Human-Interactive Mixed Reality System for Entity Recognition

Laura Ribeiro
University of Luxembourg
Kirchberg Luxembourg
laura.ribeiro@uni.lu

Holger Voos
University of Luxembourg
Kirchberg Luxembourg
holger.voos@uni.lu

Jose Luis Sanchez-Lopez
University of Luxembourg
Kirchberg Luxembourg
joseluis.sanchezlopez@uni.lu

Abstract

Reliable perception is essential for collaborative robots operating safely in shared human environments. However, automated entity detection systems still produce errors that degrade a robot's understanding of its surroundings. We present a Human-in-the-Loop (HITL) framework that enables user operators to validate and correct entity recognition and detection through an interactive Mixed Reality (MR) application and interface. Detected entities are visualized as aligned holograms, allowing users to confirm or remove them through intuitive, gesture-based spatial interactions. Our proposed method demonstrates that this shared environment and its interaction approach are functional and effective for correcting detections in real-time. By integrating the HITL approach, our system evaluation produces a more accurate representation of the shared environment and establishes the foundation for future extensions, including safer and more effective human–robot interaction and collaboration.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; • **Computer systems organization** → *Robotics*; *External interfaces for robotics*.

Keywords

Mixed Reality, Human-in-the-Loop, Entity Recognition, Entity Detection, Human-Robot Interaction, Situational Awareness

ACM Reference Format:
Laura Ribeiro, Holger Voos, and Jose Luis Sanchez-Lopez. 2026. Human-Interactive Mixed Reality System for Entity Recognition. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794401

1 Introduction

Collaborative robots (cobots) represent the idea of robots working alongside humans in a safe workspace. Cobots operating in shared environments require comprehensive situational awareness to ensure safe task completion [20]. Achieving this awareness requires continuous observation and interpretation of the surroundings [3], which is fundamental for addressing the safety-critical challenges of Human-Robot Interaction (HRI) and collaboration, such as trajectory planning [5] and search and rescue operations [18].

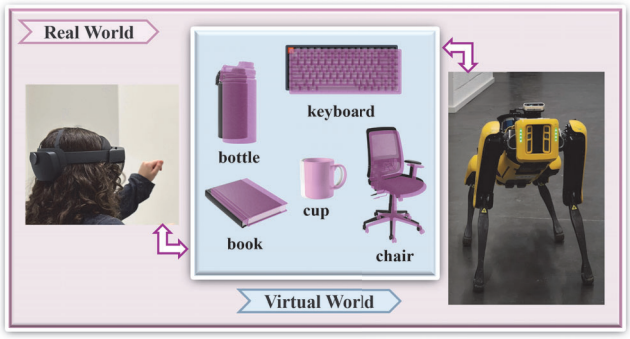

Figure 1: The overall idea behind the HIMER approach for shared environment understanding, where humans are exchanging knowledge regarding entity recognition and detection to robots, through a shared environment.

To understand the surroundings, entity recognition and detection are essential components of situational awareness. Identifying objects in a robot's field of view allows it to make decisions [16], particularly in dynamic environments such as manufacturing facilities and logistics warehouses [21]. However, automated detection systems often produce errors, including false positives, misclassifications, and missed detections, which can compromise safety and task performance.

Addressing these limitations, an approach is to include a Human-in-the-Loop (HITL) by integrating human perception into the robot's situational awareness process. Research has shown that human feedback improves obstacle detection and avoidance capabilities [9], while bidirectional knowledge exchange leverages the complementary strengths of humans and robots [6].

Effective HITL systems require a shared representational space for knowledge to be exchanged between agents. Digital twins provide this foundation by maintaining virtual representations of the workspace and detected entities [21].

Mixed Reality (MR) or Augmented/Virtual Reality devices, such as HoloLens 2 and Meta Quest 3, offer an approach for semantic knowledge exchange between humans and robots by enabling the creation of a shared digital representation combining the real and virtual world.

We proposed **HIMER**, a **H**uman-**I**nteractive **M**ixed Reality System for **E**ntity **R**ecognition, which enables effective real-time correction of entity detections. Figure 1 illustrates the system concept. The proposed work aims to explore humans' knowledge of scene understanding, as humans can naturally recognize entities, understand their semantics, and know their spatial relationships.

For this purpose, the main contributions of this work are to allow users to validate and correct autonomous detections, remove

HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794401

false positives through intuitive spatial interactions, and further create a refined environmental model that is accessible to the robot. This establishes a common understanding of the shared workspace, enabling safer and more effective HRI and collaboration.

2 Related Works

Recent advances in MR have enabled autonomous real-time object detection in dynamic environments. Several works have integrated You Only Look Once (YOLO) [12] as a base detection method with HoloLens for label visualization and object tracking [1, 8, 17], establishing the foundation for automated detection in MR systems. These approaches have been extended to domain-specific applications, including situational awareness with advanced machine learning models [21, 22] and simulated industrial environments [7].

Beyond detection, hybrid approaches have explored leveraging user context. DeepMix [10] uses user movement to fine-tune detection on planar surfaces, while [2] integrates BIM (Building Information Modeling) with MR for context-aware furniture management. In dynamic scenes with object interactions, such as furniture, additional challenges arise, including tracking lost and found [4] objects.

Recent work explores human input to enhance detection accuracy. In SLAM systems, [13] uses human feedback to improve object detection algorithms, and [15] enriches semantic information at the object level, though neither implements extended reality interfaces for direct interaction. While [19] implements semantic information with HoloLens for HRI, it lacks objects in the SLAM system. For HRI applications in general tasks, [11] proposes a multimodal approach using HoloLens and LLMs with feedback mechanisms.

While existing systems demonstrate robust automated detection and recognition, they lack mechanisms for users to interactively validate and correct detection errors in real-time in MR environments. Our HIMER approach addresses this gap by implementing a gesture-based MR interface that enables users to confirm or remove detected entities, to refine the robot's environmental understanding in HRI scenarios.

3 System Overview

This section describes the conceptual idea and workflow of HIMER's architecture for real-time entity detection, human validation, and shared knowledge representation using the MR interface, as presented in Figure 2.

HIMER employs a human-interactive spatial annotation paradigm to create a semantically rich situational awareness system. Rather than requiring robots to independently and autonomously interpret complex human workspaces, HIMER enables humans to validate their environmental understanding through intuitive MR interfaces. We organize the system description into three components: (1) environment understanding and perception, followed by entity detection, (2) shared environment for visualization and HITL validation, and (3) the robot integration for collaborative task execution.

3.1 Real to Virtual World

Figure 2 illustrates the system pipeline, beginning with the user wearing the HoloLens 2 device to capture live images of the environment. The HoloLens 2 main camera is an RGB camera that allows capturing video at up to 1080p resolution at 30 frames per second [14]. To maintain system responsiveness, the Unity application samples frames at a reduced resolution and a fixed frame rate and transmits them to an external processing module.

The processing module is dedicated to performing entity detection and segmentation. For each received frame, the model outputs segmentation masks, bounding boxes, class labels, and confidence scores for all detected entities. These detection results are published back to the Unity application, completing the perception loop from the physical environment to its digital representation.

3.2 Shared Environment

The Unity application transforms detection results into spatially registered holographic overlays aligned with physical objects. Upon receiving segmentation masks and bounding boxes, the system performs geometric back-projection to recover 3D object poses in world coordinates using the inverse pinhole camera model. Depth information for each detected entity is acquired from the HoloLens 2's time-of-flight sensor [14], while camera intrinsic parameters are obtained from the device's camera API. This spatial registration process (as in Figure 2) ensures that virtual annotations align with their physical counterparts in terms of position, width, and height.

The application employs passthrough video rendering to maintain visual continuity with the physical environment while overlaying holographic representations of detected entities. As the user moves through the workspace, images are continuously captured and processed, with holographic overlays rendered in real-time at the spatial locations of detected objects. Each holographic representation displays the associated class label and remains anchored to its detected 3D position as the user navigates the environment.

The validation interface enables users to review and correct detection errors through gesture-based interaction. Users can examine each detected entity's holographic overlay and either confirm correct detections or remove false positives. For example, if there should be one cellphone and one mouse identified in the scene, rather than recognizing the mouse twice, the user can easily remove the extra mouse from the listing.

3.3 Human-Robot Collaboration

This subsection represents the next step in the HIMER architecture and ensures its potential application. Once the user completes the validation process, the information needs to be transmitted to the robot system, improving its situational awareness and enabling informed task execution, and fed to the identification model.

Each group contains the entity's 3D pose, semantic class, segmentation geometry, and validation metadata. After establishing the coordinate transformation between the mixed-reality device world frame and the robot's base frame, robotic systems can leverage this human-curated environmental representation in multiple directions:

- The semantic knowledge provides the robot with reliable information in dynamic spaces. Human validation ensures that

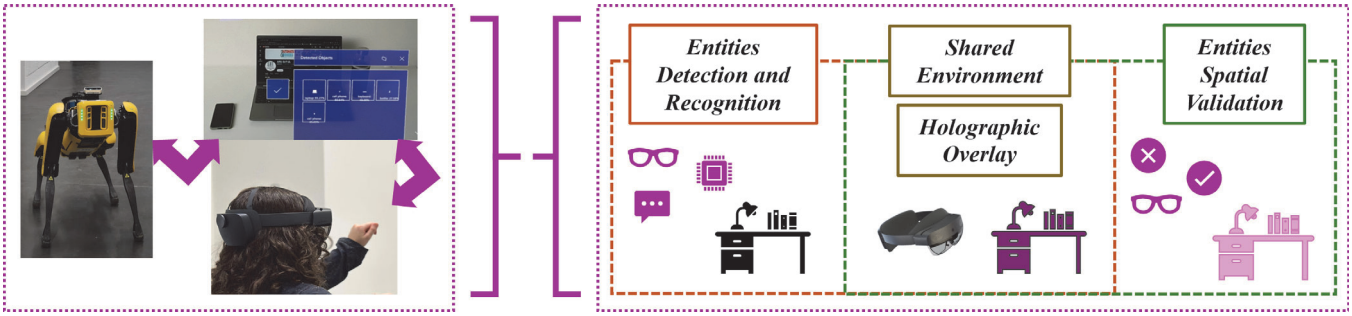

Figure 2: HIMER system architecture showing the human-in-the-loop workflow. On the left, the user shares the real world with the robot to collaboratively complete tasks. The user wearing the mixed-reality device has access to the Unity application. The right side shows the system internals. The user scans the environment to perform entity detection and recognition using the RGB camera of the HoloLens 2. In the shared environment, the user sees the holographic overlay of all entities. In the glasses, the user can perform a validation to accept or remove detected entities in the virtual world.

ambiguous or difficult-to-detect obstacles (e.g., transparent surfaces, reflective objects) are correctly represented.

- Shared semantic labels establish a common vocabulary for human-robot interaction, allowing natural-language and high-level commands and actions, such as "pick up the purple mug near the workbench". This could bridge the semantic gap between human spatial descriptions and robot world models.
- The robot can reason about object arrangements, spatial relationships, and available workspace regions based on human situational awareness knowledge. This enables higher-level task planning that respects human organizational preferences.

4 Results and Discussion

The implemented system represents a step forward in the integration of human knowledge through a high-level exchange of information. As mentioned in Section 3.2, the architecture utilizes the MR device, and the entity segmentation module employs a YOLOv8 [12] model specialized in segmentation. All components, including the Unity application and entity segmentation module, are implemented as ROS 2 nodes. The application displays the segmentation mask in real time as the user moves.

As shown in Figure 3, the user executes a grab gesture to pause the scanning of the environment and validate the detections. A menu displays the currently detected entities and their corresponding hologram representations. The menu remains open until the user finishes reviewing the elements. The user can remove detection errors from the model and confirm the selection. This process publishes the confirmed list of entities, after which environment monitoring resumes.

Figure 3 and Figure 4 demonstrate the system in a typical office environment. In Figure 3, the left panel shows the detection interface after the user pauses environment scanning, displaying four detected entities: a cup (94.33% confidence), TV (79.01%), mouse (58.95%), and bottle (50.74%). Each detection appears both as a thumbnail in the menu panel and as a spatially registered holographic overlay aligned with its physical counterpart in the workspace. The right panel illustrates the correction process, where the user has removed the mouse detection, which was identified as a false positive. On the other hand, Figure 4 illustrates the holographic overlay for human perception in the application, displaying four detected entities: a laptop (95.41% confidence), a keyboard (37.68%), a cell phone (35.81%), and a bottle (41.60%). Since the laptop is not separated from the keyboard, the user can easily remove the misclassification.

This architecture offloads intensive processing from the device, making the application more responsive. A key trade-off is the number of frames to utilize and the image resolution, as the entity segmentation component must maintain a consistent processing rate to handle all frames published by the Unity application.

The system's frame sampling rate and image resolution directly impact both detection quality and computational requirements. We sample at 10 frames per second at 896x504 resolution, balancing detection coverage with processing throughput. Higher frame rates would enable faster workspace scanning but require proportionally greater computational resources, while lower rates risk missing transient objects or requiring slower user movement.

This approach is particularly valuable in scenarios where the environment is dynamic, object recognition is ambiguous (e.g., similar-looking tools), or human verification is required (e.g., custom entities, safety applications).

5 Conclusion and Future Work

HIMER addresses a fundamental challenge in human-robot collaboration: how can humans efficiently communicate spatial and semantic information about their environment to robotic agents?

Our proposed system successfully enables a shared environment in which users can view and interact with object detections through mixed reality. The system accurately displays the detections as holograms, and the hand-gesture interaction mechanism reliably captures and removes entities with minimal effort. The results demonstrate that the validation is both feasible and effective. A core aspect of HIMER is sharing the environmental model with a

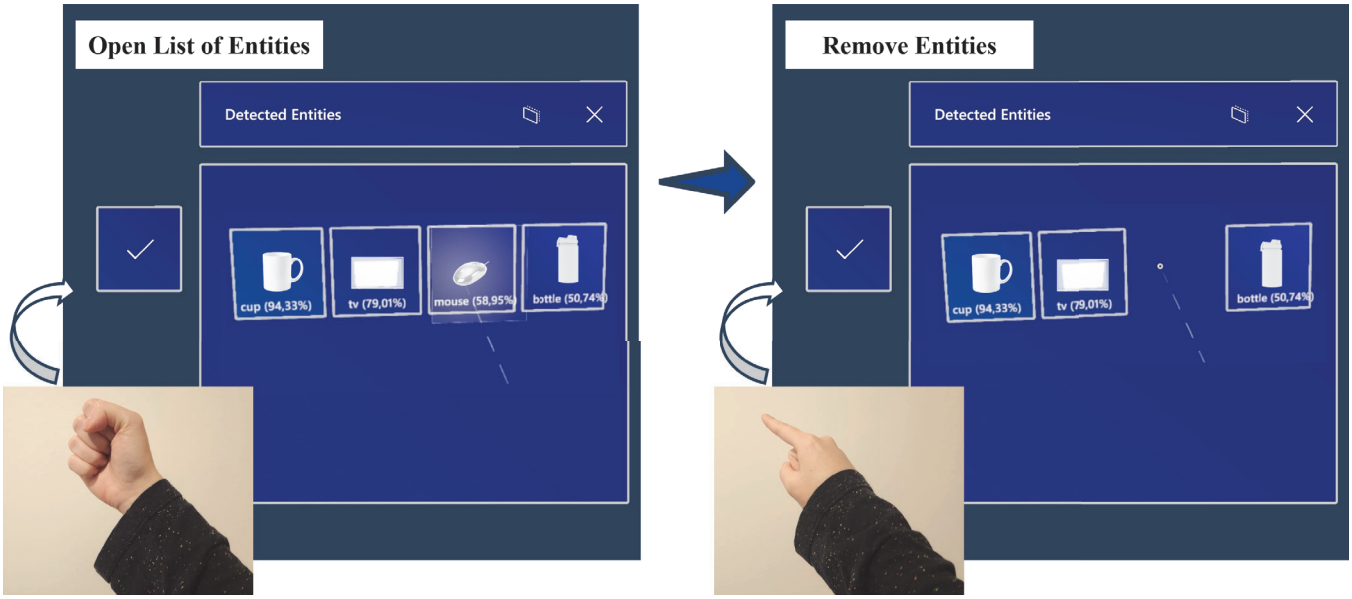

Figure 3: Human validation interface for entity detection correction. Left: The user opens the detected entities menu, which displays four objects (a cup, a TV, a mouse, and a bottle) along with their confidence scores and corresponding holographic overlays in the 3D space. Right: User performing gesture-based removal of a false positive detection (mouse), with the updated entity list showing only three validated objects.

Figure 4: Example of the qualitative results of the HIMER approach. Holograms of the respective detected entities overlay the physical objects, and through the MR interface, the user can see in real-time.

robot, allowing it to perceive what the human sees and demonstrating the potential for real applications. This approach will support diverse collaborative tasks and facilitate safer HRI.

Currently, the system only supports visual verification of detections; future work includes extending the system to provide feedback to the detection model. Additional planned improvements include more user operations, such as editing or adding new objects directly in MR during active scanning, and enriching the scene with more semantic information.

Acknowledgments

This research was funded by the Luxembourg National Research Fund (FNR) under the projects MR-Cobot (Ref. 18883697/MR-Cobot) and the DEUS (Ref. C22/IS/17387634/DEUS), at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), with the Automation and Robotics Research Group (ARG).

References

[1] Haythem Bahri, David Krcmarik, and Jan Koci. 2019. Accurate Object Detection System on HoloLens Using YOLO Algorithm. In *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*. IEEE, Athens, Greece, 219–224. doi:10.1109/ICCAIRO47923.2019.00042

[2] Haythem Bahri, David Krcmarik, Reza Moezzi, and Jan Kočí. 2019. Efficient Use of Mixed Reality for BIM system using Microsoft HoloLens. *IFAC-PapersOnLine* 52, 27 (2019), 235–239. doi:10.1016/j.ifacol.2019.12.762

[3] Hriday Bavle, Jose Luis Sanchez-Lopez, Claudio Cimarelli, Ali Tourani, and Holger Voos. 2023. From SLAM to Situational Awareness: Challenges and Survey. *Sensors* 23, 10 (May 2023), 4849. doi:10.3390/s23104849

[4] Tjark Behrens, René Zurbrügg, Marc Pollefeys, Zuria Bauer, and Hermann Blum. 2025. Lost & Found: Tracking Changes From Egocentric Observations in 3D Dynamic Scene Graphs. *IEEE Robotics and Automation Letters* 10, 4 (April 2025), 3739–3746. doi:10.1109/LRA.2025.3544518

[5] Raúl Calderón-Sesmero, Jaime Duque-Domingo, Jaime Gómez-García-Bermejo, and Eduardo Zalama. 2024. Development of a Human–Robot Interface for Cobot Trajectory Planning Using Mixed Reality. *Electronics* 13, 3 (2024). doi:10.3390/electronics13030571

[6] Kishan Chandan, Vidisha Kudalkar, Xiang Li, and Shiqi Zhang. 2021. ARROCH: Augmented Reality for Robots Collaborating with a Human. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Xi'an, China, 3787–3793. doi:10.1109/ICRA48506.2021.9561144

[7] Juan M. Deniz, Andre S. Kelboucas, and Ricardo Bedin Grando. 2024. Real-time Robotics Situation Awareness for Accident Prevention in Industry. In *2024 Latin American Robotics Symposium (LARS)*. IEEE, Arequipa, Peru, 1–6. doi:10.1109/LARS64411.2024.10786413

[8] Alessandro Farasin, Francesco Peciarolo, Marco Grangetto, Elena Gianaria, and Paolo Garza. 2020. Real-time Object Detection and Tracking in Mixed Reality using Microsoft HoloLens:. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, Valletta, Malta, 165–172. doi:10.5220/0008877901650172

[9] Francesco Ferracuti, Alessandro Freddi, Sabrina Iarlori, Andrea Monteriù, Karameldeen Ibrahim Mohamed Omer, and Camillo Porcaro. 2022. A human-in-the-loop approach for enhancing mobile robot navigation in presence of obstacles not detected by the sensory set. *Frontiers in Robotics and AI* 9 (2022), 909971. doi:10.3389/frobt.2022.909971

[10] Yongjie Guan, Xueyu Hou, Nan Wu, Bo Han, and Tao Han. 2022. DeepMix: mobility-aware, lightweight, and hybrid 3D object detection for headsets. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. ACM, Portland Oregon, 28–41. doi:10.1145/3498361.3538945

[11] Bryce Ikeda, Maitrey Gramopadhye, LillyAnn Nekervis, and Daniel Szafir. 2025. MARCER: Multimodal Augmented Reality for Composing and Executing Robot Tasks. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Melbourne, Australia, 529–539. doi:10.1109/HRI61500.2025.10974232

[12] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLOv8*. Retrieved November 24, 2025 from https://github.com/ultralytics/ultralytics

[13] Yulong Li, Yunzhou Zhang, Bin Zhao, Zhiyao Zhang, You Shen, Tengda Zhang, and Guolu Chen. 2024. HSS-SLAM: Human-in-the-Loop Semantic SLAM Represented by Superquadrics. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Abu Dhabi, United Arab Emirates, 10469–10475. doi:10.1109/IROS58592.2024.10801400

[14] Microsoft. 2023. *About HoloLens 2*. Retrieved November 22, 2025 from https://learn.microsoft.com/en-us/hololens/hololens2-hardware

[15] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. 2019. QuadricSLAM: Dual Quadrics From Object Detections as Landmarks in Object-Oriented SLAM. *IEEE Robotics and Automation Letters* 4, 1 (2019), 1–8. doi:10.1109/LRA.2018.2866205

[16] Shuvo Kumar Paul, Mircea Nicolescu, and Monica Nicolescu. 2023. Enhancing Human–Robot Collaboration through a Multi-Module Interaction Framework with Sensor Fusion: Object Recognition, Verbal Communication, User of Interest Detection, Gesture and Gaze Recognition. *Sensors* 23, 13 (2023), 5798. doi:10.3390/s23135798

[17] Ridho Fathoni Prasetya, Sritrusta Sukaridhoto, Maretha Ruswiansari, Rizqi Putri Nourma Budiarti, Ahmad Fatrian Romadhoni, Cahyo Arissabarno, and Agus Prayudi. 2023. Implementation of Object Recognition Integrated with Mixed Reality. In *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE, Palembang, Indonesia, 366–371. doi:10.1109/EECSI59885.2023.10295859

[18] Christopher Reardon, Kevin Lee, and Jonathan Fink. 2018. Come See This! Augmented Reality to Enable Human-Robot Cooperative Search. In *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, Philadelphia, PA, 1–7. doi:10.1109/SSRR.2018.8468622

[19] Laura Ribeiro, Muhammad Shaheer, Miguel Fernandez-Cortizas, Ali Tourani, Holger Voos, and Jose Luis Sanchez-Lopez. 2025. *Human Interaction for Collaborative Semantic SLAM using Extended Reality*. arXiv:2509.14949 https://arxiv.org/abs/2509.14949

[20] Nicole Robinson, Brendan Tidd, Dylan Campbell, Dana Kulić, and Peter Corke. 2023. Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review. *ACM Transactions on Human-Robot Interaction* 12, 1 (March 2023), 1–66. doi:10.1145/3570731

[21] Yuk Ming Tang, Wei Ting Kuo, and C.K.M. Lee. 2023. Real-time Mixed Reality (MR) and Artificial Intelligence (AI) object recognition integration for digital twin in Industry 4.0. *Internet of Things* 23 (Oct. 2023), 100753. doi:10.1016/j.iot.2023.100753

[22] Mikołaj Łysakowski, Kamil Żywanowski, Adam Banaszczyk, Michał R. Nowicki, Piotr Skrzypczyński, and Sławomir K. Tadeja. 2023. Real-Time Onboard Object Detection for Augmented Reality: Enhancing Head-Mounted Display with YOLOv8. In *2023 IEEE International Conference on Edge Computing and Communications (EDGE)*. IEEE, Chicago, IL, USA, 364–371. doi:10.1109/EDGE60047.2023.00059

Received 2025-12-08; accepted 2026-01-12