

Integrating Building Information Modeling and Panoramic Structure-from-Motion for Accurate Camera Pose Estimation

Max Jwo Lem Lee¹, Weisong Wen¹, Stephen Ling Ming Au²

¹ Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University (PolyU), Hong Kong

² MTECH Engineering Co.,Ltd, Hong Kong

Abstract

In this study, we present a novel approach for combining Building Information Modeling (BIM) and panoramic photogrammetry-based Structure-from-Motion (SfM) to achieve accurate camera pose estimation in architectural scenes. The fusion of BIM and SfM information addresses the limitations of individual methods: the former offers global positioning, but it suffers from suboptimal accuracy; while the latter provides accurate relative positioning, it lacks scaling and global positioning. Our method consists of four key steps: (1) computationally efficient global positioning of panorama images in the BIM model using indoor semantic skymasks to generate probability distributions, (2) relative positioning estimation from the panoramic SfM process, (3) rough alignment of the SfM reconstruction with the BIM positioning using generalized Procrustes analysis (GPA), (4) refinement of the camera pose using non-linear least-squares optimization. We evaluate the performance of our proposed method using the real-world dataset of panoramic images capturing architectural scenes and compare the refined camera poses with ground truth. The results demonstrate camera positioning accuracy of fewer than 0.6 meters when compared to using BIM or panoramic SfM individually. This research highlights the potential benefits of fusing SfM and BIM modalities, paving the way for more accurate and efficient camera pose estimation pipelines in the architecture, engineering, and construction (AEC) domain.

Keywords

Panorama, BIM, Localization, SfM

1. Introduction

The burgeoning interest in construction automation, virtual reality-enabled scene navigation, sophisticated facility management, and as-built documentation has catalyzed the development of a myriad of techniques in computer vision, photogrammetry, and Building Information Modeling (BIM) [1-3]. Visual documentation is critical in construction projects for monitoring site conditions and progress. While images alone may not suffice, BIM can provide a detailed digital representation of the building or structure, allowing construction teams to identify discrepancies between planned designs and actual construction [4]. Integrating images and BIM models enables effective communication and collaboration between stakeholders, improving decision-making and problem-solving. Recent research has shown that such integration can enhance the accuracy and completeness of construction documentation [5]. Precise camera pose estimation constitutes a pivotal aspect of these applications, as it substantially influences the integration quality between captured images and BIM models [6]. However, current approaches do not adequately address the challenges in automating camera pose estimation for panoramic images, leaving a research gap in achieving successful BIM model integration.

Panoramic imaging has witnessed a surge in popularity in recent years, attributed to its capacity to encapsulate an extensive field of view within a single image, rendering it particularly apt for


Proceedings of the Work-in-Progress Papers at the 13th International Conference on Indoor Positioning and Indoor Navigation (IPIN-WiP 2023), September 25 - 28, 2023, Nuremberg, Germany

✉ maxjl.lee@connect.polyu.hk (M. J. L. Lee); welson.wen@polyu.edu.hk (W. Wen); stephenau@mtech.com.hk (S. Au)

ORCID iD 0000-0002-5524-6724 (M. J. L. Lee); 0000-0003-4158-0913 (W. Wen)

© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

architectural scenarios. Nonetheless, automating camera pose estimation for panoramic images engenders distinct challenges that must be addressed to ensure successful BIM model integration. Structure-from-Motion (SfM) represents a widely employed photogrammetric technique for deducing camera poses and reconstructing 3D scenes from an unordered set of panorama images. Despite progress in panoramic SfM, issues pertaining to scale ambiguity and restrictions on relative positioning may impede the applicability of this method.

Several studies have explored the integration of BIM and photogrammetric techniques for 3D reconstruction and as-built documentation. For example, [7] proposed a framework for integrating BIM and laser scanning data to generate accurate and detailed as-built models. Similarly, [8] developed a method that combines BIM and photogrammetry to create a 3D model of a building interior for virtual reality applications. However, these studies do not specifically address the challenges of automating camera pose estimation for panoramic images in the context of BIM integration.

Concurrently, BIM has emerged as an indispensable tool within the architecture, engineering, and construction (AEC) fields, providing an exhaustive digital representation of a building's physical and functional characteristics [9]. BIM models offer a global coordinate system, along with invaluable geometric and semantic information to direct and constrain the camera pose estimation procedure. However, the limited level of detail in BIM models renders attaining high-precision positioning both arduous and computationally demanding. Previous studies on BIM-based visual positioning systems (VPS) have exclusively explored virtual environments, which do not accurately represent real-world conditions, as they lack scene changes due to dynamic objects [10].

In this paper, we propose a method that leverages static objects exclusively for BIM-based positioning, incorporating the layout and static semantic objects such as doors and windows. Additionally, we introduce a novel approach for positioning panoramic images by employing geo-tagged indoor semantic skymasks, building upon our previous research on skymask matching-aided positioning in urban canyons [11]. These indoor semantic skymasks offer a detailed representation of a 3D location's spatial layout within an indoor environment, encompassing both geometric and static semantic information. This approach enhances the computational efficiency of candidate-based matching as opposed to traditional image-based comparisons [10, 11]. Following this, we utilize panoramic photogrammetry based SfM to achieve accurate relative positioning. Our method aims to overcome the limitations inherent in each individual technique by harnessing the complementary information provided by both BIM and SfM. We contend that the integration of panoramic BIM and SfM will lead to more precise and reliable camera pose estimation, benefiting a wide array of applications within the AEC domain, including but not limited to:

- **Construction automation:** Accurate camera pose estimation facilitates streamlined construction processes by enabling precise alignment of the physical building with its digital representation.
- **Virtual reality-enabled navigation:** Realistic virtual tours of architectural spaces can be generated by integrating accurate camera pose estimation with immersive virtual reality experiences.
- **Facility management:** Maintaining and updating building documentation becomes more efficient and accurate through precise camera pose estimation, allowing facility managers to better track building conditions and plan renovations.
- **As-built documentation:** By combining accurate camera pose estimation with BIM models, architects and engineers can create more reliable as-built documentation for future reference and regulatory compliance.

The integration of BIM and panoramic SfM for accurate camera pose estimation, as proposed in this research, offers a multitude of advantages over conventional methods that solely employ either panoramic BIM or SfM. These benefits can be encapsulated within the following points:

- **Enhanced Accuracy:** By amalgamating information derived from BIM global positioning and panoramic SfM relative positioning, our methodology transcends the limitations inherent to each individual technique, culminating in superior camera pose estimation precision.
- **Scale Recovery:** By integrating the BIM model, replete with accurate scale information, our approach effectively recovers the appropriate scale of the SfM reconstruction.
- **Improved Computational Speed:** Our method extracts geometric and semantic information to generate pre-computed indoor semantic skymasks for positioning, significantly reducing computation time compared to visual positioning systems (VPS) that depend on whole-image comparisons.

However, it is important to note that this approach has some potential weaknesses:

- **Dependence on static objects:** The proposed method leverages static objects such as doors and windows for camera pose estimation, which may not be sufficient in all scenarios.
- **Limited level of detail:** The level of detail in BIM models may not always be sufficient for high-precision positioning, which could limit the effectiveness of the proposed method.

2. The proposed integration of BIM and SfM for accurate camera pose estimation

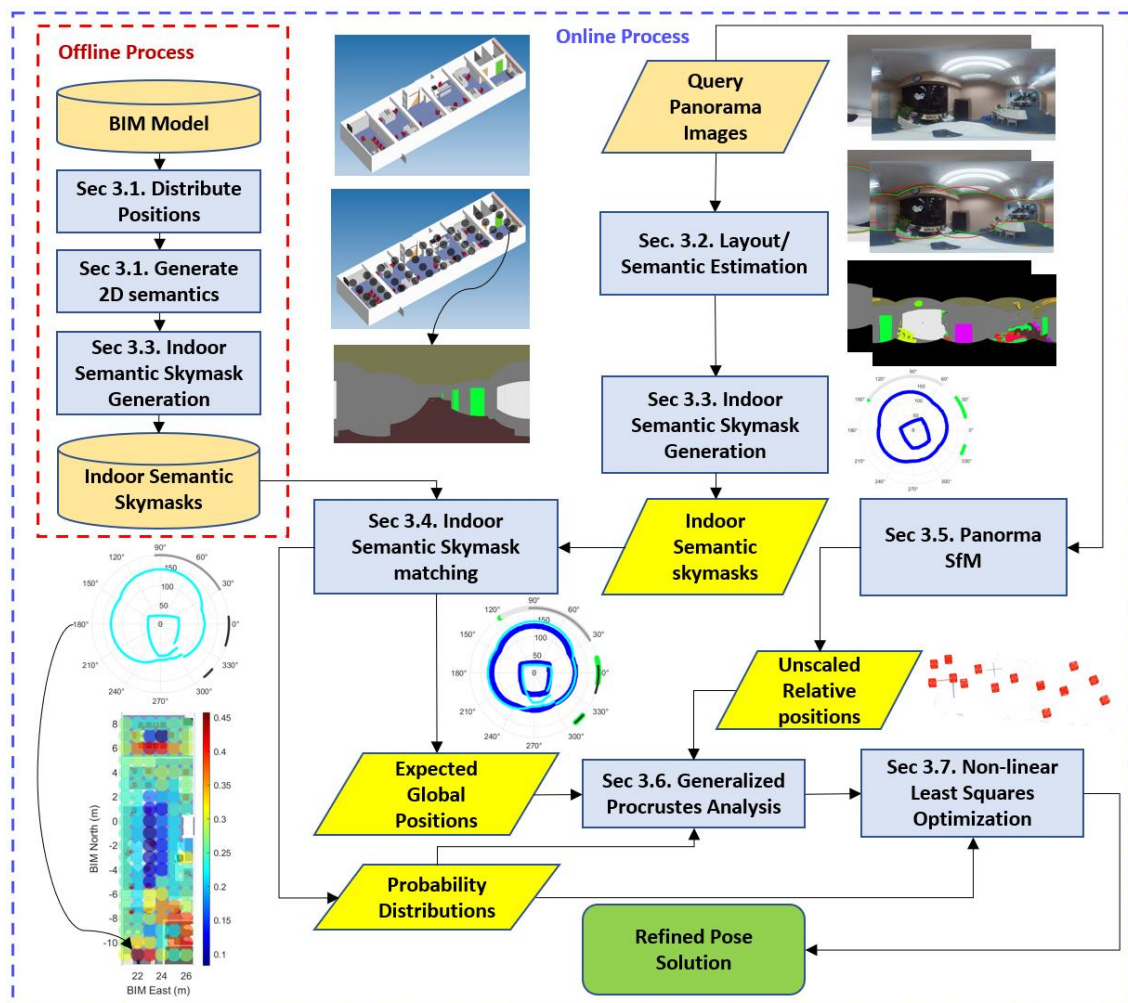


Figure 1: Flowchart of the proposed integration of BIM and panoramic SfM for accurate camera pose estimation

The proposed algorithm can be divided into two main stages in Fig. 1: an offline process and an online process. During the offline stage, the Building Information Modeling (BIM) model is processed to extract two-dimensional (2D) semantics pertaining to its environmental surroundings within the panorama frame at each position, as detailed in Sub-Section 3.1. The extracted 2D semantics is subsequently used to create an indoor semantic skymask, which stores the elevation profile of the upper and lower boundaries of walls at each azimuth, alongside binary indicators denoting the presence or absence of doors and windows at each azimuth (Sub-Section 3.3). These indoor semantic skymasks are stored in a database for utilization during the online stage.

In the online stage, panoramic images are processed using deep learning models to extract layout and semantics (Sub-Section 3.2), which are then employed to generate query indoor semantic skymasks (Sub-Section 3.3). The proposed method comprises two main components: 1) Indoor semantic skymask matching, wherein each query indoor semantic skymask is matched with candidate indoor semantic skymasks to produce a probability distribution and its corresponding expected global position (Sub-Section 3.4); and 2) SfM is applied to generate unscaled relative positioning estimates (Sub-Section 3.5). The estimated unscaled relative 2D positions serve as input for the Generalized Procrustes Analysis (GPA), which approximately aligns them with the expected global positions of the query images (Section 3.6). Finally, the positions are refined through non-linear least squares optimization, using their respective probability distributions (Sub-Section 3.7).

The paper is organized as follows: Section 3 introduces the methodology for integrating BIM and SfM for accurate camera pose estimation. Section 4 presents the experiment setup and results. Section 5 presents the conclusion and section 6 details future works, respectively.

3. Methodology

3.1. Distribute Candidate Positions and Generate 2D Semantics

The initial step in our method entails obtaining the BIM model of the indoor environment. In this study, we made use of the BIM model of an office, which was provided in the Industry Foundation Classes (IFC) open-source format. A BIM model encompasses all information related to the building, including its physical attributes. For instance, a door within a BIM model would already be labeled "door" in the IFC format, which can then be utilized to generate semantics. This study employs the ADE20k classes to categorize objects in the BIM model [8]. Owing to the Level of Detail 2 (LOD2) nature of our BIM model, we focused on the semantics of "ceiling," "wall," "floor," "door," and "window" while excluding all other dynamic objects. This exclusion is crucial, as dynamic objects can be relocated in the real environment, and their presence may lead to erroneous camera pose estimation.

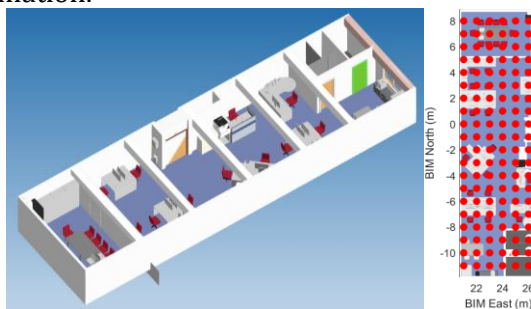


Figure 2: Candidates with 1m separation in the BIM model

Firstly, candidate positions c are spread across the BIM model with 1-meter separation and 1.8m in height as shown in Fig. 2. We assume that the height of the query images is 1.8m. The following are defined:

$$\begin{aligned}
\mathbf{c} &= [x, y] \\
\mathbf{C} &= \{\mathbf{c}_0 \cdots \mathbf{c}_s\} \\
\text{Img}_{\mathbf{c}}^{\text{seg}} &= \text{SEG}(\mathbf{u}, \mathbf{v})
\end{aligned} \tag{1}$$

Where \mathbf{c} is a two-dimensional position, and the subscript s is the index of \mathbf{C} , which are all the candidate positions inside the BIM model. Position \mathbf{c} is extracted from database \mathbf{C} , where $\mathbf{c} \in \mathbf{C}$. SEG is the function that assigns each pixel (\mathbf{u}, \mathbf{v}) an indexed number to represent a class. A segmented image for a candidate position is denoted as $\text{Img}_{\mathbf{c}}^{\text{seg}}$ as shown in Fig. 3.

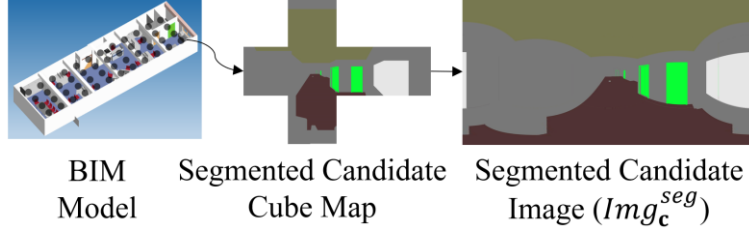


Figure 3: Semantic extraction from candidate position

3.2. Layout and semantic estimation

Given a query panorama image, we perform layout estimation and semantic segmentation. This can be achieved using LGT-net [12] and Segformer [13], respectively.

First, we employ the LGT-net model to estimate the layout of the given panoramic image. LGT-net is a state-of-the-art deep learning model specifically designed for the task of layout estimation. The output of the LGT-net model is a set of elevation angles that describe the estimated upper and lower layout.

$$\mathbf{I}_k^{\text{upper}}[\alpha], \mathbf{I}_k^{\text{lower}}[\alpha] = \text{LGT}(\text{Img}_k) \tag{2}$$

Where the subscript k represents the index of the query images. α represents the azimuth angle (0 to 359 degrees). $\mathbf{I}_k^{\text{upper}}[\alpha]$ and $\mathbf{I}_k^{\text{lower}}[\alpha]$ represents the elevation angle (0 to 180 degrees) for each azimuth angle (0 to 359 degrees), respectively for the upper and lower layout as shown in Fig. 4.

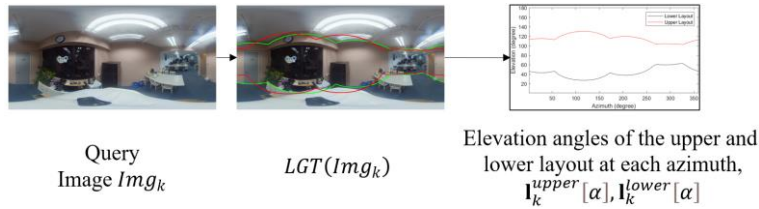


Figure 4: Layout elevation extraction from query image

Next, we apply the Segformer model to perform semantic segmentation on the panoramic image. Segformer is a semantic segmentation model that can efficiently and accurately segment the input image into various semantic classes, such as walls, floors, ceilings, furniture, and other objects based on the ADE20k dataset [13]. The output of the Segformer model is a pixel-wise label map that provides a semantic understanding of the scene, and can be formulated as:

$$\text{Img}_k^{\text{seg}} = \text{SEGFORMER}(\text{Img}_k) \tag{3}$$

Where SEGFORMER is the semantic segmentation, and $\text{Img}_k^{\text{seg}}$ is the segmented query image of index k as shown in Fig. 5.

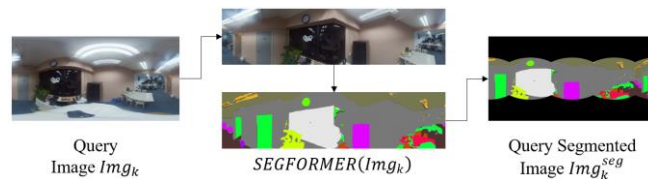


Figure 5: Segmented image extraction from query image

3.3. Indoor Semantic Skymask Generation

The indoor semantic skymask is a 360x4 matrix that serves as a detailed representation of an indoor environment 3D location's spatial layout. It encodes the elevation angle (0 to 180 degrees) for each azimuth angle (0 to 359 degrees). The matrix contains:

1. $\mathbf{layout}^{upper}[\alpha]$ - Elevation angles of the upper layout at each azimuth (column 1)
2. $\mathbf{layout}^{lower}[\alpha]$ - Elevation angles of the lower layout at each azimuth (column 2)
3. $\mathbf{d}[\alpha]$ - Binary indicators for the presence or absence of doors at each azimuth (column 3)
4. $\mathbf{w}[\alpha]$ - Binary indicators for the presence or absence of windows at each azimuth (column 4)

A segmented candidate image Img_c^{seg} can be converted to the elevation angles of the upper layout $\mathbf{l}_c^{upper}[\alpha]$ and lower layout $\mathbf{l}_c^{lower}[\alpha]$ by masking the ceiling and floor labels and extracting the boundaries at each azimuth angle as shown in Fig. 6.

The binary indicators for the “door” $\mathbf{d}[\alpha]$ and “windows” $\mathbf{w}[\alpha]$ can be extracted by observing if the respective labels are present at the azimuth angle.

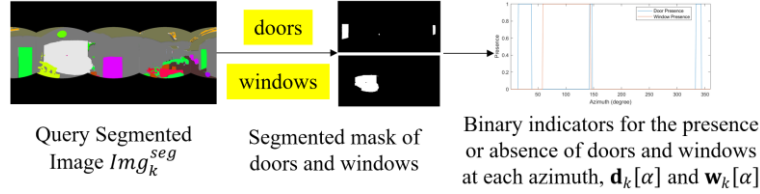


Figure 6: Binary indicator extraction from query segmented image

The layouts and binary indicator are extracted and visualized on an indoor semantic skymask as shown in Fig. 7.

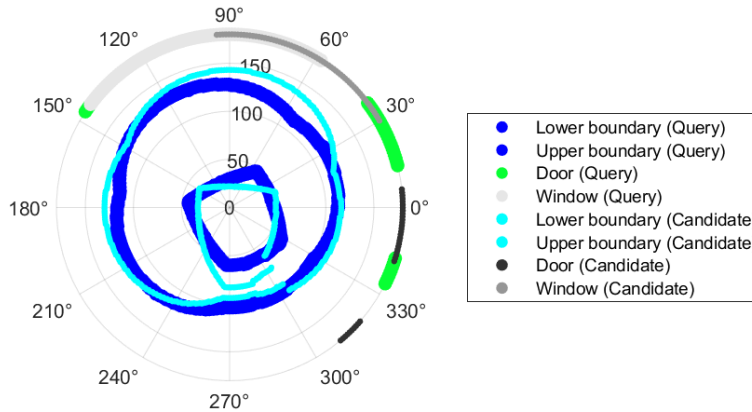


Figure 7: Query and candidate indoor semantic skymask visualization

3.4. Indoor Semantic Skymask Matching

To find the similarity between two indoor semantic skymasks, we perform the following steps:

3.4.1. Calculate the normalized circular cross-correlation of the layouts

For a pair of layouts, such as the upper layout of a query and candidate image, we calculate their normalized circular cross-correlation $\mathbf{r}_{c,k}[m]$ as follows:

$$\mathbf{r}_{c,k}^{upper}[m] = \frac{1}{N} \sum_{\alpha=0}^{N-1} \frac{\mathbf{l}_c^{upper}[\alpha] \cdot \mathbf{l}_k^{upper}[(\alpha + m) \bmod N]}{\sqrt{\left(\sum_{n=0}^{N-1} \mathbf{l}_c^{upper^2}[\alpha]\right) \cdot \left(\sum_{n=0}^{N-1} \mathbf{l}_k^{upper^2}[\alpha]\right)}} \quad (4)$$

$$\mathbf{r}_{c,k}^{lower}[m] = \frac{1}{N} \sum_{\alpha=0}^N \frac{\mathbf{I}_c^{lower}[\alpha] \cdot \mathbf{I}_k^{lower}[(\alpha + m) \bmod N]}{\sqrt{\left(\sum_{n=0}^{N-1} \mathbf{I}_c^{lower^2}[\alpha]\right) \cdot \left(\sum_{n=0}^{N-1} \mathbf{I}_k^{lower^2}[\alpha]\right)}}$$

Here, $\mathbf{r}_{c,k}^{upper}[m]$ is the m -point circularly shifted version of the upper layout boundary $\mathbf{I}_k^{upper}[\alpha]$ with respect to $\mathbf{I}_c^{upper}[\alpha]$. Where N is 360. The cross-correlation measures the similarity between the elevation profiles as a function of the circular shift m in the azimuth domain.

3.4.2. Calculate the weighted average normalized cross-correlation

We compute the weighted average normalized cross-correlation $\mathbf{r}_{c,k}[m]$ using the results from the upper and lower layouts:

$$\mathbf{r}_{c,k}[m] = 0.5 \cdot \mathbf{r}_{c,k}^{upper}[m] + 0.5 \cdot \mathbf{r}_{c,k}^{lower}[m] \quad (5)$$

The maximum value of $\mathbf{r}_{c,k}[m]$ signifies the highest similarity between the two boundaries, and the corresponding shift $m_{c,k}$ provides the optimal alignment for query k with respect to candidate \mathbf{c} .

3.4.3. Compare the door and window semantics

For each query k with the corresponding shift $m_{c,k}$, we compare the door and window semantics using binary comparison. We align the candidate semantics with the query semantics using the shift $m_{c,k}$:

$$\begin{aligned} \mathbf{d}_k^{m_{c,k}}[\alpha] &= \mathbf{d}_k[(\alpha + m_{c,k}) \bmod N] \\ \mathbf{w}_k^{m_{c,k}}[\alpha] &= \mathbf{w}_k[(\alpha + m_{c,k}) \bmod N] \end{aligned} \quad (5)$$

Next, we perform binary comparison:

$$\begin{aligned} \mathbf{s}_{\mathbf{d}_{c,k}}[\alpha] &= \mathbf{d}_c[\alpha] \cdot \mathbf{d}_k^{m_{c,k}}[\alpha], \text{ for } \alpha \text{ in } [0, N] \\ \mathbf{s}_{\mathbf{w}_{c,k}}[\alpha] &= \mathbf{w}_c[\alpha] \cdot \mathbf{w}_k^{m_{c,k}}[\alpha], \text{ for } \alpha \text{ in } [0, N] \end{aligned} \quad (6)$$

These values equal 1 when both semantics are present at the same azimuth angle and 0 otherwise. $\mathbf{s}_{\mathbf{d}_{c,k}}[\alpha]$ and $\mathbf{s}_{\mathbf{w}_{c,k}}[\alpha]$ represent the score of the door and window semantics respectively.

3.4.4. Calculate the semantics similarity score

We determine the overall similarity score by summing the door and window similarity values and normalizing by the total number of azimuth angles N :

$$\begin{aligned} seg_sim_{c,k} &= \frac{\sum(\mathbf{s}_{\mathbf{d}_{c,k}}[\alpha]) + \sum(\mathbf{s}_{\mathbf{w}_{c,k}}[\alpha])}{(2 \cdot N)}, \\ &\text{for } \alpha \text{ in } [0, N] \end{aligned} \quad (7)$$

The overall semantic similarity score $seg_sim_{c,k}$ ranges from 0 to 1, with 1 representing the highest similarity between the query and candidate indoor semantic skymasks.

3.4.5. Combine the layout boundary and semantic similarity scores for candidate selection

In order to determine the most likely candidate location for each query image k and candidate vector \mathbf{c} , we combine the maximum layout similarity, denoted as $\max(\mathbf{r}_{c,k}[m])$, and the overall semantic similarity score, represented by seg_sim . The combination is achieved by assigning weights to each component, as shown in the following equation:

$$sim_{c,k} = (0.3 \cdot \max(\mathbf{r}_{c,k}[m])) + (0.7 \cdot seg_sim_{c,k}) \quad (8)$$

The weighting for combining both layout and semantic similarity scores is estimated by comparing the combined similarity that yields the best performance on a set of 10 images, with respect to their ground truth locations. The combined similarity score $sim_{c,k}$ for each candidate \mathbf{c} and the corresponding shift $m_{c,k}$ is used to distribute a likelihood heatmap \mathbf{H}_k , as illustrated in Figure 8.

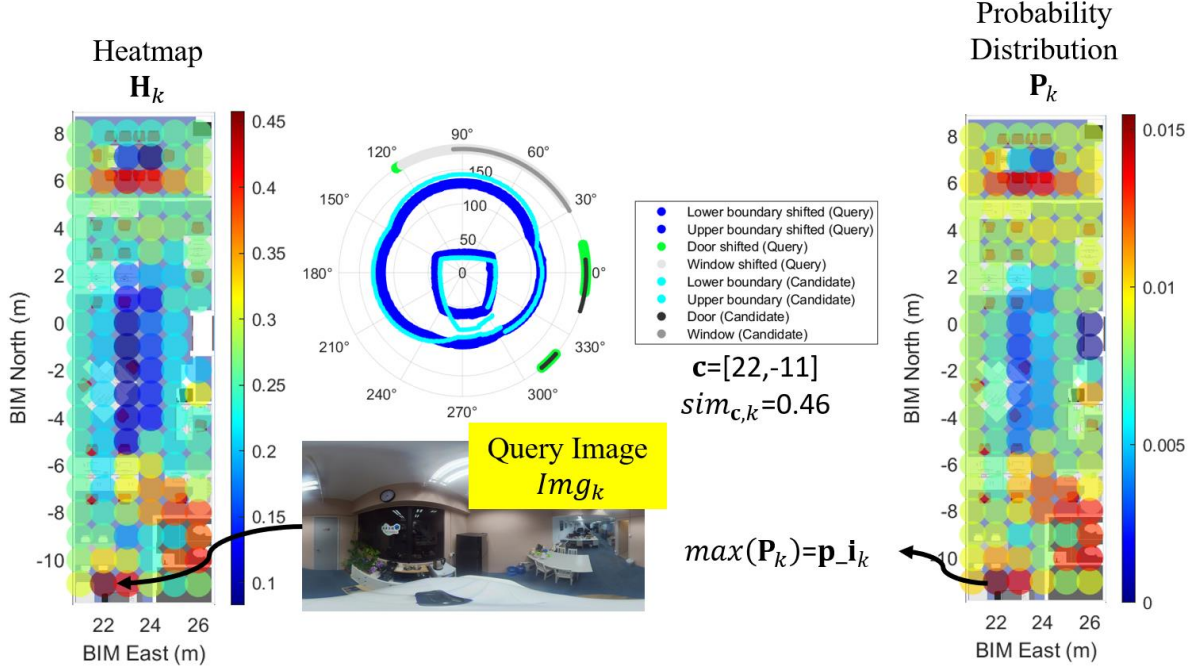


Figure 8: Likelihood heatmap of a query image (left), alignment of query indoor semantic skymask to candidate semantic skymask (top), probability distribution of a query image (right)

The heatmap \mathbf{H}_k represents the likelihood that a query image k is located at the candidate \mathbf{c} location.

3.4.6. Probability distribution

To calculate the probability distribution, we first compute the sum of likelihood values for each query image k across all candidate positions:

$$sum_k = \sum_{s=1}^N \mathbf{H}_{c_s,k} \quad (9)$$

Where s is the index of the candidates and N is the total number of candidates. Next, we derive the probability at each candidate position \mathbf{c} in the probability distribution \mathbf{P}_k :

$$\mathbf{P}_{c,k} = \frac{\mathbf{H}_{c,k}}{sum_k} \quad (10)$$

The resulting 2D array \mathbf{P}_k represents the probability distribution corresponding to the likelihood heatmap \mathbf{H}_k . Each value $\mathbf{P}_{c,k}$ indicates the probability of query k occurring at position \mathbf{c} .

The maximum probability $max(\mathbf{P}_k)$ corresponds to the expected global position \mathbf{c} of query image k . We denote this as \mathbf{p}_{i_k} .

3.5. Panorama SfM

In this study, we employed OpenMVG (Open Multiple View Geometry) software for panorama Structure from Motion (SfM) to estimate the relative camera pose of each image [14]. OpenMVG

estimates the position of each camera used to capture the input images relative to a local coordinate system with ambiguous scaling.

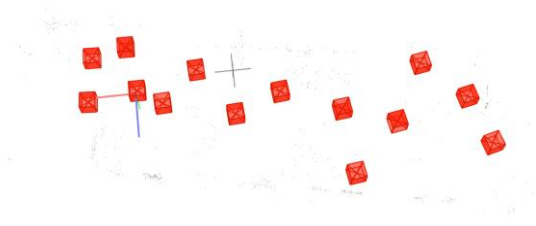


Figure 9: Relative positioning of cameras in OpenMVG

$$\begin{aligned} \mathbf{p}_s &= [x, y] \\ \mathbf{P}_S &= \{\mathbf{p}_{s_0} \cdots \mathbf{p}_{s_k}\} \end{aligned} \quad (11)$$

Where \mathbf{p}_s is a two-dimensional SfM estimated position, and the subscript k is the query index of \mathbf{P}_S . Position \mathbf{p}_s is extracted from database \mathbf{P}_S , where $\mathbf{p}_s \in \mathbf{P}_S$.

3.6. Generalized Procrustes Analysis

The key part of the GPA algorithm is the computation of the optimal similarity transformation. Let \mathbf{P}_I denote the set of positions and \mathbf{P}_S denote the set of positions to be aligned. The optimal rotation matrix \mathbf{R} , scaling factor s , and translation vector \mathbf{t} are computed.

$$\begin{aligned} \mathbf{P}_G &= s \cdot \mathbf{R} \cdot \mathbf{P}_S + \mathbf{t} \\ \mathbf{P}_G &= \{\mathbf{p}_{g_0} \cdots \mathbf{p}_{g_k}\} \\ \mathbf{p}_g &= [x, y] \end{aligned} \quad (12)$$

The GPA algorithm iteratively aligns \mathbf{P}_S to the global reference frame by computing the optimal similarity transformations and updating the reference shape until convergence. The transformed positions are denoted as \mathbf{P}_G .

3.7. Non-linear least squares optimization

Let's denote the set of query image positions from GPA algorithm as \mathbf{p}_{g_k} , where $k = 1, 2, \dots, n$, and each query image position \mathbf{p}_{g_k} has an associated spatial probability distribution \mathbf{P}_k . Additionally, we know the relative positions between the landmarks from SfM as \mathbf{s}_{kl} , where $\mathbf{s}_{kl} = \mathbf{p}_{g_k} - \mathbf{p}_{g_l}$.

Let the transformation be represented by a rotation matrix R , a scaling factor s , and a translation vector t . The transformed query image positions can be written as q_k , where $q_k = s \cdot R \cdot \mathbf{p}_{g_k} + t$. The non-linear least squares optimization problem can be formulated as:

$$\min_{R, s, t} \sum_{k=1}^n \sum_{l=1, l \neq k}^n \mathbf{D}_{kl}(\mathbf{q}_k, \mathbf{q}_l, \mathbf{s}_{kl}) \quad (13)$$

Where $\mathbf{D}_{kl}(\mathbf{q}_k, \mathbf{q}_l, \mathbf{s}_{kl})$ is a function that computes the squared difference between the transformed positions \mathbf{q}_k and \mathbf{q}_l and their expected relative positions \mathbf{s}_{kl} , weighted by the spatial probability distribution:

$$\mathbf{D}_{kl}(\mathbf{q}_k, \mathbf{q}_l, \mathbf{s}_{kl}) = \mathbf{P}_k(\mathbf{q}_k) \cdot \mathbf{P}_l(\mathbf{q}_l) \cdot \|\mathbf{q}_k - \mathbf{q}_l - \mathbf{s}_{kl}\|^2 \quad (14)$$

The optimization problem is solved using the Levenberg-Marquardt algorithm, and the final positions are denoted as $\mathbf{Q} = \{\mathbf{q}_0 \cdots \mathbf{q}_k\}$.

4. Experiment setup and results

4.1. Dataset and Preprocessing

In this section, we describe the dataset used to evaluate our proposed method for indoor camera pose estimation. The dataset consisted of 14 panoramic images and the corresponding BIM model of an office. The images were captured using an Insta360 ONE X2 camera mounted on a tripod and covered a spatial extent of approximately $120m^2$. The ground truth location of the images was established by aligning them with the BIM model via visual overlap. The BIM models were generated from architectural plans and as-built drawings, providing Level of Detail 2 (LOD2) geometric and semantic information about the built environment.

To evaluate the performance of our proposed method, we compared it against two other methods: ground truth (aligned via visual overlap) and indoor semantic skymask matching. The ground truth method served as a reference for assessing the accuracy of the other methods, while the indoor semantic skymask matching method represented a state-of-the-art approach for indoor camera pose estimation. Our proposed method combined indoor semantic skymask matching with panoramic photogrammetry-based Structure-from-Motion (SfM) techniques to achieve more accurate and reliable camera pose estimation in architectural scenes. By comparing the performance of these methods, we aimed to demonstrate the effectiveness of our proposed approach and its potential for various applications, such as virtual reality simulations, robotics, and 3D modeling.

4.2. Results and Analysis

The performance metrics presented in Table I and the accompanying statements provide important information about the accuracy of camera pose estimation using different methods in architectural scenes. The exact position and orientation are critical for various applications, such as virtual reality simulations, robotics, 3D modeling, and building inspection. In addition, for building inspection and review applications, it is important to ensure that the captured images cover the entire building area with sufficient overlap for accurate reconstruction. The precision requirements may vary depending on the specific application, but in general, the camera pose estimation must be accurate enough to enable reliable localization and navigation in the environment.

In our study, we evaluated the performance of two methods for camera pose estimation: indoor semantic skymask matching and BIM and SfM integration. The results showed that the indoor semantic skymask matching method had an average positioning error of 1.28 meters to the ground truth, which indicates relatively poor accuracy and might not be sufficient for some applications. In contrast, the BIM and SfM integration method achieved a mean distance error of 0.59 meters with 5.32° heading accuracy, which represents a significant improvement over the indoor semantic skymask matching method. The standard deviation of the errors was also lower for the BIM and SfM integration method, indicating more consistent performance across different scenes.

Table I. Accuracy of the proposed BIM and SfM Integration

Method	2D Position		Heading	
	Mean (m)	SD (m)	Mean ($^\circ$)	SD ($^\circ$)
Indoor semantic skymask matching	1.28	1.05	15.43	18.84
Proposed BIM and SfM Integration	0.59	0.31	5.32	5.09

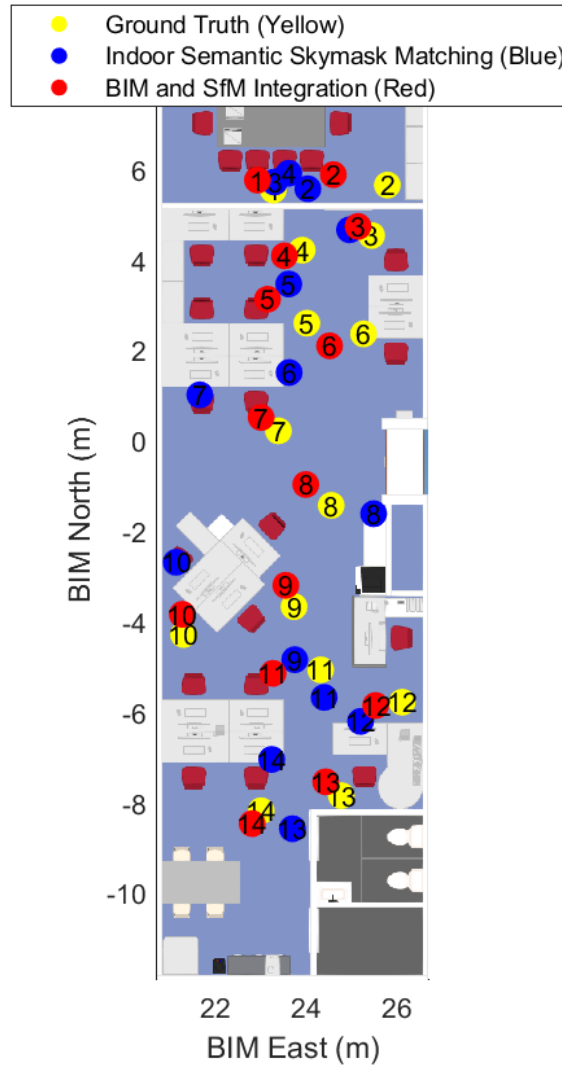


Figure 10: Positioning results of the proposed integration of BIM and panoramic SfM method

Overall, these results demonstrate that fusing BIM and SfM information can lead to more accurate and reliable camera pose estimation in architectural scenes. The precision requirements may vary depending on the specific application, but our approach has shown promising results and could be further optimized for specific use cases.

5. Conclusion

In this paper, we propose a novel approach for accurate camera pose estimation in architectural scenes by leveraging the complementary strengths of Building Information Modeling (BIM) and panoramic photogrammetry-based Structure-from-Motion (SfM). Our method fuses BIM's global positioning capabilities with SfM's precise relative positioning to overcome the limitations associated with each technique when used in isolation. Specifically, our proposed approach utilizes global positioning information from the BIM model to guide the camera pose estimation process, providing a global coordinate system and geometric information to constrain the relative positioning of panoramic images. This information is critical for achieving accurate and precise camera pose estimation and improving the integration quality between captured images and BIM models. By combining the benefits of both BIM and panoramic SfM, we aim to overcome the limitations inherent in each individual technique and achieve superior performance in camera pose estimation for architectural scenes.

The proposed pipeline consists of four key steps, including global positioning using indoor semantic skymask matching, relative positioning estimation from panoramic SfM, rough alignment using generalized Procrustes analysis, and refinement through non-linear least-squares optimization. The measured heading and point positioning are within 0.6m positioning accuracy according to the results performed in an office. The contributions of the proposed method are:

- The formulation of positioning as an indoor semantic skymask problem enables us to apply an existing wide variety of advanced matching metrics to this problem.
- Detection and exclusion of dynamic objects to prevent false measurements.
- By integrating the BIM model, replete with accurate scale information, our approach effectively recovers the appropriate scale of the SfM reconstruction.

Considering the preliminary results presented in this paper, we believe the proposed method can provide accurate positioning and heading estimation to support various indoor applications. Furthermore, this research has significant implications for the AEC domain, as the enhanced accuracy and efficiency offered by our method can lead to considerable advancements in various applications, such as augmented reality, facility management, and construction monitoring.

6. Future Works

Several potential future developments on the proposed method are suggested.

- **Simultaneous differential rendering and Factor Graph Optimization:** The proposed method uses SfM to estimate the relative position of the images and semantic skymask matching to estimate the global position. Factor graph optimization can then be applied to optimize the pose of the images simultaneously to maximize the overall indoor semantic matching score for all the images. Future work will explore the use of more advanced optimization techniques, such as bundle adjustment, to further improve accuracy.
- **Fault detection:** A simulation platform will be created to allow the injection of faults, such as occlusions, lighting changes, and object movement, and to evaluate the relationship between fault semantic indoor matching estimation and the accuracy of positioning. The proposed method will be evaluated under different fault scenarios to assess its robustness and effectiveness in real-world construction projects.
- **Dynamic update:** Dynamic updates will be introduced to automatically update the BIM model when changes are detected in the images, improving its real-time performance. Future work will explore the use of machine learning techniques, such as deep learning, to enable more accurate and efficient dynamic updates, especially when dealing with complex and dynamic scenes.

7. References

- [1] A. H. H. A. Ahmed, "Survey on indoor positioning applications based on different technologies," presented at the 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), Karachi, Pakistan, 2018.
- [2] C. Marouane, M. M. Feld, and M. Werner, "Visual positioning systems — An extension to MoVIPS," presented at the 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Busan, Republic of Korea, 2014.
- [3] M. Shu, G. Chen, Z. Zhang, and L. Xu, "Accurate Indoor 3D Location Based on MEMS/Vision by Using A Smartphone," presented at the 2022 IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN), Beijing, People's Republic of China, 2022.
- [4] J. Xue, X. Hou, and Y. Zeng, "Review of image-based 3D reconstruction of building for automated construction progress monitoring," *Applied Sciences*, vol. 11, no. 17, p. 7840, 2021.

- [5] A. J. Conde, J. Sanz-Calcedo, and A. Reyes-Rodríguez, "Use of BIM with photogrammetry support in small construction projects. Case study for commercial franchises," *JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT*, vol. 26, pp. 513-523, 06/08 2020, doi: 10.3846/jcem.2020.12611.
- [6] G. Xiaodong, H. Jiwei, L. Siyu, L. Jianhua, and D. Mingyi, "Indoor localization method of intelligent mobile terminal based on BIM," in *2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)*, 22-23 March 2018 2018, pp. 1-9, doi: 10.1109/UPINLBS.2018.8559731.
- [7] T. A. Nguyen, P. T. Nguyen, and S. T. Do, "Application of BIM and 3D Laser Scanning for Quantity Management in Construction Projects," *Advances in Civil Engineering*, vol. 2020, p. 8839923, 2020/12/28 2020, doi: 10.1155/2020/8839923.
- [8] T. Sun, Z. Xu, J. Yuan, C. Liu, and A. Ren, *Virtual Experiencing and Pricing of Room Views Based on BIM and Oblique Photogrammetry*. 2017.
- [9] M. J. L. Lee and L.-T. Hsu, "A Feasibility Study on Smartphone Localization using Image Registration with Segmented 3D Building Models based on Multi-Material Classes," presented at the 2021 International Technical Meeting of The Institute of Navigation, 2021.
- [10] M. J. L. Lee, H. Y. Ho, L.-T. Hsu, and S. L. M. Au, "BIPS: Building Information Positioning System," presented at the BIPS: Building Information Positioning System, Lloret de Mar, Spain, 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/9662575/>.
- [11] M. J. L. S. Lee, H.-F. Ng, and L.-T. Hsu, "Skymask Matching Aided Positioning Using Sky-Pointing Fisheye Camera and 3D City Models in Urban Canyons," *Sensors*, 2020, doi: 10.3390/s22176533.
- [12] Z. Jiang, Z. Xiang, J. Xu, and M. Zhao, *LGT-Net: Indoor Panoramic Room Layout Estimation with Geometry-Aware Transformer Network*. 2022.
- [13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077-12090, 2021.
- [14] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "Openmvg: Open multiple view geometry," in *Reproducible Research in Pattern Recognition: First International Workshop, RRPR 2016, Cancún, Mexico, December 4, 2016, Revised Selected Papers 1*, 2017: Springer, pp. 60-74.