buildings

MDPI

Article

Drift-Free BIM Alignment for Mixed Reality Visualization Through Image Style Transfer and Feature Matching

Mohamed Zahlan Abdul Muthalif [1,*], Davood Shojaei [1], Kourosh Khoshelham [1] and Debaditya Acharya [2]

[1] Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia; shojaeid@unimelb.edu.au (D.S.); k.khoshelham@unimelb.edu.au (K.K.)
[2] School of Science, RMIT University, 124 La Trobe St, Melbourne, VIC 3000, Australia; debaditya.acharya@rmit.edu.au
* Correspondence: mabdulmuthal@student.unimelb.edu.au

Abstract

Accurate localization is a persistent challenge for Mixed Reality (MR) applications in the construction industry, where reliable alignment between digital building models and physical environments is critical. Commercial MR devices such as the Microsoft HoloLens rely on Visual-Inertial Simultaneous Localization and Mapping (VISLAM) for pose estimation, but accumulated drift over extended trajectories and visually ambiguous indoor spaces often reduces localization accuracy. This paper presents a complementary localization refinement methodology that integrates HoloLens spatial tracking with image style transfer and geometry-based pose estimation for Building Information Modeling (BIM)-aligned MR visualization. Image style transfer is used to reduce appearance discrepancies between real-world images and synthetic BIM renderings, improving feature correspondence for geometric alignment. Pose refinement is then applied using feature matching and Perspective-n-Point (PnP) estimation to mitigate accumulated drift when sufficient visual evidence is available. The method is evaluated using 1408 image pairs captured along an indoor trajectory, demonstrating improved BIM alignment, significantly reducing accumulated drift to 1–2 pixels. The proposed approach supports more reliable MR visualization for construction-related tasks such as inspection, coordination, and spatial decision-making.

Keywords: localization; Mixed Reality; image style transfer; BIM; HoloLens

Check for updates

Academic Editor: Antonio Caggiano

Received: 31 January 2026
Revised: 15 February 2026
Accepted: 18 February 2026
Published: 20 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Mixed Reality (MR) is a technology that facilitates the integration of physical environments with virtual elements, thereby creating immersive user experiences. MR enables the interaction between real and virtual components, leading to a seamless blend of the two realms [1]. These technologies are increasingly applied in industries such as education [2], tourism [3], navigation [4], military [5], and construction [6,7], where accurate visual representation and manipulation of digital data are crucial.

Building Information Modeling (BIM) plays a vital role in enhancing the effectiveness of MR in the construction industry [8–10]. BIM serves as a digital representation of the physical and functional characteristics of a building, allowing for enhanced visualization and facilitating better decision-making and project management [11,12]. Integrating BIM with MR allows for more intuitive and real-time interaction between the virtual and the real world [13]. This combination enables more effective visualization of hidden elements [14,15] and facilitates tasks like progress tracking, maintenance, and scenario simulation in construction [16,17].

A key requirement for MR visualization of BIM geometries is accurate estimation of the MR camera pose, which involves determining the position and orientation of the devices within an indoor space. The absence of Global Navigation Satellite System (GNSS) signals indoors complicates this process, prompting research into alternative methods that can provide reliable, real-time localization without relying on GNSS [18].

To address these challenges, infrastructure-based techniques such as WiFi, Bluetooth, ultrasound, and ultra-wideband (UWB) have been developed. These systems estimate position based on metrics like signal strength and time-of-flight, but require considerable infrastructure investments, which may not always be practical [19]. As a result, there is growing interest in infrastructure-independent methods that do not depend on additional hardware.

Infrastructure-independent methods, like visual-odometry, utilize solely visual observations to estimate the movement of a device along the trajectory [20]. This method relies heavily on the quality of the images, and any degradation in image clarity or detail can significantly impact the accuracy of the motion estimates [21]. Another popular infrastructure-independent method is Simultaneous Localization and Mapping (SLAM), which is a process by which the device constructs a map of an unknown environment while simultaneously determining its position within that map, using sensor data like cameras, LiDAR, or inertial measurements [22]. However, these methods suffer from the accumulation of errors that can arise along the trajectory over time and the distance from the initialization of the device [17,22,23].

Model-based localization methods have gained increasing attention for their ability to align camera poses using digital representations such as BIM. These approaches offer an infrastructure-independent solution by leveraging pre-existing 3D models of the environment to estimate camera positions without the need for physical markers or external hardware [24–28]. However, the practical deployment of these systems faces persistent limitations. The disparity in visual appearance between synthetic BIM renderings and real-world camera images caused by lighting variations, lack of texture in BIM, and differences in environmental conditions often leads to inaccurate feature matching. Additionally, indoor environments with symmetrical architectural layouts can introduce ambiguity in pose estimation. These challenges become more pronounced in large-scale or dynamic settings, where visual drift and cumulative error along the device's trajectory compromise localization accuracy and reliability.

In recent years, domain adaptation techniques like Cycle-Consistent Generative Adversarial Network (CycleGAN) have gained popularity in addressing visual mismatches between synthetic BIM renderings and real-world images. By translating synthetic images into photorealistic styles and vice versa, image feature correspondence is improved, which enhances camera pose estimation accuracy [29–31]. However, existing CycleGAN-based approaches face several limitations. Their performance often degrades in visually uniform or repetitive environments, where the lack of strong visual cues hinders accurate matching. Additionally, artifacts from GAN training, such as texture inconsistencies or noise, can introduce distortions that compromise localization precision. Moreover, these methods typically focus on the image translation task in isolation, without integrating geometric alignment procedures such as Perspective-n-Point (PnP), which limits their effectiveness in practical AR/MR applications requiring precise spatial alignment.

Recognizing these limitations, the aim of this study is to develop and evaluate a complementary localization refinement approach for BIM-aligned MR visualization that reduces accumulated drift in HoloLens-based MR systems without replacing native SLAM tracking. Rather than attempting to solve indoor localization solely through appearance-based domain adaptation or feature correspondence, the proposed method builds upon the

Microsoft HoloLens VISLAM system as a continuous localization backbone and applies BIM-guided pose refinement opportunistically when sufficient geometric and visual cues are available. Image style transfer using CycleGAN is employed to reduce appearance discrepancies between real-world images and synthetic BIM renderings, thereby improving correspondence reliability, while geometric pose estimation using PnP is used to correct accumulated drift. By focusing on drift mitigation and alignment refinement, rather than absolute localization or feature-only tracking, the proposed approach explicitly acknowledges the challenges of low-texture and symmetric indoor environments and demonstrates how BIM-informed correction can significantly improve MR alignment when visual evidence supports it.

The contributions of this paper are as follows:

- A complementary BIM-guided drift-refinement framework for MR localization, which integrates HoloLens VISLAM tracking with BIM-based geometric alignment, CycleGAN-driven domain adaptation, and PnP-based pose estimation, demonstrating that accumulated drift in MR device trajectories can be substantially reduced through BIM-guided pose refinement, resulting in near-zero residual reprojection error.
- A unified pipeline that combines domain adaptation and geometric pose refinement, where CycleGAN is employed to mitigate appearance discrepancies between real and synthetic images, not as a standalone registration solution, but as an enabling component that improves correspondence reliability within a geometry-based refinement process.
- An opportunistic refinement strategy that applies BIM-based correction only when sufficient visual correspondences are available, while relying on VISLAM to maintain continuous localization in symmetric or low-texture environments where correspondence-based methods alone typically fail.
- A comprehensive experimental evaluation along a full indoor trajectory, demonstrating significant reprojection error reduction when refinement is applicable, together with explicit analysis of excluded frames and practical limitations related to feature availability, BIM accuracy, and computational constraints.
- A clear discussion of deployment considerations and limitations, positioning the proposed approach as a refinement layer that enhances existing MR localization systems rather than replacing them, and outlining pathways toward automated initialization and real-time implementation.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive review of related literature; Section 3 outlines the proposed methodology; Section 4 provides a detailed account of the experimental procedures; Section 5 presents the results and discussions. Section 6 concludes the findings, limitations and directions for future research.

2. Related Works

This section reviews the existing research on indoor localization for Augmented Reality (AR) and MR applications in the built environment, with a particular focus on BIM-assisted alignment and drift mitigation. The literature is organized into four thematic strands to progressively narrow the research scope. First, marker-based and markerless indoor localization techniques are reviewed to establish the foundations of infrastructure-dependent and infrastructure-independent tracking approaches. Second, model-based BIM-assisted localization methods are examined, highlighting their ability to improve pose estimation accuracy as well as their limitations in symmetric or visually repetitive environments. Third, domain adaptation and learning-based localization approaches are discussed, with emphasis on the use of generative and adversarial models to reduce ap-

pearance discrepancies between synthetic BIM renderings and real-world imagery. Finally, the identified research gap is summarized, motivating the need for a hybrid localization strategy that combines continuous sensor-based tracking with opportunistic BIM-guided pose refinement to suppress accumulated drift in practical MR scenarios.

2.1. Marker-Based and Markerless Indoor Localization

Early indoor localization approaches relied on marker-based techniques, where physical markers are placed within the environment to establish fixed reference points for camera pose estimation [32,33]. While such methods offer reliable accuracy under controlled conditions, they require manual deployment and maintenance of markers and are sensitive to occlusion, damage, and environmental changes. These limitations significantly restrict their scalability and practicality in operational construction environments.

To overcome these constraints, markerless localization approaches have been widely adopted. Markerless methods estimate camera pose using natural environmental features such as textures, edges, and geometric structures [34,35]. Visual odometry, SLAM, VISLAM, and RGB-D SLAM have become foundational techniques in AR and MR applications, enabling simultaneous camera tracking and map construction without prior infrastructure [21,22,36]. These methods have proven to be effective solutions in MR applications. SLAM enables the simultaneous construction of maps while tracking the camera's position in real-time [22]. This capability is particularly crucial in scenarios where continuous localization and map-building are essential [21]. However, SLAM systems can encounter challenges, particularly issues of drift, where small errors accumulate over time, resulting in significant misalignment between the virtual and real environments [37].

Recent work in [38] proposes a seamless, infrastructure-less AR registration framework targeting facility management tasks across mixed indoor and outdoor environments. The method integrates multiple AR registration engines, combining GNSS-Real-Time Kinematic (RTK) and vision-based techniques within a cloud platform that dynamically selects the most suitable registration strategy depending on environmental conditions. The system eliminates the need for manual setup or pre-aligned markers and was evaluated on a university campus, achieving AR overlay discrepancies of approximately 0.08–0.09 m in heterogeneous and unprepared environments.

While this approach demonstrates strong robustness at a global scale and across indoor-outdoor transitions, it primarily addresses coarse alignment and scenario-level registration rather than continuous indoor drift correction. As such, it offers a complementary perspective to BIM-based pose refinement methods, which focus on improving localization accuracy over extended indoor trajectories where accumulated drift remains a critical challenge.

2.2. Model-Based BIM-Assisted Localization

Model-based localization methods leverage prior knowledge of the built environment to improve pose estimation accuracy. BIM provide rich geometric representations that can be used as spatial references for aligning real-world sensor data. For instance, Acharya and Ramezani [25] introduced the BIM-Tracker model, which aligns real-time camera views with BIM to provide accurate pose estimation without the need for map-building during runtime. Similarly, Mahmood and Han [27] improved localization accuracy by integrating point cloud data with BIM models. The integration of SLAM with model-based tracking has further extended its applications, particularly in AR presentations for indoor construction sites. One significant advancement is BIM-PoseNet, developed by Acharya and Khoshelham [24], which utilizes synthetic images generated from 3D indoor models to estimate camera pose. However, this approach encountered challenges in

environments with symmetrical features, leading to localization ambiguities. To address these limitations, Acharya and Singha Roy [39] enhanced BIM-PoseNet by incorporating recurrent deep networks that leverage image sequences, thereby improving error reduction and robustness in complex environments through the use of temporal data. Similarly, Sattler and Zhou [40] analyzed Convolutional Neural Network (CNN)-based Absolute Pose Regression (APR) methods, noting that these models tend to approximate poses rather than accurately generalizing to real-world environments and have poor accuracy in a dynamic environment. Ha and Kim [41] encountered similar limitations in their work while addressing indoor localization by matching indoor real images to BIM images using VGG-16 features. While effective, reliance on single images limited performance, especially in dynamic environments. Radanovic and Khoshelham [13] developed an end-to-end CNN that used real and synthetic BIM image pairs to estimate the 6 DoF (Degrees of Freedom) relative camera pose.

Building on these model-based localization efforts, recent studies have shifted attention toward pose refinement and application-driven alignment, emphasizing the importance of correcting accumulated errors even when initial tracking is available. A recent study in focusing on camera pose refinement for precise BIM alignment in MR visualization [42] emphasized the importance of post hoc pose correction to improve registration accuracy between BIM and MR content. This work highlights that even when initial tracking is available, refinement against BIM geometry is essential for achieving precise alignment, reinforcing the relevance of drift correction strategies rather than purely end-to-end pose estimation. Similarly, an autonomous MR framework for real-time construction inspection was presented in [43], demonstrating how MR systems can support on-site inspection and progress monitoring without extensive manual intervention. While these systems showcase practical MR applications, they implicitly rely on stable and accurate localization, and they do not directly address long-term drift accumulation or BIM-guided pose refinement.

Despite these advances, several challenges remain unresolved. Model-based and deep learning approaches continue to struggle in environments with repetitive architectural features, where similarities can lead to localization ambiguities. Additionally, the application of these techniques in large-scale construction projects remains an area ripe for further research [17,28]. In addition, the above studies highlight deep learning's potential to improve indoor localization while revealing ongoing challenges, such as discrepancies between real-world environments and BIM-rendered geometry. Variations in furniture, lighting, and other dynamic elements can affect the accuracy of alignment, signaling the need for more refined pose estimation techniques, particularly in highly dynamic indoor environments where geometric changes are frequent.

2.3. Domain Adaptation and Learning-Based Localization

The domain adaptation techniques like CycleGAN have emerged as important tools for bridging the gap between BIM and real-world images. Domain adaptation is crucial to enhancing the accuracy of pose estimation when models trained on synthetic data are applied to real environments. Zhu and Park [44] introduced CycleGAN, which addresses this challenge by transforming synthetic BIM images into photorealistic versions, minimizing visual differences between synthetic and real images. This transformation improves feature correspondence and camera pose estimation accuracy across domains.

Recent studies have demonstrated the effectiveness of GAN-based and adversarial domain adaptation strategies in built environment applications beyond localization. For example, Wang [45] proposed a hybrid Synthetic Minority Over-sampling Technique (SMOTE) with Transfer Conditional Wasserstein Generative Adversarial Network (Trans-CWGAN)

framework to address data imbalance in real operational air handling unit fault detection, showing that adversarial generative models can improve robustness and generalization under challenging real-world conditions. Although focused on fault diagnosis rather than localization, this work reinforces the suitability of GAN-based domain adaptation for handling domain shift in operational building systems, supporting its application in MR-based BIM alignment scenarios.

Recent work by Chen and Li [30] demonstrated CycleGAN's effectiveness in indoor localization. By converting BIM renderings into photorealistic images, their method achieved a camera pose accuracy of 1.38 m and 10.1°, significantly reducing the visual gap between synthetic and real images. However, deep learning methods like CycleGAN still face limitations in uniform architectural environments, where the lack of distinctive features makes it difficult to generate detailed images [31].

Beyond GAN-based translation, domain adaptation has also been explored using transformer-based architectures. Wang [46] introduced a transformer-based domain adaptation framework for automated detection of exterior cladding materials in street view imagery, demonstrating that aligning feature distributions across domains significantly improves generalization under domain shift. Although applied to material classification rather than pose estimation, this study provides strong evidence that domain adaptation can effectively mitigate appearance variation across datasets, reinforcing its relevance to BIM-MR image alignment where synthetic and real domains differ substantially.

Sufiyan and Win [47] approached the problem differently by introducing a deep CNN-based workflow for indoor localization using 360-degree panoramic images. Their approach leveraged synthetic data generated from photogrammetry, Open Street Map (OSM), and 3D building models to create comprehensive datasets, leading to improved localization accuracy. Similarly, Hong and Park [48] utilized CycleGAN to enhance scene understanding in indoor facility management. However, like many others, they encountered noise pattern issues during GAN training, affecting the quality of the synthetic data, highlighting the need for better GAN stabilization techniques to ensure higher-quality datasets.

To address some of these challenges, Chen and Yang [49] proposed the CycleGAN-Swin Transformer-SRPnP framework, which optimized global image retrieval and 2D-to-3D image coordinate detection. This approach improved computation time and enhanced robustness against noise and motion blur, common challenges in indoor environments. CycleGAN played a crucial role in reducing visual discrepancies between BIM renderings and real images, further improving localization accuracy.

Acharya and Tatli [31] also investigated synthetic-to-real (S2R-PoseNet) and real-to-synthetic (R2S-PoseNet) adaptation strategies for indoor pose regression. Their findings revealed that real-to-synthetic adaptation outperformed synthetic-to-real adaptation, reducing artifacts from motion blur and incomplete data coverage. This shift emphasizes a growing trend in real-to-synthetic domain adaptation, which simplifies visual matching and reduces the need for highly detailed BIM models, thereby improving localization accuracy in complex environments.

While the aforementioned studies demonstrate the growing success of domain adaptation techniques in improving camera pose estimation, several limitations persist. Model-based approaches such as BIM-PoseNet and CNN-based regressors often struggle in geometrically repetitive indoor environments, where ambiguous visual cues lead to mislocalization and drift accumulation over time. These errors are further exacerbated by discrepancies between synthetic BIM visuals and real-world images, as well as by changes in indoor scenes due to lighting, occlusion, or layout variations. While CycleGAN-based methods have helped bridge the visual domain gap, they have been found to produce unsta-

ble outputs in texture-sparse or visually uniform environments, limiting their effectiveness in challenging conditions.

2.4. Identified Research Gap

Despite significant advances in BIM-assisted and learning-based indoor localization, several limitations remain unresolved. Many existing model-based and deep learning-based indoor localization approaches [42,43,50] remain sensitive to the visual discrepancies between synthetic BIM renderings and real-world images, particularly in environments with uniform textures or repetitive architectural patterns. These methods typically depend on reliable feature extraction or direct image-to-image regression and therefore struggle when distinctive visual cues are scarce. Such limitations result in pose ambiguities and drift accumulation, especially in long trajectories or narrow corridors where geometric features are minimal. Accordingly, there is a growing need for hybrid localization strategies that combine the strengths of sensor-based tracking and model-based refinement.

To address the inherent challenges of low-texture and repetitive indoor environments, this research leverages the spatial mapping and tracking capabilities of the Microsoft HoloLens, whose VISLAM and RGB-D SLAM systems provide a robust localization backbone. These sensing modalities help mitigate pose ambiguity in visually uniform regions, ensuring continuity of tracking even when reliable keypoint extraction is not possible. Because the HoloLens VISLAM maintains tracking performance even in textureless or symmetric spaces, the proposed pipeline can rely on this continuous baseline while applying BIM-guided refinement only when sufficient visual features are present. Building on this foundation, the method integrates CycleGAN-based domain adaptation to minimize the visual gap between real and synthetic images, followed by KAZE-based correspondence extraction and geometric pose estimation using PnP. Rather than depending solely on feature matching, as many prior model-based methods do, the hybrid workflow refines the initial HoloLens pose only where the visual evidence enables reliable correction. In doing so, the approach effectively suppresses accumulated drift and enhances pose accuracy without requiring dense or distinctive textures, clarifying that the contribution lies in drift mitigation rather than resolving the fundamental limitations of correspondence-based localization in highly uniform environments.

By explicitly combining sensor-based tracking with model-based refinement, the proposed workflow positions itself as a complementary enhancement to existing MR localization frameworks. It is not intended to replace end-to-end learned pose regressors or achieve fully reliable localization in all scenes, but rather to provide a robust mechanism for correcting accumulated drift when geometric cues are present. This framing ensures realistic expectations of performance while highlighting the practical value of the method in real-world MR applications where drift is inevitable, and complete reliance on feature correspondences is impractical.

3. Methodology

The proposed workflow enhances indoor localization accuracy by refining HoloLens VISLAM poses through image-based alignment with a BIM-derived virtual environment. The core idea is to exploit geometric correspondences between real images and their synthetic BIM counterparts to estimate camera pose within the BIM coordinate frame. However, substantial appearance differences between real-world images and BIM renderings introduce mismatches that can degrade correspondence quality. To mitigate this discrepancy, the pipeline employs CycleGAN-based domain adaptation to translate real images into a BIM-like style, thereby improving the consistency of visual features across domains. Feature correspondences are then extracted and used within a PnP-based geometric solver to

refine the initial VISLAM pose. As illustrated in Figure 1, the methodology comprises four interconnected stages that include BIM generation, image acquisition, domain adaptation and feature matching, and error analysis. Each stage contributes to progressively reducing the domain gap and improving alignment between the physical and virtual environments.

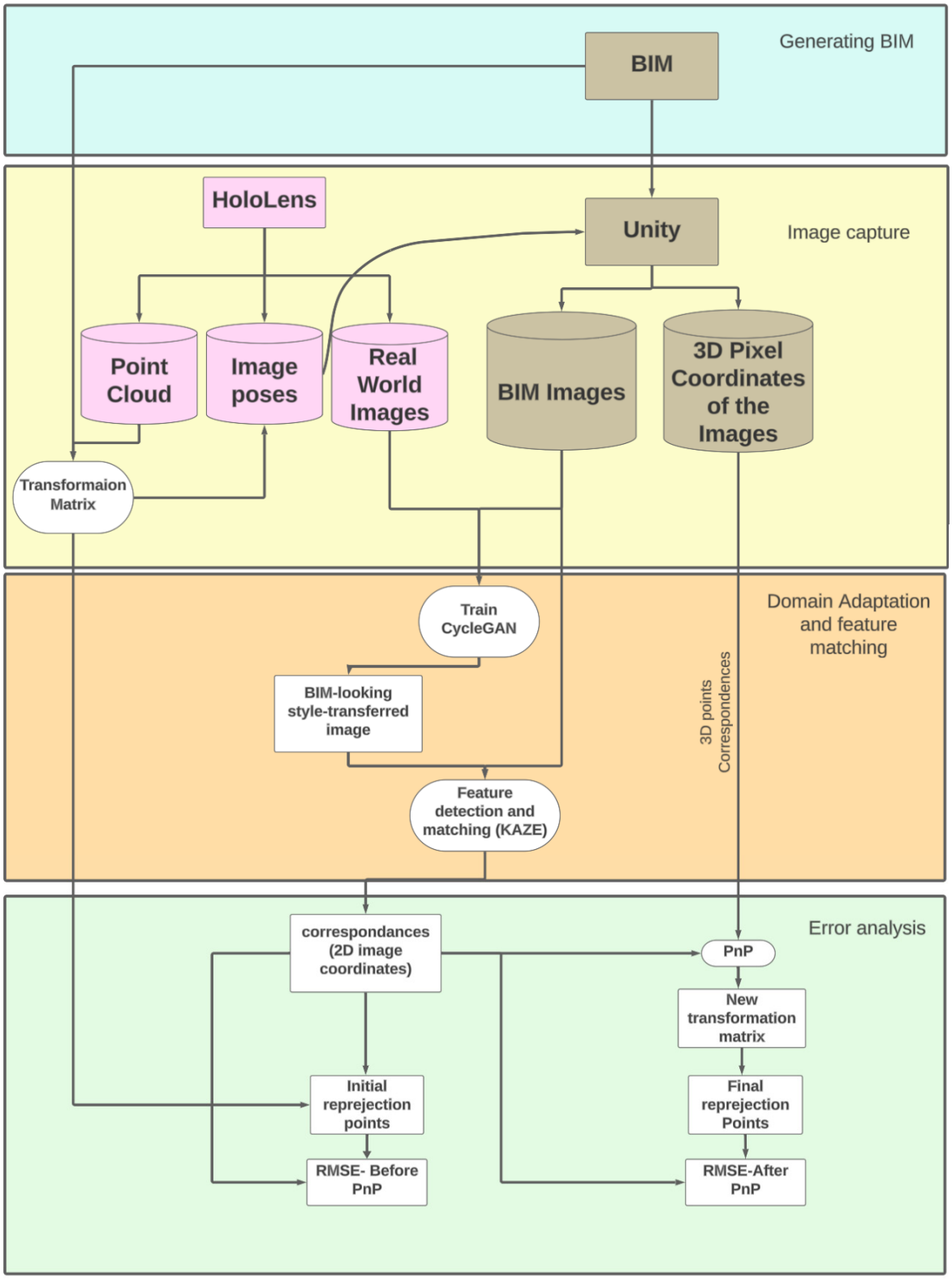


Figure 1. Methodology of the proposed approach, including four stages: BIM Generation, Image Capture, Domain Adaptation and Feature Matching, and Error Analysis.

3.1. Generating BIM

A geometrically reliable BIM is essential for generating synthetic views that can serve as a reference for camera pose refinement. The BIM used in this study was constructed in Autodesk Revit based on a dense point cloud acquired using the GeoSLAM Zeb Horizon mobile laser scanning system. This dataset enabled detailed reconstruction of architectural and structural elements with an accuracy of approximately 1 to 3 cm, ensuring that walls, doors, windows, and other key features were represented with sufficient precision for spatial analysis. To further validate geometric fidelity, critical dimensions such as inter-wall distances were manually verified with an Electro Distance Measurement device, providing an additional check on the consistency between the digital model and the physical environment. Although these verification steps were taken to improve geometric fidelity, residual discrepancies between the BIM and the physical environment may still exist and can influence the accuracy of subsequent pose refinement.

Following the modeling process, the BIM was imported into the Unity platform to generate synthetic images aligned with the HoloLens trajectory. Unity was selected for its ability to efficiently render complex indoor environments and provide pixel-level geometric information [51]. The model was optimized to support real-time rendering by approximating lighting conditions and removing furnishings and nonstructural objects that were not essential for pose estimation. The resulting environment corresponds to a Level of Detail (LoD) 300 BIM, which offers sufficient structural detail for evaluating spatial alignment while maintaining manageable computational complexity (Figure 2).

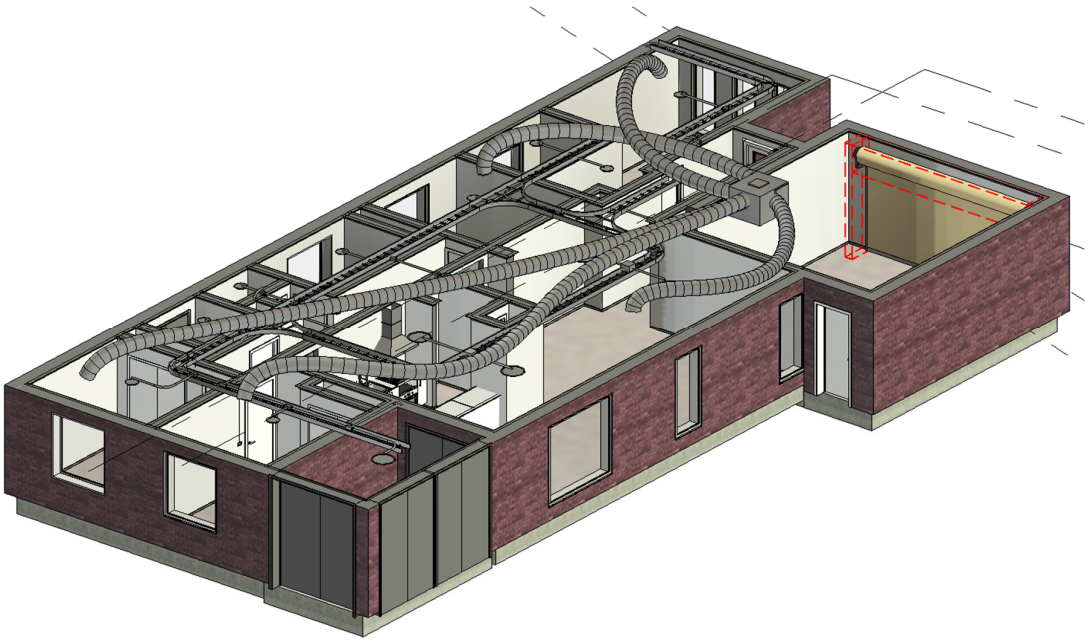


Figure 2. Generated BIM-based on a point cloud captured by a mobile laser scanning system.

3.2. Image Capture

The image acquisition stage involves generating paired real and synthetic datasets that form the foundation for evaluating pose refinement. Real-world images were collected along a predefined indoor trajectory using the Microsoft HoloLens, which provides RGB images, depth information, and associated camera poses estimated through its internal Visual Inertial SLAM system. This dataset structure follows the framework introduced by Ungureanu and Bogo [52], in which each captured image is accompanied by pose estimates derived from depth sensors and grayscale tracking cameras.

Corresponding synthetic images were generated in Unity by rendering the BIM from viewpoints that match the HoloLens trajectory. In addition to the rendered imagery, Unity provided pixel-level 3D coordinates that serve as geometric ground truth for subsequent pose correction. These coordinates enable direct comparison between the estimated and true projections of BIM points, making them essential for quantitative evaluation of reprojection error.

In total, 1408 real images and their BIM-generated counterparts were acquired for analysis (Figure 3). This paired dataset supports both domain adaptation through CycleGAN and the assessment of pose refinement accuracy through feature correspondence and geometric alignment.



(**a**) (**b**)

Figure 3. (**a**) Sample BIM images along the trajectory (**b**) Corresponding real images.

3.3. Domain Adaptation and Feature Matching

To reduce the visual discrepancy between real-world HoloLens images and synthetic BIM renderings, CycleGAN was employed. CycleGAN is well-suited for unpaired image-to-image translation, enabling domain adaptation without requiring exact correspondence between input datasets. In this study, the network was trained on unpaired sets of BIM renderings and HoloLens captures to generate style-transferred images that preserve geometric structure while approximating the textures and illumination characteristics of real scenes [44]. This transformation enhances appearance consistency and improves the likelihood of obtaining stable image correspondences.

Following domain adaptation, feature correspondences were extracted between the CycleGAN-translated images and their corresponding BIM images using the KAZE feature detector. KAZE was selected after preliminary empirical trials indicated that it performs robustly under the nonlinear intensity variations introduced by the style-transfer process. Its nonlinear scale-space construction makes it more resilient to intensity distortions compared to traditional methods [53,54]. While alternative descriptors such as SIFT, ORB, and SURF could also be applied, the objective of this work was to establish a feasible and effective pipeline for drift refinement rather than to conduct a comprehensive comparison of feature extraction algorithms. We acknowledge that KAZE, like all keypoint-based approaches, remains limited in environments with uniformly textured or symmetric surfaces, and this inherent constraint contributed to the set of image pairs for which correspondences could not be reliably established.

To obtain refined camera poses, 2D-3D correspondences derived from the matched keypoints were used as input to a PnP solver, which computes the rotation and translation, aligning the HoloLens camera frame with the BIM coordinate system [55]. The resulting transformation enables accurate projection of BIM-derived 3D points into the corresponding 2D image plane and forms the basis for correcting accumulated drift in the HoloLens trajectory.

It is important to note that the proposed pipeline does not rely solely on feature matching for continuous localization. The HoloLens VISLAM system provides stable tracking even in textureless or symmetric regions where keypoint-based methods struggle, ensuring continuity when insufficient features are available. The refinement step is therefore applied only when the visual evidence supports reliable correction. This hybrid design strengthens localization performance while acknowledging the practical limitations of both deep learning-based image translation and traditional keypoint detection methods, which we discuss further in the Limitations and Discussion sections.

It should be emphasized that the proposed refinement process is applied opportunistically. Feature matching and PnP-based pose correction are performed only when sufficient and reliable correspondences are detected. In scenes with limited texture or high symmetry, where correspondence extraction is unreliable, the refinement step is skipped. During these periods, the HoloLens VISLAM system continues to provide baseline pose tracking, ensuring continuity of localization without interruption.

3.4. Error Analysis

The final stage of the methodology evaluates the accuracy of camera pose refinement by quantifying the alignment between the real and synthetic image domains. Using the transformation estimated by the PnP algorithm [56], 3D points from the BIM were reprojected into the 2D image plane and compared with their corresponding feature locations in the CycleGAN-translated HoloLens images. The discrepancy between these points was measured using the Root Mean Square Error (RMSE) which provides a quantitative indicator of alignment accuracy and is computed as the Euclidean distance between projected and observed feature coordinates [57].

Two error metrics were used to assess the contribution of the proposed pipeline. The first, referred to as RMSE-before, represents the reprojection error associated with the initial HoloLens VISLAM pose prior to any domain adaptation or PnP refinement. The second metric, RMSE-after, represents the error following the application of CycleGAN-based image translation and PnP-based pose correction. A substantial reduction in RMSE-after relative to RMSE-before indicates that the combined domain adaptation and geometric refinement stages successfully mitigate accumulated drift and improve spatial correspondence between the virtual and physical environments.

This comparison provides direct evidence of the effectiveness of the proposed workflow in improving camera pose accuracy and supports its suitability for MR localization tasks where reliable alignment between BIM and real-world imagery is essential. It is important to note, however, that reprojection RMSE in this study is used to assess geometric consistency rather than absolute localization accuracy. The 2D correspondences employed in the RMSE computation are obtained through feature matching between CycleGAN-translated images and BIM renderings and may therefore contain inaccuracies.

To reduce the influence of unreliable correspondences, only feature matches that satisfy geometric constraints during the PnP estimation are retained, and image pairs with insufficient or unstable matches are excluded from the evaluation. Accordingly, the reported RMSE reflects the degree to which the refined camera pose is internally consistent with the BIM geometry and image observations. A reduction in RMSE after

pose refinement, therefore, indicates effective suppression of accumulated drift relative to the BIM reference, rather than absolute positional accuracy with respect to an external ground-truth coordinate system.

4. Experiments

This study was conducted to evaluate the consistency, and effectiveness of the proposed localization enhancement methodology within a controlled offline setting. MATLAB (version R2023b) served as the primary computational environment due to its versatile image processing, computer vision, and mathematical analysis capabilities. All datasets, including those captured from Unity, HoloLens, and CycleGAN, were carefully imported and organized within MATLAB to enable an integrated and iterative experimental workflow.

4.1. HoloLens Data Acquisition

Before initiating formal image acquisition, the head-mounted HoloLens moved along a predefined trajectory within a residential indoor environment. This preliminary phase was essential for allowing the HoloLens to build a consistent spatial understanding of the environment, thereby minimizing tracking drift and improving the accuracy of pose estimation during actual image capture. The goal was to establish a stable operational context, which is critical for reliable data acquisition in real-world MR scenarios. It is important to note that it is not always possible during practical situations.

Following this initialization, the HoloLens device captured 1454 real-world RGB images at a consistent frame rate of 30 frames per second (fps). The trajectory followed by the operator is illustrated in Figure 4, with the start and end locations designated as point "A." The chosen path covered varied lighting, geometry, and material conditions within the indoor environment.

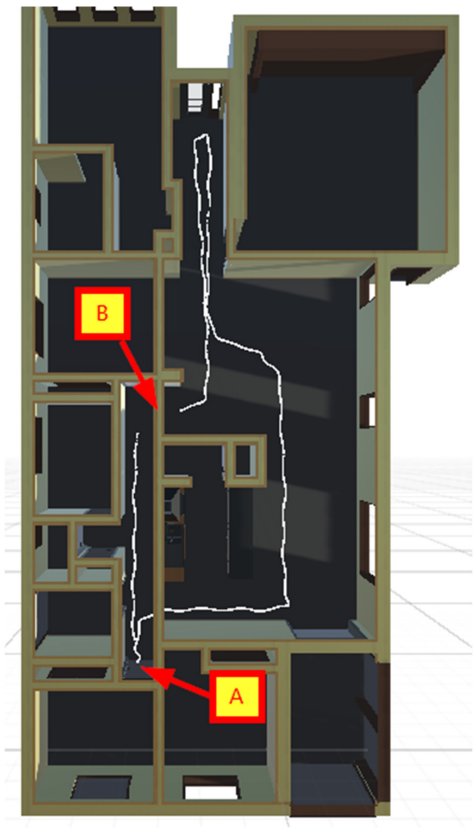


Figure 4. HoloLens Trajectory (white points) with the start and end locations as point "A" and two abrupt turns over a short distance as point "B".

A significant challenge was encountered in a narrow corridor denoted by "B" in Figure 4. This section involved two abrupt turns over a short distance, which posed difficulties for the HoloLens in maintaining accurate pose estimation. Consequently, 46 images from this section were deemed unreliable due to incorrect or missing pose data. Despite multiple attempts to re-capture data in this specific corridor, the localization failures persisted, and the associated frames were ultimately excluded from the final dataset.

The dataset, post-processing, included RGB images, corresponding camera pose information in the HoloLens local coordinate system, and spatially contextualized point cloud segments generated from the HoloLens' internal depth sensors. These elements together formed a foundational multimodal dataset necessary for subsequent alignment, synthetic data generation, and evaluation stages.

4.2. Registration of HoloLens Point Cloud with BIM

It was crucial to align the HoloLens and BIM coordinate systems to capture BIM images within Unity. This alignment step provided the spatial transformation required to convert real-world camera poses into the coordinate space used by the BIM, thereby ensuring consistency between synthetic and real-world datasets.

The initial step in this alignment process involved merging several segmented point clouds obtained from HoloLens image segments into a single cohesive 3D point cloud. This comprehensive point cloud was imported into CloudCompare software, (version v2.12 alpha), which was used to perform a two-step registration process. The first step involved a coarse alignment using point-pair registration, allowing rough alignment based on manually selected reference features. The second step involved fine-tuning through ICP registration, which minimized the Euclidean distance between corresponding point features in the merged HoloLens point cloud and the BIM-derived point cloud (Figure 5). The final registration achieved an error with an RMSE of 0.024. The transformation matrix, denoted as T_{HC}, which accurately mapped HoloLens spatial data into the BIM coordinate system:

$$T_{HC} = \begin{bmatrix} 1.000 & 0.006 & -0.006 & -1.0395 \\ -0.035 & -0.001 & -1.000 & 0.479 \\ -0.007 & 1.000 & 0.000 & 1.1738 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

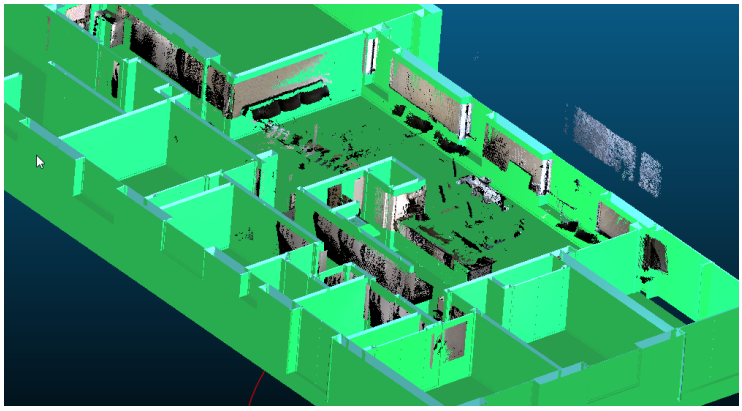


Figure 5. Registered point cloud (color) to BIM (green).

This matrix served as a critical spatial bridge, enabling the transformation of all HoloLens poses into the BIM's reference frame for direct comparison and data fusion. Although this step is performed manually in this work, a range of algorithms exists that can facilitate automated registration in real-time applications

4.3. Unity Data Preparation and Image Capture

After generating the BIM, it was subsequently imported into Unity for the purpose of capturing BIM imagery. To ensure that the BIM and real images have identical geometry, the same intrinsic camera settings as those of the HoloLens RGB camera used to capture real-world images were integrated into the virtual camera in Unity. The series of BIM images was captured in Unity as explained in Section 3.2.

4.4. CycleGAN Training for Domain Adaptation

In order to reduce domain discrepancies between real and synthetic images, a CycleGAN model was trained for unpaired image-to-image translation [44]. The dataset comprised BIM-rendered images and HoloLens images with no one-to-one pairing; these were split in a 9:1:1 ratio into training, validation, and test subsets. Following the original architecture, we used generators based on nine residual-block architectures and discriminators employing the PatchGAN paradigm. Instance normalization was applied consistently across both generators and discriminators to stabilize style transfer and preserve structure. The model's training objective balanced three loss components: adversarial loss to encourage realism in translated images, cycle-consistency loss to enforce that mapping there and back returns the original image, and identity loss to prevent unnecessary style shifts when input already lies in the target domain. During training, various checkpoints were evaluated using the validation set to assess the trade-off between visual realism and structural fidelity. Ultimately, the model at epoch 200 produced the best style-transferred results, renderings that most closely matched real-world textures while maintaining the geometric integrity of BIM structures (Figure 6a,b). As a note, the trained model was observed to suppress non-BIM semantic content, such as furniture and pendant lights, since these elements do not appear in the BIM domain, resulting in style-transferred images where structural geometry is preserved while nonstructural objects are visually de-emphasized.



(**a**) (**b**)

Figure 6. (**a**) Real images, (**b**) CycleGAN style-transferred images.

4.5. Image Rescaling

Although the original image resolution for both BIM and HoloLens datasets was 760 × 428 pixels, CycleGAN's architecture resized all training inputs to 256 × 256 pixels. To restore spatial consistency, a rescaling operation was conducted using factors of 2.968 (width) and 1.672 (height), bringing the generated CycleGAN outputs back to their

original resolution. During the rescaling, the nearest-neighbor interpolation resampling is employed (Figure 7).

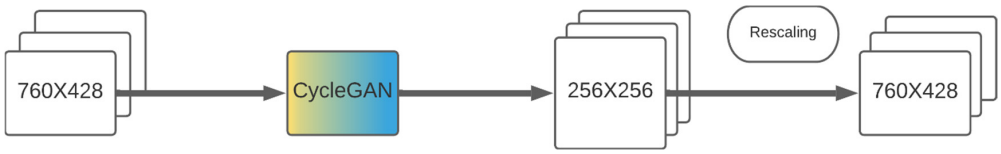


Figure 7. Image Rescaling.

4.6. Image Matching

The rescaled CycleGAN-transformed images and their corresponding BIM images underwent feature matching using the KAZE [54] algorithm implemented in MATLAB using "detectKAZEFeatures" function. The images were first converted to grayscale, and keypoints were extracted using KAZE, known for its robustness to non-linear illumination and scale changes. The descriptors were matched between image pairs, and the matched keypoints were visualized and color-coded for interpretability (Figure 8).

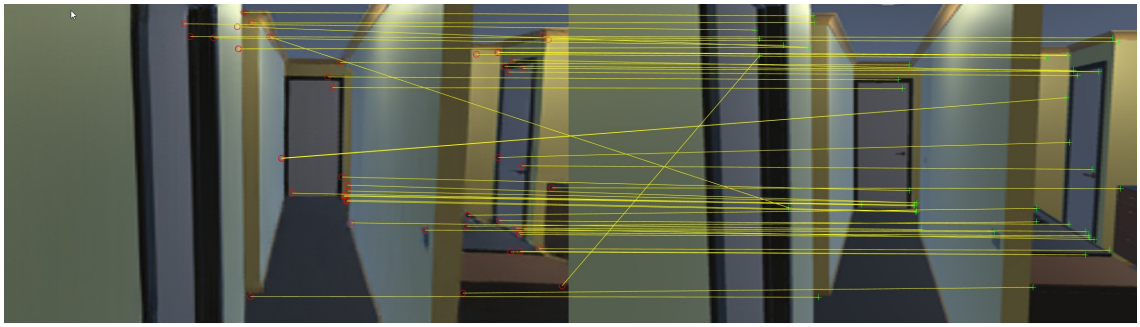


Figure 8. Keypoints ((**Left**): CycleGAN style transferred image; (**Right**): Corresponding BIM image).

4.7. PnP Pose Estimation

To refine the estimated camera poses and correct accumulated drift, the PnP algorithm was employed to compute the transformation between the image space and the BIM's 3D coordinate system. Specifically, the "estimateWorldCameraPose" function in MATLAB was used to solve the PnP problem by aligning 2D image coordinates extracted from CycleGAN images with their corresponding 3D points from the BIM, as explained in Section 3.2.

The 3D spatial coordinates were retrieved from a pre-generated dataset of BIM exported during Unity rendering, while the 2D image coordinates were extracted through feature matching as outlined in Section 3.3. These correspondences were passed to the PnP solver, which uses the Perspective-Three-Point (P3P) algorithm as its underlying method. The P3P approach provides an efficient closed-form solution, especially suitable when at least four 2D-3D point correspondences are available.

To enhance robustness against mismatches and noise, the M-estimator Sample Consensus (MSAC) [18,58] was used to reject outlier correspondences with reprojection errors exceeding 2 pixels. The MSAC implementation involved a maximum of 2000 iterations and a 99% confidence level, ensuring reliable pose estimation even in the presence of challenging visual conditions or erroneous matches.

4.8. Reprojection of 3D Points

To validate the improvement in localization accuracy, the reprojection of 3D BIM points onto the 2D image plane has been carried out before and after applying the PnP algorithm.

This comparison has been used to quantify the drift errors present in the initial HoloLens poses and to demonstrate the refinement achieved through the proposed method.

The initial camera poses R_{tran}, T_{tran} have been extracted from HoloLens tracking data and used to project known 3D BIM coordinates into the 2D image plane, resulting in the initial set of reprojected points. The corrected camera poses $R_{cam,}$ T_{cam} have been estimated through the CycleGAN-enhanced PnP algorithm based on matched 2D-3D keypoint correspondences. Both sets of projections have been computed using the intrinsic parameters of the HoloLens RGB camera, calibrated before experimentation.

The 2D correspondences have been extracted from CycleGAN-translated images using geometric feature matching techniques (as described in Section 3.3). These 2D image points have been compared with the reprojected BIM points derived from both the initial and corrected poses.

As illustrated in Figure 9, green points denote the projections based on the refined pose, representing the expected location of features in the absence of drift. In contrast, red points have represented the projections from the initial HoloLens poses, highlighting the effect of accumulated drift.

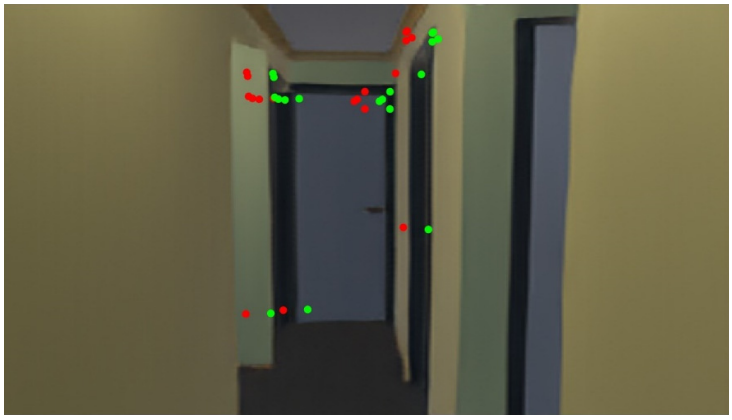


Figure 9. Reprojected points, initial reprojected points (red), refined reprojected points (green).

4.9. Error Evaluation

The accuracy of the camera pose refinement has been quantitatively evaluated by computing RMSE between the 2D image correspondences and the reprojected points generated using both initial and corrected poses using the following formulas. EMSE-before has been calculated for the initial HoloLens poses, whereas the RMSE-after has been derived using the refined values. The value is in image pixels.

$$\text{RMSE-before} = \sqrt{\frac{1}{N}\sum_{1=1}^{N}\left\|P_i^C - P_i^{before}\right\|^2} \quad (1)$$

$$\text{RMSE-after} = \sqrt{\frac{1}{N}\sum_{1=1}^{N}\left\|P_i^C - P_i^{after}\right\|^2} \quad (2)$$

where P_i^{after} is the reprojected point using the estimated camera pose ($R_{cam,}$ T_{cam}), P_i^{before} is the reprojected point using the HoloLens pose (R_{tran}, T_{tran}), N is the number of inlier points/correspondence in each image pair, and P_i^C is the corresponding 2D image point.

These RMSE values have been used to assess the geometric accuracy of the alignment process and to validate the impact of the proposed method in correcting accumulated drift. The 2D correspondences have been treated as ground truth, and reductions in RMSE have indicated improved localization performance.

The evaluation has confirmed that the proposed CycleGAN-enhanced pose refinement pipeline significantly reduced trajectory drift and improved spatial alignment between real and virtual environments across the 1408 tested image pairs.

5. Results and Discussion

The comprehensive evaluation of the proposed methodology was conducted systematically, repeating the entire process for all 1408 captured image pairs. To ensure statistical significance and enhance the reliability of the findings, a MATLAB-based computational workflow was executed iteratively 100 times for each image pair. This thorough approach facilitated the calculation of the RMSE for each pair, effectively capturing the average reprojection error between the initial and refined camera poses.

Figure 10 illustrates the distribution of RMSE values for each image pair, providing a comparative analysis of pose estimation accuracy across the entire dataset. The red line in the graph represents the RMSE prior to the application of the PnP, with values ranging approximately from 1 to 90 pixels, indicating a substantial degree of drift. In contrast, the blue line illustrates the RMSE after the implementation of the PnP, showcasing a remarkable reduction in error to a range of 1 to 2 pixels. The vertical axis is scaled logarithmically to improve visibility, allowing a clearer comparison of the differences between the two phases of the methodology and highlighting the effectiveness of the pose refinement process.

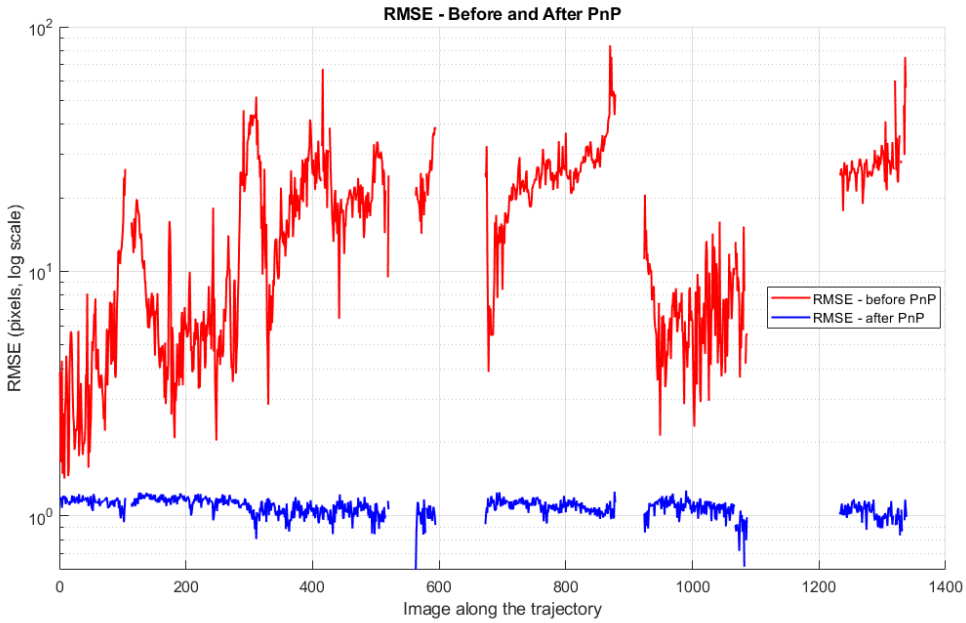


Figure 10. RMSE in each image pair along the trajectory of the camera.

Gaps in the RMSE curves correspond to image pairs for which refinement could not be performed, because the feature-matching stage did not yield a sufficient number of reliable correspondences. These cases predominantly occurred in low-texture or symmetric environments where correspondence extraction remains inherently difficult. However, unlike prior BIM-MR localization methods that completely lose tracking under such conditions, the HoloLens VISLAM system maintained a usable trajectory throughout these intervals. These unrefined poses appear as gray points in Figure 11, indicating that only the drift-correction step was unavailable, while baseline tracking remained intact. The only exception is Location B in Figure 4, where coarse alignment was intentionally omitted to demonstrate an extreme failure case.

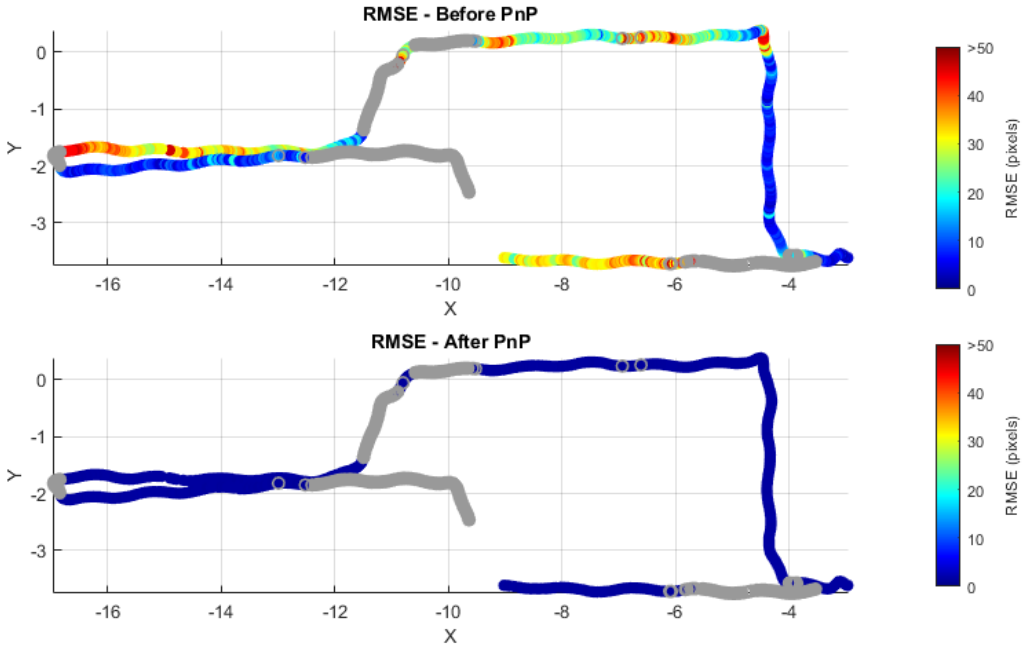
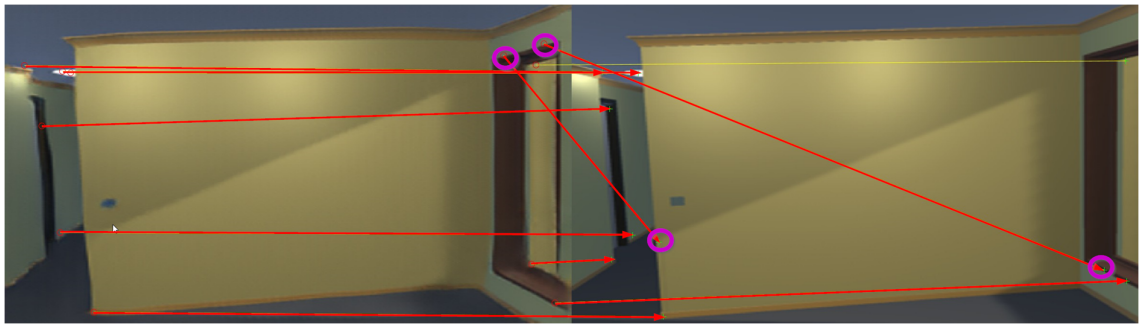


Figure 11. Colorized initial and final RMSE along the trajectory before and after PnP.

To visualize the spatial distribution of these outcomes, Figure 11 presents a colorized RMSE map along the camera trajectory. The plot shows both initial and refined reprojection errors at their corresponding spatial locations, making the impact of drift correction more interpretable within the context of the scene. The gray points denote positions where refinement could not be applied due to insufficient correspondence, yet the HoloLens continued to provide stable tracking with last known drift correction applied. This behavior contrasts with the substantially larger initial drift observed at the same locations prior to applying the proposed refinement pipeline. Together, these results demonstrate that while the refinement process is limited by the availability of geometric cues, the overall system maintains localization continuity and effectively mitigates drift whenever conditions permit.

It is important to note that not all 1408 image pairs were included in the RMSE analysis. Specifically, 398 image pairs were excluded due to an insufficient number of reliable feature correspondences required for accurate PnP estimation. A minimum threshold of 10 inlier correspondences was established based on empirical tuning, ensuring a balance between analytical coverage and pose estimation accuracy. Lowering this threshold increased the total number of usable image pairs, but it also led to a higher occurrence of erroneous or spurious feature matches, thereby compromising the reliability of the estimated poses. This trade-off is illustrated in Figure 12, where an example of an image pair with erroneous correspondences is shown, emphasizing the necessity of enforcing a minimum inlier constraint. Importantly, the image exclusions do not represent localization failure. As illustrated by the uncorrected trajectory points, the HoloLens VISLAM system remained active during these intervals, and only the refinement component was unavailable. This behavior contrasts with correspondence-only localization methods, which typically experience complete tracking failure under similar conditions. Additionally, reprojection RMSE is reported in pixel units, it provides a meaningful indicator of pose consistency with respect to the BIM-derived geometric reference. For typical indoor environments and camera-to-surface distances of approximately 2–5 m, a reprojection error of 1–2 pixels corresponds to a translational misalignment on the order of a few millimeters to centimeters, depending on camera intrinsics and viewing geometry. Accordingly, the observed reduction from tens of pixels to approximately 1–2 pixels indicates substantial suppression of accumulated drift

relative to the BIM model, even though absolute translation and rotation errors in metric units were not explicitly computed in this study.



(**a**) (**b**)

Figure 12. (**a**) CycleGAN image, (**b**) BIM image, erroneous correspondence (circled), arrows indicate the matched points.

Further, the environment was segmented into distinct sections to facilitate a region-specific analysis. Section A, located at the beginning of the trajectory, demonstrated consistently low RMSE values, indicating high accuracy in localization during the initial phase (Figure 13).

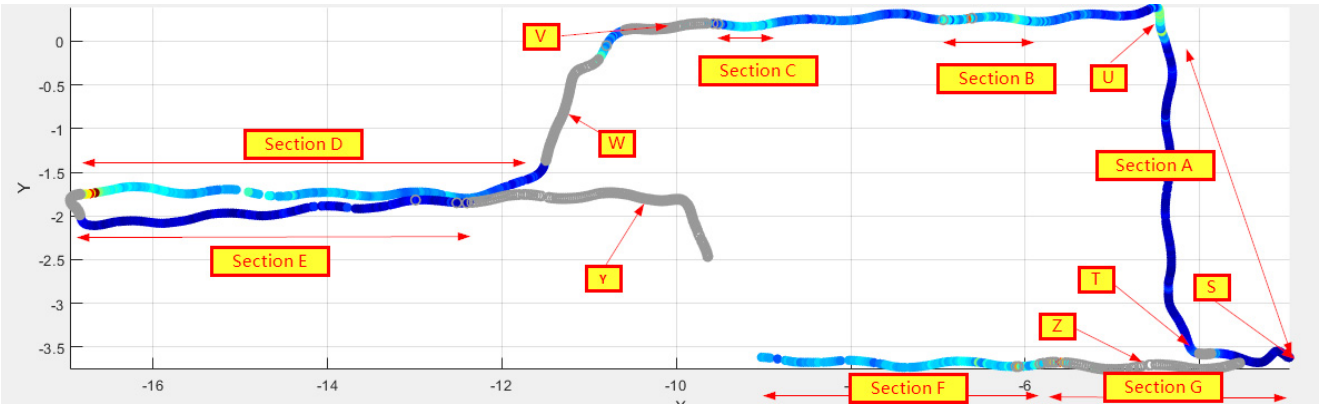


Figure 13. Section analysis.

Minimal reprojection error observed in Section A (Figure 14a) in the start of the trajectory. However, in sections involving turns, such as Turning Points "T" and "U", a noticeable increase in RMSE was observed (Figure 14b,c). This rise in error is attributed to motion-induced blur and reduced image sharpness, which affected the CycleGAN-generated imagery and led to compromised localization from the HoloLens.

Sections B and C, characterized by a wider hallway and fewer distinctive features, showed progressively increasing RMSE values. The larger scale and uniform textures of these regions posed challenges for HoloLens mapping and further contributed to the accumulation of pose estimation errors (Figure 14d,e).

In Section D, RMSE continued to increase due to compounding trajectory errors. Nevertheless, a sharp reduction in RMSE occurred at the start of Section E. This improvement resulted from the camera's ability to view extended spatial features, allowing the relocalization process to self-correct based on the broader field of view and increased environmental cues (Figure 14f,g). Conversely, Section E's confined geometry limited feature visibility, preventing effective relocalization (Figure 14h,i).

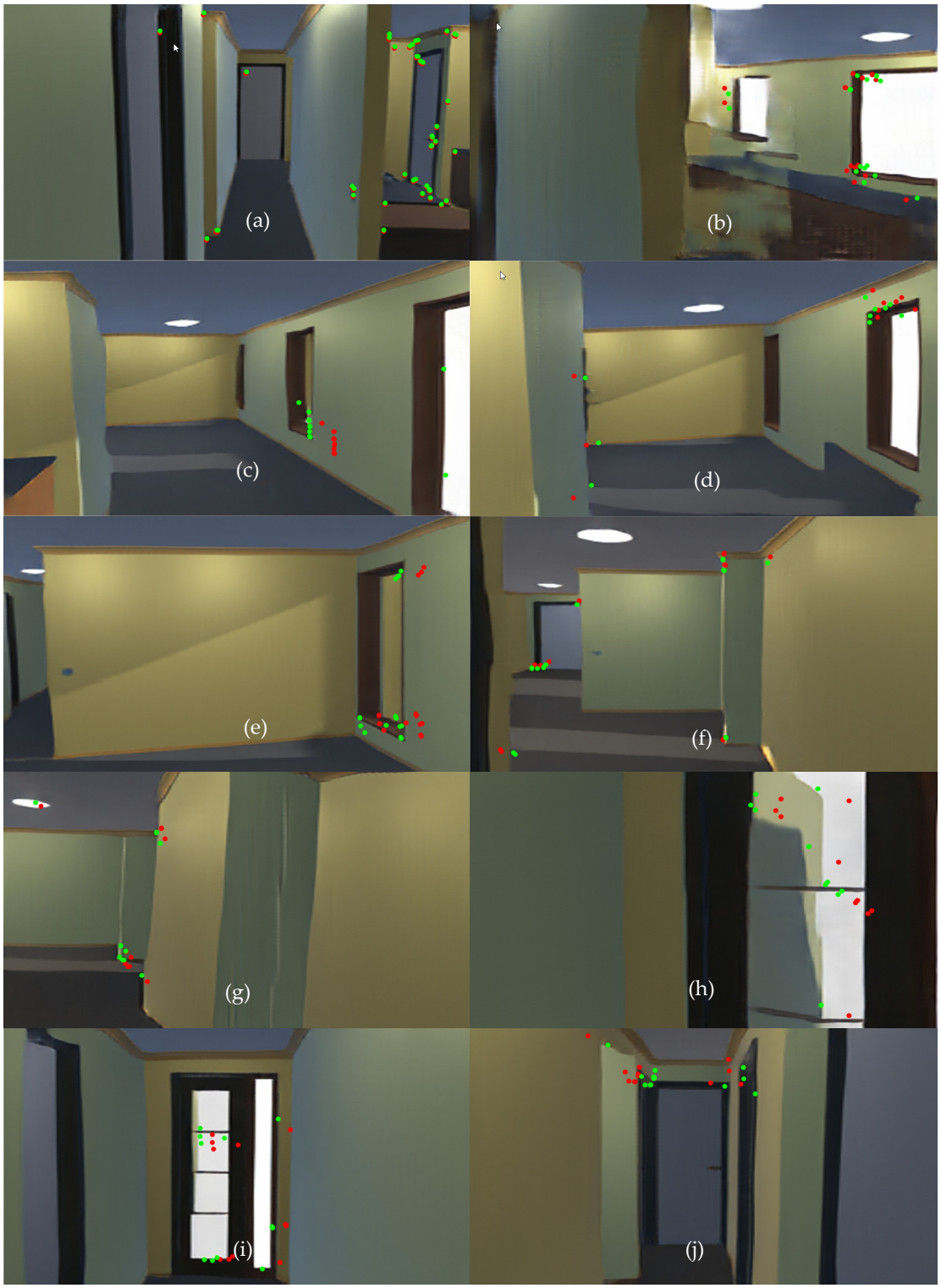


Figure 14. (**a**) Start of the trajectory (S), (**b**) Turning Point (T), (**c**) Turning Point (U), (**d**) Middle of section B, (**e**) Middle of Section C, (**f**) Middle of Section D, (**g**) Middle of Section E, (**h**) End of Section D (**i**), the beginning of Section E, (**j**) Middle of Section F, initial reprojected points (red), refined reprojected points (green).

Section F presented some of the highest RMSE values across the entire dataset. This trend is associated with the prolonged accumulation of errors due to the drift and the limited number of distinctive features available for accurate relocalization (Figure 14j). The final segment, spanning image pairs from index 1351 to 1408, was particularly problematic.

Many of these frames were excluded from RMSE calculations due to insufficient feature correspondences, often because the camera's view was dominated by homogeneous elements such as plain doors or featureless walls (Figure 15).

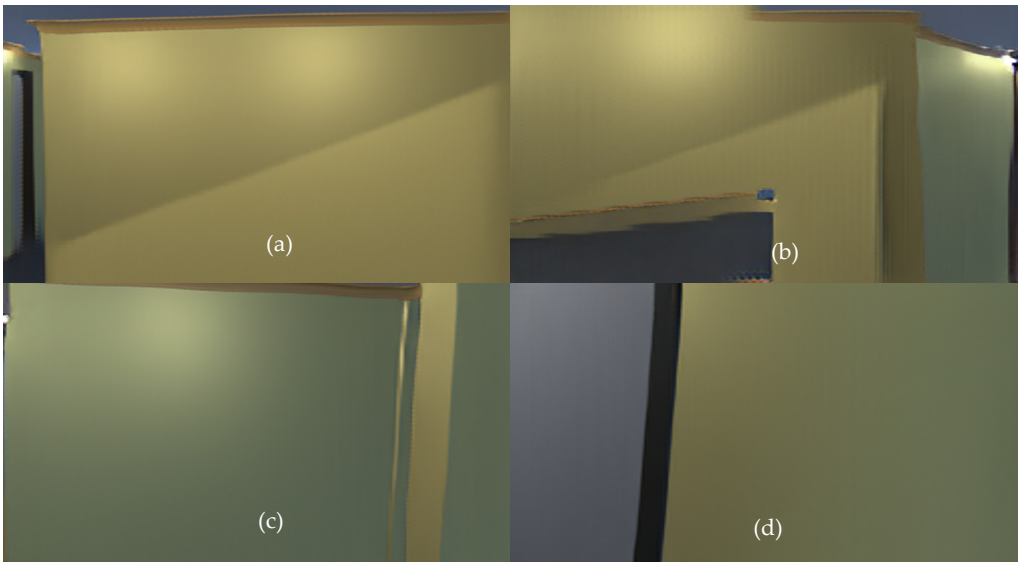


Figure 15. Removed image pairs from calculating RMSE: (**a**) location V (**b**) location W (**c**) location Y (**d**) location Z.

Overall, correspondence failures predominantly occur in three scene categories:

- Textureless or visually uniform corridors,
- Repetitive architectural layouts, and
- High-motion turning regions.

Importantly, correspondence failure in these areas did not equate to complete localization failure. While BIM-guided refinement could not be applied, the HoloLens VISLAM system continued to provide a continuous and operational pose estimate, with drift increasing gradually rather than abruptly. Operationally problematic localization, defined as sudden pose jumps or loss of tracking, was observed only in rare cases involving rapid motion or extreme lack of visual cues. This analysis confirms that the proposed framework functions as a selective drift-refinement mechanism, enhancing localization when visual conditions permit while safely deferring to the VISLAM baseline in feature-sparse environments.

To further quantify the distribution of pose estimation accuracy, we generated a Cumulative Distribution Function (CDF) plot of RMSE-before (red line) and RMSE-after (blue line), as illustrated in Figure 16. The x-axis is scaled logarithmically to enhance the visual representation of both lines on the graph. This plot provides an aggregated statistical perspective, indicating that only 60% of RMSE values are around 20 pixels, while 90% of RMSE values exceed 30 pixels before implementing the PnP method. However, after applying PnP, all 100% of RMSE values are less than 2 pixels. These results validate the effectiveness of the proposed localization refinement framework and underscore its robustness in significantly reducing drift errors that can accumulate along the trajectory over distance and time.

In this study, drift refers to the accumulation of pose estimation error over time in the HoloLens VISLAM trajectory, manifested as increasing misalignment between real-world images and the BIM-derived virtual scene. Rather than measuring drift directly in metric position or orientation units, drift correction is evaluated through reprojection error, which reflects the consistency between estimated camera pose and BIM geometry. A trajectory segment is considered effectively drift-corrected when reprojection RMSE is reduced to

approximately 1–2 pixels following pose refinement, indicating that accumulated drift relative to the BIM reference has been suppressed. Accordingly, the term "drift-free" is used in a relative sense to denote negligible residual drift with respect to the BIM model, rather than absolute elimination of localization error in physical space.

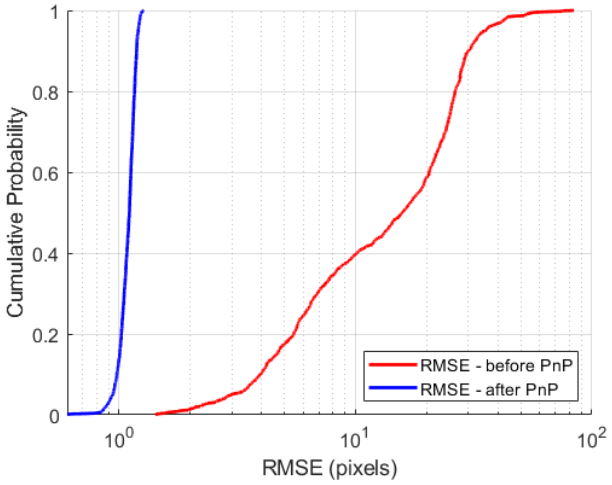


Figure 16. CDF plot of alignment errors before and after PnP.

6. Conclusions

This research presented a hybrid localization refinement framework that integrates HoloLens VISLAM tracking with BIM-based geometric alignment, CycleGAN-driven domain adaptation, and feature-based pose estimation. The primary contribution lies in demonstrating that accumulated drift in MR device trajectories, one of the most common limitations in extended MR operation, can be substantially reduced by leveraging style-transfer, feature-matching and PnP-based correction. Across 1408 image pairs, the proposed workflow consistently reduced reprojection error from tens of pixels to below two pixels whenever reliable feature correspondences were available. As a result, the approach improves spatial alignment between real and virtual environments and enhances the reliability of MR visualization for construction-related applications.

Unlike previous BIM-MR localization approaches that depend exclusively on feature correspondences and therefore fail in symmetric or textureless areas, our system maintains tracking through the HoloLens VISLAM, which remains active even when few or no geometric features are detected. The refinement stage then corrects accumulated drift only when sufficient geometric cues exist, ensuring that the system benefits from both continuous sensor-based tracking and BIM-informed correction. This structure clarifies the scope of the contribution and highlights how BIM-assisted refinement enhances, rather than competes with, current SLAM-based MR pipelines.

Beyond the technical contributions and quantitative improvements demonstrated, the outcomes of this study also have clear practical implications for multiple stakeholder groups. For MR system developers, the proposed framework illustrates how BIM-guided pose refinement can be integrated with existing VISLAM pipelines to mitigate accumulated drift without replacing native tracking mechanisms. For construction practitioners and facility managers, improved alignment stability enables more reliable MR visualization for inspection, coordination, and spatial decision-making tasks, reducing the need for repeated manual calibration during on-site operations. From a research perspective, the experimental analysis and identified limitations provide insight into the conditions under which domain adaptation and feature-based refinement are effective, highlighting opportunities to advance hybrid localization strategies that balance continuous sensor-based tracking

with selective model-based correction. Collectively, these outcomes position the proposed method as a practical enhancement to current MR localization workflows rather than a standalone localization solution.

While the results show strong potential, several limitations still exist. Firstly, the CycleGAN model was trained on a relatively small, fixed dataset, and its performance in dynamic or cluttered construction environments has not yet been tested. Second, although the KAZE feature extractor proved empirically effective under the nonlinear intensity variations introduced by style transfer, its performance was not evaluated using quantitative metrics such as repeatability, inlier ratio, or matching precision, and it remains limited in low-texture or highly repetitive environments, as is the case for all keypoint-based methods. Similarly, while quantitative image-quality metrics for CycleGAN could provide complementary insight, these analyses were not included because the focus of this research is on improving localization accuracy rather than assessing the standalone performance of the style-transfer or feature-extraction components. Incorporating extensive metric-based evaluations for CycleGAN and KAZE would broaden the scope of the study and risk diverting attention from its central objective, which is to demonstrate the effectiveness of the proposed drift-refinement pipeline. For this reason, both components were evaluated indirectly through their contribution to reducing reprojection error in the final pose estimation stage, and more comprehensive quantitative evaluations are identified as important directions for future work.

Third, coarse alignment between BIM and HoloLens point clouds relied on manual selection of reference points. Although this step served only as initialization, operator variability may influence the starting transformation, and automated registration strategies such as ICP-based registration or learning-based alignment could replace this step without altering the core drift refinement pipeline would be beneficial in future implementations. Fourth, the current pipeline was executed offline using MATLAB, which limits its immediate applicability for real-time MR deployment due to the computational constraints of the HoloLens and similar mobile devices. Fifth, absolute translation and rotation errors were not reported because reliable ground-truth poses in metric space were not available for the entire trajectory, and the focus of this work was on relative drift correction with respect to the BIM reference rather than absolute localization accuracy. And finally, the proposed BIM-guided refinement framework assumes that the available BIM reasonably represents the physical environment. While the model used in this study was constructed from laser scanning data and manually verified, a formal sensitivity analysis quantifying the impact of BIM geometric errors on pose refinement accuracy was not conducted. As a result, the reported improvements reflect drift reduction relative to the available BIM accuracy rather than absolute localization performance.

These limitations outline several important directions for future research. A primary priority is expanding CycleGAN training to include more diverse and dynamic construction datasets, accompanied by quantitative evaluations of image translation quality using complementary metrics alongside reprojection error. Further improvements may be achieved by incorporating learning-based keypoint detection or multi-view geometric constraints to enhance correspondence extraction in texture-poor and repetitive environments. Replacing manual coarse alignment with automated registration strategies suitable for on-device execution will also improve scalability and practical deployment. In addition, benchmarking the proposed framework against emerging BIM-MR localization approaches and conducting systematic evaluations of BIM geometric error sensitivity, including controlled perturbations and as-built deviations, would provide deeper insight into performance robustness.

Although reprojection RMSE serves as a meaningful indicator of geometric alignment consistency, future evaluations could be strengthened through independent ground-truth validation, such as manually verified correspondences, fiducial targets, or controlled benchmark datasets to quantify absolute pose accuracy. Addressing these research directions will support the development of a fully integrated, real-time localization enhancement framework that advances MR visualization reliability, construction progress monitoring, and spatial decision-making in complex built environments.

Author Contributions: Conceptualization, M.Z.A.M., D.S., K.K. and D.A.; methodology, M.Z.A.M.; software, M.Z.A.M.; validation, M.Z.A.M., D.S. and D.A.; formal analysis, M.Z.A.M.; investigation, M.Z.A.M.; resources, D.S. and K.K.; data curation, M.Z.A.M.; writing—original draft preparation, M.Z.A.M.; writing—review and editing, D.S., K.K. and D.A.; visualization, M.Z.A.M.; supervision, D.S. and K.K.; project administration, D.S.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the University of Melbourne (Application Reference: 644655. 2020). The authors did not receive funding to cover article processing charges.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request. Source code is publicly available at https://github.com/Mabdulmuthal/MR-localization (accessed on 17 February 2026).

Acknowledgments: The authors express their sincere appreciation to the University of Melbourne for providing access to facilities, software, and computational resources that supported this research.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

AR	Augmented Reality
APR	Absolute Pose Regression)
AEC	Architecture, Engineering and Construction
BIM	Building Information Modeling
CDF	Cumulative Distribution Function
CNN	Convolutional Neural Network
CycleGAN	Cycle-Consistent Generative Adversarial Network
DoF	Degrees of Freedom
FPS	Frame per Second
GNSS	Global Navigation Satellite System
LoD	Level of Detail
MR	Mixed Reality
MSAC	M-Estimator Sample Consensus
OSM	Open Street Map
PnP	Perspective-n-Point
P3P	Perspective-Three-Point
RMSE	Root Mean Square Error
RTK	Real-Time Kinematic
R2S-PoseNet	real-to-synthetic PoseNet
SLAM	Simultaneous Localization and Mapping
SMOTE	Synthetic Minority Over-sampling Technique
S2R-PoseNet	synthetic-to-real PoseNet
Trans-CWGAN	Transfer Conditional Wasserstein Generative Adversarial Network
UWB	Ultra-Wideband
VISLAM	Visual Inertial Simultaneous Localization and Mapping
VR	Virtual Reality

References

1. Muthalif, M.; Shojaei, D.; Khoshelham, K. A review of augmented reality visualization methods for subsurface utilities. *Adv. Eng. Inform.* **2022**, *51*, 101498. [CrossRef]
2. Osadchyi, V.; Valko, N.; Kuzmich, L. Using augmented reality technologies for STEM education organization. In *Proceedings of the International Conference on Mathematics, Science and Technology Education, Kryvyi Rih, Ukraine, 15–17 October 2020*; Journal of Physics: Conference Series; IOP Publishing: Bristol, UK, 2021.
3. Gharaibeh, M.K.; Gharaibeh, N.K.; Khan, M.A.; Abu-Ain, W.A.K.; Alqudah, M.K. Intention to Use Mobile Augmented Reality in the Tourism Sector. *Comput. Syst. Sci. Eng.* **2021**, *37*, 187–202. [CrossRef]
4. Liu, B.; Ding, L.; Wang, S.; Meng, L. Designing Mixed Reality-Based Indoor Navigation for User Studies. *KN—J. Cartogr. Geogr. Inf.* **2022**, *72*, 129–138. [CrossRef]
5. Livingston, M.A.; Ai, Z.; Karsch, K.; Gibson, G.O. User interface design for military AR applications. *Virtual Real.* **2010**, *15*, 175–184. [CrossRef]
6. Bouchlaghem, D.; Shang, H.; Whyte, J.; Ganah, A. Visualisation in architecture, engineering and construction (AEC). *Autom. Constr.* **2005**, *14*, 287–295. [CrossRef]
7. Shin, D.H.; Dunston, P.S. Identification of application areas for Augmented Reality in industrial construction based on technology suitability. *Autom. Constr.* **2008**, *17*, 882–894. [CrossRef]
8. Irizarry, J.; Karan, E.P.; Jalaei, F. Integrating BIM and GIS to improve the visual monitoring of construction supply chain management. *Autom. Constr.* **2013**, *31*, 241–254. [CrossRef]
9. Volk, R.; Stengel, J.; Schultmann, F. Building Information Modeling (BIM) for existing buildings—Literature review and future needs. *Autom. Constr.* **2014**, *38*, 109–127. [CrossRef]
10. Garbett, J.; Hartley, T.; Heesom, D. A multi-user collaborative BIM-AR system to support design and construction. *Autom. Constr.* **2021**, *122*, 103487. [CrossRef]
11. Li, X.; Yi, W.; Chi, H.-L.; Wang, X.; Chan, A.P. A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Autom. Constr.* **2018**, *86*, 150–162. [CrossRef]
12. Alizadehsalehi, S.; Hadavi, A.; Huang, J.C. From BIM to extended reality in AEC industry. *Autom. Constr.* **2020**, *116*, 103254. [CrossRef]
13. Radanovic, M.; Khoshelham, K.; Fraser, C.S.; Acharya, D. Continuous BIM Alignment for Mixed Reality Visualisation. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *X-1/W1-2023*, 279–286. [CrossRef]
14. Abdul Muthalif, M.Z.; Shojaei, D.; Khoshelham, K. Interactive Mixed Reality Methods for Visualization of Underground Utilities. *PFG—J. Photogramm. Remote Sens. Geoinf. Sci.* **2024**, *92*, 741–760. [CrossRef]
15. Muthalif, M.Z.A.; Shojaei, D.; Khoshelham, K. Resolving Perceptual Challenges of Visualizing Underground Utilities in Mixed Reality. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *XLVIII-4/W4-2022*, 101–108. [CrossRef]
16. Albahbah, M.; Kıvrak, S.; Arslan, G. Application areas of augmented reality and virtual reality in construction project management: A scoping review. *J. Constr. Eng. Manag. Innov.* **2021**, *4*, 151–172. [CrossRef]
17. Hsieh, C.-C.; Chen, H.-M.; Wang, S.-K. On-site Visual Construction Management System Based on the Integration of SLAM-based AR and BIM on a Handheld Device. *KSCE J. Civ. Eng.* **2023**, *27*, 4688–4707. [CrossRef]
18. Ramezani, M.; Acharya, D.; Gu, F.; Khoshelham, K. Indoor Positioning by Visual-Inertial Odometry. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *IV-2/W4*, 371–376. [CrossRef]
19. Williams, G.; Gheisari, M.; Chen, P.-J.; Irizarry, J. BIM2MAR: An Efficient BIM Translation to Mobile Augmented Reality Applications. *J. Manag. Eng.* **2015**, *31*, A4014009. [CrossRef]
20. Ramezani, M.; Khoshelham, K.; Fraser, C. Pose estimation by Omnidirectional Visual-Inertial Odometry. *Robot. Auton. Syst.* **2018**, *105*, 26–37. [CrossRef]
21. Qin, J.; Li, M.; Liao, X.; Zhong, J. Accumulative Errors Optimization for Visual Odometry of ORB-SLAM2 Based on RGB-D Cameras. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 581. [CrossRef]
22. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
23. Acharya, D. Visual Indoor Localisation Using a 3D Building Model. Ph.D. Thesis, University of Melbourne, Melbourne, VIC, Australia, 2020.
24. Acharya, D.; Khoshelham, K.; Winter, S. BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 245–258. [CrossRef]
25. Acharya, D.; Ramezani, M.; Khoshelham, K.; Winter, S. BIM-Tracker: A model-based visual tracking approach for indoor localisation using a 3D building model. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 157–171. [CrossRef]
26. Chen, K.; Chen, W.; Li, C.T. A BIM-based location aware AR collaborative framework for facility maintenance management. *J. Inf. Technol. Constr.* **2019**, *24*, 360–380.

27. Mahmood, B.; Han, S.; Lee, D.-E. BIM-Based Registration and Localization of 3D Point Clouds of Indoor Scenes Using Geometric Features for Augmented Reality. *Remote Sens.* **2020**, *12*, 2302. [CrossRef]
28. Vermandere, J.; Bassier, M.; Vergauwen, M. Two-Step Alignment of Mixed Reality Devices to Existing Building Data. *Remote Sens.* **2022**, *14*, 2680. [CrossRef]
29. Chen, J.; Li, S.; Lu, W.; Liu, D.; Hu, D.; Tang, M. Markerless Augmented Reality for Facility Management: Automated Spatial Registration based on Style Transfer Generative Network. In *Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC), Dubai, United Arab Emirates, 2–4 November 2021*; International Association for Automation and Robotics in Construction (IAARC): Oulu, Finland, 2021.
30. Chen, J.; Li, S.; Liu, D.; Lu, W. Indoor camera pose estimation via style-transfer 3D models. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 335–353. [CrossRef]
31. Acharya, D.; Tatli, C.J.; Khoshelham, K. Synthetic-real image domain adaptation for indoor camera pose regression using a 3D model. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 405–421. [CrossRef]
32. Saito, S.; Hiyama, A.; Tanikawa, T.; Hirose, M. Indoor Marker-based Localization Using Coded Seamless Pattern for Interior Decoration. In *Proceedings of the 2007 IEEE Virtual Reality Conference, Charlotte, NC, USA, 10–14 March 2007*; IEEE: New York, NY, USA, 2007.
33. Einizinab, S.; Khoshelham, K.; Winter, S.; Christopher, P. Offset-Based Marker Placement for BIM Alignment in Mixed Reality. In *Proceedings of the 2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW), Kuala Lumpur, Malaysia, 8–11 October 2023*; IEEE: New York, NY, USA, 2023.
34. Abhishek, M.T.; Aswin, P.; Akhil, N.C.; Souban, A.; Muhammedali, S.K.; Vial, A. Virtual Lab Using Markerless Augmented Reality. In *Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Wollongong, NSW, Australia, 4–7 December 2018*; IEEE: New York, NY, USA, 2018.
35. Scargill, T. Context-Aware Markerless Augmented Reality for Shared Educational Spaces. In *Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Bari, Italy, 4–8 October 2021*; IEEE: New York, NY, USA, 2021.
36. Jinyu, L.; Bangbang, Y.; Danpeng, C.; Nan, W.; Guofeng, Z.; Hujun, B. Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality. *Virtual Real. Intell. Hardw.* **2019**, *1*, 386–410. [CrossRef]
37. Hansen, L.H.; Fleck, P.; Stranner, M.; Schmalstieg, D.; Arth, C. Augmented Reality for Subsurface Utility Engineering, Revisited. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 4119–4128. [CrossRef]
38. Messi, L.; Spegni, F.; Vaccarini, M.; Corneli, A.; Binni, L. Seamless Augmented Reality Registration Supporting Facility Management Operations in Unprepared Environments. *J. Inf. Technol. Constr.* **2024**, *29*, 1156–1180. [CrossRef]
39. Acharya, D.; Roy, S.S.; Khoshelham, K.; Winter, S. A Recurrent Deep Network for Estimating the Pose of Real Indoor Images from Synthetic Image Sequences. *Sensors* **2020**, *20*, 5492. [CrossRef] [PubMed]
40. Sattler, T.; Zhou, Q.; Pollefeys, M.; Leal-Taixe, L. Understanding the limitations of CNN-based absolute camera pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; IEEE: New York, NY, USA, 2019.
41. Ha, I.; Kim, H.; Park, S.; Kim, H. Image-based Indoor Localization using BIM and Features of CNN. In *Proceedings of the 35th International Symposium on Automation and Robotics in Construction (ISARC), Berlin, Germany, 20–25 July 2018*; IAARC Publications: Waterloo, ON, Canada, 2018; pp. 1–4.
42. Einizinab, S.; Khoshelham, K.; Winter, S.; Christopher, P. Camera Pose Refinement for Precise BIM Alignment in Mixed Reality Visualization. *J. Comput. Civ. Eng.* **2025**, *39*, 04025072. [CrossRef]
43. Boan, T.; Jiajun, L.; Bosché, F. Autonomous Mixed Reality Framework for Real-Time Construction Inspection. *J. Inf. Technol. Constr. (ITcon)* **2025**, *30*, 852–874.
44. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017.
45. Wang, S. A hybrid SMOTE and Trans-CWGAN for data imbalance in real operational AHU AFDD: A case study of an auditorium building. *Energy Build.* **2025**, *348*, 116447. [CrossRef]
46. Wang, S. Domain adaptation using transformer models for automated detection of exterior cladding materials in street view images. *Sci. Rep.* **2025**, *16*, 2696. [CrossRef]
47. Sufiyan, D.; Win, L.S.T.; Win, S.K.H.; Tan, U.-X.; Foong, S. Direct Aerial Visual Localization using Panoramic Synthetic Images and Domain Adaptation. In *Proceedings of the 2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Boston, MA, USA, 15–19 July 2024*; IEEE: New York, NY, USA, 2024.
48. Hong, Y.; Park, S.; Kim, H. Synthetic data generation for indoor scene understanding using BIM. in ISARC. In *Proceedings of the 37th International Symposium on Automation and Robotics in Construction (ISARC), Kitakyushu, Japan, 27–28 October 2020*; IAARC Publications: Waterloo, ON, Canada, 2020.

49. Chen, H.; Yang, H.; Chen, J.; Zhang, S.; Jing, X. *Bim Aided Indoor Camera Pose Estimation Based on Cross-Domain Image Retrieval*; SSRN 4913115; SSRN: Rochester, NY, USA, 2024.
50. Alnajjar, O.; Atencio, E.; Turmo, J. A systematic review of lean construction, BIM and emerging technologies integration: Identifying key tools. *Buildings* **2025**, *15*, 2884. [CrossRef]
51. Büyüksalih, G.; Kan, T.; Özkan, G.E.; Meriç, M.; Isın, L.; Kersten, T.P. Preserving the Knowledge of the Past Through Virtual Visits: From 3D Laser Scanning to Virtual Reality Visualisation at the Istanbul Çatalca İnceğiz Caves. *PFG—J. Photogramm. Remote Sens. Geoinf. Sci.* **2020**, *88*, 133–146. [CrossRef]
52. Ungureanu, D.; Bogo, F.; Galliani, S.; Sama, P.; Duan, X.; Meekhof, C.; Stühmer, J.; Cashman, T.J.; Tekina, B.; Schönberger, J.L.; et al. Hololens 2 research mode as a tool for computer vision research. *arXiv* **2020**, arXiv:2008.11239. [CrossRef]
53. Tareen, S.A.K.; Saleem, Z. A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. In *Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 3–4 March 2018*; IEEE: New York, NY, USA, 2018.
54. Zhang, P.; Yan, X. Application of Improved KAZE Algorithm in Image Feature Extraction and Matching. *IEEE Access* **2023**, *11*, 122625–122637. [CrossRef]
55. Wu, Y.; Hu, Z. PnP problem revisited. *J. Math. Imaging Vis.* **2006**, *24*, 131–141. [CrossRef]
56. Gao, X.-S.; Hou, X.-R.; Tang, J.; Cheng, H.-F. Complete solution classification for the perspective-three-point problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 930–943.
57. Lepetit, V.; Fua, P. *Monocular Model-Based 3D Tracking of Rigid Objects*; Now Publishers Inc.: Delft, The Netherlands, 2005.
58. Aijazi, A.K.; Malaterre, L.; Trassoudaine, L.; Chateau, T.; Checchin, P. Automatic Detection and Modeling of Underground Pipes Using a Portable 3D LiDAR System. *Sensors* **2019**, *19*, 5345. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.