DOI: 10.1111/mice.12714

ORIGINAL ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Indoor camera pose estimation via style-transfer 3D models

Junjie Chen[1,2] | **Shuai Li**[1] | **Donghai Liu**[3] | **Weisheng Lu**[2]

[1] Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, Tennessee, USA

[2] Department of Real Estate and Construction, The University of Hong Kong, Hong Kong, China

[3] State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Tianjin, China

Correspondence
Shuai Li, Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, 851 Neyland Dr., Knoxville, TN, USA.
Email: sli48@utk.edu

Funding information
National Science Foundation, Grant/Award Numbers: 1850008, 2038967

Abstract

Many vision-based indoor localization methods require tedious and comprehensive pre-mapping of built environments. This research proposes a mapping-free approach to estimating indoor camera poses based on a 3D style-transferred building information model (BIM) and photogrammetry technique. To address the cross-domain gap between virtual 3D models and real-life photographs, a CycleGAN model was developed to transform BIM renderings into photorealistic images. A photogrammetry-based algorithm was developed to estimate camera pose using the visual and spatial information extracted from the style-transferred BIM. The experiments demonstrated the efficacy of CycleGAN in bridging the cross-domain gap, which significantly improved performance in terms of image retrieval and feature correspondence detection. With the 3D coordinates retrieved from BIM, the proposed method can achieve near real-time camera pose estimation with an accuracy of 1.38 m and 10.1° in indoor environments.

1 | INTRODUCTION

Robust and accurate positioning information is the cornerstone of many location-based services in indoor built environments. For example, construction robots need to locate themselves indoors to perform necessary actions such as progress monitoring (Asadi et al., 2019; Hamledari et al., 2017). In facility management based on augmented reality (AR) (Palmarini et al., 2018), the position and orientation of a device holder serve as prerequisites for 3D registration into a relevant virtual model. Customers in large commercial buildings rely on positioning information for wayfinding. Despite its importance, however, indoor localization is more challenging than outdoor localization because global navigation satellite system (GNSS) signals are unavailable in the bounded spaces inside a structure (J. Z. Liang et al., 2015; Winter et al., 2019). Existing indoor localization techniques—enabled by radio frequency identification (RFID), Wi-Fi, Bluetooth, or ultra-wideband (UWB)—depend on the deployment of signal emission infrastructure (Acharya, Khoshelham, et al., 2019). The dependency on external infrastructure makes these techniques difficult to scale up to a wide range of applications, hindering broader progress in this area (Winter et al., 2019). The challenge and dilemma confronting indoor localization were well depicted by Ellard (2009) in his book subtitled "Why we can find our way to the moon, but get lost in the mall."

In light of this deficiency, researchers have sought to develop infrastructure-independent indoor localization techniques based on machine vision. One stream of such works attempts to estimate camera posture using visual odometry (VO) (Nister et al., 2004) and simultaneous localization and mapping (SLAM) (Davison, 2003). VO estimates local camera motion by detecting and processing consistent feature correspondences from sequential camera frames, whereas SLAM constructs a global map

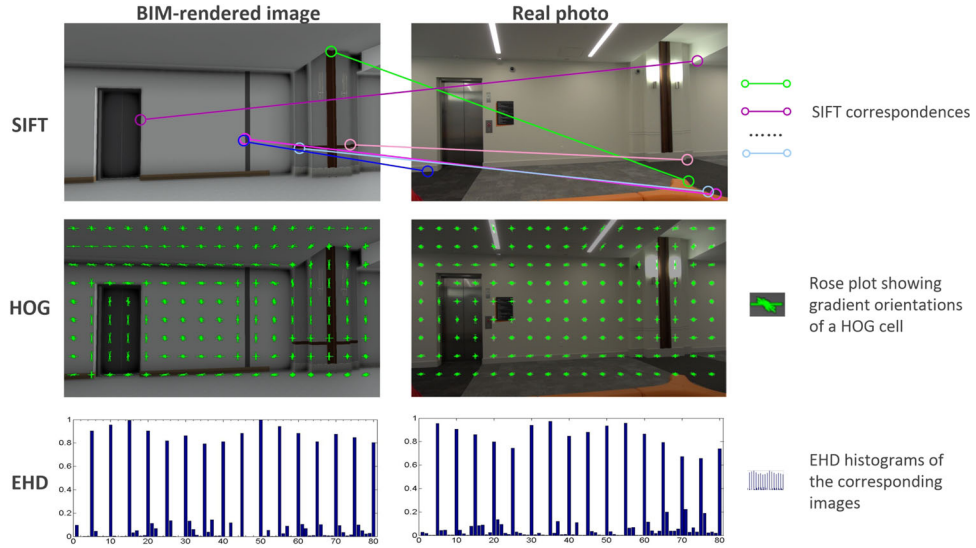

FIGURE 1 Cross-domain gap between visual representations displayed by BIMs and real-life scenarios

of the environment and simultaneously determines the camera position and orientation with respect to the map (Yousif et al., 2015). However, using either VO or SLAM alone fails to provide absolute positioning information with regard to a global reference frame. To achieve global localization, the initial location of the camera is required (Acharya, Khoshelham, et al., 2019). Another problem involves the issue of drifting because both VO and SLAM techniques derive the current state of camera posture by summing incremental movement based on the previous state. These drifting issues can lead to large deviations from the estimated positions (Winter et al., 2019).

Approaches based on image retrieval address the drawbacks of VO and SLAM by providing the absolute coordinates of cameras in the global reference frame. With these approaches (J. Z. Liang et al., 2015; Ravi et al., 2006), the absolute location of a camera is determined by comparing the similarity between an input query image and a database of georeferenced or geotagged images. The position information associated with the most similar image is retrieved as the camera location of the current query image. Further developments in image retrieval approaches have attempted to register or compare the query image with 3D point clouds (Svärm et al., 2017) or RGB-D images (Shotton et al., 2013). In these methods, a camera pose with six degrees of freedom (6DoF) can be estimated using an analytical solution based on photogrammetry. However, image retrieval approaches and their variations require a large database of photographs with known positions or a pre-existing map of the entire indoor space, which is highly labor intensive and time consuming to acquire (Acharya, Khoshelham, et al., 2019; Ha et al., 2018; Piasco et al., 2018).

The proliferation and adoption of building information models (BIMs) provide 3D spatial models for constructed facilities that contain both the visual appearance (i.e., images) and spatial layout (i.e., position) of indoor environments. Because information retrieval from BIMs is substantially more convenient than in-situ data collection and mapping, the rich information archived in BIMs can be exploited as a spatial registration landmark to enhance indoor localization techniques. However, there is a cross-domain gap (Ha et al., 2018; Shrivastava et al., 2011) between BIM renderings with plain texture and real-life photographs with vivid texture, as demonstrated in Figure 1. Given a BIM-rendered image and a photograph of the same indoor scenario, the computer failed to detect correct scale-invariant feature transform (SIFT) correspondences (Lowe, 2004) and identified different histogram of oriented gradient (HOG) (Dalal & Triggs, 2005) and edge histogram descriptor (EHD) (Won et al., 2002) distribution patterns for different domains. This phenomenon indicates that the computer perceives BIM renderings and photographs in different ways, which make it difficult to directly perform similarity comparisons for image retrieval (Lenjani et al., 2020) or registration based on SIFT, HOG, or other image descriptors (Ha et al., 2018). To address this cross-domain gap, some pioneering research efforts have attempted to enable indoor localization using readily available information from BIMs (Acharya, Khoshelham, et al., 2019; Acharya, Ramezani, et al., 2019; Asadi et al., 2019; Ha et al., 2018). However, existing studies have either directly obtained the location through a simple image retrieval approach or have used a regression model to estimate indoor location. Localization based on image retrieval is subject to viewpoint change (Piasco et al., 2018). Regression-enabled approaches suffer from the black-box

nature of deep learning models, the localization rationale of which is uninterpretable. In addition, the highest localization accuracy—achieved by Acharya, Khoshelham, et al. (2019)—was still lower than the baseline performance when real-life images were used for both training and testing. Thus, the negative effects of the cross-domain gap persist and addressing the problem of how to enable effective indoor localization remains an open challenge.

To overcome this issue, this study developed a photogrammetry-based indoor localization approach that converts texture-less BIMs into a style with a photorealistic texture and retrieves spatial information for camera pose estimation. Recently emerging generative adversarial networks (GANs) (Goodfellow et al., 2014) provide the technical tools for this conversion. A study by Hong et al. (2020) demonstrated that CycleGAN (Zhu et al., 2017), a variation of GAN, could synthesize vivid and plausible indoor photographs based on BIM-rendered images of the same scene. However, the study was intended for indoor scene understanding. Much remains unclear regarding how the synthetic photographs after style transfer can facilitate cross-domain information retrieval and how such style-transfer renderings from BIMs can be used to enable indoor localization.

The contributions of this study are threefold. First, CycleGAN was trained to transform BIM-rendered images without texture into photorealistic images with a vivid texture. The experimental results demonstrate that the synthetic photographs after style transfer can be better used for image retrieval based on global features and SIFT correspondence detection, which provides a basis for extracting spatial information from BIMs to estimate camera poses.

Second, a photogrammetry-based indoor localization method was proposed and tested using the spatial information retrieved from the style-transferred BIM. Inputting the retrieved spatial information and solving the perspective-*n*-point (P*n*P) problem, the proposed method can yield a 6DoF camera pose with an average accuracy of 1.38 m and 10.1° in less than 1 s, eliminating the onerous process of indoor pre-mapping required by many existing methods.

Third, the study evaluated the influence of the features used and database size on image retrieval accuracy, as well as the effects of the number of key points on camera pose estimation. Through this analysis, the following findings were obtained: (1) Compared with other global image features, such as HOG and color histograms, EHD shows the most promising performance in retrieving synthesized indoor photographs from a large database; and (2) the precision of the estimated camera pose fluctuated with the number of key points, and the best performance was observed when the value equaled 10.

2 | RELATED WORK

2.1 | Vision-based localization

Robust and accurate object localization is an important premise for implementing effective construction management (Fang et al., 2020) and facility maintenance (X. Liang, 2019; Pan & Yang, 2020). Compared with traditional localization techniques (e.g., RFID, Bluetooth, Wi-Fi, UWB), vision-based approaches stand out for their cost-effectiveness, accessibility, and ability to function without deploying external signal emission or reception infrastructure. These merits have drawn many researchers to devise robust localization techniques based on machine vision. Fang et al. (2020) proposed an approach to locating construction resources from monocular videos based on deep semantic segmentation models and humans' prior knowledge. Li et al. (2018) developed a vision-based method for the simultaneous detection and localization of concrete defects. In Wang et al. (2017), an automated algorithm was developed to estimate the positions of rebars in reinforced precast concrete for construction quality control.

Another stream of work has attempted to realize the self-localization of subjects (humans or robots) in indoor environments with camera-captured visual information. SLAM (Davison, 2003) and VO (Nister et al., 2004) are two classical visual algorithms for robot self-localization and navigation. Both approaches determine a robot's position in an incremental manner by processing feature correspondences between sequential camera frames using photogrammetry. Thus, SLAM and VO can only yield relative position coordinates with regard to an initial location (Acharya, Khoshelham, et al., 2019) and usually suffer from drifting issues (Winter et al., 2019). To obtain subjects' absolute locations in a global reference frame, many vision-based algorithms have been proposed, typically involving the registration of newly captured monocular (Dong et al., 2015; J. Z. Liang et al., 2015; Lu & Kambhamettu, 2014; Ravi et al., 2006) or omnidirectional (Murillo et al., 2007; Rituerto et al., 2010) photographs to a collection of indoor photographs with known positions or 3D point clouds of indoor environments. When indoor photographs with known positions are used as a registration target, the localization problem is essentially converted to a problem of content-based image retrieval (Li et al., 2018; J. Park et al., 2018; Ravi et al., 2006). This approach uses the position of the retrieved photograph as the location of the input query image; thus, the localization accuracy heavily relies on the quantity of pre-collected photographs and is subject to viewpoint change (Piasco et al., 2018). Approaches based on 3D point clouds (J. Z. Liang et al., 2015; Lu & Kambhamettu, 2014; Svärm et al.,

2017) can provide the 6DoF camera posture of the query image through a rigorous analytical calculation based on photogrammetry. Some newly proposed strategies have considered camera pose calculation as a regression problem that aims to reconstruct the mapping relationship between photographs (Kendall et al., 2015) or RGB-D images (Shotton et al., 2013) and their camera poses using machine learning models. One common drawback of the above approaches is that they all require pre-mapping of the indoor environment to obtain either photographs with known positions, point clouds, or RGB-D images. Such pre-mapping efforts are laborious, expensive, and time consuming (Acharya, Khoshelham, et al., 2019; Ha et al., 2018; Piasco et al., 2018).

Compared with in-situ mapping, capturing renderings with known geo-spatial information from 3D models is cost effective and can be fully automated. With the proliferation of BIMs, pioneering research has been conducted to utilize the visual and spatial information from 3D BIM models to enable indoor localization. Ha et al. (2018) proposed a novel indoor localization approach based on image retrieval that compared deep feature maps extracted by convolutional neural networks (CNNs). The deep features successfully bridged the cross-domain gap and led to high matching accuracy. Asadi et al. (2019) developed an approach to inferring camera positions through perspective matching between image frames and the corresponding visual views displayed by BIMs. Acharya, Khoshelham, et al. (2019) used a CNN to train a camera pose regression model with synthetic images obtained from a 3D indoor model. The regression model achieved an accuracy of 2 m with real photographs. However, due to the black box nature of deep learning, the localization rationale of regression models is difficult for humans to interpret (Rudin, 2019). In contrast, our approach uses the style-transfer technique to transform BIM-rendered images into geo-registered photorealistic images and estimates camera pose via analytic solutions based on classical photogrammetry theory.

2.2 | Cross-domain style transfer

Style transfer aims to transfer the appearance or visual style of images from one domain to another. Traditionally, this has been done using image analogy techniques (Hertzmann et al., 2001), which train a machine learning model on a collection of image pairs from two domains (e.g., photographs and paintings) to learn the transformation mapping. However, the method has had limited practical application due to the difficulty of preparing the dataset of image correspondences. Neural style transfer (NST), which does not require cross-domain image pairs, was developed to address these limitations (Gatys et al., 2016; Johnson et al., 2016). NST disentangles the content of an image from its style. By minimizing both content loss and style loss, a synthetic image can be obtained with the desired texture applied onto the source image.

Style transfer can also be formulated as the problem of image-to-image translation (Isola et al., 2017; Zhu et al., 2017), which has traditionally been tackled with separate, domain-specific machinery (Buades et al., 2005; Fergus et al., 2006). Inspired by recently proposed GAN techniques, Isola et al. (2017) presented a general-purpose approach to image-to-image translation using conditional adversarial networks. Based on their work (Isola et al., 2017), Zhu et al. (2017) proposed CycleGAN, an image-to-image translation framework that can be trained on a collection of unpaired images from two domains. CycleGAN is a simple but powerful technique that aims to learn the mapping between the source and target domain and has shown promising performance in various application scenarios, such as style transform and object transfiguration. Compared with other style-transfer techniques, CycleGAN can achieve state-of-the-art performance in image-to-image translation without paired training samples from two domains. The high quality of the translation results and less strict requirements for training data make CycleGAN a viable solution for our application: BIM-to-real style transfer.

BIM-rendered images and indoor photographs are data from two different domains with varying styles. Because of the cross-domain nature, there is a perception gap for computers with regard to effectively interpreting visual information from BIMs. Arguably, this cross-domain perception gap can be addressed with CycleGAN by transferring the visual information in the two involved domains into an identical one. In civil engineering, researchers have explored GANs' capability to generate plausible new samples as a promising data augmentation technique for improving performance in structural defect detection (Gao et al., 2019; Maeda et al., 2020), construction resource management (Bang et al., 2020), and flood depth estimation (S. Park et al., 2021). However, CycleGAN has rarely been applied for style transfer in civil engineering contexts, with only a few exceptions (Hong et al., 2020; Pouyanfar et al., 2019). Hong et al. (2020) performed a pilot study of BIM-to-real style transfer, which demonstrated the feasibility of using CycleGAN to synthesize photorealistic images from given BIM-rendered images. However, it remains unclear whether the results of transformation can be better exploited to enable image retrieval or registration with algorithms such as SIFT. It is also unclear how style transfer can be incorporated into the pipeline of indoor localization.

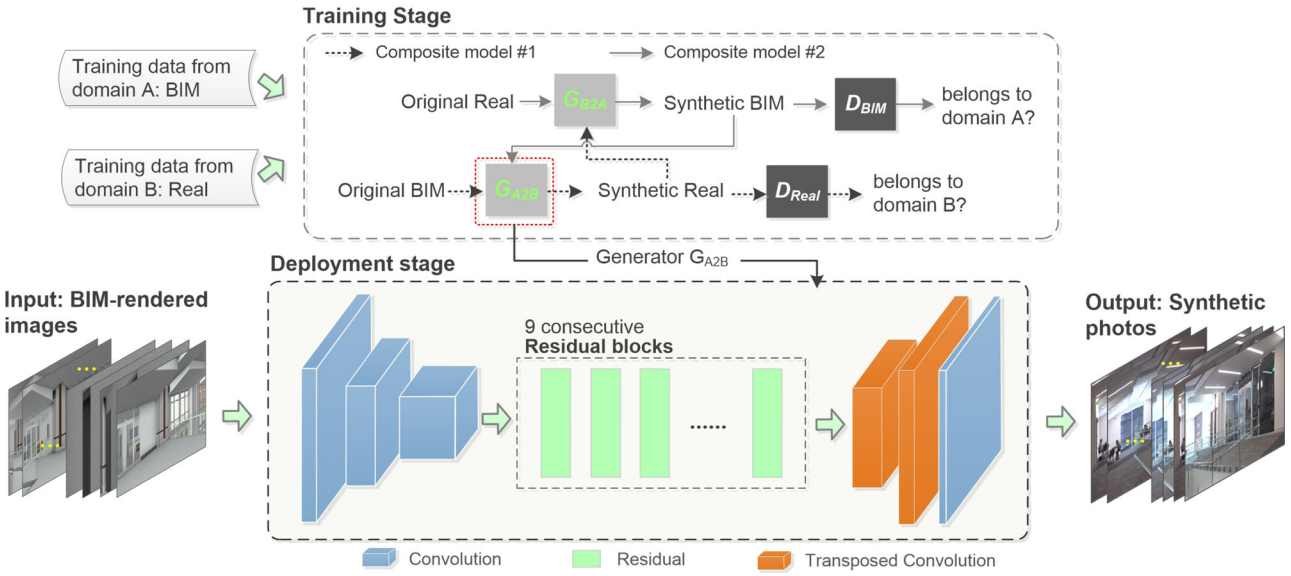

FIGURE 2 Schematic diagram of BIM style transfer based on CycleGAN model

2.3 | Knowledge gaps

The above literature review revealed the following knowledge gaps. First, existing visual indoor localization methods based on BIM use either image retrieval or regression techniques to obtain position or camera pose. Localization based merely on image retrieval is subject to viewpoint change (Piasco et al., 2018) and the influence of database scale. Regression-based approaches suffer from the black box nature of deep learning, which could affect users' confidence in deploying such solutions (Rudin, 2019). Second, although a previous study (Hong et al., 2020) demonstrated the efficacy of CycleGAN in BIM-to-real style transfer, much remains unclear regarding how synthetic photographs can be used to enable indoor localization.

Our proposed method innovatively leverages 3D style-transferred BIMs for accurate indoor camera pose estimation, which computationally integrates cross-domain image matching based on deep learning and photogrammetry-based localization. Information retrieval is realized by addressing the cross-domain gap with CycleGAN style transfer, and indoor localization is achieved through an analytical solution to ensure interpretability and accuracy. The proposed method provides a new venue for computationally exploiting the visual and spatial information in BIMs and leverages this new stream of data to improve indoor localization that is not amenable to many existing solutions. The new method could also enable a variety of important applications, such as visual SLAM for robots and AR in indoor environments.

3 | METHODOLOGY

3.1 | BIM style transfer with CycleGAN

Due to the cross-domain gap, rendered images captured from BIMs must be transferred to images with photorealistic texture before they can be used for indoor localization. As shown in Figure 2, CycleGAN, a GAN for image-to-image translation, is trained to enable cross-domain style transfer. Similar to other GANs, CycleGAN comprises generators and discriminators. Generator G_{A2B} takes BIM images (i.e., domain A) as the input and aims to transform them into photographs with realistic textures (i.e., domain B). Discriminator D_{Real} is designated to determine whether a given image is from domain B (i.e., whether it is a "real" photograph). G_{A2B} and D_{Real} together form composite model #1, where the generator and discriminator are adversarial from a game theory perspective. However, training the model to map from domains A to B can only ensure the synthesis of photorealistic images; it does not guarantee that individual inputs and outputs can be paired in a meaningful way (Zhu et al., 2017). Therefore, the concept of cycle consistency is introduced by adding another composite model (composite model #2 in Figure 2) into the framework. This model includes a generator referred to as G_{B2A} and a discriminator termed D_{BIM}. G_{B2A} is used to translate G_{A2B}'s output back to the source domain, which produces a synthetic version of the original input BIM image. This synthetic BIM image should be as similar as possible to the original to ensure that G_{A2B}'s output directly corresponds to the input (Brownlee, 2019;

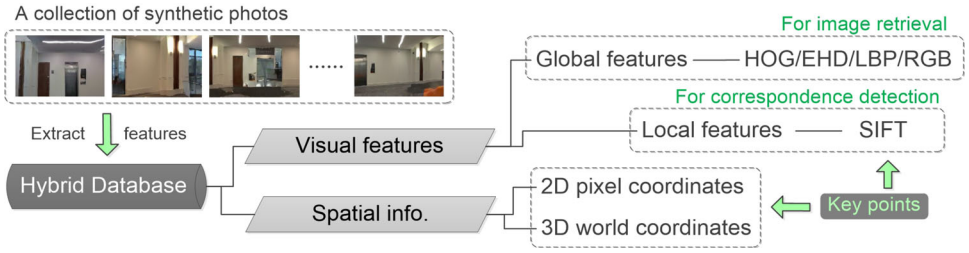

FIGURE 3 Compositions of the constructed visual-spatial hybrid database

Lee, 2017). Composite models #1 and #2 are trained simultaneously on unpaired images from both domains.

After training, G_{A2B} is used separately for style transfer. First, a collection of BIM-rendered images is captured from a BIM corresponding to the indoor environment where later localization takes place. The image collection is then input to G_{A2B}, which processes and translates the BIM-rendered images into synthetic photographs with vivid textures using its internal network structure. The network first encodes the input in three convolutional layers, performs transformation in nine consecutive residual blocks, and finally, after decoding the transformation results, obtains a collection of synthetic photographs corresponding to the original input.

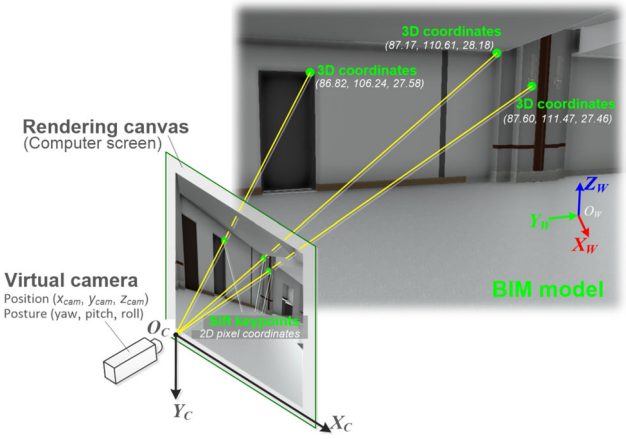

FIGURE 4 Acquisition of 3D world coordinates of key points from BIM

3.2 | Constructing a visual-spatial hybrid database

After style transfer, a collection of photorealistic images of designated indoor environments can be obtained. From visual perspectives, the images present vivid textures and ambience that resemble the nature of real-life scenarios, which makes them easy for computers to process using methods similar to those used to process real photographs. Meanwhile, the images maintain spatial layout and connection with the original BIM; thus, corresponding 3D coordinates can be retrieved from BIM for arbitrary 2D pixels in the images. In this stage, a database was constructed by extracting both visual features and spatial information from the photorealistic image collection.

Figure 3 shows compositions of the visual-spatial hybrid database. The visual features stored in the database include two aspects—global and local features—which are later used for image retrieval and correspondence detection. Global image features characterize the overall patterns of an image in terms of color, texture, or shape; some examples include HOG, EHD, local binary pattern (LBP) (Pietikäinen et al., 2000), and RGB color histograms. Because synthetic images after style transfer inevitably suffer from blurring or distortion at the local scale, the characteristics of global features make them more suitable than local features for initial image retrieval. Local image features focus on a pattern or distinct structure found in local areas, such as a point, edge, or small image patch. A representative algorithm here is SIFT (Lowe, 2004), which detects local features invariant to scale and rotation and robust to viewpoint changes. In this study, SIFT is used to detect scale-invariant key points (KPs) on each image from the photorealistic image collection. Each extracted SIFT feature is connected to its KP ID and image ID so that the corresponding KP and image can be retrieved when required in later processing steps.

The other important part of the hybrid database is the spatial information of KPs extracted by the SIFT algorithm, which includes the 2D pixel coordinates of KPs on their respective images and their 3D world coordinates in the BIM. Because the synthetic photographs are one-to-one correspondences of their original BIM images, they can be directly related to the 3D BIM, where the 3D spatial geometry is predetermined and can be considered prior knowledge. Thus, obtaining 3D world coordinates of the KPs becomes trivial, as shown in Figure 4. With known camera position and posture in BIM and a KP's pixel coordinates on the rendering in question, a radial line can be drawn from the camera through the KP. The position where the line first clashes with an element is the 3D coordinates

of the KP in the real world. In implementation, this can be realized using off-the-shelf BIM tools. This study uses the *clientToWorld*() function (Autodesk, 2020a) provided by Autodesk Forge Viewer to obtain KPs' 3D world coordinates.

3.3 | Image retrieval and correspondence detection

The visual-spatial hybrid database can be used as a basis for inferring the camera pose of a new image (referred to as a query image) captured in the same indoor environment. The query image is registered into the database through two steps: image retrieval and correspondence detection.

First, N_{img} images that are most similar to the query image are retrieved from the database by comparing their global features, such as HOG, EHD, LBP, and/or RGB histograms. The metric used for comparison is cosine similarity, which measures similarity according to the angle between two feature vectors and thus does not require normalization (Ha et al., 2018; Qian et al., 2004). Let $\mathbf{V}_{\text{query}}$ and $\mathbf{V}_{\text{train}}$ respectively denote the global features of the query image and an arbitrary image from the database (referred to as a trained image). Their cosine similarities can then be calculated as follows:

$$similarity = \frac{\mathbf{V}_{\text{query}} \cdot \mathbf{V}_{\text{train}}}{\|\mathbf{V}_{\text{query}}\|\|\mathbf{V}_{\text{train}}\|} \tag{1}$$

In a positive feature space, cosine similarity is in a range of [0, 1], where a similarity score of 1 indicates that two vectors are identical, and 0 indicates they are completely dissimilar. The number of retrieved images N_{img} and the specific global features to use are not yet determined in this stage; they are investigated and specified through empirical analysis in Section 4.

Second, SIFT correspondences are detected between the query image and each of the N_{img} retrieved images. In typical implementations, a k-d tree structure is often used for SIFT correspondence detection to expedite the searching process in a huge database (J. Z. Liang et al., 2015). However, because the image retrieval operation in the first step dramatically narrows the search range (from thousands of images to N_{img} images), we used brute force searching for correspondence detection in this study. The SIFT features detected from the query images are successively compared with those from the N_{img} retrieved images. To ensure the quality of detection, cross-checking techniques (OpenCV, 2021a) or ratio tests (Lowe, 2004) can be implemented for outlier exclusion. The top N_{sift} nearest neighbors from each of the retrieved images are extracted as SIFT correspondences. As a result, for a query image, a total of $N_{\text{sift}} \times N_{\text{img}}$ SIFT correspondences are detected from the database.

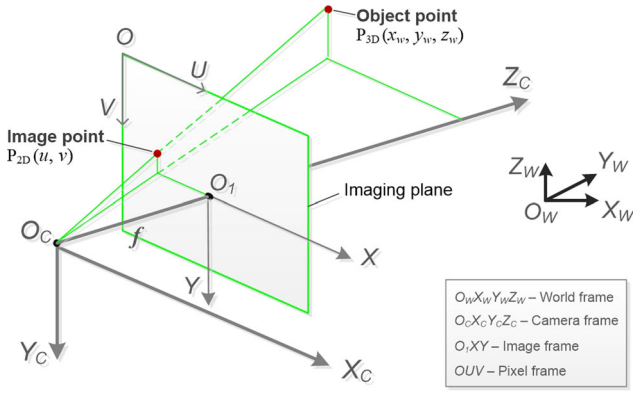

FIGURE 5 Coordinate transformation relationship under pinhole camera model

3.4 | Camera pose estimation

Through the correspondence relationship and indexing by image ID and KP ID, the actual 3D world coordinates of KPs on the query image can be obtained from the visual-spatial hybrid database. This gives $N_{\text{sift}} \times N_{\text{img}}$ pairs of 2D–3D coordinate correspondences. With n 2D points on an image and their corresponding 3D coordinates in the world frame, the estimation of camera pose is a typical PnP problem (Fischler & Bolles, 1981).

Figure 5 shows how a 3D object point in the world coordinate frame is related to its 2D counterpart on the imaging plane under the pinhole camera (or perspective projection) model. Note that for convenience of expression, the imaging plane has been placed in front of the camera, which is mathematically equal to the actual model with the plane behind the camera. The projection from the 3D object point to the 2D image point can be formulated as a series of coordinate transformations involving the world coordinate frame $O_WX_WY_WZ_W$, camera coordinate frame $O_CX_CY_CZ_C$, image coordinate frame O_1XY, and pixel coordinate frame OUV. The transformation between $O_WX_WY_WZ_W$ and $O_CX_CY_CZ_C$ can be expressed as in Equation (2):

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{2}$$

where $[x_c, y_c, z_c, 1]^{\text{T}}$ and $[x_w, y_w, z_w, 1]^{\text{T}}$ are homogeneous coordinates of the object point in camera frame $O_CX_CY_CZ_C$ and world frame $O_WX_WY_WZ_W$, respectively, and R and T

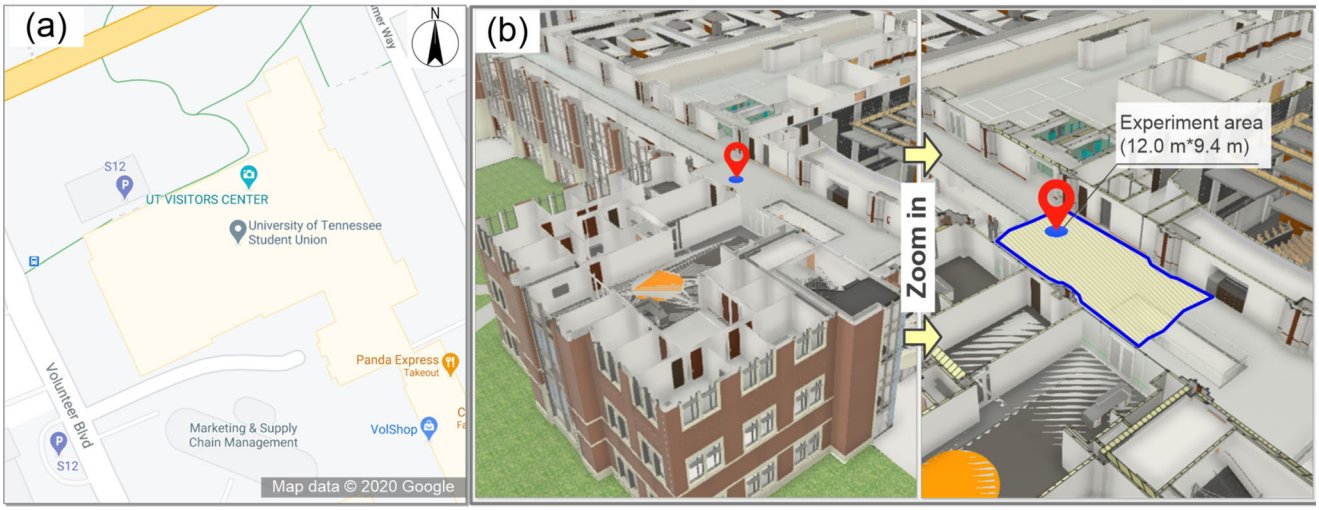

FIGURE 6 (a) Street map of Student Union at UTK campus; (b) Location of experiment site in Student Union complex

are the rotation matrix (3×3) and translation vector (3×1), respectively, from the world frame to the camera frame.

The object coordinates in the camera frame are projected onto the imaging plane according to Equation (3):

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f/dx & 0 & u_0 & 0 \\ 0 & f/dy & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \tag{3}$$

where $[u, v, 1]^T$ represents the 2D pixel coordinates (in homogeneous form) of the object point after projection; f is the focal length of the camera lens; dx and dy are the size of an individual pixel on the image sensor (e.g., a charge-coupled device [CCD]) on the respective U and V axes; and (u_0, v_0) are the coordinates of the camera optical center in the pixel frame OUV.

By combining Equations (2) and (3), the mapping relationship between a 3D world point and its image point can be expressed as follows:

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f/dx & 0 & u_0 & 0 \\ 0 & f/dy & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{4}$$

where K, termed the intrinsic matrix, only correlates to the internal structure of a camera and can be determined by camera calibration (Zhang, 2000) and $[R|T]$, referred to as the extrinsic matrix, represents the camera pose with respect to the world frame. To solve $[R|T]$ (i.e., camera pose), at least four correspondences with known 2D (i.e., $[u, v, 1]^T$) and 3D coordinates (i.e., $[x_w, y_w, z_w, 1]^T$) are required.

The required 2D–3D correspondences can be obtained from the database, as mentioned at the beginning of this section. However, due to deviations in SIFT KPs or even incorrect correspondence detection, there will inevitably be noise in the data. To address this issue, random sample consensus (RANSAC) is used in conjunction with the PnP solution to make the estimated pose more robust to outliers (OpenCV, 2021b).

4 | EXPERIMENTAL STUDIES

Experimental studies were carried out at a campus building at the University of Tennessee, Knoxville (UTK) to investigate the performance of the proposed indoor localization approach. As shown in Figure 6, an experiment area was set up at the northwest corner of the third floor of the Student Union complex, which covers an area of 12.0 m × 9.4 m. The BIM of the building has a level of development of 350 and was created with Autodesk Revit. Training of the BIM-to-real style-transfer model was performed on a Dell Precision 5820 Tower workstation equipped with an Intel® Core™- i9-7920X CPU, 32 GB RAM, and a NVIDIA GeForce GTX 1080 Ti GPU. Image retrieval, correspondence detection, and camera pose estimation were run on an ASUS VivoBook S15 laptop with an Intel® Core™- i7-8550U processor and a NVIDIA GeForce MX150 GPU.

4.1 | Data collection schemes

Data were collected from the experiment area for two purposes: training a CycleGAN model for style transfer

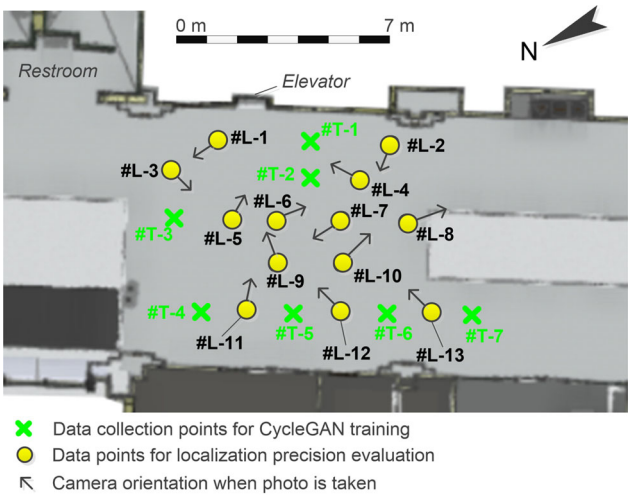

FIGURE 7 Data collection scheme in the experiment area

and evaluating the precision of the indoor localization approach.

To train CycleGAN for BIM-to-real transfer, it was necessary to prepare a dataset consisting of images from both domains (i.e., the real and the BIM). The real photographs were collected with a digital video (DV) camera (SONY HDR-CX760V). The DV camera was used to record videos of the experiment area at seven designated locations, indicated with green crosses in Figure 7. The DV camera was supported by a tripod to maintain its stability. When collecting data at each point, the camera was designated to spin 360° around the central vertical axis of the tripod to record a video of the surrounding indoor scenario. Seven 2–3-minute videos were obtained. From these videos, static image frames were extracted at three different time intervals: 1.5, 0.75, and 0.375 s. This resulted in approximately 100, 200, and 400 real images, respectively, for each video (or at each data collection point). The BIM images were rendered using a web-based BIM displayer termed Autodesk Forge Viewer, which provides a function—*getScreenShot*()—that allows users to obtain screenshot images of the BIM scenario displayed in the viewer (Autodesk, 2020b). Screenshots of the surrounding BIM scenario were generated at the seven data collection locations using the *getScreenShot*() function. The quantity of screenshots at each location must be approximately equivalent to that of the real images (i.e., 100, 200, and 400).

According to the quantity of images at each location, three different datasets were obtained: DS-7-100, DS-7-200, and DS-7-400. The "7" indicates that seven locations were designated for data collection, and the last number represents the approximate quantity of images collected at each location. Each of the three datasets was randomly split into a training set and a test set. Table 1 lists the details of the three datasets, where "trainA/testA" and "trainB/testB" denote the training and test sets of domains A (i.e., BIM) and B (i.e., Real), respectively. Additionally, to ensure the model's generalizability to spatial variation, extra samples were acquired at the neighborhood (0.5–1.3 m) of each data collection point. The extra samples formed the "testA_add (BIM)" and the "testB_add (Real)" subset in Table 1. Images from both the BIM and Real domains in each dataset were resized to the resolution of 256×144. The total number of the original and extra test samples accounts for approximately 10% of the corresponding dataset, which guarantees the reliability of subsequent evaluation. If not specified in the remaining of the study, we use "test set/samples" to denote the original test set/samples, while "test_add set" to represent the set of extra test samples.

TABLE 1 Details of the collected datasets

Split	Datasets		
	DS-7-100	DS-7-200	DS-7-400
trainA (BIM)	681	1362	2614
trainB (Real)	502	1039	2709
testA (BIM)	25	50	107
testB (Real)	42	50	101
testA_add (BIM)	63	107	192
testB_add (Real)	46	101	203

Abbreviation: BIM, building information model.

To evaluate indoor localization precision, 13 indoor photographs were collected at 13 different locations with different camera orientations, indicated with yellow dots in Figure 7. The camera used was a SONY HDR-CX760V. To avoid potential evaluation biases, the 13 photographs were taken on a different day and under varying lighting conditions than the training data. The collection points were designed to be scattered across different locations in the experiment area and to avoid overlap with the training data's collection points (the green crosses in Figure 7). In addition, the camera orientations (the arrows in Figure 7) with which the photographs were taken should be as diverse as possible to ensure that the evaluation covers photographs of the indoor space from different angles.

4.2 | Evaluation of style transfer results

Using the collected datasets summarized in Table 1, CycleGAN models were trained with the following hyperparameters. Batch size, number of training epochs, and initial learning rate were set to 1, 200, and 0.0002, respectively. After 100 epochs of the training, the learning rate began to decay. The gradient penalty weight, cycle loss weight, and identity loss weight were specified as 10, 10, and 0, respectively. The pool size used to store fake samples

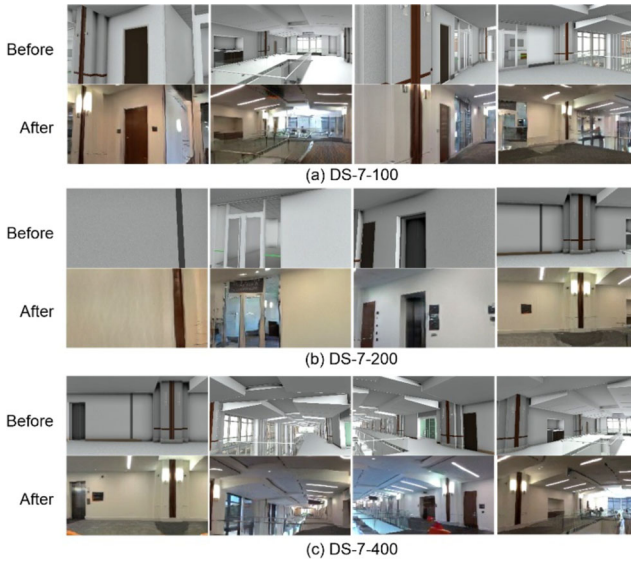

FIGURE 8 Examples of style transfer results based on CycleGAN

was 50. After the training, three models were obtained based on DS-7-100, DS-7-200, and DS-7-400. As shown in Figure 8, despite some blurring and distortion in local areas, quite plausible images with photorealistic textures were generated.

Two approaches were adopted to investigate to what extent the readability of the synthetic images was improved for both humans and computers. Human perception was assessed using a visual Turing test (Salimans et al., 2016). Fake samples in both the testA and testA_add sets generated by the three models, along with 151 authentic photographs of the same indoor scenario, were outsourced to workers on Amazon Mechanical Turk to be rated as real or fake. Only a single image was presented at a time. Within 15 seconds, workers were required to determine whether the presented image was real or fake and, in the latter case, to judge its level of similarity (75, 50, 25, or 0%) to an authentic image. Each image was rated by multiple workers (50, in our case). As the rating process continues, workers might become both fatigued and more proficient at performing the task, which could influence their performance. To minimize such influences, the image sequence was randomized for each worker. In this way, the objectiveness of the evaluation can be ensured by averaging the rates given by many different workers:

$$SS = \frac{\sum_{i=1}^{N} S_i}{N} \tag{5}$$

where S_i is the similarity level of a sample rated by worker i, of which the values include 100 (real), 75, 50, 25, and 0% (fake); N is the total number of recruited workers who rated the sample; and SS is the similarity score of the sample.

Figure 9a shows average SS and its standard deviation of the original, extra, and all test samples in different datasets. The average SS of the "Test_add" set, of which samples were collected from different locations than the training data, is basically at the same level with the "Test_original" set, demonstrating the models' generalizability to spatial variation. Considering all the test samples, DS-7-100 elicited the highest average SS of the synthetic photographs (0.58), whereas the average SS of the 151 authentic photographs was 0.78. Figures 9b and c show the detailed vote distribution for DS-7-100 and the Real image dataset. For DS-7-100, a significant proportion (23%) of votes were distributed at the "100%" (i.e., real) level. For the Real dataset, nearly 50% of votes were distributed over the similarity levels of 75, 50, 25, and 0% (i.e., fake). In a study that attempted to synthesize road pothole images with GANs (Maeda et al., 2020), 50% of synthetic samples were mistaken for real ones, whereas 76% of real samples were considered fake. Compared with pothole images with rather simple textures, synthesizing images of indoor environments is more difficult because the pattern of texture, illumination, color, and shape features is more complex and diverse. Hence, although there is space for further improvement, the synthetic indoor photographs obtained in our study were already satisfactorily capable of confusing human observers.

For indoor localization, a more important question is whether style transfer can improve the machine readability of cross-domain visual information from BIMs. To answer this question, five typical algorithms for image feature extraction were used: SIFT, HOG, EHD, LBP, and RGB color histograms (RGB for short). Other than the synthesized test samples offered by the three datasets in Table 1, another two datasets—consisting of entirely BIM-rendered images and real photographs, respectively—were also prepared as baselines for direct comparison. Samples in the five datasets (three synthetic, one BIM, and one real) were successively compared with a large collection of real photographs of the experiment area for similarity measurement. The top 20 matches for each sample were taken, and the similarity values of such matches of all samples from the same dataset were averaged to serve as a metric of machine readability of the corresponding dataset. The analysis results are listed in Table 2. For SIFT, similarity was measured by Euclidean distance. The SIFT feature distances of synthetic photographs provided by DS-7-100, DS-7-200, and DS-7-400 were significantly lower than those of BIM-rendered images, indicating a higher likelihood of identifying correct correspondences between synthesized and real photographs. For the other four features, similarity was measured by cosine similarity. Table 2 indicates the higher similarity among samples from the three synthetic photograph datasets than the original

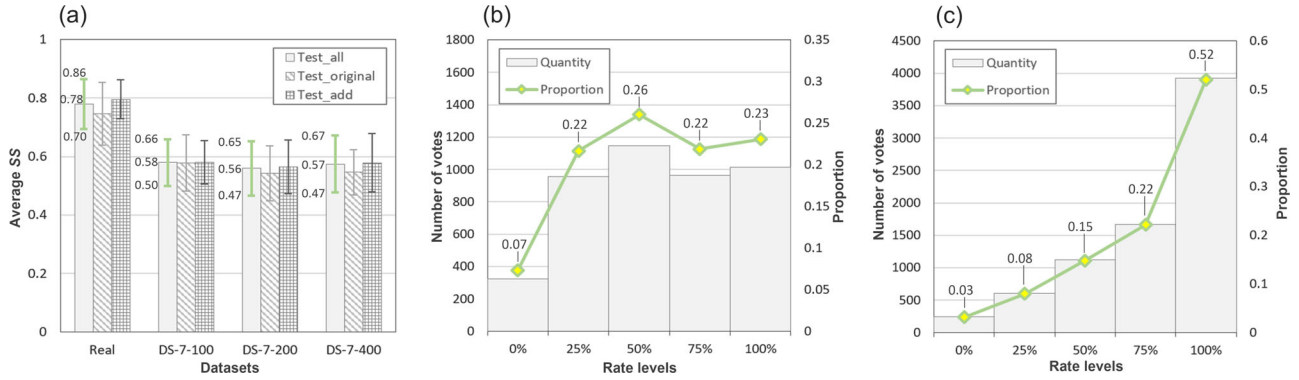

FIGURE 9 Results of the visual Turing test: (a) Average similarity scores of all test samples in different datasets; vote distribution over different similarity levels on (b) DS-7-100 and (c) the collection of 151 real samples

TABLE 2 Improvement of machine readability of synthesized photographs by five typical computer algorithms

Algorithm	Metrics	Dataset				
		BIM	Syn-DS-7-100	Syn-DS-7-200	Syn-DS-7-400	Real
SIFT	Euclidean distance	136.9	95.0	93.9	80.0	71.8
HOG	Cosine similarity	0.810	0.841	0.837	0.838	0.855
EHD	Cosine similarity	0.912	0.931	0.930	0.932	0.945
LBP	Cosine similarity	0.970	0.996	0.995	0.996	0.999
RGB	Cosine similarity	0.700	0.890	0.900	0.903	0.939

Note: "Syn-DS-7-xxx" stands for synthesized test samples from the respective datasets.
Abbreviations: BIM, building information model; EHD, edge histogram descriptor; HOG, histogram of oriented gradient; LBP, local binary pattern; RGB, red green blue; SIFT, scale-invariant feature transform.

BIM-rendered images. The quantitative analysis is consistent with the examples given by Figure 10. After style transfer, the synthetic versions of the original BIM-rendered images present a realistic texture similar to that of photographs, leading to more robust SIFT correspondence detection performance and better consistency of HOG gradient orientation. This finding demonstrates that style transfer based on CycleGAN can significantly facilitate the exploitation of cross-domain visual information by computer algorithms, thus paving the way for retrieving spatial coordinates from BIM to enable indoor localization.

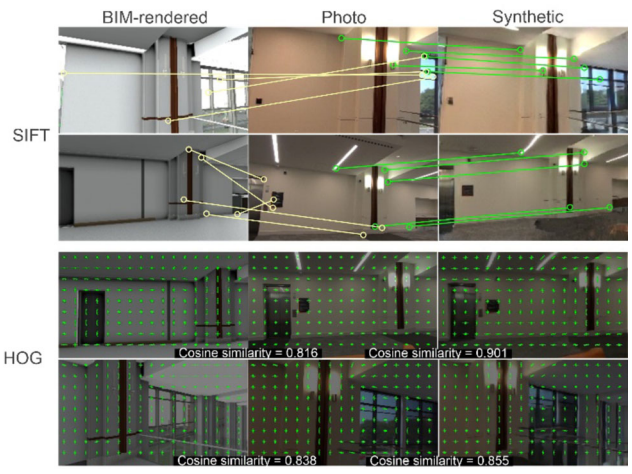

FIGURE 10 Examples showing how synthetic images can be better exploited by algorithms such as SIFT and HOG

4.3 | Analysis of image retrieval results

With the trained CycleGAN models, all BIM-rendered images (including both training and test samples) in the respective datasets in Table 1 were transformed into images with a photorealistic texture, which then served as a basis for constructing the visual-spatial hybrid databases.

When retrieving similar matches for an input query image from the databases, different global image features (e.g., HOG, EHD, LBP, and RGB) can be used for similarity comparison. The number of retrieved images, N_{img}, is undetermined. In addition, different datasets provide different numbers of synthetic images with varying levels of fidelity, which can lead to different retrieval performance. In this section, the influence of different combinations of the above factors (i.e., features, N_{img}, and datasets) was analyzed.

Table 3 lists the accuracy and computation time when retrieving N_{img} similar matches of the 13 indoor

TABLE 3 Image retrieval performance from different synthetic datasets with different features and N_{img}

Features	N_{img}	Syn-DS-7-100		Syn-DS-7-200		Syn-DS-7-400	
		Accuracy (%)	Time (s/per)	Accuracy (%)	Time (s/per)	Accuracy (%)	Time (s/per)
HOG	2	53.8	0.070	69.2	0.126	84.6	0.217
	5	44.6		63.1		83.1	
	10	38.5		55.4		73.1	
EHD	2	61.5	0.177	73.1	0.160	100.0	0.187
	5	52.3		73.8		89.2	
	10	50.8		67.7		82.3	
LBP	2	0.0	0.153	3.8	0.140	7.7	0.150
	5	3.1		12.3		7.7	
	10	3.1		12.3		7.7	
RGB	2	38.5	0.032	53.8	0.040	57.7	0.064
	5	32.3		52.3		58.5	
	10	31.5		50.0		50.0	

Notes: 1. "Syn-DS-7-xxx" stands for the respective datasets of synthesized samples. 2. "s/per" stands for "second/per image."
Abbreviations: EHD, edge histogram descriptor; HOG, histogram of oriented gradient; LBP, local binary pattern; RGB, red green blue.

TABLE 4 Intrinsic and lens distortion parameters of the used camera

Items	Intrinsic parameters	Distortion parameters
Values	f/dx: 1470.4; f/dy: 1458.7; u_0: 964.4; v_0: 526.8	–0.0679; 0.2366; 0; 0

photographs (yellow dots in Figure 7) with varying features from different datasets. The table reveals a general trend that retrieval accuracy increases with a decrease in N_{img} and growth in dataset size. With an identical N_{img} (e.g., 2) and dataset (e.g., Syn-DS-7-400), retrieval based on feature EHD achieved the highest accuracy, followed by HOG and RGB. Retrieval based on LBP never exceeded an accuracy of 20%, indicating that this feature can hardly retrieve meaningful results. Figure 11 shows the retrieval performance resulting from the four features with examples taken at #L-7 and #L-9. Incorrectly retrieved images bring in outliers in later SIFT correspondence detection, which can then undermine indoor localization performance. Hence, when computation complexity is within an acceptable range, the retrieval accuracy should be as high as possible. The highest accuracy was observed when N_{img} was 2 and EHD and Syn-DS-7-400 were used. The computation time under these settings was 0.187 s per image, which is adequately efficient for real-time applications. As a result, the settings $N_{img} = 2$, EHD, and Syn-DS-7-400 were adopted for indoor localization evaluation.

4.4 | Evaluation of localization performance

The 13 indoor photographs used for localization performance evaluation were collected using a SONY HDR-CX760V camera. The intrinsic parameters of the camera were obtained using Zhang's camera calibration method (Zhang, 2000), as shown in Table 4.

The performance of camera pose estimation was evaluated in terms of localization and orientation errors. The former refers to the Euclidean distance (m) between the estimated camera location and its ground-truth location; the latter is defined as the angle (°) between two vectors representing the estimated and ground-truth look-at direction of the camera. From the top two images retrieved from Syn-DS-7-400, SIFT correspondences can be detected to obtain multiple pairs of 2D–3D coordinate correspondences, based on which the camera pose of the query image can be estimated. However, the number of extracted SIFT correspondences per image, N_{sift}, can influence performance: a larger N_{sift} might bring more outliers and increase computation complexity, whereas a small N_{sift} might provide too few correct correspondences to ensure estimation accuracy.

Figure 12 shows the camera pose estimation performance of all 13 indoor photographs under different N_{sift} values (5, 10, 15, and 20). As can be observed from Figures 12a and b, most samples (dashes in gray) were distributed below localization and orientation errors of 2.5 m and 15°. Average accuracy fluctuates with changes in N_{sift}, and the highest accuracy was observed when N_{sift} was set to 10. The computation time gradually grows with the increase of N_{sift}, but the highest (0.66 s) is still

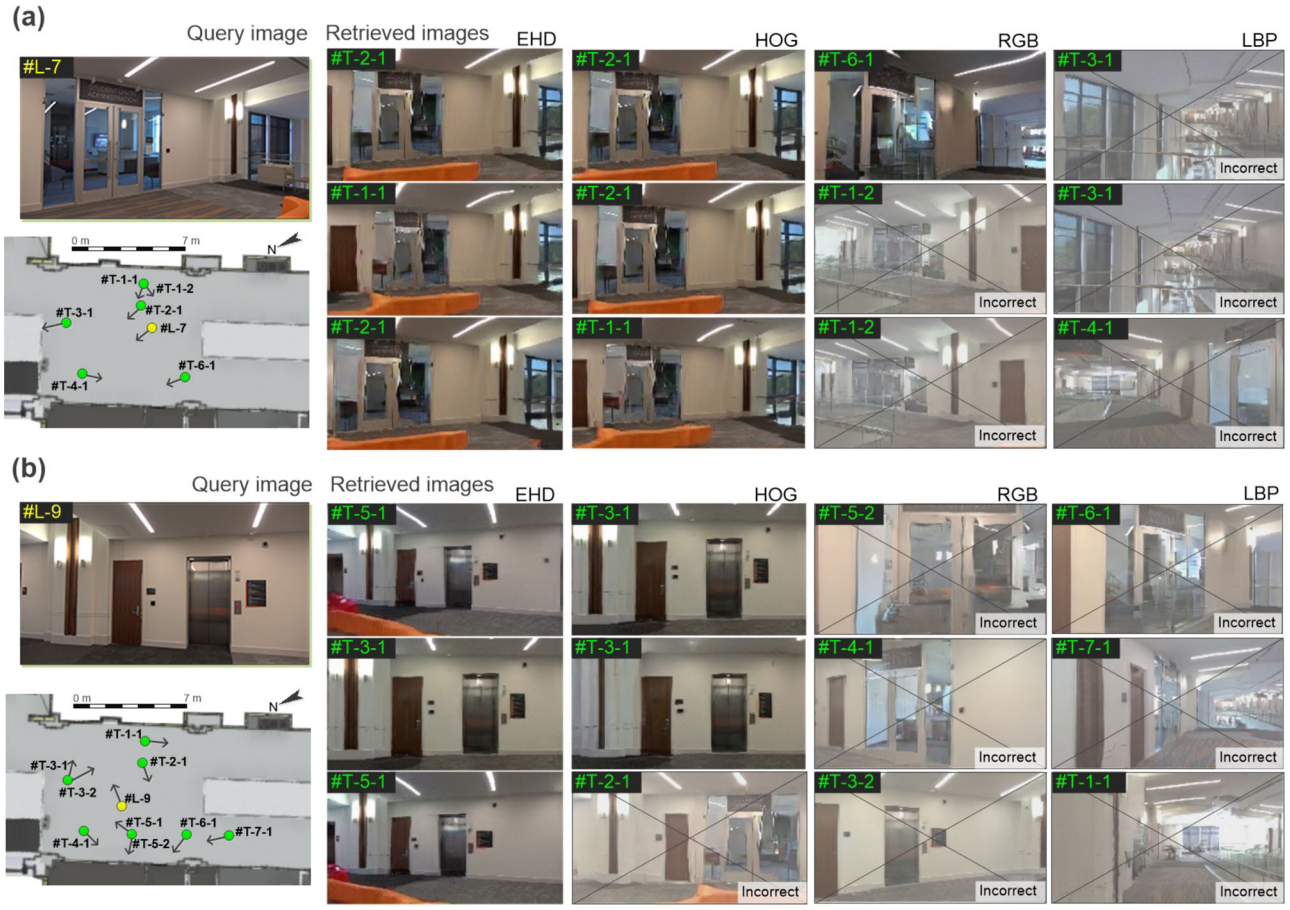

FIGURE 11 Top three matches from Syn-DS-7-400 for query images captured at (a) #L-7 and (b) #L-9

within an acceptable range. Figure 13 shows the estimated camera poses and their deviations from the ground truth when N_{sift} is 10. In Figure 13, dots with different colors are 3D point clouds corresponding to the SIFT features used to estimate the respective camera pose. Table 5 lists the details of estimation errors when $N_{\text{sift}} = 10$. The average localization and orientation errors are 1.38 m and 10.1°, respectively.

5 | DISCUSSION

5.1 | Mapping-free indoor localization with high precision

The proposed approach achieved camera localization with an average accuracy of 1.38 m and 10.1° in a 12.0 m × 9.4 m (112.8 m^2) indoor space. With a newly captured photograph, the strategy can yield its corresponding camera pose in less than 1 s (0.187 s for image retrieval and another 0.52 s for P*n*P problem solving). Compared with existing vision-based indoor localization methods, the advantages of this approach are as follows. First, it does not require a tedious and comprehensive pre-mapping of indoor environments to obtain visual assets (in the form of either photographs or point clouds) with known positions. Rather, such visual-spatial information is directly retrieved from a readily available 3D model after style transfer. Thus, the laborious, costly, and time-consuming pre-mapping process can be omitted, making the proposed approach scalable to a wider range of applications.

Second, it provides an analytical solution to estimate the 6DoF camera pose using the visual-spatial information retrieved from BIMs. Localization based merely on image retrieval is highly influenced by viewpoint change (Piasco et al., 2018) and dataset scale and might fail to provide camera posture. Regression-based approaches are subject to the black box nature of deep learning (Rudin, 2019). This study directly extracted 3D spatial coordinates from BIMs via SIFT correspondences between the query image and photorealistic BIM renderings after style transfer. With the retrieved 2D–3D correspondences, the camera pose is estimated by solving a classical P*n*P problem in photogrammetry theory.

Third, it can achieve near-real-time (less than 1 s) indoor localization with high precision. Its performance

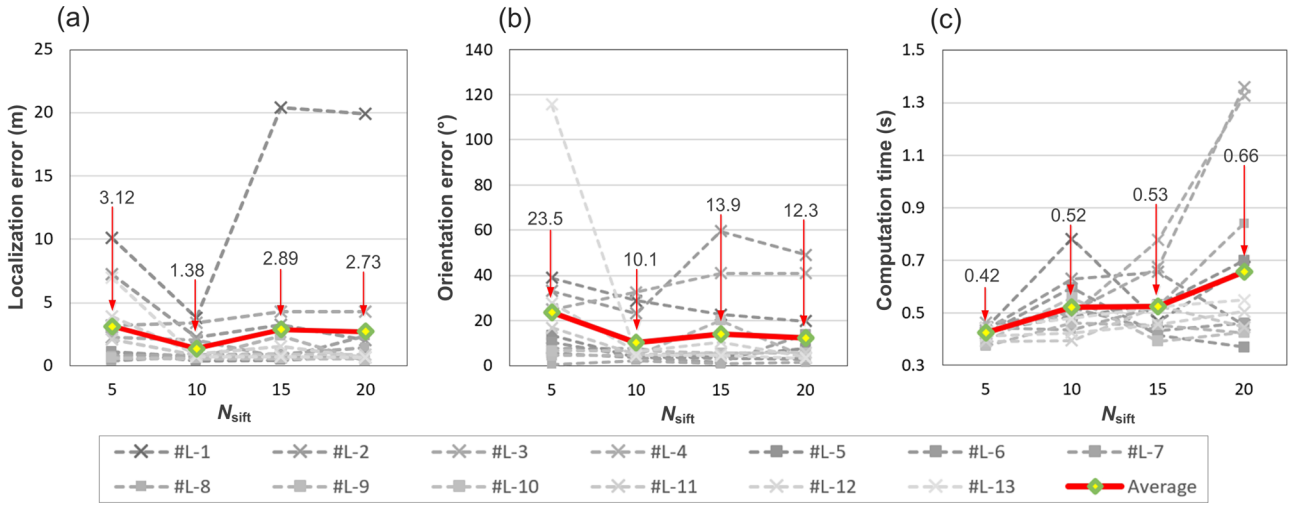

FIGURE 12 Camera pose estimation performance under different N_{sift} values: (a) Localization error, (b) orientation error, and (c) computation time

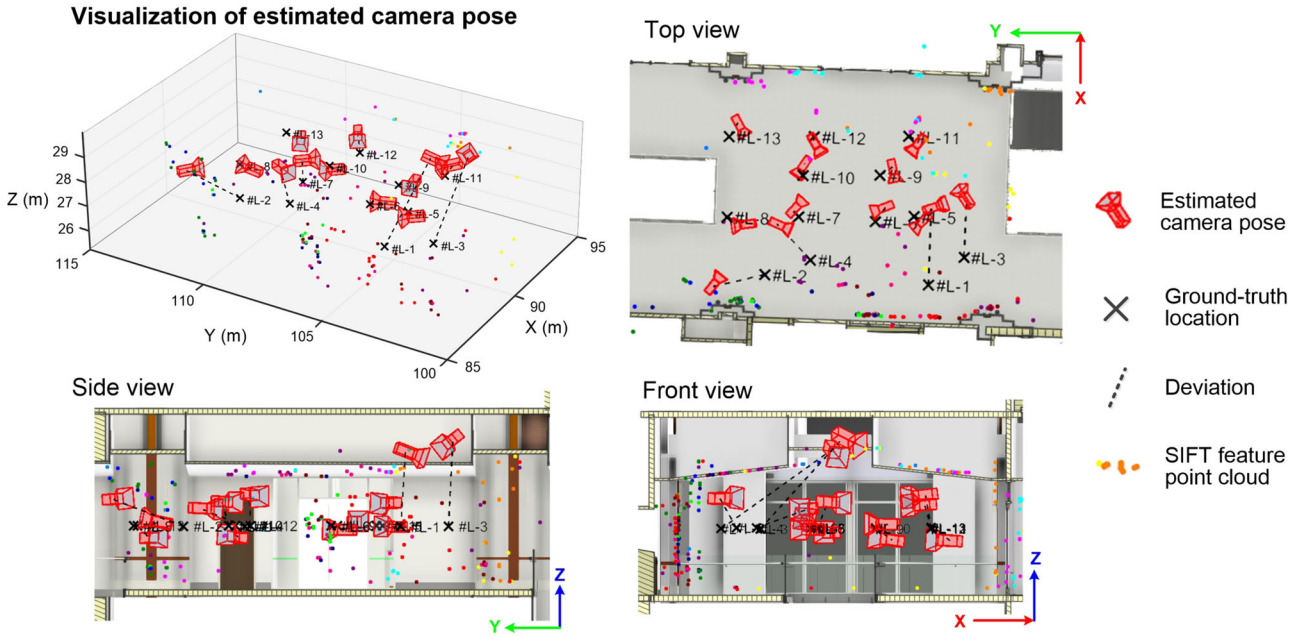

FIGURE 13 Camera poses estimated using the proposed approach ($N_{sift} = 10$)

exceeds or is comparable to state-of-the-art BIM-enabled indoor localization algorithms (Acharya, Khoshelham, et al., 2019; Acharya et al., 2020; Zhao et al., 2020), the localization and orientation errors of which range from 1.6 m to 2.0 m and 7° to 11°, respectively.

The exploitation of 3D models for localization is hindered by a cross-domain gap between virtual and real visual content. To address the cross-domain gap, an image-to-image translation technique (CycleGAN) was adopted to transfer a photorealistic texture style onto BIM renderings. Our study, through comprehensive quantitative analysis, demonstrates that the synthetic photographs generated by CycleGAN can significantly improve SIFT correspondence detection performance and maintain better similarity with authentic photographs regarding global features such as HOG and EHD. This finding provides a new direction for exploiting the visual-spatial information from virtual 3D models for indoor localization.

In contrast to the current vision-based approaches that rely on extensive mapping, our methodology only needs to collect photographs from a small cluster of points in the indoor space (e.g., seven points in an area

TABLE 5 Details of pose estimation errors ($N_{sift} = 10$)

ID	Localization error (m)				Orientation error (°)
	ΔX	ΔY	ΔZ	Dist.[a]	
#L-1	3.07	−0.12	2.20	3.78	28.7
#L-2	−0.45	2.06	0.80	2.26	23.0
#L-3	2.30	−0.12	2.54	3.43	32.5
#L-4	1.45	1.32	0.53	2.03	7.8
#L-5	−0.13	−0.40	−0.02	0.42	3.5
#L-6	0.59	−0.42	−0.11	0.73	3.8
#L-7	0.29	0.20	0.72	0.80	3.3
#L-8	−0.27	−0.83	0.11	0.88	2.2
#L-9	0.34	−0.49	0.12	0.61	4.4
#L-10	0.47	−0.07	−0.33	0.58	6.5
#L-11	−0.19	−0.44	0.67	0.82	4.9
#L-12	−0.16	0.01	0.91	0.93	6.2
#L-13	0.53	−0.34	−0.36	0.72	4.5
Mean				1.38	10.1

[a]Euclidean distance.

of 112.8 m^2 in Figure 7) without necessarily knowing the positions of those photographs. The effort required for in-situ data collection is, therefore, substantially reduced. In addition, a style-transfer BIM has the potential to enable more applications than indoor localization. It can serve as a source for AR device registration, thus saving effort in manually deploying markers. The synthetic BIM renderings can also provide training samples with annotated labels for facility segmentation, which has been demonstrated in a pilot study (Hong et al., 2020). In Chen (2020), CycleGAN demonstrated its potential to render BIMs with photorealistic textures for the production of indoor panoramas. As common enablers of the above application scenarios, the effort required to prepare the style-transfer renderings is, thus, justified because they can be reused for different purposes to add more values.

5.2 | Robustness to condition variation

This section discusses the potential influences of lighting and weather conditions and indoor layout deviation, as well as the generalizability of the proposed approach to such factors.

Illumination variation, which is caused by changing lighting or weather conditions, is an important influencing factor in performing computer vision tasks. In our experiments, the indoor photographs used for CycleGAN training and camera pose estimation were collected in the morning and afternoon, respectively. This difference was reflected in part of the collected data. As shown in Figure 14, in contrast to images captured in the morning,

FIGURE 14 Photographs taken under different lighting conditions

photographs captured in the afternoon showed light beams cast from the west onto the floor. Because this issue was not experienced in the training data, it might influence performance. Figure 15 compares the results of SIFT correspondence detection and pose estimation between photographs taken at #L-12 and #L-13, one with cast light issues and another without. The abnormal lighting condition confused the computer, leading to five incorrect SIFT correspondences detected from two retrieved images. However, because of the existence of many other correct correspondences, RANSAC successfully excluded those outliers in the pose estimation stage, resulting in precision comparable to (if not better than) that for the image with normal light conditions.

Another important factor is the layout variation in indoor environments. Such variation is typically due to a deviation between the as-planned BIM and the actual as-built condition. For example, the BIM used in our study does not contain sign boards (Figure 16a, left) or furniture (Figure 16a, middle) and has a special painted pattern on the wall (Figure 16a, right). However, CycleGAN

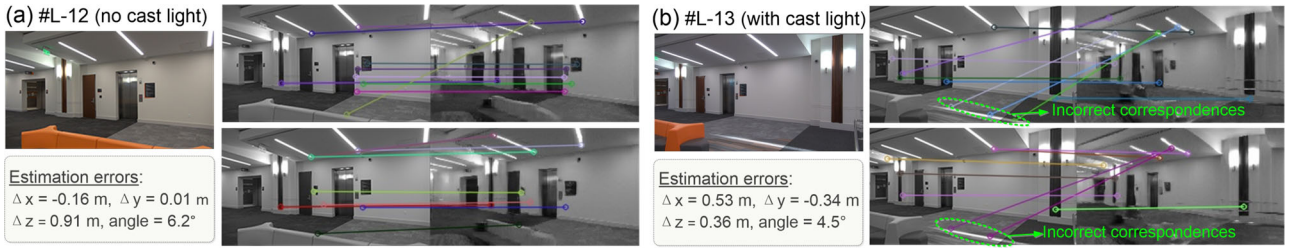

FIGURE 15 Robustness of the proposed approach to illumination variation

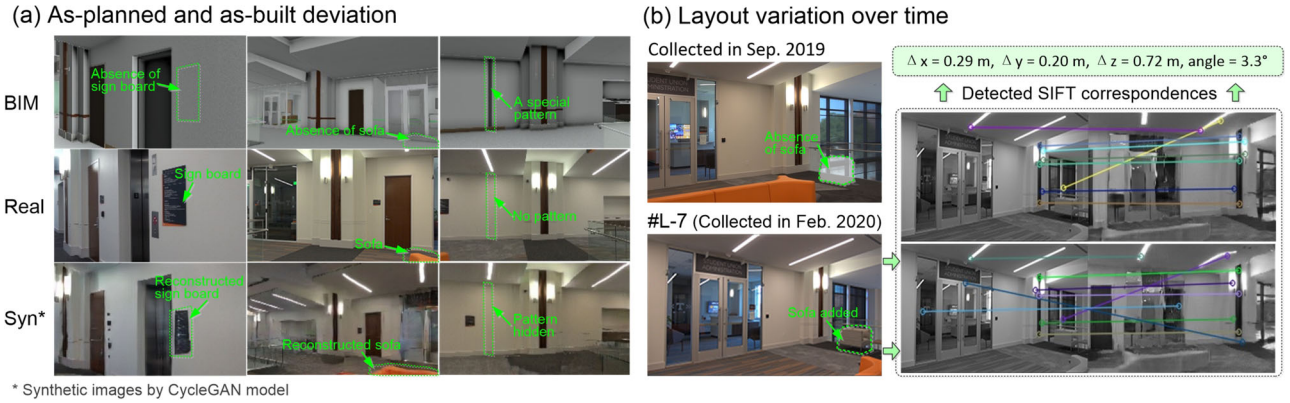

FIGURE 16 Robustness of the proposed approach to layout variation

learned the mapping between these differences and successfully reconstructed synthetic images with the sign boards and furniture added and the wall pattern hidden (last row in Figure 16a). The deviation between the as-planned and as-built was, therefore, compensated for with the style-transfer techniques. The indoor layout variation also occurs between different time nodes. As shown in Figure 16b, a sofa was newly added in the corner when the query image at #L-7 was taken. This change did not undermine the localization performance. Instead, many SIFT correspondences were detected, primarily because SIFT features were extracted on a local basis and, thus, are robust to occlusion. The newly added furniture might occlude features hidden behind it, but it cannot prevent distinct features on other parts of the image from being detected. As a result, the robustness and reliability of the estimated pose were assured (errors within 1 m and 5°), regardless of the occlusion and layout changes occasioned by newly added furniture.

5.3 | Limitations and future work

Further research and developments are required to address the following limitations

1. Despite overall high performance, the precision of estimated camera poses at #L-1 and #L-3 is not satisfactory. As shown in Figure 17, this deviation might be related to a failure of the CycleGAN model to synthesize the relatively complex scenes depicted in the photographs in question. For #L-1, the photo involves a large glass curtain wall, which makes scenery outside

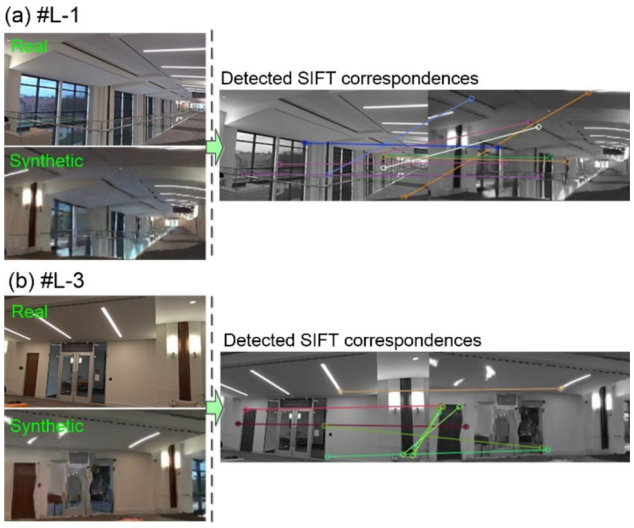

FIGURE 17 Erroneous cases caused by insufficient fidelity in synthesizing complex scenes

the building visible. In its synthetic counterpart, although CycleGAN performed quite well in generating vivid content in other parts of the scene (e.g., ceiling, floor, wall lamp), it failed to plausibly reconstruct the glass curtain wall, resulting in large blurry areas. For #L-3, the glass entrance in the photograph presents a complex texture, with both reflections of light outside the room and items inside the room cast through. The complexity of such glass material is difficult for the model to mimic, and it generated a correspondence with distortion and blurring. The insufficient fidelity of synthetic images then led to many incorrect SIFT correspondences, which were beyond RANSAC's compensation capability for accurate pose estimation. To address this limitation, future research should seek to further improve the fidelity of synthesized images by testing different model structures and emerging supervised learning techniques (Ahmadlou & Adeli, 2010; Alam et al., 2020; Pereira et al., 2020; Rafiei & Adeli, 2017). In real-life applications, reasonable user interfaces can be developed that direct users not to use photographs with complex patterns such as glass doors or windows for localization purposes.

2. Similar to other visual approaches, the proposed indoor localization method is subject to the impact of uniform design (Winter et al., 2019). For example, different floor levels or rooms within the same building usually have a similar design style or even nearly identical layouts, which makes it challenging to identify their differences using only a visual perspective. To filter out ambiguities, user interaction or sensors from other modalities (e.g., barometers, accelerometers, gyroscopes) can be incorporated in future development to provide more information on the subject's location at different scales and from different angles. Indoor scene understanding could also be incorporated to improve performance. With techniques such as object detection and semantic segmentation, it is viable to identify which equipment, facilities, or furniture appear in a photograph and to understand their spatial relationships. By comparing such information with the prior knowledge of indoor layout extracted from BIMs, it is possible to approximately infer where a photograph was taken. This approach could significantly narrow the search range for subsequent image retrieval, thus improving both accuracy and efficiency.

6 | CONCLUSIONS

Cost-effective and reliable indoor localization solutions are important for location-based services in built environments. Existing vision-based approaches require dedicated effort and costly equipment to pre-map the indoor scene of interest, making them difficult to scale up for many applications. Exploiting readily available 3D models such as BIMs, this research proposed a mapping-free visual methodology for estimating 6DoF camera poses based on photogrammetry. CycleGAN was used to transfer georeferenced BIM renderings to photorealistic images, which narrows the cross-domain gap between virtual and real content to enable effective image retrieval and feature correspondence detection. The detected feature correspondences enable the retrieval of the spatial coordinates from BIMs to solve the P*n*P problem and estimate camera poses from the images. The results of experiments conducted in an indoor space demonstrated the effectiveness and efficiency of the proposed method. The main findings are summarized as follows:

1. The synthetic photographs generated by CycleGAN can be exploited to better detect feature correspondence with camera-captured images, thus enabling more accurate retrieval of spatial information stored in BIMs for indoor localization.
2. Among the four investigated features, EHD demonstrated the highest accuracy in image retrieval from a database of synthesized photographs, and the retrieval accuracy increased with the database size.
3. The proposed approach achieved state-of-the-art performance (localization and orientation errors of 1.38 m and 10.1°, respectively) for camera pose estimation in indoor environments.

ACKNOWLEDGMENT

This research was supported by the U.S. National Science Foundation (NSF) through grant #1850008 and #2038967. The authors would like to thank the anonymous reviewers for their valuable comments, and relevant PhD students from the University of Tennessee, Knoxville for their help and support in data collection.

REFERENCES

Acharya, D., Khoshelham, K. & Winter, S. (2019). BIM-Posenet: indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *150*, 245–258.

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

Acharya, D., Singha Roy, S., Khoshelham, K. & Winter, S. (2020). A recurrent deep network for estimating the pose of real indoor images from synthetic image sequences. *Sensors*, *20*(19), 5492.

Ahmadlou, M. & Adeli, H. (2010). Enhanced probabilistic neural network with local decision circles: a robust classifier. *Integrated Computer-Aided Engineering*, *17*(3), 197–210.

Alam, K. M. R., Siddique, N. & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*(12), 8675–8690.

Asadi, K., Ramshankar, H., Noghabaei, M. & Han, K. (2019). Real-time image localization and registration with BIM using perspective alignment for indoor monitoring of construction. *Journal of Computing in Civil Engineering*, *33*(5), 04019031.

Autodesk. (2020a). Forge Viewer Documentation: Autodesk. Viewing.Viewer3d.Clienttoworld. https://forge.autodesk.com/en/docs/viewer/v2/reference/javascript/viewer3d/#clienttoworld-clientx-clienty-ignoretransparent

Autodesk. (2020b). Forge Viewer Documentation: Autodesk. Viewing.Viewer3d.Getscreenshot. https://forge.autodesk.com/en/docs/viewer/v7/reference/Viewing/Viewer3D/#getscreenshot-w-h-cb-overlayrenderer

Bang, S., Baek, F., Park, S., Kim, W. & Kim, H. (2020). Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques. *Automation in Construction*, *115*, 103198.

Brownlee, J. (2019). *A gentle introduction to cyclegan for image translation*. https://machinelearningmastery.com/what-is-cyclegan/

Buades, A., Coll, B. & Morel, J.-M. (2005). A non-local algorithm for image denoising. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, pp. 60–65.

Chen, J. (2020). *Indoor panorama synthesized in real time with cyclegan*. https://www.youtube.com/watch?v=n5WcJ-DLB9Y-feature=youtu.be

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, pp. 886–893.

Davison. (2003). Real-time simultaneous localisation and mapping with a single camera. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, Nice, France, pp. 1403–1410.

Dong, J., Xiao, Y., Noreikis, M., Ou, Z. & Ylä-Jääski, A. (2015). Imoon: Using smartphones for image-based indoor navigation. *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, Seoul, South Korea, pp. 85–97.

Ellard, C. (2009). *You are here: why we can find our way to the moon, but get lost in the mall*. Anchor.

Fang, Q., Li, H., Luo, X., Li, C. & An, W. (2020). A sematic and prior-knowledge-aided monocular localization method for construction-related entities. *Computer-Aided Civil and Infrastructure Engineering*, *35*(9), 979–996.

Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T. & Freeman, W. T. (2006). Removing camera shake from a single photograph. *ACM Transactions on Graphics*, *25*(3), 787–794.

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Gao, Y., Kong, B. & Mosalam, K. M. (2019). Deep leaf-bootstrapping generative adversarial network for structural image data augmentation. *Computer-Aided Civil and Infrastructure Engineering*, *34*(9), 755–773.

Gatys, L. A., Ecker, A. S. & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA (pp. 2414–2423).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems- Volume 2,* MIT Press, Montreal, Canada, (pp. 2672–2680*)*.

Ha, I., Kim, H., Park, S. & Kim, H. (2018). Image retrieval using BIM and features from pretrained VGG network for indoor localization. *Building and Environment*, *140*, 23–31.

Hamledari, H., McCabe, B. & Davari, S. (2017). Automated computer vision-based detection of components of under-construction indoor partitions. *Automation in Construction*, *74*, 78–94.

Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B. & Salesin, D. H. (2001). Image analogies. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 327–340). Association for Computing Machinery.

Hong, Y., Park, S. & Kim, H. (2020). Synthetic data generation for indoor scene understanding using BIM. *Proceedings of the 37th International Symposium on Automation and Robotics in Construction (ISARC)*, (pp. 334–338). International Association for Automation and Robotics in Construction (IAARC)

Isola, P., Zhu, J., Zhou, T. & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5967–5976.

Johnson, J., Alahi, A. & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision* (pp. 694–711). Springer.

Kendall, A., Grimes, M. & Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-Dof camera relocalization. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, (pp. 2938–2946).

Lee, H. (2017). *Gan Lecture 2 (2017): Cyclegan*. https://www.youtube.com/watch?v=9N_uOIPghuo

Lenjani, A., Yeum, C. M., Dyke, S. & Bilionis, I. (2020). Automated building image extraction from 360° panoramas for postdisaster evaluation. *Computer-Aided Civil and Infrastructure Engineering*, *35*(3), 241–257.

Li, R., Yuan, Y., Zhang, W. & Yuan, Y. (2018). Unified vision-based methodology for simultaneous concrete defect detection and geolocalization. *Computer-Aided Civil and Infrastructure Engineering*, *33*(7), 527–544.

Liang, J. Z., Corso, N., Turner, E. & Zakhor, A. (2015). Image-based positioning of mobile devices in indoor environments. In Choi, J. & Friedland, G. (Eds.), *Multimodal location estimation of videos and images* (pp. 85–99). Springer International Publishing.

Liang, X. (2019). Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with bayesian optimization. *Computer-Aided Civil and Infrastructure Engineering*, *34*(5), 415–430.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, G. & Kambhamettu, C. (2014). Image-based indoor localization system based on 3D SFM model. *IS&T/SPIE Electronic Imaging*, SPIE, 90250H, San Francisco, CA, USA.

Maeda, H., Kashiyama, T., Sekimoto, Y., Seto, T. & Omata, H. (2020). Generative adversarial network for road damage detection. *Computer-Aided Civil and Infrastructure Engineering*, *36*(1), 47–60.

Murillo, A. C., Guerrero, J. J. & Sagues, C. (2007). Surf features for efficient robot localization with omnidirectional images. *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, Rome, Italy (3901–3907).

Nister, D., Naroditsky, O. & Bergen, J. (2004). Visual odometry, *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Washington, DC, USA.

OpenCV. (2021a). *BFMatcher class reference*. https://docs.opencv.org/master/d3/da1/classcv_1_1BFMatcher.html

OpenCV. (2021b). *Camera calibration and 3D reconstruction: SolvePnPRansac*. https://docs.opencv.org/master/d9/d0c/group__calib3d.html#ga50620f0e26e02caa2e9adc07b5fbf24e

Palmarini, R., Erkoyuncu, J. A., Roy, R. & Torabmostaedi, H. (2018). A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, *49*, 215–228.

Pan, X. & Yang, T. Y. (2020). Postdisaster image-based damage detection and repair cost estimation of reinforced concrete buildings using dual convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, *35*(5), 495–510.

Park, J., Cai, H. B. & Perissin, D. (2018). Bringing information to the field: automated photo registration and 4d BIM. *Journal of Computing in Civil Engineering*, *32*(2), 04017084.

Park, S., Baek, F., Sohn, J. & Kim, H. (2021). Computer vision-based estimation of flood depth in flooded-vehicle images. *Journal of Computing in Civil Engineering*, *35*(2), 04020072.

Pereira, D. R., Piteri, M. A., Souza, A. N., Papa, J. P. & Adeli, H. (2020). FEMa: a finite element machine for fast learning. *Neural Computing and Applications*, *32*(10), 6393–6404.

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

Pietikäinen, M., Ojala, T. & Xu, Z. (2000). Rotation-invariant texture classification using feature distributions. *Pattern Recognition*, *33*(1), 43–52.

Pouyanfar, S., Tao, Y., Sadiq, S., Tian, H., Tu, Y., Wang, T., Chen, S. & Shyu, M. (2019). Unconstrained flood event detection using adversarial data augmentation. *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, pp. 155–159.

Qian, G., Sural, S., Gu, Y. & Pramanik, S. (2004). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. *Proceedings of the 2004 ACM Symposium on Applied Computing*, Nicosia, Cyprus, pp. 1232–1237.

Rafiei, M. H. & Adeli, H. (2017). A new neural dynamic classification algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 3074–3083.

Ravi, N., Shankar, P., Frankel, A., Elgammal, A. & Iftode, L. (2006). Indoor localization using camera phones, *Seventh IEEE Workshop on Mobile Computing Systems & Applications (WMCSA'06 Supplement)*, Orcas Island, WA, USA, p. 49.

Rituerto, A., Puig, L. & Guerrero, J. J. Visual slam with an omnidirectional camera. *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, pp. 348–351.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. & Chen, X. (2016). Improved techniques for training GANs. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Barcelona, Spain, (pp. 2234–2242).

Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A. & Fitzgibbon, A. (2013). Scene coordinate regression forests for camera relocalization in RGB-D images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, pp. 2930–2937.

Shrivastava, A., Malisiewicz, T., Gupta, A. & Efros, A. A. (2011). Data-driven visual similarity for cross-domain image matching, *ACM Transactions on Graphics*, *30*(6), 154.

Svärm, L., Enqvist, O., Kahl, F. & Oskarsson, M. (2017). City-scale localization for cameras with known vertical direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(7), 1455–1461.

Wang, Q., Cheng, J. C. P. & Sohn, H. (2017). Automated estimation of reinforced precast concrete rebar positions using colored laser scan data. *Computer-Aided Civil and Infrastructure Engineering*, *32*(9), 787–802.

Winter, S., Tomko, M., Vasardani, M., Richter, K. F., Khoshelham, K. & Kalantari, M. (2019). Infrastructure-independent indoor localization and navigation. *ACM Computing Surveys*, *52*(3) 1–24.

Won, C. S., Park, D. K. & Park, S. J. J. E. j. (2002). Efficient use of MPEG-7 edge histogram descriptor. *ETRI Journal*, *24*(1), 23–30.

Yousif, K., Bab-Hadiashar, A. & Hoseinnezhad, R. (2015). An overview to visual odometry and visual slam: applications to mobile robotics. *Intelligent Industrial Systems*, *1*(4), 289–311.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330–1334.

Zhao, H., Acharya, D., Tomko, M. & Khoshelham, K. (2020). Indoor lidar relocalization based on deep learning using a 3D model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B1*, 541–547.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2223–2232.

How to cite this article: Chen J, Li S, Liu D, & Lu W. (2021). Indoor camera pose estimation via style-transfer 3D models. *Comput Aided Civ Inf*, , 1–19. https://doi.org/10.1111/mice.12714