

XR-assisted semantic scene understanding of building interiors via AI segmentation and photogrammetric reconstruction

Bryan G. Pantoja-Rosero¹

¹Digital Intelligence for Structures and Construction Laboratory (DISC),
School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

bryan.pr@ntu.edu.sg

Abstract -

Semantic understanding of indoor building environments is essential for digitalization, facility management, robotics, and human-computer interaction. However, existing workflows typically decouple data acquisition, semantic interpretation, and 3D reconstruction into offline processes, limiting on-site verification and interaction. This paper presents an extended reality assisted framework for semantic understanding and digitalization of indoor scenes by integrating on-device artificial intelligence based semantic segmentation with photogrammetric reconstruction. Image frames captured using a mixed reality headset are processed in near real time using a lightweight YOLOv11-seg model to obtain object-level detections and segmentation masks. The same image data are subsequently used for Structure-from-Motion reconstruction, where 2D semantic information is spatially propagated into a 3D point cloud and mesh, yielding semantically enriched 3D scene models. These models are visualized and interactively validated by the user within the extended reality environment, enabling immediate assessment of digitalization completeness. The proposed framework demonstrates how extended reality can function as an active tool for near real-time semantic perception, 3D modeling, and user-guided scene digitalization.

Keywords -

Extended reality; Semantic segmentation; Structure-from-Motion; Building digitalization; 3D semantic modeling; Interior building modeling; Deep learning; Photogrammetry; Mixed Reality; Edge computing

1 Introduction

The semantic understanding of indoor building environments is a fundamental requirement for a wide range of applications, including building documentation, facility management, digital twins, robotics, and human-computer interaction [1, 2]. Accurately identifying and localizing building elements and objects within interior spaces enables more efficient digitalization workflows and supports downstream tasks such as asset inventory, spatial analysis, and interaction-aware planning [3]. De-

spite recent advances in computer vision, most existing semantic understanding approaches rely on offline image processing pipelines, where data are collected on site and analyzed afterward on external computing platforms [4]. This separation between data acquisition and semantic interpretation limits responsiveness, hinders on-site verification, and often requires repeated visits when additional information is needed or errors are discovered during post-processing.

Recent progress in artificial intelligence (AI), particularly deep learning-based semantic and instance segmentation, has significantly improved the automated recognition of objects and scene elements from images [5, 6]. These methods have demonstrated strong performance in both outdoor and indoor environments, enabling pixel-level delineation of semantic classes and supporting dense scene understanding [7]. However, most AI-driven semantic segmentation workflows remain inherently two-dimensional, operating on individual images or image sequences without explicit spatial grounding [8]. As a result, semantic information is often disconnected from the three-dimensional (3D) structure of the environment, limiting its interpretability and usefulness for spatial reasoning, digital modeling, and interactive applications [9].

In parallel, photogrammetric reconstruction techniques such as Structure-from-Motion (SfM) have become well established for generating accurate 3D representations of indoor scenes from image collections [10, 11]. While these methods provide detailed geometric models, they typically lack semantic richness, producing unstructured point clouds or meshes that require additional processing to identify meaningful objects or building elements [12, 13]. Bridging the gap between 2D semantic perception and 3D geometric reconstruction remains a key challenge, particularly in scenarios where rapid feedback, validation, and interaction with the reconstructed scene are required [14, 15].

Extended Reality (XR) technologies—including augmented, virtual, and mixed reality—offer a unique opportunity to address these challenges by tightly coupling perception, computation, and visualization within a single interactive platform [16, 17]. XR head-mounted devices

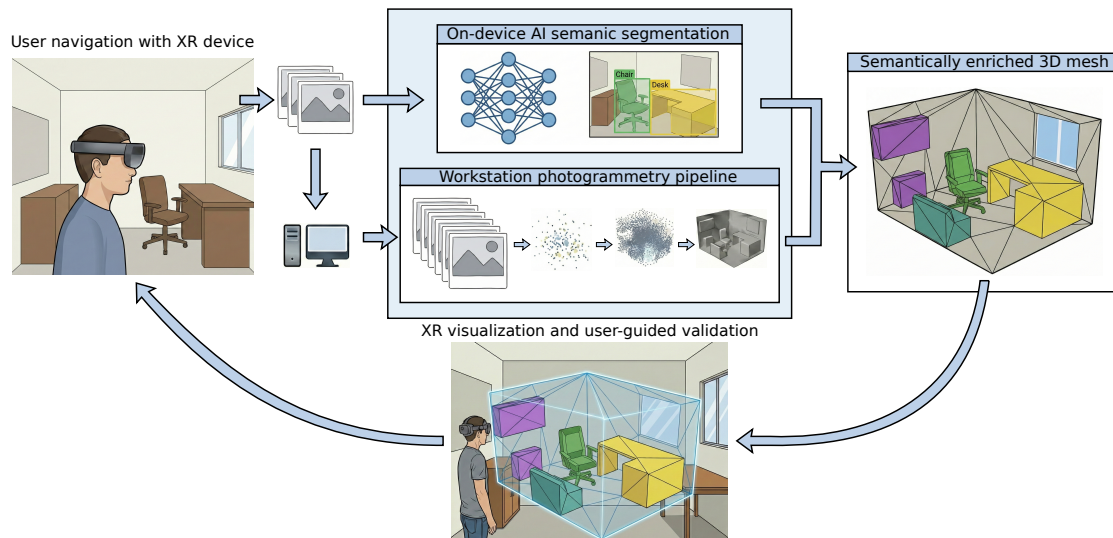


Figure 1. Closed-loop XR-assisted pipeline for semantic scene understanding. On-device semantic segmentation is performed on image frames acquired with a mixed reality headset, while photogrammetric reconstruction is executed on an external workstation. Semantic information is propagated into a 3D reconstruction and visualized in XR for interactive validation.

enable hands-free, spatially aware interaction with both physical and digital content, allowing users to observe, verify, and contextualize AI-generated information directly within the environment being analyzed [18]. Compared to conventional mobile devices or laptops, XR systems provide an immersive interface that supports near real-time understanding of spatial relationships, immediate validation of captured data, and intuitive exploration of semantic and geometric representations [19]. These capabilities are particularly beneficial for indoor scene digitalization, where occlusions, clutter, and complex layouts make purely image-based verification difficult.

Despite these advantages, current XR-based applications for semantic scene understanding remain limited. Many systems focus primarily on visualization, while AI-based semantic interpretation is performed offline or relies on cloud-based processing, introducing latency and restricting near real-time interaction [20, 21]. Furthermore, semantic outputs are often displayed in 2D overlays without enforcing spatial consistency across views or integrating semantic labels into a coherent 3D scene representation [22]. This limits the ability to verify semantic accuracy on site and reduces the usefulness of XR as an active tool for scene digitalization rather than a passive visualization interface.

To address these limitations, this paper proposes an XR-assisted framework for semantic understanding of indoor building scenes that integrates AI-based segmentation with photogrammetric reconstruction. In the proposed approach, images acquired using a mixed reality (MR) headset are processed using a deep learning-based segmentation model to identify semantic elements within

the scene. The same image set is then used for SfM reconstruction to generate a 3D representation of the environment. By associating 2D semantic predictions with the reconstructed geometry, semantic information is spatially propagated into the 3D scene, enabling the creation of semantically enriched 3D models. These models are subsequently visualized and explored within the XR environment, allowing users to inspect, validate, and interact with semantic scene representations directly on site.

The contributions of this work are threefold:

- An XR-assisted pipeline for indoor semantic scene understanding, combining AI-based image segmentation and photogrammetric reconstruction within a unified workflow.
- A method for spatially consistent projection of 2D semantic information into 3D reconstructed scenes, enabling semantically enriched models of indoor environments.
- An interactive XR visualization framework that supports near real-time verification and exploration of semantic 3D scenes, facilitating immediate validation of digitalization results.

By shifting XR from a passive visualization tool to an active component of semantic scene understanding and digitalization, this work demonstrates the potential of XR-enabled AI systems to support more efficient, reliable, and interactive indoor environment modeling. Given the scope of this paper, the emphasis is placed on demonstrating the feasibility and end-to-end integration of the proposed workflow rather than on a comprehensive quantitative benchmarking of its components. The proposed frame-

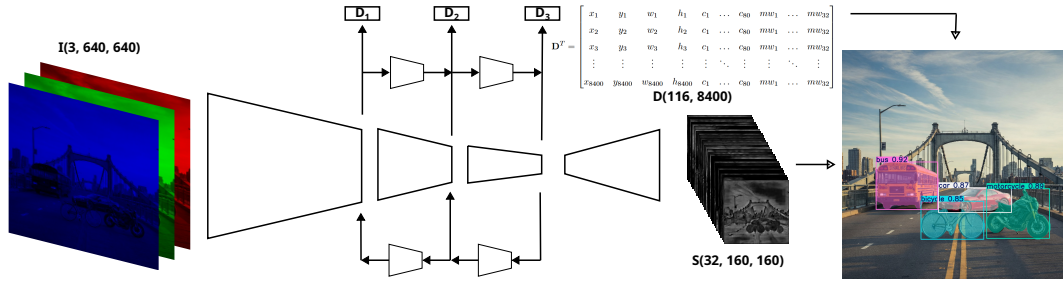


Figure 2. Overview of the YOLOv11-seg architecture used for on-device semantic perception, illustrating the processing pipeline from RGB image input to joint object detection and instance-level segmentation outputs.

work lays the groundwork for future extensions toward native 3D semantic segmentation, language-grounded scene queries, and tighter integration with building information models and digital twins.

2 Methodology

This work proposes an XR-assisted framework for semantic understanding and digitalization of indoor building scenes by integrating on-device AI-based semantic segmentation with photogrammetric reconstruction. The methodology is designed to enable near real-time semantic perception together with spatially consistent 3D modeling and on-site verification through XR. Building upon previous research integrating XR and AI for real-time infrastructure inspection [18], this work extends the approach from damage assessment to semantic scene understanding and 3D digitalization of indoor environments.

The overall pipeline is illustrated conceptually in Figure 1 and follows a closed-loop workflow. While the user navigates an indoor environment wearing an XR device (in this work, a Microsoft HoloLens 2 MR headset), image frames are continuously acquired and processed in near real time using an embedded deep learning model for semantic segmentation. In parallel, selected image frames are stored and transmitted to an external workstation, where a SfM pipeline is executed offline to reconstruct a three-dimensional representation of the scene. Semantic information extracted from the 2D images is subsequently associated with the reconstructed geometry, enabling the generation of a semantically enriched 3D model. This model is then transferred back to the XR device and visualized within the XR environment, allowing the user to interact with, inspect, and validate the reconstructed scene.

2.1 System overview

The proposed framework consists of three tightly coupled components:

1. XR-based semantic perception: A MR headset is used as both an image acquisition and visualiza-

tion device. Image frames captured by the headset's RGB camera are processed on-device using a lightweight AI model based on the YOLOv11-seg architecture, pre-trained on the COCO dataset for semantic and instance-level object segmentation. During operation, the model performs near real-time inference, generating bounding boxes and segmentation masks for common indoor objects such as furniture and equipment. These predictions are visualized directly in the XR environment through spatially aligned 3D text labels, bounding boxes, and image-aligned planes that overlay segmented regions onto the physical scene.

2. Photogrammetric reconstruction and semantic mapping: In parallel with near real-time XR perception, image data are transmitted to an external computer, where a SfM pipeline is executed to reconstruct a 3D representation of the environment in the form of a point cloud or mesh. Point feature correspondences used during the SfM process are exploited to establish spatial relationships between 2D image pixels and 3D points. By leveraging the semantic segmentation masks generated during inference, semantic labels are propagated from the image domain into the reconstructed geometry, enabling object-level segmentation of the 3D scene.
3. XR-based visualization and validation of semantic 3D models: Once the reconstruction and semantic mapping are completed, the semantically enriched 3D model is transferred back to the XR device. The model is then visualized within the XR environment, where users can manipulate, inspect, and validate both geometry and semantic labels. This enables immediate assessment of digitalization completeness and semantic correctness. If inconsistencies or missing regions are identified, additional images can be acquired on site, closing the loop between perception, reconstruction, and validation.

A key feature of the proposed methodology is its iterative and user-in-the-loop nature. Rather than treating digitalization as an offline, post-processing task, the frame-

work is intended to allow users to continuously assess reconstruction quality and semantic coverage during acquisition. This reduces the risk of incomplete or erroneous models and minimizes the need for repeated site visits. Once a scene is considered satisfactorily digitized, the workflow could be extended to adjacent rooms or spaces, supporting scalable indoor environment digitalization.

2.2 XR-based image acquisition and on-device semantic segmentation

Near real-time image acquisition and semantic perception are performed directly on a MR headset, which serves as both the sensing and visualization platform. During operation, the headset's RGB camera continuously captures image frames from the user's viewpoint as they navigate the indoor environment. These frames are processed on-device to enable near real-time semantic understanding of the scene without reliance on cloud-based computation. Performing inference directly on the device enables immediate feedback during acquisition, avoids dependency on network connectivity, and supports interactive, user-guided data collection within the XR environment.

Semantic perception is achieved using the YOLOv11-seg architecture (Figure 2), which performs joint object detection and instance-level segmentation in a single inference pass. In this work, a pre-trained YOLOv11-seg model trained on the COCO dataset is employed without additional fine-tuning. The COCO dataset provides a diverse set of common object categories typically found in indoor environments, such as chairs, desks, monitors, and other furniture or equipment, making it suitable for demonstrating semantic scene understanding in building interiors. By relying on a pre-trained model, the focus of this study is placed on system integration and spatial semantic mapping rather than model development or training.

For each processed image frame, the model outputs a set of object detections represented by bounding boxes, class labels, confidence scores, and corresponding pixel-wise segmentation masks. These outputs are generated in the 2D image domain but are immediately translated into spatially aligned visual elements within the XR environment. Bounding boxes are rendered as three-dimensional wireframe overlays aligned with the camera view, while object class labels are displayed as 3D text annotations positioned relative to the detected objects (Figure 3). This enables users to directly perceive semantic information in context, without interrupting their interaction with the physical environment.

In addition to bounding boxes and labels, segmentation results can be visualized through image-aligned 3D planes instantiated in the XR scene (Figure 4). Each plane is textured with the corresponding image frame, where seg-

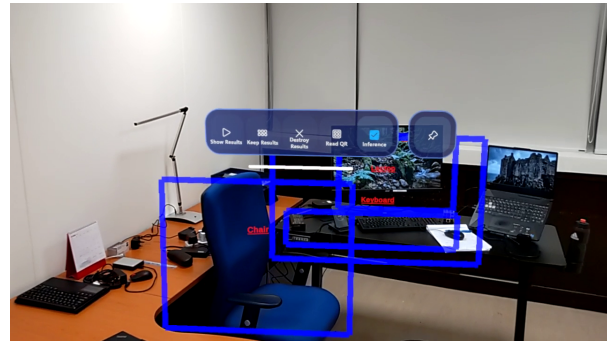


Figure 3. YOLOv11-seg detection outputs visualized in the XR environment.

mented object regions are highlighted according to their predicted classes. These planes are spatially aligned with the real-world background, creating the perception that the segmentation results are embedded within the physical scene. This visualization strategy allows users to observe semantic boundaries while preserving spatial context.

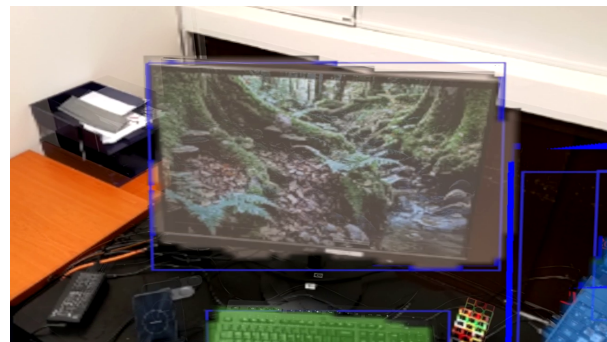


Figure 4. YOLOv11-seg segmentation results visualized as 3D textured planes in XR.

To avoid visual clutter and maintain usability, segmentation planes are not instantiated continuously for every processed frame. Instead, their creation is triggered explicitly by the user, enabling selective visualization of segmentation results when needed. Multiple planes corresponding to different viewpoints can be instantiated sequentially, supporting inspection of semantic predictions from multiple perspectives. This interaction-driven visualization approach leverages the strengths of XR by combining near real-time perception with user-controlled exploration of semantic information.

2.3 Photogrammetric reconstruction and semantic mapping

Photogrammetric reconstruction is employed to generate a three-dimensional representation of the indoor environment using the image data acquired during XR-based operation. Image frames captured by the MR device are transmitted to an external workstation, where, using Mesh-

room [11], a SfM pipeline executed to estimate camera poses and reconstruct the scene geometry in the form of a sparse point cloud, followed by a dense point cloud and surface mesh. The reconstructed geometry provides a spatial backbone onto which semantic information extracted from 2D images can be mapped.

A key aspect of the proposed framework is the association of 2D semantic segmentation results with the reconstructed 3D scene. During the SfM process, distinctive point features are detected and matched across multiple images to establish correspondences and triangulate 3D points. Because these features originate from known pixel locations in the input images, their projections can be directly related to the semantic segmentation masks generated during on-device inference. For each reconstructed 3D point, its originating 2D feature locations are checked against the corresponding segmentation masks to determine whether the point belongs to a specific semantic object class.

Using this association, semantic labels are propagated from the image domain into the 3D reconstruction. Points whose corresponding 2D features fall within segmented object regions are assigned the semantic class of that object. This results in a semantically labeled point cloud (Figure 5), where each point carries both geometric and semantic information. When multiple image observations are available for a given 3D point, semantic labels are aggregated to improve robustness and reduce the impact of occasional misclassifications.

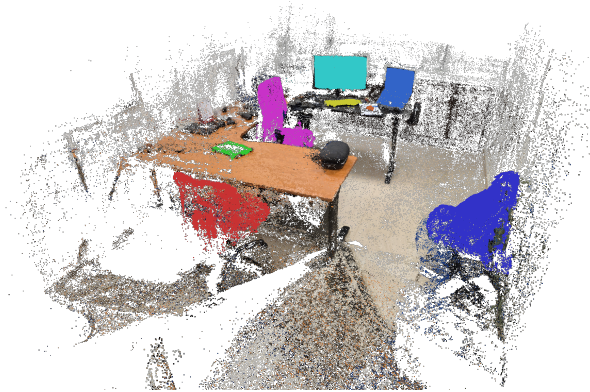


Figure 5. Semantically segmented 3D point cloud generated from 2D YOLOv11-seg predictions.

To generate a semantically segmented surface representation, the labeled point cloud is further linked to the reconstructed mesh. For each mesh vertex, nearby reconstructed points are queried to determine its most likely semantic class. Mesh triangles whose vertices are consistently associated with the same semantic label are then identified and highlighted as belonging to a specific object or scene element (Figure 6). This process yields a

semantic mesh in which object-level regions are explicitly delineated, enabling clear visualization and interaction at the surface level rather than only at sparse point locations.

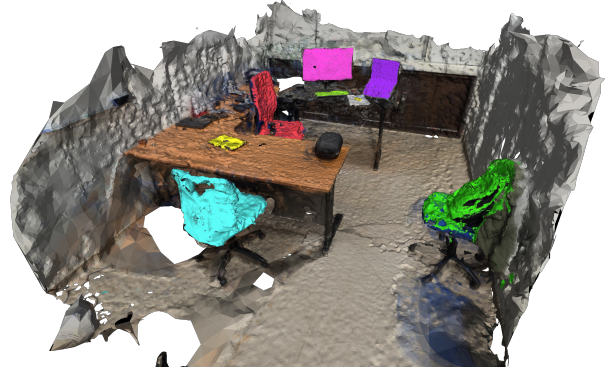


Figure 6. Semantically segmented 3D mesh derived from the segmented point cloud.

The resulting semantically enriched 3D model preserves both, geometry and object-level semantic structure, providing a coherent representation suitable for visualization, interaction, and validation in XR environment. This representation is subsequently transferred back to the HoloLens device, where it can be inspected and manipulated by the user, closing the loop between perception, reconstruction, and semantic scene understanding. Note that this semantic propagation from 2D observations to 3D geometry builds upon well-established approaches (e.g., Semantic-Fusion [8]), and is here adopted as a component within the proposed workflow rather than introduced as a novel contribution.

2.4 XR visualization and user-guided validation of semantic 3D models

After photogrammetric reconstruction and semantic mapping are completed, the semantically enriched 3D model is transferred back to the MR device. The model is then loaded and visualized within the XR environment, allowing the user to inspect the reconstructed scene directly in situ (Figure 7). This visualization step enables the seamless integration of geometry and semantic information within the physical context of the building interior.

Within the XR environment, the reconstructed model is rendered as an interactive 3D object. Semantic information is displayed through color-coded surfaces corresponding to different object classes, enabling intuitive identification of scene elements. Users can freely manipulate the model—such as rotating, translating, or scaling it—to examine its geometry and semantic segmented objects. This interaction facilitates rapid assessment of reconstruction completeness and semantic correctness.

A central feature of the proposed framework is its user-guided validation loop. By visualizing the semantic 3D



Figure 7. Semantically enriched 3D mesh visualized in the XR environment.

model immediately after reconstruction, users can identify missing regions, insufficient coverage, or incorrect semantic assignments while still on site. If deficiencies are observed, additional image data can be acquired using the XR device and incorporated into the reconstruction process, improving both geometric fidelity and semantic completeness. This iterative process reduces the likelihood of incomplete digitalization and minimizes the need for repeated site visits.

Once the user is satisfied with the reconstructed and semantically labeled scene, the workflow can be extended to additional rooms or adjacent spaces, supporting scalable indoor environment digitalization. By enabling immediate feedback and validation, the XR-based visualization closes the loop between data acquisition, AI-driven semantic perception, and 3D reconstruction, transforming digitalization from an offline post-processing task into an interactive, near real-time process.

3 Conclusions

This paper presented an XR-assisted framework for semantic understanding and digitalization of indoor building scenes by integrating on-device AI-based segmentation with photogrammetric reconstruction. By combining near real-time semantic perception, SfM-based 3D modeling, and immersive visualization within a single workflow, the proposed approach enables users to capture, verify, and interact with semantically enriched 3D scene representations directly on site.

Unlike conventional pipelines that decouple data acquisition, semantic interpretation, and reconstruction into separate offline stages, the proposed framework tightly couples these processes through XR. On-device semantic segmentation provides immediate contextual understanding of the environment, while photogrammetric reconstruction and semantic mapping allow this information to be propagated into a spatially consistent 3D representation. The XR-based visualization and user-guided validation loop further support immediate assessment of digital-

ization completeness and semantic correctness, reducing the risk of incomplete models and minimizing the need for repeated site visits.

Despite these advantages, several limitations remain. MR headsets offer limited computational resources compared to desktop or cloud-based platforms, making the deployment of large-scale spaces or large deep learning models challenging. As a result, lightweight segmentation models are required, which may limit achievable accuracy compared to offline processing. In addition, current XR hardware still presents ergonomic constraints, including device weight and comfort during prolonged use, as well as limited battery life that restricts continuous operation time. Environmental factors also affect usability; in particular, high ambient lighting conditions—common in construction sites—can reduce hologram visibility and impact the clarity of XR visualizations. Addressing these hardware limitations is largely dependent on future advances in XR device design and is beyond the scope of this work. This study focuses on demonstrating the feasibility of the proposed integrated workflow and therefore does not include a comprehensive quantitative evaluation, including an assessment of the effectiveness of the user-guided validation loop within the XR environment. Finally, photogrammetric reconstruction and semantic 3D mapping stages are performed offline on an external workstation and depend on the selected pipeline and available computing resources.

While the present study employs a pre-trained model based on the COCO dataset to demonstrate semantic scene understanding, the proposed framework is model-agnostic and can be readily adapted to other segmentation models trained on domain-specific datasets. This flexibility enables extension to a broad range of construction-related applications, including semantic segmentation of structural elements, condition assessment and monitoring of construction activities. Future work will explore native 3D semantic segmentation, tighter integration between reconstruction and inference pipelines, and language-grounded models as interaction mechanisms that allow users to query and manipulate semantic content directly within the XR environment. Additionally, it will include a systematic evaluation of the proposed framework. This will involve assessing 2D segmentation performance, evaluating 3D semantic consistency and reconstruction completeness, and measuring end-to-end latency of the pipeline. In addition, a user study will be conducted to quantify the impact of the XR-based validation loop on usability and data acquisition efficiency. Together, these directions highlight the potential of XR as an active tool for near real-time scene understanding, digitalization, and decision support in the built environment.

Acknowledgements

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (Project ID: RS52/25 — “Spatio-temporal visual perception and AI-agent-based scene representation for robotic support in construction”).

References

- [1] B G Pantoja-Rosero, A Rusnak, F Kaplan, and K Beyer. Generation of lod4 models for buildings towards the automated 3d modeling of bims and digital twins. *Automation in Construction*, 168:105822, 2024. ISSN 0926-5805. doi:10.1016/j.autcon.2024.105822. URL <https://www.sciencedirect.com/science/article/pii/S0926580524005582>.
- [2] Bryan G Pantoja-Rosero, Radhakrishna Achanta, and Katrin Beyer. Damage-augmented digital twins towards the automated inspection of buildings. *Automation in Construction*, 150:104842, 2023. doi:10.1016/j.autcon.2023.104842. URL <https://www.sciencedirect.com/science/article/pii/S0926580523001024>.
- [3] B G Pantoja-Rosero, R Achanta, M Kozinski, P Fua, F Perez-Cruz, and K Beyer. Generating lod3 building models from structure-from-motion and semantic segmentation. *Automation in Construction*, 141, 2022. doi:10.1016/j.autcon.2022.104430. URL <https://www.sciencedirect.com/science/article/pii/S092658052200303X?via%3Dihub>.
- [4] A Chassignet and R Achanta. Partial annotations in active learning for semantic segmentation. *Automation in Construction*, 168:105828, 2024. doi:10.1016/j.autcon.2024.105828. URL <https://www.sciencedirect.com/science/article/pii/S0926580524005648>.
- [5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017. doi:10.1109/CVPR.2017.261.
- [6] Bryan G Pantoja-Rosero, D Oner, Mateusz Kozinski, Radhakrishna Achanta, Pascal Fua, Fernando Pérez-Cruz, and Katrin Beyer. Topo-loss for continuity-preserving crack detection using deep learning. *Construction and Building Materials*, 344:128264, 2022. doi:10.1016/j.conbuildmat.2022.128264.
- URL <https://www.sciencedirect.com/science/article/pii/S0950061822019250>.
- [7] Alexander Rusnak, Bryan G Pantoja-Rosero, Frédéric Kaplan, and Katrin Beyer. Automatic removal of non-architectural elements in 3d models of historic buildings with language embedded radiance fields. *Heritage*, 7(6):3332–3349, 2024. doi:10.3390/heritage7060157. URL <https://www.mdpi.com/2571-9408/7/6/157>.
- [8] John McCormac, Ankur Handa, Andrew J. Davison, and Stefan Leutenegger. Semanticfusion: Dense 3D semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4635, 2017. doi:10.1109/ICRA.2017.7989538.
- [9] Martin Runz, Maud Bufier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20, 2018. doi:10.1109/ISMAR.2018.00024.
- [10] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. doi:10.1109/CVPR.2016.445.
- [11] Carsten Griwodz, Simone Gasparini, Lilian Calvet, Pierre Gurdjos, Fabien Castan, Benoit Maugean, Gregoire De Lillo, and Yann Lanthony. Alice-vision meshroom: An open-source 3d reconstruction pipeline. *Proceedings of the 12th ACM multimedia systems conference*, pages 241–247, 2021. doi:10.1145/3458305.3478443. URL <https://dl.acm.org/doi/10.1145/3458305.3478443>.
- [12] Joseph Rowell. Semantic validation in structure from motion. *arXiv preprint*, 2023. doi:<https://doi.org/10.48550/arXiv.2304.02420>.
- [13] Abhijit Kundu, Yin Li, Fuxin Li, and James M. Rehg. Joint semantic segmentation and 3D reconstruction from monocular video. In *Computer Vision – ECCV 2014*, pages 703–718. Springer, 2014. doi:10.1007/978-3-319-10599-4_45.
- [14] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. SAM3D: Segment anything in 3D scenes. *arXiv preprint*, 2023. doi:<https://doi.org/10.48550/arXiv.2306.03908>.
- [15] Shuyuan Chen, Zhaoqing Wang, Jiepeng Wang, Yuexin Ma, Mingming Gong, Wenping Wang, and

Tongliang Liu. Panoslam: Panoptic 3D scene reconstruction via gaussian SLAM. *arXiv preprint*, 2024. doi:<https://doi.org/10.48550/arXiv.2501.00352>.

- [16] J.I Fan and K Khoshelham. Augmented reality asset tracking using hololens. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-4-2021:121–127, 2021. doi:10.5194/isprs-annals-V-4-2021-121-2021.
- [17] Juan Carlos Navares-Vázquez, Zhouyan Qiu, Pedro Arias, and Jesús Balado. Hololens 2 performance analysis for indoor/outdoor 3d mapping. *Journal of Building Engineering*, 108:112826, 2025. ISSN 2352-7102. doi:<https://doi.org/10.1016/j.jobbe.2025.112826>.
- [18] B.G. Pantoja-Rosero and S. Salamone. Integrating extended reality and ai-based damage segmentation for near real-time, traceable bridge inspections. *Automation in Construction*, 180:106567, December 2025. doi:10.1016/j.autcon.2025.106567. URL <https://doi.org/10.1016/j.autcon.2025.106567>.
- [19] Boan Tao, Frédéric Bosché, and Jiajun Li. Mixed reality-based mep construction progress monitoring: Evaluation of methods for mesh-to-mesh comparison. *Automation in Construction*, 168:105852, dec 2024. doi:10.1016/j.autcon.2024.105852.
- [20] David DL Mascareñas, JoAnn P Ballor, Oscar L McClain, Miranda A Mellor, Chih-Yu Shen, Brian Bleck, John Morales, Li-Ming R Yeong, Benjamin Narushof, Philo Shelton, Eric Martinez, Yongchao Yang, Alessandro Cattaneo, Troy A Harden, and Fernando Moreu. Augmented reality for next generation infrastructure inspections. *Structural Health Monitoring*, 20(4):1957–1979, 2021. doi:10.1177/1475921720953846.
- [21] Mahta Zakaria, Enes Karaaslan, and F Necati Catbas. Advanced bridge visual inspection using real-time machine learning in edge devices. *Advances in Bridge Engineering*, 3:27, 2022. ISSN 2662-5407. doi:10.1186/s43251-022-00073-y. URL <https://doi.org/10.1186/s43251-022-00073-y>.
- [22] Kaveh Malek, Ali Mohammadkhorasani, and Fernando Moreu. Methodology to integrate augmented reality and pattern recognition for crack detection. *Computer-Aided Civil and Infrastructure Engineering*, 38(8):1000–1019, 2023. doi:10.1111/mice.12932.