MULTI-SENSOR FUSION AND SLAM-BASED DIGITAL TWIN INTEGRATION FOR SIMULATED ACCESSIBILITY ASSESSMENTS IN COMPLEX ARCHITECTURAL ENVIRONMENTS

Luis Borunda[a]

[a]College of Architecture, Arts, and Design, Virginia Tech, USA

ABSTRACT

Ensuring accessibility in architectural environments remains a challenge, especially for visually impaired users who encounter subtle hazards like unmarked curbs, abrupt surface changes, and overhead obstructions that often go undetected. This paper introduces a simulation-based framework that detects and geolocates accessibility barriers using egocentric RGB video, GPS, and inertial data from AR glasses. Critical hazards are identified through monocular depth estimation, semantic segmentation, and 3D object detection, then anchored via Simultaneous Localization and Mapping (SLAM) trajectories and fused with OpenStreetMap data, digital models, and point clouds to improve spatial accuracy. Hazards are filtered for plausibility and consistency before being annotated and visualized in an interactive Rhino/Grasshopper-based digital twin. While the system can run on RGB and GPS data alone making it broadly sensor-agnostic and deployable on common mobile devices, SLAM data is integrated to review precision. Case studies show strong alignment with ground-truth conditions and robust integration with spatial simulation models for accessibility auditing.

Keywords: accessibility assessment, augmented reality, digital twin, computer vision, hazard detection.

1 INTRODUCTION

Accessibility in urban and architectural environments remains a critical concern, especially for individuals with visual or mobility impairments. Hazards like unmarked steps, sudden curb drop-offs, and low-hanging obstacles often go unnoticed until after construction, depending on manual audits that are time-consuming and costly. Bauer et al. [1] showed that deep learning, multisensory feedback, and affordable wearable sensors can reconstruct 3D scenes and enhance environmental awareness for visually impaired users. While white canes detect ground-level obstacles, they cannot identify overhead hazards, an area where AI-driven sensing offers key advantages. A survey of 300 blind individuals reported that $\sim$40% experienced head-level collisions annually, and 15% as often as once per month [2]. Addressing these "invisible" barriers early in the design or retrofit process is essential to creating safer, more inclusive spaces. Recent work highlights the potential of generative AI to bridge such accessibility gaps, especially in dynamic environments requiring real-time spatial understanding [3].

Advances in computer vision and augmented reality (AR) are enabling new methods for simulation-based accessibility audits. Wearable AR devices can capture egocentric video and sensor data, which AI models then process to infer scene geometry and semantics in real time. Monocular depth estimation from a single camera feed can reveal elevation changes and head-level obstacles, which our prior work [4] and other disability-centered studies have identified as critical hazards for individuals with vision impairments. Despite innovations in smart canes and AI glasses, most tools remain prototypes, often developed without co-design input and tested on blindfolded sighted users [5]. Combining depth inference with semantic

Proc. of the 2025 Annual Modeling and Simulation Conference (ANNSIM'25)
May 26-29, 2025, Universidad Complutense de Madrid, Madrid, Spain
J.L. Risco-Martín, G. Rabadi, D. Cetinkaya, R. Cárdenas, S. Ferrero-Losada, and A. Bany Abdelnabi, eds.

segmentation (e.g., sidewalks, stairs, signage) enables automatic detection of key accessibility features. In parallel, digital twin technology integrates sensor data into virtual building models, supporting interactive simulation of potential interventions.

This paper presents a multi-sensor fusion framework that links wearable AR data with digital twin environments to simulate and detect accessibility hazards. Using Meta's Project Aria research glasses, we collect synchronized video and inertial data, including Simultaneous Localization and Mapping (SLAM) 6-DoF trajectories and semi-dense point clouds [6]. The pipeline applies monocular depth estimation and semantic segmentation to identify hazards like curbs, steps, and overhead objects. Detections are localized in a Global Positioning System (GPS) based 3D coordinate frame and integrated into BIM or point cloud models to generate annotated digital twins. This allows designers to visualize hazards in context, evaluate their severity, and test remediation strategies virtually. Figure 1 illustrates our system architecture, which combines AR sensing, AI-based perception, and digital twin simulation. The smart glasses record multi-modal data during user walkthroughs. Geometry and semantics are extracted via computer vision, and SLAM anchors the detections in global coordinates. Hazards are output as 3D bounding boxes with metadata (type, confidence, timestamp) and visualized in Rhino/Grasshopper or geolocated on OpenStreetMap (OSM) tiles using only RGB and GPS. Optional SLAM-derived point clouds support spatial validation [7]. The modular pipeline ensures resilient performance, with fallback options (e.g., GPS-only or OSM overlays) when data inputs are missing or degraded.

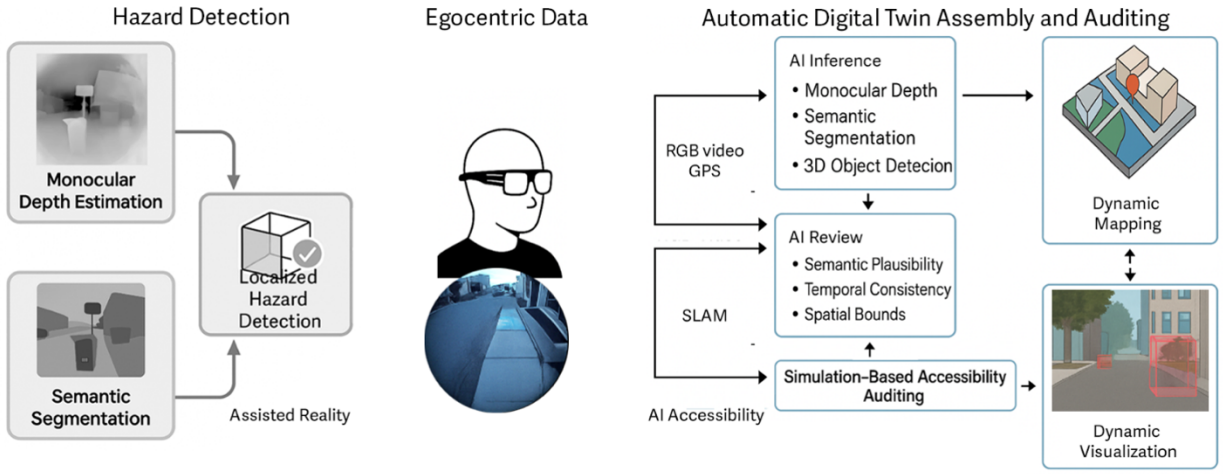


Figure 1: System architecture combining AR sensing, AI-based perception, and digital twin simulation.

This work uniquely bridges AR sensing and simulation by introducing a continuous, real-world-informed loop for accessibility assessment, moving beyond assistive navigation or static design reviews. Our main contributions include:

1. Accurate sensor fusion for detecting elevation and overhead hazards;
2. Integration with digital twins and GIS;
3. A Rhino/Grasshopper interface for spatial review;
4. A reproducible 3D hazard dataset.

We also propose fallback strategies for incomplete data and explore uses in ADA audits, urban design, and assistive navigation. The paper proceeds with related work (Section 2), methodology (Section 3), case studies (Section 4), limitations (Section 5), and broader implications (Section 6).

2 RELATED WORK

Efforts to improve environmental accessibility span multiple disciplines, including assistive technology, computer vision, and architectural simulation. We review three relevant areas: (i) wearable obstacle detection systems, (ii) vision-based hazard detection, and (iii) BIM/digital twin integration with sensor data.

Wearable Obstacle Detectors. Electronic travel aids (ETAs) have long aimed to augment the white cane and guide dog by detecting nearby obstacles using ultrasonic or infrared sensors [5]. For example, EyeMate uses ultrasonic sensors for both head- and ground-level hazard detection [8]. While effective, traditional canes cannot detect overhead objects—often leading to accidents. Newer smart glasses like Lighthouse Tech's "TAMI" use stereo vision to detect obstacles up to 3 meters away and alert users via vibration [2]. Some prototypes combine RGB-D cameras and wearable computing to detect terrain features like steps or curbs [9]. Unlike real-time navigation tools, our work repurposes AR devices for environmental mapping and simulation, targeting design-phase accessibility auditing.

Vision-Based Hazard Detection. Deep learning has greatly advanced scene understanding from standard cameras. Monocular depth estimation can now predict relative depth from a single RGB image [10], with models like MiDaS and MonoDepth2 offering strong performance even across domains [11]. Semantic segmentation enriches depth data by classifying pixels into categories like "sidewalk" or "stairs." Palafox et al. [12] combine segmentation and depth to produce semantically labeled 3D point clouds, highlighting features such as road edges or obstacles. We apply a similar multi-modal fusion to detect curbs and protrusions, improving robustness and reducing false positives. Unlike vehicle-mounted approaches for curb detection [13], our pipeline focuses on pedestrian-scale hazards via wearable AR.

BIM and Digital Twin Integration. While BIM encodes building geometry and compliance rules, it often omits temporary or as-built variations. Digital twins extend BIM by incorporating real-time sensor data. Prior work has aligned point clouds to BIM for model verification [14, 15]. Pepe et al. [14] propose a Grasshopper–Revit workflow to bridge laser scans and parametric models. We build on this using Rhino/Grasshopper as a central hub for integrating SLAM, segmentation outputs, and BIM data. We also leverage OpenStreetMap (OSM) [16] for geospatial context such as sidewalk edges or crossings. Our pipeline adds value by embedding hazard detections into these enriched models. This aligns with prior work on virtual exploration for blind users [17] and enables design teams to proactively assess accessibility barriers in both indoor and outdoor settings.

3 METHODOLOGY

We hypothesize that egocentric RGB video and GPS data, when processed through AI-based perception models, can accurately detect and localize accessibility hazards in real-world environments—even without specialized hardware like LiDAR. To test this, we developed a modular multi-sensor fusion pipeline that processes AR device recordings (e.g., smart glasses) to generate a hazard-augmented digital twin.

The pipeline (Figure 1) includes:

1. Data acquisition via smart glasses capturing timestamped RGB video and GPS;
2. SLAM-based localization to spatially anchor video frames;
3. Monocular depth estimation using transformer-based models;
4. Semantic segmentation for identifying hazard types;
5. Hazard detection and filtering based on temporal, spatial, and semantic consistency;
6. Integration with BIM, GIS, and SLAM point clouds for contextualization;
7. Digital twin visualization via Rhino/Grasshopper and online maps.

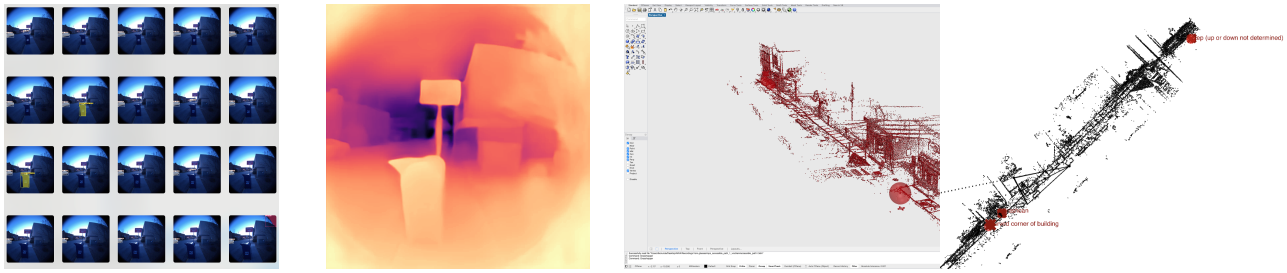
Each module includes fallback logic to maintain robustness when some inputs (e.g., GPS or BIM) are noisy or missing. We validated performance by comparing detected hazards with ground-truth annotations and cross-checking SLAM-aligned locations against OpenStreetMap (OSM). Project Aria's MPS framework [6] served as a spatial benchmark.

3.1 Data Sources

Meta's Project Aria glasses were set up to record 20 fps RGB video and 1 Hz GPS. SLAM outputs a 6-DoF trajectory and semi-dense point cloud (~30–50k features/scene) [18]. SLAM and GPS are fused to scale monocular depth to real-world units.

3.2 Monocular Depth Estimation and Object Detection

Each frame (Figure 2a) is processed using DPT-large (a ViT-based monocular depth network), producing a depth heatmap (Figure 2b). Sharp discontinuities signal hazards such as curb drops. Post-processing smooths noise and filters extreme values. Fused with SLAM, depth predictions achieve 1–2 cm reconstruction accuracy with ±5% localization error. Detected hazards include objects such as trash cans or signage, flagged at a threshold of >70% confidence.



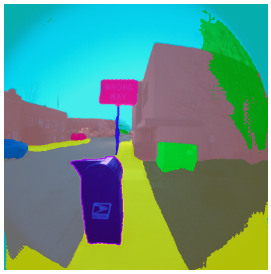
(a) Streetscape frames with YOLO segmentation. (b) Depth estimation output. (c) SLAM-based reconstruction in Rhino. (d) Top-down point cloud and hazard location.

Figure 2: RGB and GPS Hazard Identification with SLAM-based Auditing. (a) Egocentric frames processed through YOLO segmentation; (b) Monocular depth inference; (c) Integration into a 3D reconstruction in Rhino; (d) Top-down view showing hazard locations over the SLAM point cloud.

3.3 Semantic Segmentation

To improve scene understanding, we tested several modules with identifying several strengths: YOLOv8-seg (fast object detection), Google AI Studio (scene parsing directly from video), DINOv2 (unsupervised visual feature learning) and DeepLabv3+ (terrain-sensitive segmentation). Each model produces pixel-wise class labels (e.g., *road*, *sidewalk*, *person*), enabling class-specific logic for hazard detection. To ensure spatial consistency, segmentation masks are projected temporally via 3D reprojection. This allows hazard annotations to persist across frames as the user moves through the environment. Performance averaged approximately 28 FPS using an Apple M3 Pro GPU. We align the egocentric video stream with point cloud data to reconstruct the precise, geolocated device trajectory and frame configuration. Figures 2b, 3b, and 3c illustrate segmentation and depth output, aligned frame by frame with the Rhino/Grasshopper interface for spatial analysis and integration of design.

(a) Preprocessed egocentric frame.



(b) Semantic segmentation output via DINOv2.

(c) Corresponding 3D point cloud in Rhino/Grasshopper

Figure 3: Semantic Segmentation and Spatial Projection Pipeline. (a) Sample preprocessed streetscape frame; (b) Output of DINOv2 segmentation highlighting scene elements such as sidewalk, signage, and obstacles; (c) Frame-aligned point cloud, geolocalized device trajectory simulation and visualization with spatial hazard mapping in Grasshopper.

3.4 Hazard Detection and Filtering

Detected hazards are classified into three primary types:

1. **Elevation changes** — such as curbs, ramps, or steps;
2. **Overhanging obstacles** — such as low signage or protruding branches;
3. **Abrupt surface changes** — including potholes or transitions between materials.

To reduce false positives and ensure detection reliability, a multistage filtering process is applied:

- **Temporal consistency:** the hazard must be visible across multiple consecutive frames;
- **Spatial plausibility:** minor discontinuities or noise-level artifacts are excluded;
- **Semantic validation:** the detection must align with plausible scene classes (e.g., a drop must occur at a floor or sidewalk edge).

Each confirmed hazard is recorded with metadata including a 3D bounding box, estimated position, confidence score, and timestamp (Table 1). This structured output forms the basis of a reproducible hazard database, following the annotation pipeline described in [4].

Table 1: Example output from the hazard detection module showing detected objects, locations, and confidence scores. Non-hazard detections are also logged for context.

Vid	Time	Hazard	Frame Loc (x,y)	3D Loc {x,y,z}	Conf.	Notes
1	0:04	Step	(550, 300)	{38.41, 34.88, 1.39}	60%	–
1	0:40	Trashcan	(500, 400)	{-3.76, -4.04, -0.22}	75%	–
1	0:43	Corner	(600, 200)	{-6.94, -7.22, -0.39}	70%	–
1	0:53	Smartphone	(512, 321)	{-14.76, -14.88, -0.78}	95%	Not Hazard
2	0:06	Tree branches	(500, 50)	{19.76, 18.96, 1.72}	70%	Not Hazard
2	0:16	Overhang	(300, 100)	{-7.73, 8.29, 0.79}	75%	–
2	0:30	Stop Sign	(490, 270)	{-7.79, -6.81, -0.6}	90%	–
2	1:51	Smartphone	(480, 440)	{-90.47, -85.92, -3.38}	95%	Not Hazard

3.5 Integration with BIM and Point Cloud Models

The detected hazards are embedded in a 3D environment composed of GPS-aligned OpenStreetMap (OSM) data and SLAM-derived point clouds (see Figure 4). For example, curb-related detections are spatially validated against known OSM-sourced sidewalk edge geometries. Although the SLAM point clouds are semi-dense, they offer sufficient geometric detail to anchor hazard annotations with reliable accuracy. This integrated representation enables a spatially contextualized hazard map within parametric modeling environments such as Rhino/Grasshopper.

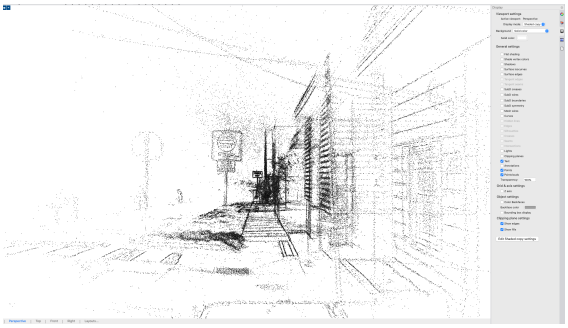
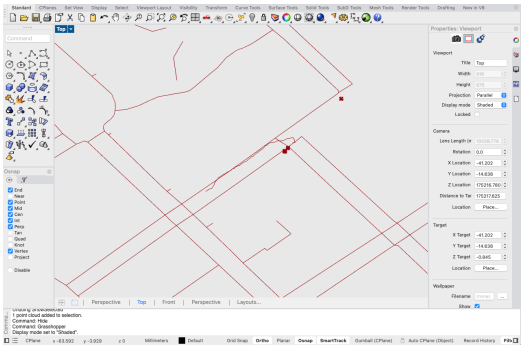


Figure 4: Integrated point cloud and hazard visualization in the Rhino-based digital twin.

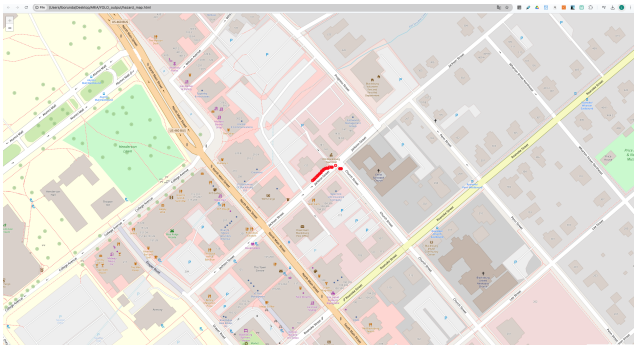
3.6 Digital Twin Visualization (Rhino/Grasshopper Interface)

We visualize hazards in Rhino 8 using custom Grasshopper scripts. Each detected hazard is rendered as a labeled 3D bounding box anchored to the SLAM-derived point cloud. Rhino's interoperability with Revit, via the Rhino.Inside framework, enables bidirectional updates—for example, inserting a ramp where a step hazard is detected.

Additionally, a Folium-based web interface overlays GPS-tagged hazards onto OpenStreetMap (OSM) tiles (Figure 5). This supports remote geovisualization and lightweight review of hazard distribution without the need for 3D modeling software.



(a) Rhino model with GPS-linked frames.



(b) OpenStreetMap overlay of geo-located hazards.

Figure 5: OpenStreetMap geo-localization model and map displaying each frame's spatial position and corresponding hazard detections.

4 EVALUATION RESULTS

We evaluated our system across four outdoor urban blocks featuring common accessibility challenges, including curb ramps, steps, overhangs, and sidewalk obstacles. Project Aria glasses recorded RGB, GPS, and SLAM data during controlled walkthroughs. Ground-truth annotations of hazards were manually documented and used for benchmarking.

Key findings include:

- **Accurate detection of physical hazards:** The system successfully identified most curb and step hazards and all overhead obstructions (e.g., signage, branches). ADA-compliant overhangs were occasionally flagged due to conservative thresholds. Missed curbs were associated with extremely low lighting. Post-processing with temporal and semantic filters removed most false positives.
- **Strong localization accuracy:** Combined depth estimation and SLAM produced <15 cm horizontal and $\sim$30 cm vertical accuracy, which is suitable for both design review and simulation purposes. GPS remained reliable in outdoor settings, while SLAM trajectories improved spatial consistency in drift-prone conditions.
- **Simulation integration:** Hazards were embedded into digital twin models via Rhino/Grasshopper and geolocated using OpenStreetMap. This supported interactive review and remediation simulation in standard design tools.
- **Robustness under partial input:** Even without semantic segmentation or with sensor/frame loss, the system remained functional, supporting fallback scenarios for lower-end devices or degraded data.
- **Cross-device compatibility:** Android phone testing confirmed the pipeline generalizes beyond AR glasses. Most phones update GPS at 1Hz. Although depth accuracy was lower, results support mobile deployment for broader use.

5 DISCUSSION AND LIMITATIONS

This research contributes a scalable, simulation-driven method for accessibility auditing through egocentric video and multi-sensor spatial data. Unlike traditional accessibility audits that depend on manual inspection, our pipeline fuses AR-derived RGB, GPS, and SLAM data with digital twin modeling, enabling dynamic evaluation both at the early design phase and during post-occupancy reviews [19].

For simulation and design workflows, the system enables early-stage hazard detection and iterative remediation planning. By integrating real-world sensor data with architectural models, designers can proactively visualize and address barriers before construction. Our results show that multi-modal sensor fusion significantly improves detection reliability compared to monocular depth-only methods [10], reducing false positives and improving spatial localization—especially in outdoor and irregular settings.

For applications in the AEC industry, the framework aligns with ongoing interest in smart city initiatives, ADA compliance, and parametric design auditing. Integration with Rhino/Grasshopper workflows ensures compatibility with existing BIM tools [14], and the outputs can inform both design professionals and urban planners. Future extensions may include simplified web or VR viewers for wider stakeholder access.

Nonetheless, several limitations remain. Monocular depth estimation, while hardware-efficient, struggles with transparent or reflective surfaces such as glass doors and windows [10]. Similarly, current segmentation models lack accessibility-specific object classes (e.g., wheelchair ramps, compliant signage), largely due to the limited representation of such features in existing datasets. Dynamic hazards—such as mov-

ing pedestrians or lighting changes—are not yet incorporated, and real-time outputs (e.g., voice or haptic feedback) are not currently implemented.

To address these challenges, future work will expand the hazard taxonomy, incorporate 3D semantic maps with frame-level temporal coherence, and explore fine-tuned, domain-specific segmentation models such as SAM. Participatory design is a critical next step: direct input from individuals with disabilities can help define hazard thresholds and interface preferences [20].

From a broader perspective, this work positions smartphones and AR wearables as potential agents for community-driven data collection. With GPS and RGB video, even low-cost devices could contribute to distributed accessibility audits. The democratization of such tools, combined with platforms like OpenStreetMap [16] and Project Aria [18], opens new possibilities for real-world validation and deployment at scale.

Overall, our system represents a resilient, sensor-flexible approach to accessibility auditing—one that bridges egocentric AI perception with architectural modeling to support more inclusive, simulation-based design practices.

6 CONCLUSION

We introduced a simulation-enabled accessibility auditing pipeline that combines AR hardware, AI perception, SLAM, and digital twin modeling. By detecting steps, curbs, and overhangs from egocentric video and localizing them in 3D, we enable proactive, design-oriented accessibility reviews. The system integrates into familiar tools (e.g., Rhino/Grasshopper, OSM) and promotes inclusive planning through simulation, moving beyond post-construction audits.

This pipeline supports a shift toward early, evidence-based accessibility assessments and contributes geolocated hazard datasets for future AI model development. By making simulated audits feasible with mobile or wearable hardware, the framework democratizes access to accessibility compliance tools. Future work includes real-time assistive features and deployment in live urban environments.

ACKNOWLEDGMENTS

We acknowledge Facebook Reality Labs for providing access to Project Aria hardware and data infrastructure. We thank the Commonwealth Cyber Initiative, the Virginia Tech Institute for Creativity, Arts and Design, and the Center for New Ventures for funding support. We are grateful to Dr. Na Meng for AI systems expertise and Professor Andrew Gipe-Lazarou for accessibility and inclusion guidance. Findings are the author's and do not reflect the views of Meta.

REFERENCES

[1] S. Bauer, D. Cheung, and R. Lutz, "A deep learning system for object detection for blind and visually impaired people," *Sensors*, vol. 19, no. 23, p. 5283, 2019.

[2] R. Manduchi and S. Kurniawan, "Mobility-related accidents experienced by people with visual impairment," *AER Journal: Research and Practice in Visual Impairment and Blindness*, vol. 4, pp. 44–54, 2011.

[3] M. Chui, J. Manyika, A. Singla, L. Clarke, A. Sukharevsky, L. Yee, and R. Zemmel, "The economic potential of generative ai: The next productivity frontier," McKinsey Global Institute, New York, Tech. Rep., 2023, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

[4] L. Borunda, A. Gipe-Lazarou, and N. Meng, "I WANT: Agency and accessibility in the age of AI," in *Proceedings of the ACSA International Conference: Inflections*, Querétaro, Mexico, June 2024.

[5] M. Hersh, "Wearable travel aids for blind and partially sighted people: A review with a focus on design issues," *Journal of Assistive Technologies*, vol. 2, no. 1, pp. 4–19, 2008.

[6] Meta Reality Labs, "Project aria slam outputs," https://facebookresearch.github.io/projectaria_tools/docs/data_formats/mps/slam, 2023.

[7] Python Visualization, "Folium documentation," https://python-visualization.github.io/folium/latest/, 2023.

[8] M. S. R. Tanveer, M. M. A. Hashem, and M. K. Hossain, "Android assistant eyemate for blind and blind tracker," in *2015 18th International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2015, pp. 266–271.

[9] A. Budrionis *et al.*, "Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review," *Assistive Technology*, vol. 34, no. 2, pp. 178–194, 2020.

[10] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022.

[11] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3828–3838.

[12] P. R. Palafox *et al.*, "Semanticdepth: Fusing semantic segmentation and monocular depth estimation for enabling autonomous driving in roads without lane lines," *Sensors*, vol. 22, no. 1, p. 253, 2022.

[13] F. Ma, P. Hou, Y. Liu, M. Liu, and J. Ma, "Annotation-free curb detection leveraging altitude difference image," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4112–4122, 2022.

[14] M. Pepe, "From point cloud to bim: A new method based on efficient point cloud simplification by geometric feature analysis and building parametric objects in rhinoceros/grasshopper software," *Applied Sciences*, vol. 12, no. 3, p. 1284, 2022.

[15] S. Halder, K. Afsari, E. Chiou, R. Patrick, and K. A. Hamed, "Construction inspection & monitoring with quadruped robots in future human-robot teaming: A preliminary study," *Journal of Building Engineering*, vol. 65, p. 105814, 2023.

[16] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[17] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators & Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997.

[18] S. Karakas *et al.*, "Aria data tools: Project aria – a platform for spatial understanding and ar research," https://github.com/facebookresearch/aria_data_tools, 2022.

[19] World Health Organization, *World Report on Vision*. Geneva, Switzerland: World Health Organization, 2019.

[20] I.-J. Kim, "Recent advancements in indoor electronic travel aids for the blind or visually impaired: A comprehensive review of technologies and implementations," *Universal Access in the Information Society*, 2023.

AUTHOR BIOGRAPHY

LUIS BORUNDA is an Assistant Professor of Advanced Building Design at Virginia Tech's School of Architecture where he leads research and teaching in computational design, accessibility, construction automation, lightweight structures, and simulation-based methods for the built environment. His research explores the intersection of architecture, artificial intelligence, and spatial computing to improve environmental accessibility and advance adaptive construction systems. His email address is lborunda@vt.edu.