false positives or "hallucinations" in the Vision–Language Model (VLM) output. Assets detected outside their designated Areas of Interest (AoIs) on the map are flagged for review, allowing location verification and record updates. These AoIs are defined within a location dictionary specifying expected positions for each asset type. By integrating ROS with the VLM, sensor data streams are fused into the topological map in real time, enabling *Spot* to localise assets accurately, cross-reference detections with the location dictionary, and minimise misclassifications through context-aware mapping.

Spot's payload capacity allowed the integration of a ZED2 stereo camera from Stereolabs. This addition was necessary because the robot's onboard cameras publish only monochrome images, whereas the VLM requires RGB input. The ZED2 also offers compatibility with NVIDIA's vision AI frameworks, enabling accelerated processing for future extensions. The camera drivers were configured to publish a 360p, 30 Hz video feed, producing a continuous stream of frames from which candidate images could be selected for analysis. Along with the video stream, the ZED2—through its ROS driver—also published IMU and odometry data, enabling estimation of *Spot*'s pose and orientation within the workshop. This multimodal data capture supported experimentation with frame-selection strategies while demonstrating that the approach can operate effectively at relatively low resolution, thereby reducing computational and bandwidth demands.

The captured video stream was subsequently encoded as a Base64 string and transmitted to an external large language model (LLM) for processing. Unlike traditional object detection pipelines that rely on locally deployed convolutional neural networks (CNNs) or other deep learning classifiers [2], this approach leveraged the *Spot* platform's networking capabilities to offload computation to a remote LLM.

To achieve that, an API bridge was developed to evaluate specific metrics before uploading, as shown in Figure 3. These approaches looked for a delta in the variables listed in the table and determined the appropriate frame to assess. This led to other benefits paired with a reduced computational load on the system.

VLMs trained on large-scale image–text datasets can deliver superior zero-shot performance, but such models require substantial computational resources to run locally [19]. To mitigate this constraint, images were offloaded to an online API, allowing inference to be performed remotely. This approach enabled the use of a more advanced identification model without increasing the onboard computational load. Offloading inference also provided the flexibility to experiment with multiple frame-selection strategies, facilitating the identification of the most suitable frames for analysis.

We evaluated three complementary heuristics to decide whether a new frame warrants analysis:

1) **Temporal spacing** — enforce a minimum time interval Δt between successive evaluations.
2) **Odometry delta** — trigger only when the robot's displacement exceeds a threshold Δs.
3) **Pixel-wise change** — compute grayscale and HSV differences; proceed if the mean absolute error exceeds τ_p.

A. Use-Case Setup

This section details the use-case setup for implementing *Spot*'s autowalk and developing the contextual understanding of the environment detailed above. The API Bridge uploads the image with the following prompt:

> "You are an asset-managing robot that is very specific about what its looking for. Evaluate these frames for the following objects(model_id-label): 1-Laptops, 2-3D Printer, 3-Storage Units, 4-Robots, 5-Desktops, 6-Tablets, 7-Tools, 8-Miscellaneous. List the objects you see in the frames along with their model_id and assign a random asset_tag. Provide the output in the form of a CSV table with the following columns: name, asset_tag model_id(integer), notes. Provide exactly four columns of data."

The API Response is then exported as a CSV, and *Spot* resumes its route. This CSV column is then uploaded to the SnipeIT API where the dashboard is prepared. This keeps the tasks in a fixed format and informs the users about the misplaced items in the space. This lets the user list the various items that *Spot* has analysed.

Misplaced items are identified by cross-referencing detections against the location dictionary; in the example shown in Figure 5, the asset "ROBOT" was found in Zone 8 rather than its designated location. Such cases are flagged as "deployed," and the dashboard is updated accordingly. Equipped with this information, operators can target their search to the identified zone, aided by the captured frame in which the item was

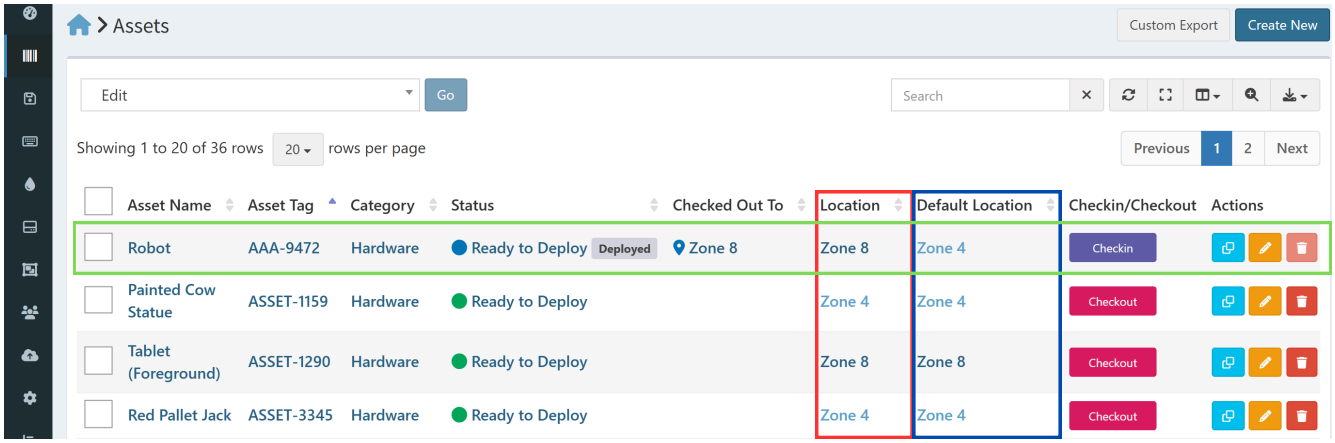

Fig. 5. Snipe-IT asset-management dashboard populated with detections generated by *Spot*. Each entry includes an autogenerated asset name, asset tag, category, status, and associated location metadata. Assets identified outside their default zones (e.g., "Zone 8" instead of "Zone 4") are flagged for operator review.

detected, allowing them to visually confirm and locate the asset on the workshop floor.

III. Results and Analysis

Figure 5 presents the Snipe-IT asset-management dashboard populated with detections generated by the proposed pipeline. Each record contains an autogenerated label, location coordinates, and associated metadata, providing operators with actionable inventory information. The pipeline integrates seamlessly with the dashboard interface and creates structured, searchable asset entries without the need for manual tagging or data entry. This demonstrates its potential for deployment in diverse industrial layouts where rapid configuration and minimal operator workload are priorities.

Spatial coverage of the test environment was assessed using the robot's recorded trajectory and occupancy-grid mapping, as shown in Figure 6. The plotted path reveals both high-density traversal zones and unvisited regions, enabling segmentation into asset-based AoIs consistent with industrial space-planning practice. This analysis is important for assessing the completeness of the asset survey and identifying spatial blind spots that could affect localisation performance.

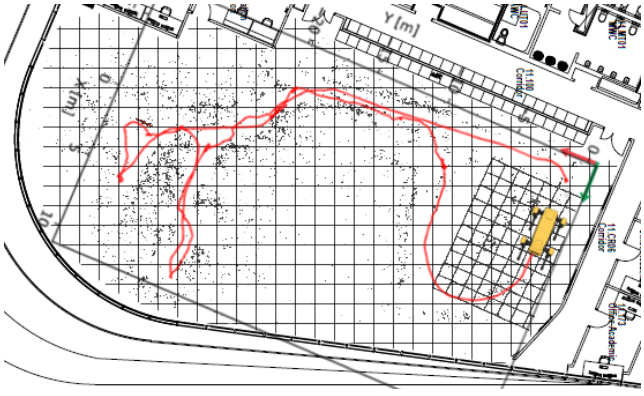

Fig. 6. Pose trace and occupancy grid generated during Spot's autonomous survey of the workshop environment. The red trajectory represents the robot's recorded path, while the overlaid occupancy grid (1 m × 1 m cells) indicates surveyed and unsurveyed regions. The map aligns with the global coordinate frame used for waypoint planning.

Table I summarises the performance of four frame-sampling heuristics implemented within the API bridge to reduce redundant inference calls. The Δ-Pose (5 %) method achieved the highest coverage (64.0 %) and a low false-identification rate (12.68 %), at an API cost equivalent to the time-sampled baseline (5 s). The Δ-Greyscale (20 %) heuristic reduced frames analysed to 2.19 % with a moderate drop in coverage (42.7 %) and lower false IDs (13.07 %). In contrast, Δ-HSV (20 %) minimised API usage to 1.17 % of frames but at the expense of coverage (22.7 %) and increased false IDs (18.46 %). These results indicate that motion- and pose-based triggers preserve spatial coverage more effectively than purely colour-based change detection in this environment, while still constraining operating costs.

Overall, the combination of coverage analysis and cost benchmarking demonstrates that selective frame evaluation can substantially reduce cloud-inference expenses while maintaining acceptable spatial coverage and detection reliability. These findings validate the proposed pipeline's viability for

TABLE I
Frames analysed, coverage, false identification, assets, and cost.

Approach	Frames (%)	Cov. (%)	False ID (%)	Assets	Cost ($)
Time-sampled (5 s)	3.09	60.0	22.00	179	0.18
Δ-Greyscale (20%)	2.19	42.7	13.07	153	0.22
Δ-HSV (20%)	1.17	22.7	18.46	65	0.06
Δ-Pose (5%)	3.29	64.0	12.68	205	0.18

Notes. Frames = proportion of captured images dispatched for inference analysis. Cov. (%) = observed floor area coverage. False ID = percentage of alerts requiring operator override (lower is better). Cost in USD ($) of the API usage (inference) excluding infrastructure costs. Inference in these runs used OpenAI's `o4-mini`.

deployment in cost-sensitive industrial contexts, provided that sampling heuristics are tuned to the specific environmental dynamics and asset distribution patterns of the target facility. While occlusion was qualitatively observed to affect certain detections in cluttered areas, the proportion of detection failures specifically attributable to occlusion was not quantified in this study. Attempts to resolve this issue remain a fundamental locus of computer vision research as explained in section I.

IV. CONCLUSION

This research offers a glimpse into the future of asset management by demonstrating how an autonomous quadruped equipped with vision–language perception can integrate with existing asset-management platforms to provide real-time, tag-free localisation and classification of factory assets. Building on the challenges in asset management identified by Alhadi et al. [2] and Saxena et al. [1], this work leveraged recent advances in imaging and transformer-based analysis to produce structured, actionable outputs, exemplified in Figure 5. These results show that mobile vision systems can maintain human-supervised inventory tracking in industrial environments.

Despite these promising outcomes, several limitations warrant consideration. First, the visual analysis pipeline incurs a pay-per-call cost (Table I), which scales with the number of frames processed and must be factored into operational budgets. Second, the study was conducted in a single workshop environment, limiting the generalisability of the results to other facility types. Third, vision-only sensing is susceptible to occlusion, lighting variation, and environmental noise, which can affect detection reliability. Asset scale also presents challenges, as extremely small or large items may not be consistently recognised or categorised correctly. Furthermore, *Spot* performs its tally offline between shifts, meaning that mid-shift changes to asset locations are not immediately reflected in the database. The *Spot* platform itself carries a significant capital cost, and the introduction of autonomous robots into the workplace may encounter resistance from operators due to perceived intrusiveness or workflow disruption.

Future work should address these limitations by validating the approach across multiple factory types to assess cross-domain robustness; refining frame-evaluation heuristics to minimise cloud-processing costs; exploring low-cost sensing platforms to improve economic viability; studying operator workflows to ensure the system integrates smoothly into daily operations; and investigating human–robot interaction factors to improve acceptance and safety. Addressing these areas will help translate the demonstrated potential into scalable, sustainable asset-management solutions for a broad range of industrial contexts.

REFERENCES

[1] N. Saxena and B. Sarkar, "Random misplacement and production process reliability: A sustainable industrial approach to deal with the discrepancy and deficiency," vol. 19, no. 7, pp. 4844–4873, Sat Jul 01 04:00:00 UTC 2023. [Online]. Available: https://www.aimsciences.org/en/article/doi/10.3934/jimo.2022151

[2] A. Alhadi, B. Dr Tom, and R. Yacine, "Enhancing asset management: Integrating digital twins for continuous permitting and compliance - A systematic literature review," vol. 99, p. 111515.

[3] F. Faiz, V. Le, and E. K. Masli, "Determinants of digital technology adoption in innovative SMEs," vol. 9, no. 4, p. 100610. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2444569X24001495

[4] S. Sepasgozar, A. Khan, K. Smith, J. Romero, X. Shen, S. Shirowzhan, H. Li, and F. Tahmasebinia, "BIM and Digital Twin for Developing Convergence Technologies as Future of Digital Construction," vol. 13, no. 2, p. 441. [Online]. Available: https://www.mdpi.com/2075-5309/13/2/441

[5] S. Krishnan and R. X. Mendoza Santos, "Real-Time Asset Tracking for Smart Manufacturing," in *Implementing Industry 4.0: The Model Factory as the Key Enabler for the Future of Manufacturing*, C. Toro, W. Wang, and H. Akhtar, Eds. Springer International Publishing, pp. 25–53. [Online]. Available: https://doi.org/10.1007/978-3-030-67270-6_2

[6] K. Staniec, M. Kowal, S. Kubal, and P. Piotrowski, "TrackMe—A Hybrid Radio-Optical System for Assets Localization in Industry 4.0 Plants," vol. 69, no. 2, p. navi.524. [Online]. Available: http://navi.ion.org/lookup/doi/10.33012/navi.524

[7] J. A. Fisher and T. Monahan, "Evaluation of real-time location systems in their hospital contexts," *International Journal of Medical Informatics*, vol. 81, no. 10, pp. 705–712, 2012. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2012.07.001

[8] R. C. Shit, S. Sharma, D. Puthal, P. James, B. Pradhan, p. u. family=Moorsel, given=Aad, A. Y. Zomaya, and R. Ranjan, "Ubiquitous Localization (UbiLoc): A Survey and Taxonomy on Device Free Localization for Smart World," vol. 21, no. 4, pp. 3532–3564.

[9] L. Zhao, H. Huang, C. Su, S. Ding, H. Huang, Z. Tan, and Z. Li, "Block-Sparse Coding-Based Machine Learning Approach for Dependable Device-Free Localization in IoT Environment," vol. 8, no. 5, pp. 3211–3223. [Online]. Available: https://ieeexplore.ieee.org/document/9178320/?arnumber=9178320

[10] Y. N. Naggar, A. H. Kassem, and M. S. Bayoumi, "A Low Cost Indoor Positioning System Using Computer Vision," vol. 11, no. 4, p. 8. [Online]. Available: https://www.mecs-press.org/ijigsp/ijigsp-v11-n4/v11n4-2.html

[11] F. Graf, J. Lindermayr, B. Graf, W. Kraus, and M. F. Huber, "HIPer: A Human-Inspired Scene Perception Model for Multifunctional Mobile Robots," vol. 40, pp. 4668–4683.

[12] A. Bouman, M. F. Ginting, N. Alatur, M. Palieri, D. D. Fan, T. Touma, T. Pailevanian, S.-K. Kim, K. Otsu, J. Burdick, and A.-a. Agha-mohammadi. Autonomous Spot: Long-Range Autonomous Exploration of Extreme Environments with Legged Locomotion. [Online]. Available: http://arxiv.org/abs/2010.09259

[13] Matterport3D: Learning from RGB-D Data in Indoor Environments. [Online]. Available: https://niessner.github.io/Matterport/

[14] A. Zhang, S. Li, J. Wu, S. Li, and B. Zhang, "Exploring Semantic Information Extraction From Different Data Forms in 3D Point Cloud Semantic Segmentation," vol. 11, pp. 61 929–61 949.

[15] Autonomy Technical Summary — Spot 4.1.1 documentation.

[16] Spot ROS Driver Usage — Spot ROS User Documentation 0.0.0 documentation. [Online]. Available: https://www.clearpathrobotics.com/assets/guides/melodic/spot-ros/ros_usage.html

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. [Online]. Available: http://arxiv.org/abs/1706.03762

[18] N. Nikolakis, P. Catti, L. Fabbro, and K. Alexopoulos, "Adapting Vision Transformers for Cross-Product Defect Detection in Manufacturing," vol. 253, pp. 2693–2702. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050925003370

[19] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," vol. 46, no. 8, pp. 5625–5644. [Online]. Available: https://ieeexplore.ieee.org/document/10445007/