Single-image to model registration for semantic enrichment of indoor BIM

Dorota Włodarczyk , Małgorzata Jarząbek-Rychard

Institute of Geodesy and Geoinformatics, Wrocław University of Environmental and Life Sciences, Poland – dorota.wlodarczyk@upwr.edu.pl, malgorzata.jarzabek-rychard@upwr.edu.pl

Keywords: indoor mapping, single-image registration, Mask2Former, panoptic segmentation, line segments, Structured3D

Abstract

Effective integration of geometric and semantic data within Building Information Models (BIM) is essential for the efficient life cycle management of modern facilities. However, maintaining accurate as-is BIM models for existing buildings remains a significant challenge, as manual updates are labour-intensive and full 3D reconstruction is often impractical for incremental changes. In such cases, image-based approaches offer a fast and flexible alternative, but require reliable alignment of 2D imagery with existing BIM geometry. To address this challenge, this study introduces a streamlined pipeline for semantic enrichment that uses a single-image visual localisation approach to directly align 2D imagery with existing BIM geometry. The proposed method integrates transformer-based panoptic segmentation (Mask2Former) with a closed-form Perspective-n-Line solver to estimate 6-degrees-of-freedom (6-DoF) camera poses. The novelty of the proposed approach lies in the explicit use of semantic information as a geometric constraint to guide the selection of 2D–3D correspondences for pose estimation. Semantic labels are utilised to filter line correspondences, ensuring that only stable architectural boundaries (e.g., walls, floors, and ceilings) are used in the registration process. Such semantic filtering stabilises correspondence selection, effectively mitigating pose ambiguity in repetitive indoor layouts or scenes where structural elements are partially obscured by furniture and clutter. Experimental results confirm high accuracy, achieving a median position error of 9.84 cm and an orientation error of 1.05° in complex indoor environments. This robust registration framework provides a reliable foundation for the downstream semantic enrichment and digital twin updates.

1. Introduction

The Building Information Model (BIM) provides a digital object-oriented description of a facility, in which geometry is linked to computable semantic attributes (Mirarchi et al., 2024). The as-is BIM model is the most current representation of the object and accompanies the building throughout its operations and maintenance. In practice, the as-is BIM model is obtained by updating the as-built model and keeping the as-is state up to date whenever the real-world physical object changes. For most existing buildings, however, the models are often outdated, and their updates remain time-consuming, labour-intensive, and prone to errors (Schönfelder et al., 2023). Recent research on building model updating has primarily focused on geometric reconstruction, while semantic attributes and relations have received comparatively less attention (Mirarchi et al., 2024). This imbalance motivates the development of methods that enrich as-is models by extracting component attributes and structure-aware relations.

Methods for maintaining as-is BIM models can be categorized based on the update objective and the available sensor data. When extensive structural changes occur or no prior model exists, full reconstruction is typically employed, generating new geometry from dense LiDAR point clouds (Collins et al., 2024; Schönfelder et al., 2023). For incremental updates, differential change detection registers new scans against the existing BIM to identify and correct geometric discrepancies (Meyer et al., 2022; Tran and Khoshelham, 2019). While these approaches primarily address geometric consistency, many practical applications require updating not only geometry but also semantic information, like object types or functional relations within the model (Mirarchi et al., 2024). Although such semantic enrichment can be performed using point clouds, it often relies on complementary data sources when a complete re-scan is impractical or prior clouds are unavailable, particularly when the model was originally developed from CAD designs or archival plans (Gülch and Obrock, 2020). In these cases, imagery provides a more cost-effective and flexible alternative for updating the semantic layer, as it avoids the high equipment expenses and labour-intensive registration associated with 3D scanning (Jarząbek-Rychard and Maas, 2022). Image to model registration allows semantic data extracted from single-view imagery to be reprojected onto BIM elements without the need for full-scale 3D reconstruction (Acharya et al., 2019).

This registration process relies on precise camera localisation, often formulated as 6-degrees-of-freedom (6-DoF) pose estimation from a single image (Gaudillière et al., 2019). While existing BIM models provide valuable geometric and semantic constraints that can reduce correspondence ambiguity during image-to-model alignment (Jarząbek-Rychard and Maas, 2022; Gaudillière et al., 2019), this task remains challenging in indoor environments. In particular, occlusions, varying lighting conditions, and similar architectural patterns often support multiple plausible pose hypotheses for visually similar architectural elements (Gard et al., 2024; Taira et al., 2019). Conventional feature-based (Taira et al., 2019) and regression-based (Acharya et al., 2019; Li et al., 2018) localisation methods often become unreliable in indoor environments, particularly in low-texture scenes where unique visual cues are limited. In such settings, purely geometric matching can become unstable when structural redundancy, partial visibility, or weak texture reduces the distinctiveness of 2D-3D correspondences (Sarlin et al., 2021; Li et al., 2018). Yet, a key limitation persists: semantic outputs from scene understanding are rarely utilised as geometric priors to stabilise 2D-3D correspondence selection, leading to poor pose accuracy in repetitive or cluttered indoor layouts (Gard et al., 2024; Sarlin et al., 2021; Taira et al., 2019). This gap highlights the need for semantically informed selection strategies that mitigate spatial ambiguities and improve the robustness of subsequent geometric alignment.

The objective of this study is, therefore, to enhance the robustness of single-image-to-BIM registration by integrating structure-aware constraints extracted from the semantic layout of indoor

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-575-2026 | © Author(s) 2026. CC BY 4.0 License.

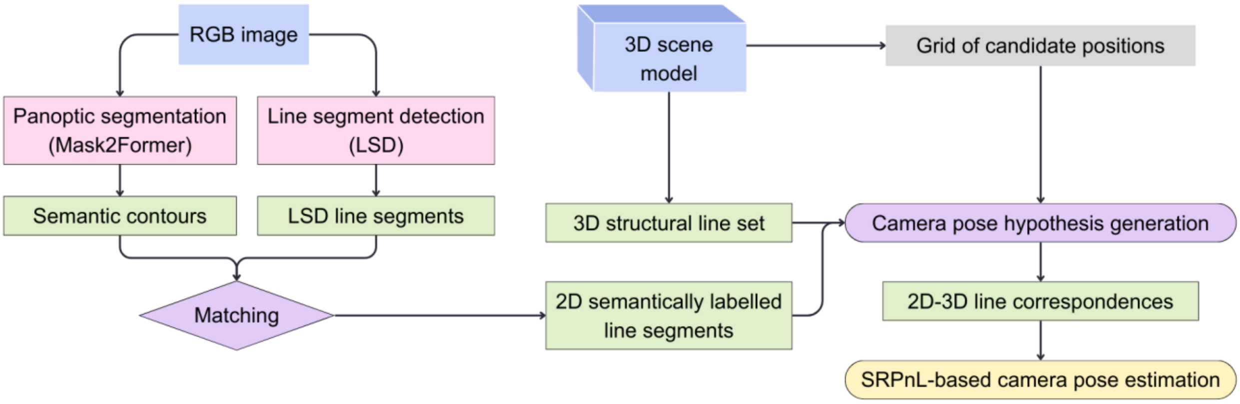

Figure 1. Processing pipeline: from RGB input to estimated camera pose.

scenes. To achieve this, we employ transformer-based panoptic segmentation (Cheng et al., 2022) to generate detailed panoptic maps, which are subsequently converted into structural edge constraints for 2D-3D alignment. The novelty of the proposed approach lies in the explicit use of panoptic-derived semantic contours as geometric priors to stabilise 2D-3D correspondence selection. By bridging the gap between scene understanding and geometric localisation, this method ensures that only stable architectural boundaries are used to recover the camera pose in complex indoor environments. The primary contributions of this research are as follows: (i) development of an end-to-end pipeline that aligns interior imagery with existing BIM models through deep panoptic segmentation and geometric reasoning; (ii) introduction of a method for generating structure-aware edge constraints to mitigate ambiguity in repetitive layouts; (iii) validation of the accuracy and efficiency of this registration approach in complex indoor scenes. By evaluating the resulting image-to-model registration, the study assesses its potential to support future automation of semantic enrichment and updates in digital as-is BIM models.

2. Method

The proposed workflow (Figure 1) integrates semantic scene understanding with geometry-based registration to recover camera poses in indoor BIM environments. The pipeline combines parallel image analysis, semantic-geometric matching, hypothesis generation from the BIM model, and final 2D-3D alignment for pose estimation.

2.1 Panoptic segmentation and post-processing

The panoptic segmentation stage is based on the Mask2Former architecture (Cheng et al., 2022). This approach utilises a ResNet-50 backbone (He et al., 2016) to facilitate multiscale feature extraction and incorporates weights originally obtained from pretraining on the COCO panoptic segmentation dataset (Kirillov et al., 2019). To further adapt this pretrained network to the specific semantic and geometric characteristics of indoor spatial layouts, we performed additional fine-tuning on the Structured3D dataset (Zheng et al., 2020). Following this adaptation, the optimised model is employed to generate panoptic maps that provide a unified representation of the scene, integrating both pixel-level class assignments (semantic segmentation) and instance-level descriptors (instance segmentation). To address void pixels left by ambiguous contours and ensure topological consistency, an iterative 8-neighbour majority fill is applied with a deterministic tie-break criterion, replacing the panoptic map once convergence is reached (Figure 2a).

From the filled panoptic map, we extract semantic contours (Figure 2b), retaining only structural elements (e.g., walls, floors, ceilings, windows, doors) to obtain a compact description of the room envelope. The filtering process identifies all inter-segment boundaries and preserves only those where both adjacent labels belong to the structural class set, effectively excluding edges associated with furniture or temporary objects. These contours are represented as polylines labelled with semantic adjacency information (e.g., wall–ceiling junctions), providing context to reduce matching ambiguity in repetitive layouts. Due to raster discretization, semantic contours may be spatially coarse; thus, we refine them using the Line Segment Detector (LSD, Grompone von Gioi et al., 2010). LSD extracts high-precision line segments directly from the original RGB image (Figure 2c). These segments are then aligned with the semantically labelled polylines. Matched segments inherit the semantic labels and are merged into longer, more stable primitives (Figure 2d). This refinement reduces fragmentation and enhances robustness for the subsequent 2D-3D matching process.

2.2 Camera pose estimation

Correspondences are established between labelled 2D line sets and 3D structural primitives derived from the BIM model. Camera pose (comprising the rotation matrix and translation vector) is estimated using the Perspective-n-Line (PnL) formulation (Xu et al., 2017). The selection of a PnL-based approach is motivated by the specific challenges of indoor environments, where structural contours are frequently fragmented or partially obscured by furniture and decor. Unlike Perspective-n-Point methods, which require matching point correspondences with identical coordinates in both 2D and 3D, the PnL approach leverages the geometric incidence between 2D segments and 3D lines (Přibyl et al., 2017; Xu et al., 2017). Because the registration relies on the underlying line equations (the infinite line relationship) instead of exact segment endpoints, the solver can recover the 6-DoF pose even when only disconnected fragments of a wall or ceiling junction are visible (Přibyl et al., 2017; Wang et al., 2019).

By integrating semantic labels into the matching process, the search space is restricted to semantically consistent pairs, ensuring, for instance, that 3D ceiling-wall contours are matched exclusively with 2D lines labelled as "ceiling-wall." This semantic filtering suppresses geometrically plausible but

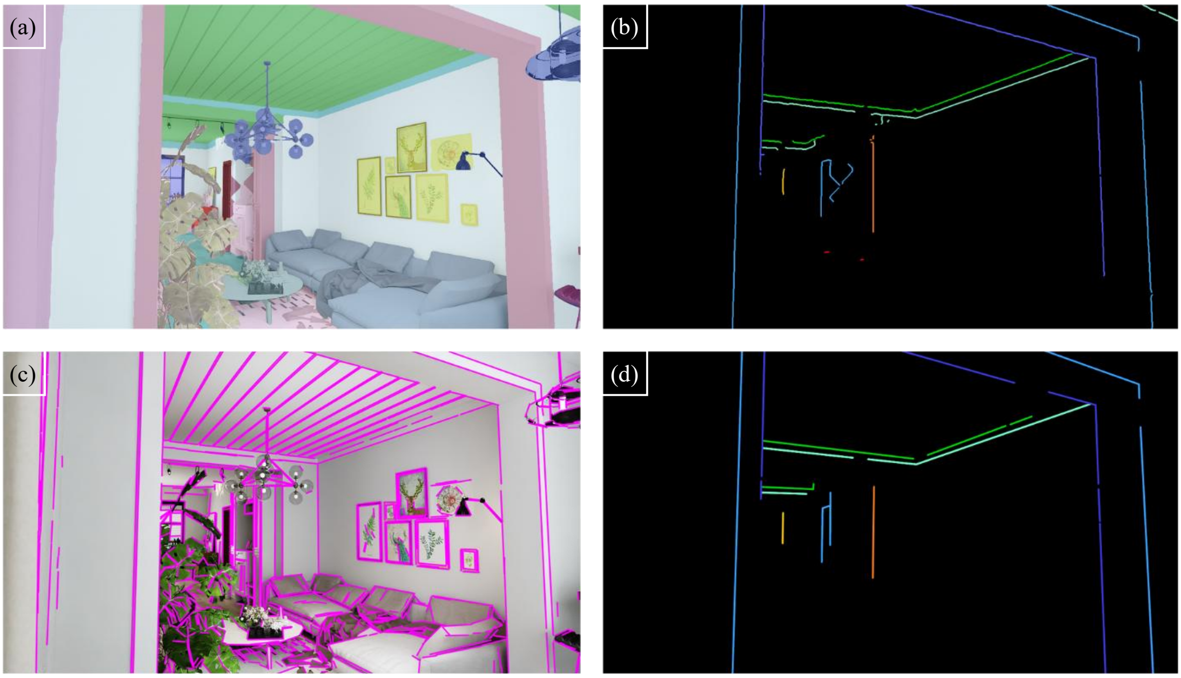

Figure 2. Extraction of semantically labelled structural edges: (a) panoptic map derived from Mask2Former, after void-filling, (b) labelled structural contours extracted from the panoptic map, (c) line segments detected from the RGB image via LSD, (d) final semantically labelled line primitives after matching.

semantically inconsistent matches, thus achieving a more reliable 6-DoF localisation even in low-texture or ambiguous indoor scenes (Jarząbek-Rychard and Maas, 2022).

The pose recovery process follows a coarse-to-fine strategy. First, a discrete hypothesis generation phase defines a regular grid over the room layout at a fixed height. To account for images captured from adjacent spaces, the grid extends beyond the room boundaries. Candidate locations are sampled every 0.5 m across this grid, with yaw angles discretized at 2° increments. Pitch and roll remain fixed assuming an upright camera with known intrinsics.

For each hypothesis, 3D structural primitives from the BIM are projected onto the image plane and matched against the extracted 2D line set. A list of compatible 2D candidates is constructed for each 3D line based on geometric and semantic similarity. The final correspondence set is resolved through a global one-to-one assignment, in which each 2D and 3D line is used once. Matches are retained only if they satisfy both geometric and semantic criteria. Configurations with fewer than 4 matched lines are discarded to ensure numerical stability in the solver. Moreover, within this set, there must be at least 2 horizontal and 1 vertical line to avoid nearly parallel lines, thereby ensuring sufficient geometric diversity for stable pose estimation. It should be noted that these grid-based hypotheses are used solely to guide correspondence filtering and are not provided as an initial pose for the final solver.

The final 6-DoF camera pose is recovered using the SRPnL solver (Wang et al., 2019), chosen for its computational efficiency and robustness against measurement noise. Unlike iterative approaches that may converge to local minima, this closed-form formulation provides a mathematically stable mapping from 2D-3D line correspondences to the camera pose (Wang et al., 2019; Xu et al., 2017). The solver's linear complexity is particularly advantageous for indoor scenes where numerous structural boundaries are extracted, ensuring high performance even as the number of line segments increases (Wang et al., 2019).

2.3 Ablation analysis

To quantify the impact of semantic constraints on registration stability, we conduct an ablation study comparing the proposed method against two reduced variants:

a) Full Semantic (SEM): The proposed baseline, which uses all available semantic adjacency labels (walls, floors, ceilings, windows, and doors) to restrict the matching search space.
b) Reduced Semantic (SEM-RED): An ablation variant where windows and doors are excluded from the process. Matching is restricted to the most stable room-envelope boundaries (walls, floor, ceiling) to evaluate the influence of less stable contours that are prone to distortion, e.g., windows obscured by curtains or door frames.
c) Geometry-only (GEOM): A second ablation variant where semantic labels are entirely ignored. Correspondences are established using only geometric line properties, representing a traditional PnL-based registration approach.

In the semantic variants, horizontal structural constraints (floor and ceiling) are derived directly from the contour labels, whereas vertical ones (wall–wall junctions) are identified analogously. In the GEOM variant, these must be inferred solely from the 2D image configuration (line orientation). This design allows for analysis of how semantic specificity and label-restricted matching contribute to the overall robustness of the registration pipeline.

2.4 Evaluation metrics

The proposed method is evaluated using a two-stage protocol that focuses on the quality of the semantic predictions and the precision of the resulting single-image registration to the BIM model.

For panoptic outputs, we report the Panoptic Quality (PQ) metric (Kirillov et al., 2019) computed as the product of two components: (i) Segmentation Quality (SQ), which measures shape fidelity as the average Intersection over Union (IoU) for all correctly matched segments; and (ii) Recognition Quality (RQ), which assesses instance-level detection performance after one-to-one matching, increasing with correct matches and decreasing with misses and duplicates. Furthermore, we report the mean Intersection over Union (mIoU) to evaluate the pixel-level semantic labelling accuracy across all classes.

For registration, we report the accuracy of the 6-DoF camera pose recovery using translation error (Te [cm]), which measures the Euclidean distance between the estimated and ground-truth camera centres, and rotation error (Re [deg]), which quantifies the geodesic angular distance between the estimated rotation matrix and the reference orientation.

2.5 Dataset

The Structured3D dataset (Zheng et al., 2020) provides the foundation for our experimental evaluation, offering a large-scale collection of photo-realistic synthetic indoor scenes with corresponding 3D structural layouts. Following the standard protocol established by the dataset creators, the data was partitioned into 3,000 scenes for training, 250 for validation, and 250 for testing. The Mask2Former model (Cheng et al., 2022) was fine-tuned on the full training set to ensure robust performance across diverse indoor environments.

The camera pose estimation accuracy was evaluated on a targeted subset of the test data. This evaluation involved a detailed analysis of 6 representative rooms across 4 different scenes, selected to assess the registration accuracy in environments with varying levels of furniture density and structural complexity.

3. Results and discussion

On the Structured3D test dataset, the fine-tuned panoptic model achieved a PQ of 59.7, SQ of 86.5, RQ of 66.2, and an mIoU of 66.6. High SQ indicates that mask shapes align well with ground truth, while the lower RQ suggests that most residual errors come from instance recognition, such as missed or duplicated objects. For reference, the pretrained Mask2Former model adopted in this study was originally reported by Cheng et al. (2022) with 51.9 PQ and 61.7 mIoU for the corresponding configuration on COCO panoptic. After fine-tuning on Structured3D, the model achieved higher PQ and mIoU values, indicating effective adaptation to indoor scenes.

Despite the presence of numerous non-structural elements, such as furniture and decorations, the extracted structural contours remain accurate and spatially consistent (Figure 3). Their geometric quality enables reliable matching with LSD-derived

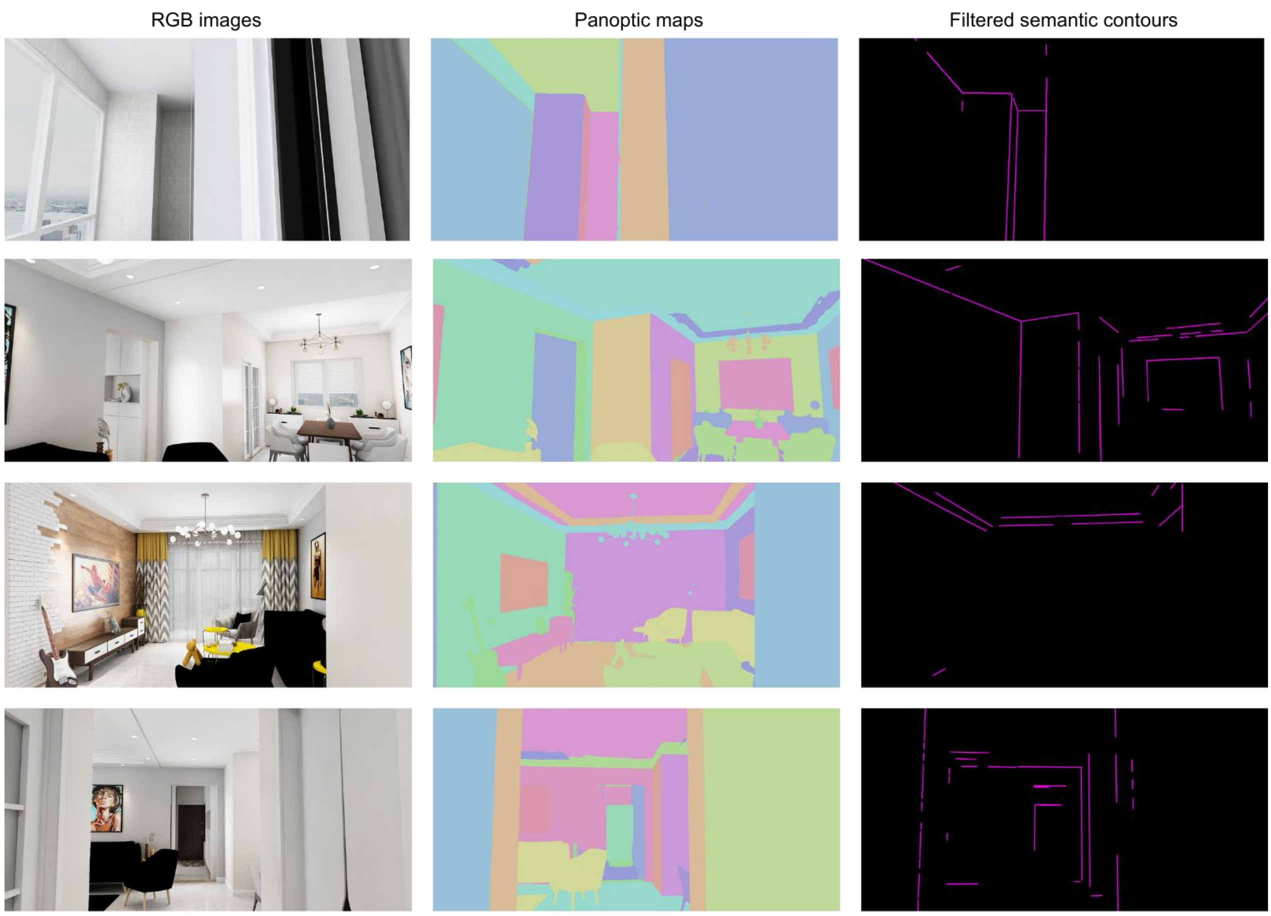

Figure 3. Examples of RGB input images, panoptic segmentation maps, and extracted semantically labelled structural line segments.

line segments, yielding semantically labelled line primitives suitable for subsequent 2D-3D registration.

The comparative analysis of the registration variants defined in Section 2.3 demonstrates a clear hierarchy, as illustrated in Figure 4 and Figure 5. Both semantic-based approaches consistently outperform the GEOM baseline, with the SEM-RED configuration achieving the highest precision and the lowest median errors for both translation and rotation. Specifically, the median Te decreased from 18.83 cm to 12.90 cm (SEM) and 9.84 cm (SEM-RED). A similar trend is observed in rotation accuracy, where the median Re was reduced from 2.19° to 1.63° (SEM) and further to 1.05° (SEM-RED).

This marked reduction in localisation and orientation errors confirms that semantic constraints effectively stabilize the selection of 2D-3D correspondences. While the SEM variant provides a significant improvement over purely geometric matching by using all semantic labels, the best performance of SEM-RED suggests that focusing on the most stable room-envelope boundaries provides an optimal balance between semantic selectivity and robust matches.

Visual examples of the projected overlays, shown in Figure 6, confirm these quantitative results. Semantic variants show improved alignment of projected 3D BIM lines compared to 3D ground-truth projection. In contrast, the GEOM scenario is more

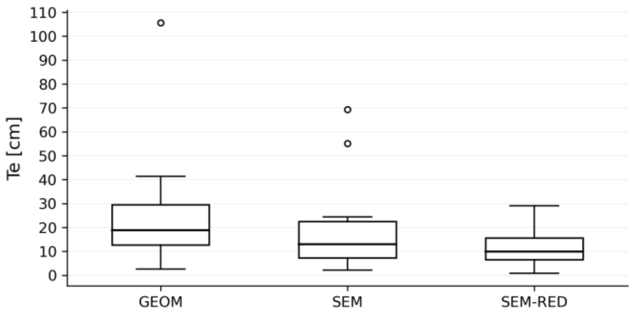

Figure 4. Distribution of translation errors (Te) for the evaluated matching variants between the estimated and reference camera localisation.

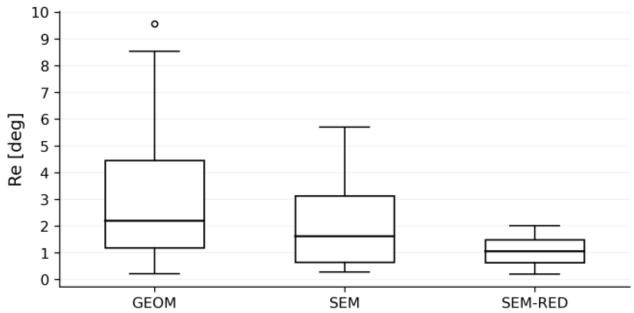

Figure 5. Distribution of rotation errors (Re) for the evaluated matching variants between the estimated and reference camera orientations.

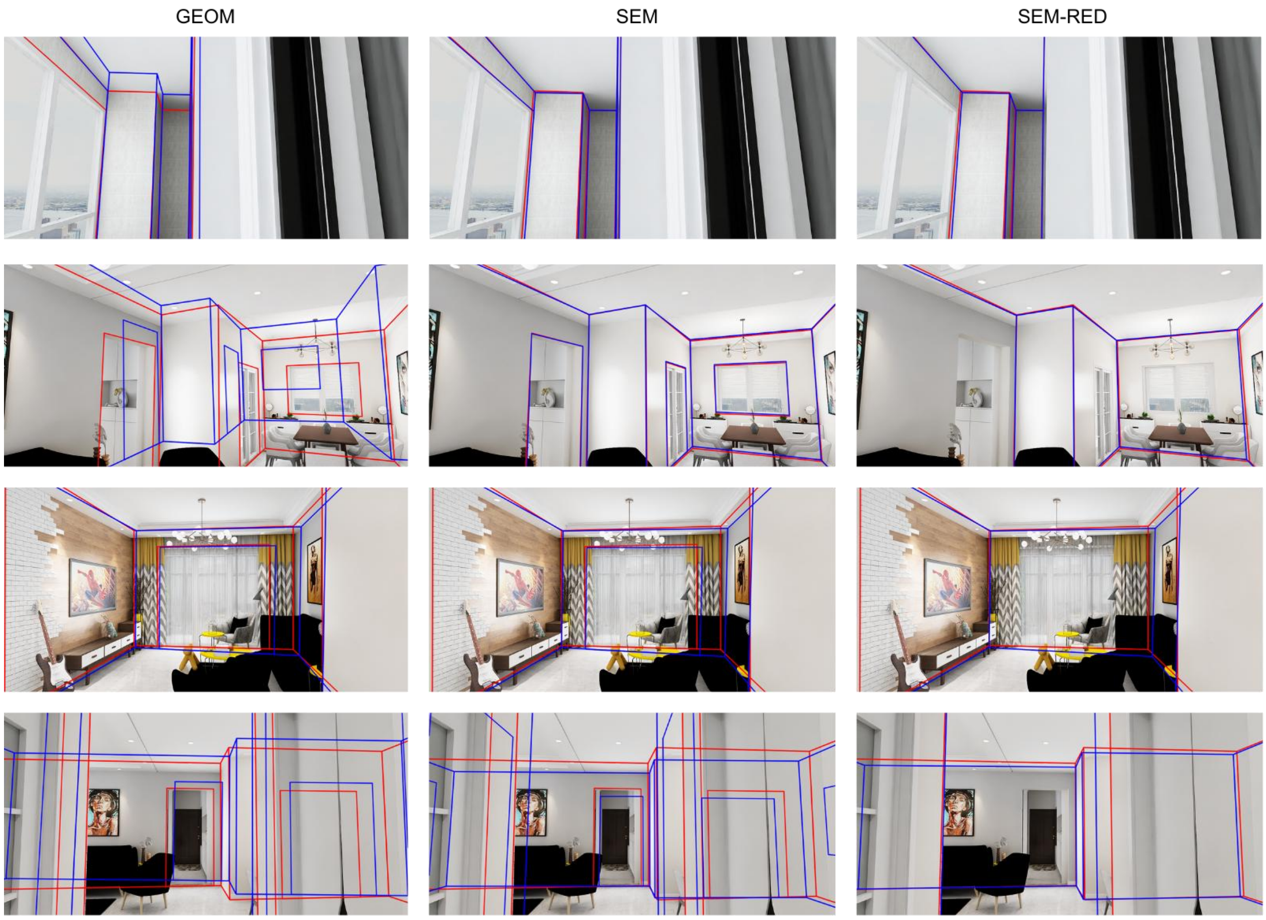

Figure 6. Examples of single-image registration to BIM results shown as room-layout reprojections on the input RGB images. Red lines represent the ground-truth room layout projection; blue lines represent the projection obtained from the estimated camera pose.

prone to misalignments from repetitive features or weak textures, often keeping implausible matches. The SEM-RED variant gives the most stable visuals by focusing on reliable room-envelope boundaries.

4. Conclusions

This study presented a method for single-image to BIM registration by integrating transformer-based panoptic segmentation with geometric pose estimation. By using semantic labels to constrain the correspondence search space, the proposed method mitigates errors from misleading matches in challenging indoor environments. The experimental results confirmed that semantic constraints significantly enhance registration stability. The ablation analysis demonstrated a clear performance hierarchy, with the reduced-semantic variant achieving the highest accuracy by prioritising stable room-envelope boundaries.

A key factor in achieving high registration accuracy (median errors of 9.84 cm and 1.05°) was the use of the state-of-the-art Mask2Former model, which enabled high-fidelity segmentation contours and improved structural edge detection. The integration of the SRPnL solver enabled efficient and stable pose recovery from dense line sets without needing an initial camera pose. The experiments confirm that incorporating semantic information into camera pose estimation yields more stable results, overcoming the limitations of purely geometric registration.

Future work will focus on two main directions. First, the method will be validated in real indoor environments using smartphone imagery to assess its robustness under practical conditions, such as variable lighting and motion blur. Second, we intend to leverage the estimated camera poses to automate the semantic enrichment of 3D models. By reprojecting identified objects and structural elements from the registered images directly into the BIM space, our system could effectively detect spatial discrepancies or missing instances, providing a foundation for the automated semantic enrichment of as-is BIM models.

Acknowledgements

The research was supported by the Wrocław University of Environmental and Life Sciences (Poland) as part of the research project no. N060/0002/24.

References

Acharya, D., Khoshelham, K., Winter, S., 2019. BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 245–258. doi.org/10.1016/j.isprsjprs.2019.02.020.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299. doi.org/10.1109/CVPR52688.2022.00135.

Collins, F. C., Noichl, F., Slepicka, M., Cones, G., Borrmann, A., 2024. 360-degree tri-modal scanning: Engineering a modular multi-sensor platform for semantic enrichment of BIM models. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1, 41–48. doi.org/10.5194/isprs-annals-X-1-2024-41-2024.

Gard, N., Hilsmann, A., Eisert, P., 2024. SPVLoc: Semantic panoramic viewport matching for 6D camera localization in unseen environments. *arXiv preprint*. doi.org/10.48550/arXiv.2404.10527.

Gaudillière, V., Simon, G., Berger, M.-O., 2019. Camera pose estimation with semantic 3D model. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4569–4576. IEEE. doi.org/10.1109/iros40897.2019.8968180.

Grompone von Gioi, R., Jakubowicz, J., Morel, J.-M., Randall, G., 2010. LSD: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4), 722–732. doi.org/10.1109/TPAMI.2008.300.

Gülch, E., Obrock, L., 2020. Automated semantic modelling of building interiors from images and derived point clouds based on deep learning methods. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 421–426. doi.org/10.5194/isprs-archives-XLIII-B2-2020-421-2020.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE. doi.org/10.1109/CVPR.2016.90.

Jarząbek-Rychard, M., Maas, H.-G., 2022. Automatic enrichment of indoor 3D models using a deep learning approach based on single images with unknown camera poses. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, VIII-5/W1-2022, 1–7. doi.org/10.5194/isprs-annals-VIII-5-W1-2022-1-2022.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P., 2019. Panoptic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9396–9405. doi.org/10.1109/CVPR.2019.00963.

Li, X., Ylioinas, J., Kannala, J., 2018. Full-frame scene coordinate regression for image-based localization. *arXiv preprint*. doi.org/10.48550/arXiv.1802.03237.

Meyer, T., Brunn, A., Stilla, U., 2022. Change detection for indoor construction progress monitoring based on BIM, point clouds and uncertainties. *Automation in Construction*, 141, 104442. doi.org/10.1016/j.autcon.2022.104442.

Mirarchi, C., Gholamzadehmir, M., Daniotti, B., Pavan, A., 2024. Semantic enrichment of BIM: The role of machine learning-based image recognition. *Buildings*, 14(4), 1122. doi.org/10.3390/buildings14041122.

Přibyl, B., Zemčík, P., Čadík, M., 2017. Absolute pose estimation from line correspondences using direct linear transformation. *Computer Vision and Image Understanding*, 161, 130–144. doi.org/10.1016/j.cviu.2017.05.002.

Sarlin, P.-E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., Sattler, T., 2021. Back to the feature: Learning robust camera localization from pixels to pose. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3246–3256. doi.org/10.1109/CVPR46437.2021.00326.

Schönfelder, P., Aziz, A., Faltin, B., König, M., 2023. Automating the retrospective generation of as-is BIM models using machine learning. *Automation in Construction*, 152, 104937. doi.org/10.1016/j.autcon.2023.104937.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Šivic, J., Pajdla, T., Torii, A., 2019. InLoc: Indoor visual localization with dense matching and view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4), 1293–1307. doi.org/10.1109/TPAMI.2019.2952114.

Tran, H., Khoshelham, K., 2019. Building change detection through comparison of a lidar scan with a building information model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W13, 889–893. doi.org/10.5194/isprs-archives-XLII-2-W13-889-2019.

Wang, P., Xu, G., Cheng, Y., Yu, Q., 2019. Camera pose estimation from lines: A fast, robust and general method. *Machine Vision and Applications*, 30(4), 603–614. doi.org/10.1007/s00138-019-01012-0.

Xu, C., Zhang, L., Cheng, L., Koch, R., 2017. Pose estimation from line correspondences: A complete analysis and a series of solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1209–1222. doi.org/10.1109/TPAMI.2016.2582162.

Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z., 2020. Structured3D: A large photo-realistic dataset for structured 3D modeling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), *Computer Vision – ECCV 2020*, pp. 519–535. Springer. doi.org/10.1007/978-3-030-58545-7_30.