Global localization for Mixed Reality visualization using wireframe extraction from images

Sajjad Einizinab[1,2,*], Kourosh Khoshelham[1,2], Stephan Winter[2], Philip Christopher[1,2]

[1] Building 4.0 CRC , Caulfield East, 3145, Victoria, Australia
[2] Department of Infrastructure Engineering, University of Melbourne, Parkville, 3010, Victoria, Australia
(einizinabs, k.khoshelham, winter, pbc)@unimelb.edu.au

Keywords: Mixed Reality, Localization, Deep Learning, Wireframe, BIM, Indoor Positioning

Abstract

Mixed Reality (MR) global localization involves precisely tracking the device's position and orientation within a digital representation, such as Building Information Model (BIM). Existing model-based MR global localization approaches have difficulty addressing environmental changes between the BIM and real-world, particularly in dynamic construction sites. Additionally, a significant challenge in MR systems arises from localization drift, where the gradual accumulation of positional errors over time can lead to inaccuracies in determining the device's position and orientation within the virtual model. We develop a method that extracts structural elements of the building, referred to as a wireframe, which are less likely to change due to their inherent permanence. The extraction of these features is computationally inexpensive enough that can be performed on MR device, ensuring a reliable and continuous global localization over time, thereby overcoming issues associated with localization drift. The method incorporates a deep Convolutional Neural Network (CNN) to extract the 2D wireframes from images. The reconstruction of 3D wireframes is achieved by utilizing the extracted 2D wireframe along with their depth information. The simplified 3D wireframe is subsequently aligned with the BIM. Real-world experiments demonstrate the method's effectiveness in 3D wireframe extraction and alignment with the BIM, successfully mitigating drift issues by 4cm in prolonged corridor scans.

1. Introduction

Mixed Reality (MR) integrates virtual objects into the real-world, allowing seamless interaction between virtual and physical elements, distinguishing itself from Augmented Reality (AR), which merely overlays virtual content on top of the real-world (Kopsida, 2018). The real-time feedback and digital interaction capabilities of MR systems make them highly suitable for many applications, such as building work inspection, ensuring that constructed elements comply with design specifications (Einizinab et al., 2023b). MR provides the means to align the Building Information Model (BIM) with the real building, simplifying the inspection of corresponding elements between BIM and the physical structure (Radanovic et al., 2023a). A precise superimposition of the BIM onto the real building for MR visualisation is essential for a reliable inspection of building works (Kopsida and Brilakis, 2017, Einizinab et al., 2023b).

In the AR/MR technology, localization of the device camera involves the estimation of its pose (position and orientation) within a reference coordinate system. Two types of localization methods exist: local and global. The local method, such as Simultaneous Localization and Mapping (SLAM) and visual odometry, estimates the pose of the camera with respect to a previous pose (Radanovic et al., 2023a). This task constructs the 3D model in reference to a local coordinate system and cannot be directly used to position the virtual model over corresponding real-world locations unless the first camera pose is determined with respect to the virtual model. However, it suffers from drift. Global localization, often referred to as the alignment between the virtual model and the real world, involves estimating the pose of the camera with respect to a global map or model of the environment. Alignment challenges are common in AR/MR, leading to spatial discrepancies between the real-world and virtual models (Kopsida, 2018). This task becomes significantly more challenging in indoor environments without access to external sensors such as Global Navigation Satellite Systems (GNSS) (Radanovic et al., 2023b).

A common strategy for the alignment involves an initial coarse alignment, which roughly overlays the digital model onto the real-world, followed by a subsequent fine alignment to refine and optimize the initial alignment (Vermandere et al., 2022). Coarse alignment is achieved through the absolute localization of the camera within the virtual model (Acharya et al., 2019a). This involves establishing basic spatial anchors using markers (Einizinab et al., 2023a) or employing marker-less methods (Radanovic et al., 2023a). Marker-based approaches can be applied in any environment, but their implementation can be expensive or impractical. In marker-less approaches, precise correspondences can directly arise from preexisting point clouds, surface models, or image datasets (Li et al., 2019, Sheik et al., 2022). By utilizing a repository of referenced images and/or scans of the facility, spatial correspondences can be established through techniques such as image feature matching or geometric feature matching (Vermandere et al., 2022). Convolutional Neural Networks (CNNs) are also proposed for the feature extraction in this context (Radanovic et al., 2023b).

A significant challenge for marker-less methods is posed by environmental changes between the reference and newly collected data by MR device, especially in dynamic scene environments. Furthermore, the considerable size of the collected data and reference datasets presents a hindrance to fast processing due to their memory and computation intensity. This also necessitates an external processing device since current MR devices have difficulty handling the extensive auxiliary data involved (Radanovic et al., 2023a). These

* Corresponding author

limitations create challenges for the continuous alignment between the virtual model and the real-world, resulting in significant misalignment, particularly when moving away from the initial localization position, because of the drift (Kopsida, 2018). Regarding the current continuous alignment approaches, they mainly rely on vision-based methods, such as projecting BIM frame lines onto real image edges (Marchand et al., 2015, Acharya et al., 2019b). These methods often require multiple renderings to solve for a single frame, and may lack robustness in environments rich with lines due to potential ambiguities between the edges. Alternatively, another approach involves regressing to a relative camera pose difference between real and synthetic BIM images (Radanovic et al., 2023b), which not only requires auxiliary synthetic BIM images but also suffers from inaccurate alignments in areas where differences exist between the BIM elements and real building. Overall, the main challenges in MR global localization involves resource-intensive requirements, environmental variations between BIM and real-world, and the localization drift caused by accumulated positional errors over time.

This paper aims to tackle these challenges by utilizing a building wireframe—a simplified representation that outlines the structural elements and spatial layout of a building. Extracting wireframe elements, which possess greater permanence, involves low computational efforts on the MR device, ensuring reliable continuous alignment. An RGB input image taken by the MR device undergoes processing with a deep CNN to generate pixel-wise junction and line heat maps. The deep learning algorithm is trained to detect a specific type of wireframes, representing corners and edges of the building. By utilizing the information from the extracted edges and junction pixels, along with data from the MR depth sensor, the wireframe representation enables the efficient and precise reconstruction of the 3D geometry of the scene, even when provided with only a single input image. This method which requires no auxiliary data, extracts building wireframes and their corresponding depth information from the MR device, enabling a fast alignment process within the device's processing unit. After an initial manual registration of BIM over the real-world, our proposed method performs fine alignment. Through continuous refinement with newly captured images, the method also effectively addresses the localization drift issues in long paths without loop closures. The effectiveness of the proposed method in performing accurate global localization, and its ability to mitigate drift was assessed using real RGB images captured by a MR device within a site equipped with a BIM model.

The remainder of this paper is structured as follows. The related works are presented in Section 2. The detailed methodology is described in Section 3. Experimentation of the proposed method along with the results and discussions are presented in Section 4. Finally, the conclusions and suggestions for the future works are outlined in Section 5.

2. Related Works

Global localization in the AR/MR systems involves estimating the pose of the AR/MR camera in the virtual model environment such as BIM (Vermandere et al., 2022). Vision-based approaches play a prominent role in global localization for AR/MR applications. These approaches, which can dynamically refine global pose estimation during the application's runtime, involve the utilization of images, surface models, or point clouds to precisely locate the MR device and align the digital model with the real-world environment (Kopsida, 2018, Blut and Blankenbach, 2021). Vision-based approaches can be categorized into feature-based, model-based, image retrieval, and image-based pose regression methods (Blut and Blankenbach, 2021, Sheik et al., 2022, Radanovic et al., 2023b).

Feature-based methods leverage correspondences established between 2D image feature points and their corresponding 3D coordinates in a reference model. The practical implementation involves generating a 3D reference scene reconstruction from the images. For localizing a query image, feature extraction is performed by identifying keypoints and their corresponding descriptions, within local neighborhoods. This involves employing a suitable feature descriptor, such as SIFT, SURF, and ORB (Vermandere et al., 2022, Radanovic et al., 2023a). Subsequently, a search for matching keypoints in the reference 3D model is conducted, typically through nearest neighbor search in the descriptor space. The camera localization is then achieved using found matches and Perspective from Points (PnP) algorithm. Additionally, a robust estimation process, such as Random Sample Consensus (RANSAC), is incorporated to eliminate incorrect correspondences and enhance the accuracy of the localization (Marchand et al., 2015, Oh and Kim, 2023). The main downside of the feature-based method is its dependence on image-based point feature descriptors, necessitating the construction of a huge database using a specific descriptor (Li et al., 2019).

Contrary to feature-based methods, model-based approaches utilize models like CAD or BIM, relying on lines or other shapes within the model for localization. The fundamental idea is to minimize the distance between detected contour points in an image and the projection of a corresponding 3D line or shape from the model (Petit et al., 2012, Acharya et al., 2019b). However, model-based approaches require an initial estimate of the position and face challenges in extracting correct and complete contour lines (Acharya et al., 2019b).

In image retrieval localization approaches, the query image is compared to a database of geo-referenced images or image features, either previously captured or generated from a BIM (Marchand et al., 2015, Mahmood et al., 2020). The main drawback of image retrieval methods includes the need for a substantial dataset and their susceptibility to viewpoint changes (Radanovic et al., 2023a). More importantly, image retrieval provides an approximate localization and cannot guarantee the alignment.

Image-based pose regression methods estimate the pose by regression from a set of images with known pose (Piasco et al., 2018). Predominantly relying on convolutional regression networks such as PoseNet (Kendall et al., 2015) and BIM-PoseNet (Acharya et al., 2019a), these approaches learn the mapping from images to their corresponding global poses (Radanovic et al., 2023b). In this approach, the presence of changes in scene geometry or virtual model elements adversely impacts the accuracy of localization (Piasco et al., 2018). While recent methods can address appearance differences between the real images and virtual model images (Acharya et al., 2022, Acharya et al., 2023), the accuracy is still insufficient for high-precision alignment. Overall, according to the literature, the current vision-based global localization approaches encounter challenges related to low accuracy, environmental changes,

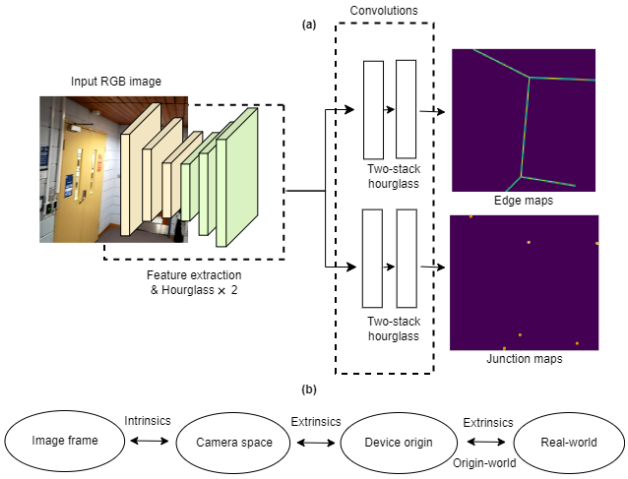


Figure 1. a) Architecture of the used neural network (Zhou et al., 2019b). b) Transformation between different frames in a MR system.

memory and computation intensity, and constraints in handling extensive auxiliary data.

Among high-level geometric features, straight lines and their junctions, forming the building wireframe, are fundamental and stable elements for assembling 3D structures and establishing correspondence with BIM elements (Zhou et al., 2019a). An optimized approach capable of extracting the building wireframe could offer a novel solution for efficiently achieving continuous global localization in MR applications (Zhou et al., 2019b). Traditional methods such as Hough transform detect lines based on local edge features (Stephens, 1991). In contrast to the wireframe representation, conventional line detection algorithms lack information about junctions and their connections, limiting their applicability in scene parsing and understanding. Recent advances in deep learning enable the extraction of high-level features from labeled data (Xue et al., 2019). In this case, Huang et al. (Huang et al., 2018) introduced a wireframe parsing task employing a deep learning-based approach. This involved the training of two distinct neural networks dedicated to predicting junction and line heat maps from an input image. Subsequently, the outputs of these two networks were combined using a heuristic wireframe fusion algorithm, ultimately yielding the final vectorized output. Additionally, Zhou et al. (Zhou et al., 2019a, Zhou et al., 2019b) presented an alternative framework characterized by a direct approach. This framework was founded on a single end-to-end trainable neural network capable of directly generating a 2D wireframe and reconstructing its 3D representation as the final output. Considering the capabilities of deep CNN approaches that directly extract the wireframe of a building as the key geometric features, we apply this methodology for global localization in AR/MR applications.

3. Method

In the global localization of MR with respect to BIM, referred to as the MR-BIM alignment process, two distinct 3D models exist, each lacking spatial reference to the other: a 3D BIM and a 3D real model generated by the MR device. In most of the AR/MR applications such as building work inspection, the 3D model created by the MR device is not needed for direct user visualization; it represents the genuine real-world environment perceived by the device holder. In contrast, the virtual model (BIM) is intended to be projected onto the real-world by the MR system and is fully visible to the device holder. Acknowledging that extracting only essential geometric features from the real-world data can expedite the alignment process and overcome associated challenges, we propose a method that exclusively identifies the wireframe of the building as significant and stable geometric features. Subsequently, the real 3D model is reconstructed using only these features, and the resulting simplified 3D model undergoes global localization with BIM.

Hence, our proposed method comprises three main stages: firstly, the extraction of a 2D wireframe from RGB images captured by the MR device; secondly, the reconstruction of a 3D representation based on the extracted 2D wireframe; and thirdly, the execution of a global localization process between the MR camera and BIM model.

3.1 2D wireframe extraction

The deep CNN method employed for extracting junctions and edges pixels is in accordance with the work conducted by Zhou et al. (Zhou et al., 2019b). Notably, our implementation differs from theirs as they utilized depth information to train the model for 3D wireframe reconstruction. In contrast, our model was not trained with depth information, as we always have access to the depth information of the extracted wireframe in our MR system.

As illustrated in Figure 1 (a), our implemented approach initiates with a neural network that takes a single image captured by the MR system as input. This network jointly predicts 2D heatmaps of lines and junctions. In the geometric wireframe W = (V, E) representing the scene, V and E denote the junctions and the lines, respectively. Specifically, E represents lines formed by the physical intersections of two planes, excluding planar textural lines, while V represents the intersections of lines among E. The approach aims to capture the global scene geometry, specifically the building wireframe, while disregarding local textural details.

For each image, the pixel-wise outputs of the implemented neural network consist of three items including junction probability J, junction offset O, and edge probability E. Among these outputs, J and E will be utilized for generating the wireframe.

The network structure is derived from the stacked hourglass network. Input training images are cropped to dimensions of 512 × 512 before entering the network. The initial feature-extracting module, comprising strided convolution layers and one max pooling layer, downsamples the feature map to 128 × 128. Subsequently, the network consists of S hourglass modules, each sequentially downsamples and upsamples the feature map. The stacked hourglass network progressively refines the output map to align with supervision from the training data. In the training phase, the main objective is to minimize the total loss, which is computed as the sum of individual losses across all training images and hourglass modules. Moreover, the individual loss for an image is calculated by comparing the output of the hourglass module with the actual ground truth representation for that image. Indeed, the total loss is the accumulation of these individual losses for all images and hourglass modules. The loss for each

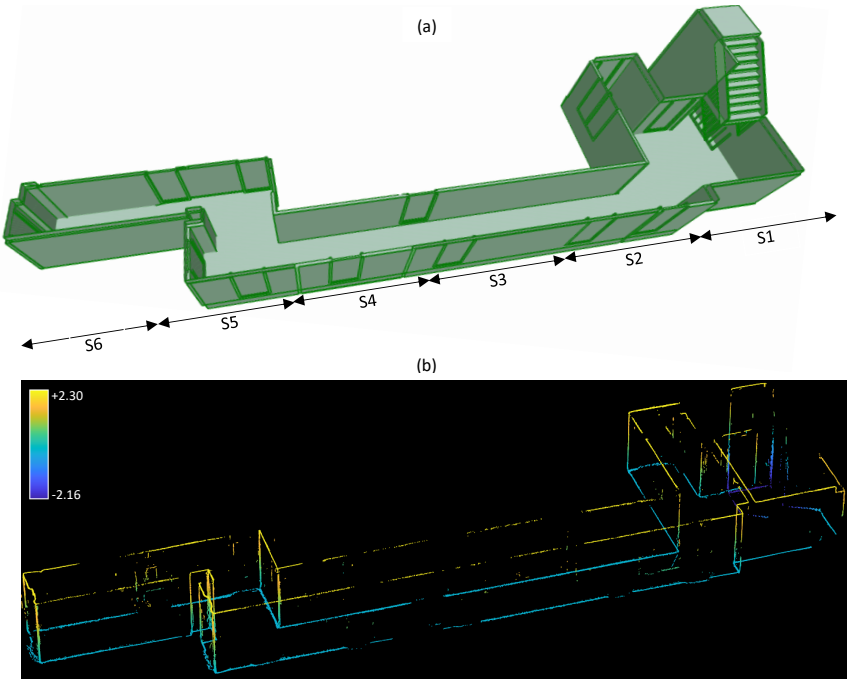


Figure 2. a) The BIM of the testing site, depicting the 3rd floor of Infrastructure Engineering Block B. Segments S1 to S6 illustrate the segmented areas designated for the experiments. b) The 3D wireframe extracted from the images (approximately 700 images). Colour represents elevation from the floor.

image (L), is a composite function, which is a weighted (λ) sum of specific loss functions including junction loss, offset loss, and edge loss:

$$L \doteq \sum_k \lambda_k L_k, \quad k \in \{J, O, E\} \tag{1}$$

For training our network, for each input image we prepared the following information, and the outputs of the neural network are image-space heatmaps of the desired wireframe.

Junction Map: The ground truth junction map is a down-sampled binary map indicating the presence of junctions in pixel locations with sub-pixel accuracy:

$$\hat{J}(p) = \begin{cases} 1 & \exists v \in V : p = \lfloor \frac{v}{4} \rfloor \\ 0 & otherwise \end{cases} \tag{2}$$

where p is the integer coordinate and v is the pixel coordinate of a junction in the input image. The network is trained to predict the junction maps using softmax cross entropy loss for each pixel. The resulting probability maps indicate the likelihood of a junction at specific locations in the input image.

Offset Map: To address precision issues caused by the lower resolution of the junction map, as it gets four times less than the resolution of the input image, an offset map is employed. The offset map stores the difference vector from the ground truth junction position to its original position with sub-pixel accuracy:

$$\hat{O}(p) = \begin{cases} \frac{v}{4} - p & \exists v \in V : p = \lfloor \frac{v}{4} \rfloor \\ 0 & otherwise \end{cases} \tag{3}$$

The offset map loss is computed only near actual junctions using the heatmap as a mask by applying a sigmoid and constant translation function to the last layer of the offset branch in the neural network. The offset map loss is normalized by the number of junctions.

Edge Map: Line positions are estimated by representing them in an edge heatmap. Ground truth lines are drawn on the edge map with an antialiasing technique for better accuracy. The edge map is defined as:

$$\hat{E}(p) = \begin{cases} \max_e 1 - dist(p, e) & \exists e \in E : dist(p, e) < 1 \\ 0 & otherwise \end{cases} \tag{4}$$

dist(p, e) denotes the shortest distance between a pixel p and the nearest line segment e. The edge map with the range between 0 and 1, represents the probability of a line close to point p. Then, treating it as a probability distribution, the network is trained using the sigmoid cross entropy loss. The resulting probability map indicates the likelihood of a line at specific points in the input image.

3.2 3D wireframe reconstruction

In the majority of MR systems, such as Microsoft HoloLens, a variety of sensors equipped with their respective coordinate frames are employed. Visible light and depth cameras stand out as the main sensor types crucial for scene reconstruction within an MR system. Each camera sensor (visible light and depth) is associated with a specific coordinate system, while the device itself maintains a unique coordinate system as the origin frame. The known intrinsic parameters of the camera sensors, estimated during calibration, in conjunction with the known pose values of the visible light and depth cameras relative to the origin coordinate system, enable the dynamic transformation of data captured by these sensors to the device's origin coordinate system during runtime. IMU sensors, including accelerometer, gyroscope, and magnetometer, contribute to determining both the relative and absolute poses of the device. Furthermore, in the high-tech MR devices such as the Microsoft HoloLens, the utilization of camera and depth sensor data enables the

execution of SLAM techniques. This allows for accurate self-positioning within the environment, ensuring that the pose of the MR device is consistently known relative to the real-world reference point.

In the MR system, visible light camera and depth sensor are associated with three primary frames: the 2D image frame, the 3D camera space, and the 3D real-world coordinate system (Figure 1 (b)). Real-world coordinate system is shared between the depth and visible cameras. The 2D wireframes extracted from the visible camera image frame cannot be directly transformed into 3D space. However, by utilizing the depth camera frame, which contains depth information, these wireframes can be converted to the 3D real-world coordinate system. In a synchronized depth and visible image pair, 3D points extracted from the depth camera are first transformed into real-world coordinates and then projected onto the visible camera frame. The matching process guarantees accurate preservation of depth values for the pre-extracted 2D wireframe, enabling its reconstruction into 3D coordinates within the real-world reference.

3.3 Global localization

After converting the 2D image space wireframes to the 3D real-world frame, two 3D models exist for the alignment: the virtual model (BIM) and the real-world reconstructed model. For precise alignment, this paper employs the Iterative Closest Point (ICP) method, which minimizes the difference between two point clouds. Assuming an initial manual alignment, ICP fine-aligns the models. Continuous refinement during runtime in a long path corridor, utilizing newly captured images, effectively overcomes misalignment issues caused by drift and refines the global pose parameters of the MR system relative to the BIM.

4. Experiments and Results

To evaluate the efficacy of our proposed method, we first implemented the 2D wireframe extraction network using PyTorch. Training was conducted on images collected by Microsoft HoloLens2 within the Block B of the Department of Infrastructure Engineering at the University of Melbourne. Utilizing sample codes from (Ungureanu et al., 2020), we extracted raw streams on HoloLens2. The method was then applied to reconstruct 3D models from the extracted wireframes, enabling continuous global localization of the MR in a BIM model. The system used for training employed a Dell XPS 15 laptop equipped with a 13th Generation Intel(R) Core(TM) i9 CPU, a NVIDIA(R) GeForce(R) RTX(TM) 4070 GPU, and 64GB (2x32GB) RAM.

4.1 Dataset and annotation

A total of 158 real RGB images, captured by the Microsoft HoloLens 2 on the third floor of Block B, were utilized for both network training and validation purposes. 70% of the images were employed for network training, while the remaining 30% were reserved for the validation. The selection of train and validation images was performed randomly. To annotate the images, an online tool called V7Darwin (V7Labs, Darwin) was employed. The junctions and edges labels to be identified in each image were generated one by one. As for the offset map, it was generated through Python coding and exported in a suitable format for the network.

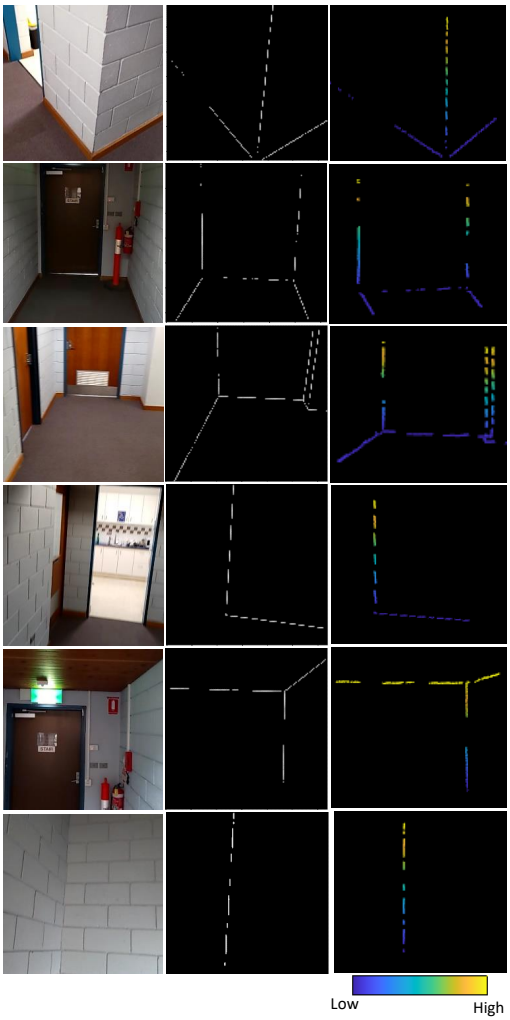


Figure 3. Real images with corresponding 2D and 3D wireframe extractions: RGB image (Left), 2D wireframe (Middle), and 3D wireframe (Right). Colour represents elevation from the floor.

Concerning the virtual model, we utilized the BIM model of the same site, as illustrated in Figure 2 (a), which has been employed in various prior studies, including (Acharya et al., 2019a, Radanovic et al., 2023b). In the alignment process, the BIM model is initially aligned roughly with the real-world manually. Subsequently, when the device captures an image, its 3D wireframe, if available, is employed to enhance the alignment through the ICP method. This refinement process can be iterated by capturing new images while moving the device until the alignment reaches completion at the end of the corridor.

4.2 2D wireframe

We adopted the network architecture proposed by (Zhou et al., 2019b) for junctions and edges detection, maintaining its structure. The backbone is a two-stack hourglass network each consists of 6 stride-2 residual blocks and 6 nearest neighbour upsamplers. Following the stacked hourglass feature extractor, different head modules are inserted for each map. Each head comprises a 3×3 convolutional layer to reduce channel numbers, followed by a 1×1 convolutional layer to compute the corresponding map. During training, the ADAM optimizer is employed with a learning rate set to 10^{-4} for four epochs. All the experiments are conducted on a single GPU, with a batch

size of 4. Loss weights are configured as λ_J = 2.0, λ_O = 0.25, and λ_E = 3.0, ensuring approximately equal loss terms ($\lambda_k L_k$ in Eq 1).

The total loss curves for the training and validation datasets are depicted in Figure 4 (a). Notably, the loss values for the validation dataset consistently exhibit a higher magnitude compared to the training dataset, indicating potential overfitting. The initial stages of training are characterized by an exponentially decreasing trend, signifying rapid convergence and adaptation to the training data. As iterations progress, both curves stabilize, resulting in a smooth and straight trend. This phenomenon suggests that the model achieves a level of convergence, demonstrating improved generalization on both the training and validation datasets.

In this study, in the process of 3D wireframe extraction, a distinction was not made between junction and edge pixels. Following the extraction of these pixels through their corresponding heatmaps, all identified interest pixels were employed uniformly for 3D wireframe extraction. Some samples of extracted 2D wireframes are shown in Figure 3 (Middle). Despite the presence of numerous other textural linear elements in the images, the network reliably extracts only the wireframe. To evaluate the extracted 2D wireframe, precision and recall, two standard metrics, were employed, computed by varying the threshold between 0.4 and 1. The comparison of precision and recall curve, as illustrated in Figure 4 (b), was conducted on the edge pixels extracted from the test images. A higher precision signifies a robust capability to accurately identify the desired edge pixels within the images. In contrast, the model exhibits low recall values, indicating its inability to extract the majority of positive true edge pixels. However, this limitation does not significantly impact our primary objective, as even sparse edge points contribute to the reconstruction of the 3D wireframe for use in our broad goal of global localization. Consequently, a threshold of 0.8 was set for acceptance, corresponding to a precision of 0.94 and a recall of 0.06, ensuring the inclusion of pixels with a high probability of belonging to true edge pixels. This strategic thresholding enhances the reliability of the reconstructed 3D model of the real-world.

4.3 3D wireframe-BIM alignment

In this section, we aim to demonstrate the efficacy of our proposed method in the global localization of the MR system within a BIM framework. Furthermore, we analyze the effectiveness of our proposed method in mitigating the drift issue encountered by the MR system along a lengthy corridor. To achieve this objective, the corridor is divided into six segments, as illustrated in Figure 2 (a), and two distinct scenarios are defined. In both scenarios, we applied our proposed method, aligning the BIM with the 3D wireframe. In the first scenario, we exclusively performed the alignment using the 3D wireframe extracted from the initial segments (Segment 1 and partially Segment 2). In the second scenario, alignment was continuously refined at each segment as we progress toward the end of the corridor (S6). Consequently, for whole site, in Scenario 1, the alignment process was performed once, while in the second scenario, we refined the alignment six times. By having these scenarios, the effectiveness of our proposed method in terms of global localization accuracy is evaluated in both scenarios. Additionally, the contrasting outcomes between scenarios highlight the significance of our proposed method in addressing the drift issue.

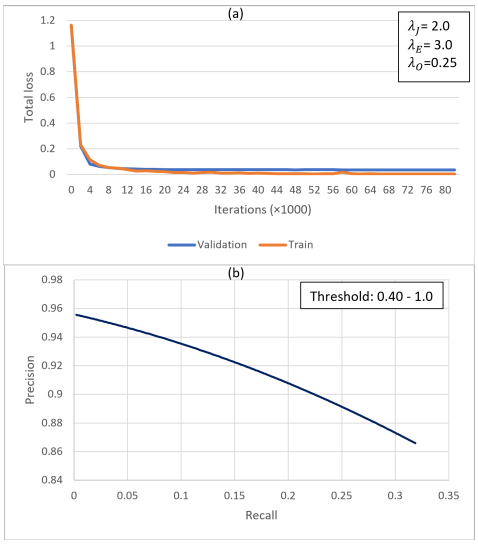


Figure 4. a) Total loss curves for training and validation datasets. b) Precision-recall curve for 2D wireframe extraction.

As illustrated in the Method section, following the extraction of the 2D wireframe from RGB images, the corresponding pixels in the depth map are identified, and a 3D wireframe is reconstructed relative to the MR coordinate system. In Figure 3 (Right), several 3D wireframes generated from 2D pixels are illustrated. Additionally, Figure 2 (b) provides a comprehensive view of the entire 3D wireframe of the testing site. In both scenarios, for the purpose of aligning the BIM model with the real-world environment, an initial manual rough alignment is conducted. Subsequently, utilizing the reconstructed 3D wireframe and their corresponding objects in the BIM, the ICP method is implemented to refine the alignment. The alignment process initiates from the right side of the corridor (S1), and by refining the registration through our defined scenarios, the corridor is scanned until reaching its end (S6).

The corridor's total length spans approximately 27 meters, resulting in each segment covering around 4.5 meters. In order to verify the capability of the method in performing global localization with limited inputs, we endeavored to utilize the minimum number of RGB images for reconstructing the 3D wireframe of each segment. Specifically, 32 images that were not included in the training set, were employed for the entire set of segments. The chosen metric for accuracy evaluation involves measuring the distance between the nearest points in the two models (BIM and real-world) after alignment. Figure 5 provides a comprehensive overview of the accuracy values for all designated segments of the corridor within the two specified scenarios.

In both scenarios, the effectiveness of the proposed method is evident. Despite various factors influencing global localization accuracy, such as the quality of the initial alignment, geometric precision of the utilized BIM, reliability of the extracted wireframes, and the convergence ability of the ICP method, the proposed method successfully aligns both the models. In the first scenario, the average alignment accuracy is approximately 10 cm in the initial segments, while an accuracy of below 15 cm on average was achieved at the end of the corridor. However, it becomes evident that as the MR system progresses towards S6, alignment accuracy decreases, leading to a wider spread in registration errors, especially in S5. The substantial accuracy

difference between the first and sixth segments in the first scenario is unexpected, as the mentioned criteria are unlikely to cause such a significant variation. If the virtual model accurately represents the real-world, even with a fine alignment in the initial segments (as we performed in Scenario 1), the registration accuracy of the end segments should logically be close to that of the initial segments. The notable difference (≈5cm) is likely attributed to the drift issue in the MR system, as no closure loop scan was performed, and SLAM localization is affected by drift in lengthy paths, such as that presents in our test site.

Scenario 2 is designed to assess the efficacy of the method in mitigating the drift issue. One of the main benefits of the proposed approach lies in its independence from auxiliary data and a high-performance processing unit.This enables the complete global localization process to be carried out exclusively within the device processing unit located on the building site, eliminating the requirement for an external processing unit. This capability facilitates continuous alignment during scanning, utilizing the BIM model as a reference. Through ongoing alignment, the global pose of the MR device can be progressively refined as it advances toward the end of the corridor during the registration process. This helps in addressing challenges arising from SLAM drift in scans without loop closures. The outcomes of the second scenario validate this assertion. As depicted in Figure 6, notably in Segments 5 and 6, the most substantial drift issues are mitigated through continuous alignment, resulting to a reduction in drift of up to 4cm. In addition, as anticipated, Scenario 2 exhibits a consistent trend in global localization accuracy values from Segment 1 to Segment 6.

The results suggest that the significant drift does not occur until Segment 4. However, a substantial difference between the two scenarios becomes more apparent in Segments 5 and 6. This discrepancy can be attributed to the presence of a turn in Segment 5, which completely alters the scanning route. It can be inferred that within the experimental site, the alignment process demonstrates reliability under the condition of a consistent route without alterations in direction or entry into new paths. The places where changes occur can be recognized as appropriate positions for refining the alignment.

5. Conclusion and Future Works

This paper addressed crucial challenges faced by model-based methods in achieving continuous global localization within MR systems, particularly in dynamic construction environments. The proposed method, focusing on the extraction and utilization of the building structural elements such as wireframes, demonstrates significant advantages. Employing deep learning, specifically a deep CNN, we successfully extracted 2D wireframes and reconstructed their 3D counterparts, enabling efficient and precise continuous global localization.

The 2D wireframe extraction network is carefully trained and evaluated, demonstrating its capability to identify edges and junctions with satisfactory precision. The application of the method in 3D wireframe extraction and subsequent alignment with BIM models is demonstrated through experiments in a real-world setting. The method effectively mitigated the drift issue encountered by the MR system during a lengthy corridor scan, depicting its adaptability in challenging environments. The results indicated that the proposed method maintains a high

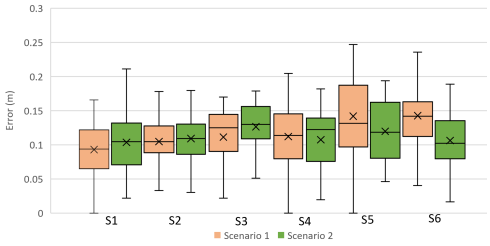
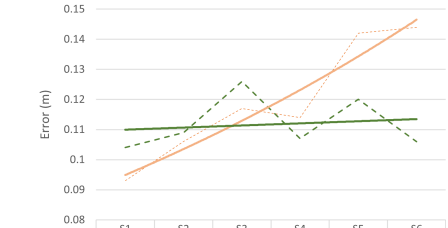


Figure 5. Localization errors for both Scenario 1 and Scenario 2.



	S1	S2	S3	S4	S5	S6
Scenario1	0.093	0.106	0.117	0.114	0.142	0.144
Scenario2	0.104	0.109	0.126	0.107	0.12	0.106

Figure 6. The average global localization errors for both alignment scenarios, with the trends representing the exponential fits applied to the data.

level of global localization accuracy, especially in scenarios involving continuous alignment.

A pivotal strength of the proposed approach lies in its independence from auxiliary data and the ability to perform the entire global localization process within the device processing unit at the construction site. The fine alignment achieved using only a single RGB image containing the building's wireframe. This eliminated the need for external processing units and enhanced the feasibility of continuous alignment. The investigation revealed a noteworthy distinction in accuracy between non-continuous and continuous alignments, attributed to a change in direction at a specific location altering the scanning route. Despite this observed difference, the overall accuracy in both scenarios did not manifest a substantial disparity. The method's efficacy is particularly evident in addressing challenges arising from SLAM drift in scans without loop closures. These benefits prove especially advantageous in applications such as building work inspections using AR/MR technology.

However, it is important to highlight a constraint associated with the proposed method's dependence on the data capturing range of the depth sensor, which necessitates attention in future investigations. One potential resolution could involve training the network to predict depth values as well. Furthermore, while the processing component of the method is highly optimized, there is a need for a more in-depth quantitative analysis regarding the time assessment of the entire alignment process. Future research directions may also focus on optimizing and validating the method in diverse construction scenarios to enhance its robustness and broaden its applicability. This study contributes to advancing global localization solutions for MR devices, offering valuable insights for both current and future research endeavors in this dynamic field.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-119-2024 | © Author(s) 2024. CC BY 4.0 License.

6. Acknowledgments

This research is supported by Building 4.0 CRC. The support of the Commonwealth of Australia through the Cooperative Research Centre Program is acknowledged.

References

Acharya, D., Khoshelham, K., Winter, S., 2019a. 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.

Acharya, D., Ramezani, M., Khoshelham, K., Winter, S., 2019b. BIM-Tracker: A model-based visual tracking approach for indoor localisation using a 3D building model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 157–171.

Acharya, D., Tatli, C. J., Khoshelham, K., 2023. Synthetic-real image domain adaptation for indoor camera pose regression using a 3D model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 405–421.

Acharya, D., Tennakoon, R., Muthu, S., Khoshelham, K., Hoseinnezhad, R., Bab-Hadiashar, A., 2022. Single-image localisation using 3D models: Combining hierarchical edge maps and semantic segmentation for domain adaptation. *Automation in Construction*, 136, 104152.

Blut, C., Blankenbach, J., 2021. Three-dimensional CityGML building models in mobile augmented reality: A smartphone-based pose tracking system. *International Journal of Digital Earth*, 14(1), 32–51.

Einizinab, S., Khoshelham, K., Winter, S., Christopher, P., 2023a. Offset-based marker placement for BIM alignment in Mixed Reality. *2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, IEEE, 3684–3688.

Einizinab, S., Khoshelham, K., Winter, S., Christopher, P., Fang, Y., Windholz, E., Radanovic, M., Hu, S., 2023b. Enabling technologies for remote and virtual inspection of building work. *Automation in Construction*, 156, 105096.

Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., Ma, Y., 2018. Learning to parse wireframes in images of man-made environments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 626–635.

Kendall, A., Grimes, M., Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-DoF camera relocalization. *Proceedings of the IEEE international conference on computer vision*, 2938–2946.

Kopsida, M., 2018. Automated progress monitoring using Mixed Reality. PhD thesis, University of Cambridge.

Kopsida, M., Brilakis, I., 2017. BIM registration methods for mobile Augmented Reality-based inspection. *eWork and eBusiness in Architecture, Engineering and Construction: ECPPM 2016*, CRC Press, 201–208.

Li, J., Wang, C., Kang, X., Zhao, Q., 2019. Camera localization for Augmented Reality and indoor positioning: a vision-based 3D feature database approach. *International journal of digital earth*.

Mahmood, B., Han, S., Lee, D.-E., 2020. BIM-based registration and localization of 3D point clouds of indoor scenes using geometric features for Augmented Reality. *Remote Sensing*, 12(14), 2302.

Marchand, E., Uchiyama, H., Spindler, F., 2015. Pose estimation for Augmented Reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12), 2633–2651.

Oh, J., Kim, H., 2023. A Camera Center Estimation Based on Perspective One Point Method. *IEEE Transactions on Intelligent Vehicles*.

Petit, A., Marchand, E., Kanani, K., 2012. Tracking complex targets for space rendezvous and debris removal applications. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 4483–4488.

Piasco, N., Sidibé, D., Demonceaux, C., Gouet-Brunet, V., 2018. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74, 90–109.

Radanovic, M., Khoshelham, K., Fraser, C., 2023a. Aligning the real and the virtual world: Mixed Reality localisation using learning-based 3D-3D model registration. *Advanced Engineering Informatics*, 56, 101960.

Radanovic, M., Khoshelham, K., Fraser, C., Acharya, D., 2023b. Continuous BIM Alignment for Mixed Reality Visualisation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 279–286.

Sheik, N. A., Deruyter, G., Velaerts, P., 2022. Automated registration of building scan with BIM through detection of congruent corner points. *The 7th International Conference on Smart City Applications*, 48, Copernicus GmbH, 179–185.

Stephens, R. S., 1991. Probabilistic approach to the Hough transform. *Image and vision computing*, 9(1), 66–71.

Ungureanu, D., Bogo, F., Galliani, S., Sama, P., Duan, X., Meekhof, C., Stühmer, J., Cashman, T. J., Tekin, B., Schönberger, J. L., Tekin, B., Olszta, P., Pollefeys, M., 2020. HoloLens 2 Research Mode as a Tool for Computer Vision Research. *arXiv:2008.11239*.

V7Labs, Darwin. V7-Ai data platform for Ml teams. `https://www.v7labs.com/`. Accessed on January 22, 2024.

Vermandere, J., Bassier, M., Vergauwen, M., 2022. Two-step alignment of Mixed Reality devices to existing building data. *Remote Sensing*, 14(11), 2680.

Xue, N., Bai, S., Wang, F., Xia, G.-S., Wu, T., Zhang, L., 2019. Learning attraction field representation for robust line segment detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1595–1603.

Zhou, Y., Qi, H., Ma, Y., 2019a. End-to-end wireframe parsing. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 962–971.

Zhou, Y., Qi, H., Zhai, Y., Sun, Q., Chen, Z., Wei, L.-Y., Ma, Y., 2019b. Learning to reconstruct 3D manhattan wireframes from a single image. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7698–7707.