DOI: 10.1111/mice.12795

RESEARCH ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Precise indoor localization with 3D facility scan data

Jiahao Xia | **Jie Gong**

Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, New Jersey, USA

Correspondence
Jie Gong, Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, NJ, 08854, USA.
Email: jg931@soe.rutgers.edu

Funding information
US National Science Foundation, Grant/Award Numbers: 1827505 and 1737533

Abstract

Visual indoor localization for smart indoor services is a growing field of interest as cameras are now ubiquitously equipped on smartphones. In this study, a hierarchical indoor localization algorithm is designed and validated based on 3D facility scan data, which are originally collected for facility modeling purposes. The study has shown promising results in indoor localization. The study also demonstrated a scalable approach to generate high-quality images with reference poses from laser scan data, opening doors to generate labeled images to train end-to-end pose regression model (i.e., PoseNet). In this regard, this study is the first attempt to leverage facility scan data, which are commonly collected for Building Information Modeling (BIM) purpose, for indoor localization. As more facilities are documented with laser scanners, our algorithm can unlock additional values of collected data for intelligent applications.

1 | INTRODUCTION

According to the National Human Activity Pattern Survey, people normally spend over 87% of their daily lives indoors (Klepeis et al., 2001). This estimate is likely much higher during this pandemic crisis. Indoor localization, which refers to the process of obtaining the poses of a device or a user under a given coordinate system (Zafari et al., 2019), is the foundation of many smart indoor services. Accurately localizing a user or device in an indoor setting has a prominent place in the health, industry, disaster and building management, and surveillance sector (Asimakopoulou & Bessis, 2011; Borrion et al., 2012; Zelenkauskaite et al., 2012). To date, indoor localization has contributed significantly to a wide array of applications including virtual reality and augmented reality, robotics, autonomous driving (Wu et al., 2018), Internet of Things (Atzori et al., 2010), smart buildings (Snoonian, 2003), and machine type communication (Taleb & Kunz, 2012).

Compared with outdoor localization, indoor localization has its unique challenges since most indoor scenes are GPS-denied environments and have to rely on techniques such as WiFi (Chen et al., 2015), ultra-wideband (Alarifi et al., 2016), radio frequency identification (Stella et al., 2012), Bluetooth (Kriz et al., 2016), ZigBee (Niu et al., 2015), ultrasonic (Hazas & Hopper, 2006), or magnetic fields (Subbu et al., 2013). Most of these techniques are built upon physics centered mechanisms in angle of arrival, time of flight (ToF), return TOF, or received strength of specific signals (Davidson & Piché, 2016; Zafari et al., 2019). A common drawback of these systems is the necessity of deploying dedicated infrastructure, making them costly or disruptive to ongoing building operations. As an alternative, visual localization does not require any infrastructure change and the only required sensors are cameras that are ubiquitously equipped on nearly all smartphones or 3D ranging sensors, which are increasingly available on many mobile devices, such as the recent Apple devices.

Given images or videos taken by a camera, visual localization can estimate 6 degree-of-freedom poses of the camera, that is, its positions and orientations. Visual localization can be categorized as image-based or structure-based methods (Sarlin et al., 2019). The key technique adopted in the image-based localization method is image retrieval (Arandjelović et al., 2016; Torii et al., 2015; Weyand et al., 2016). Recent studies on this field have shown promising robustness and efficiency, but they can only estimate approximate positions and orientations of a camera given the fact that the images in the database are often discretely distributed in space. The image retrieval-based methods try to measure the similarity between images through some similarity metrics (e.g., cosine similarity index (CSI), L1 norm distances) based on features extracted by deep neural networks (DNNs; Gálvez-López & Tardos, 2012; Lu et al., 2017). The estimated pose accuracy can be improved if the images in the database are sampled at denser space points, but this would inevitably increase the data collection cost. Recently, the development of machine learning techniques, especially DNN, provides improved solutions for many classic tasks (i.e., object detection and tracking, variables estimation) in various fields (Arabi et al., 2020; Luo et al., 2020; Ni et al., 2020; Oh et al., 2017; Rafiei & Adeli, 2016, 2018; Rodriguez Lera et al., 2019; Vera-Olmos et al., 2019; Yang et al., 2019). Nowadays, high-quality synthesized 2D images can be obtained from 3D point cloud data (Dai et al., 2020; Pittaluga et al., 2019), and they can be used to establish the image database for image-retrieval based localization. Structure-based methods are another widely accepted indoor localization solution. 3D structure model, which usually is sparse 3D point cloud reconstructed through structure from motion (SfM; Schonberger & Frahm, 2016; Snavely et al., 2008) or simultaneous localization and mapping (SLAM; Davison et al., 2007), is an important component in these algorithms. In most cases, this reconstructed point cloud is associated with feature attributions extracted from the images that are used to construct 3D model through triangulation. The features used in this process can be classified into two categories: handcrafted features and learned features. Designing handcrafted features is an important research field in the last decade, and the classical handcrafted features include scale-invariant feature transform (SIFT; Lowe, 2004), speeded up robust features (SURF; Bay et al., 2008), binary robust independent elementary features (BRIEF; Calonder et al., 2010), oriented fast and rotated BRIEF (ORB; Rublee et al., 2011), and so forth. Recently, many studies have made significant progress on learned features, especially these based on DNN, and they have been successfully applied to image matching and shown satisfying performance in terms of efficiency and effectiveness, compared to handcrafted counterparts (DeTone et al., 2018; Dusmanu et al., 2019; Sarlin et al., 2019; Zhang et al., 2020). Finally, correspondences are built among queries of 2D images and 3D point cloud based on these extracted features, and the pose of each image can be estimated through solving the perspective-n-point (PnP) problem (Fischler & Bolles, 1981; Haralick et al., 1994).

Terrestrial laser scanners can acquire 3D coordinates through emitting laser pulses and recording their reflections from the surrounding environment to obtain dense 3D measurements. The modern user-friendly laser scanners can collect dense and accurate 3D point cloud in a short period and have been widely used in structures monitor and displacement measurement (H. S. Park et al., 2007; S. Park et al., 2015; Smith et al., 2011). Light detection and ranging scanners have been used to generate 3D prior maps, and previous studies have developed visual localization methods based on comparison between queries and synthetic images from point cloud (Stewart & Newman, 2012; Wolcott & Eustice, 2014). However, the lower qualities of the synthetic images limit the accuracy of their algorithms. Most current scanners have been designed to simultaneously take photos of the surrounding environment while scanning. Combining the 2D images and 3D point cloud, they can characterize the environment with rich details. It is undeniable that the 3D model collected by terrestrial laser scanners is superior to that constructed by SfM or SLAM in terms of density and accuracy (Nouwakpo et al., 2016; Wallace et al., 2016). Considering more and more facility owners started using laser scanning technologies to create digital twins (i.e., as-built building information models) of their facilities, our hypothesis is that the facility scan data can be used to enable precise indoor localization. Consequently, we propose a computational algorithm for accurate indoor localization based on the data collected with laser scanners. The algorithm mainly consists of two hierarchical modules: coarse and fine localizations. A neural point cloud rendering model is trained to generate images database for image-retrieval based localization in the course module, and approximate poses of the query images can be obtained at this stage. The use of point cloud rendering model is to address the data gap issues with terrestrial laser scanning. More specifically, the point cloud data and panoramic images are only collected at several scanning positions, and they inevitably leave gaps in synthesized images to cover the whole study area. In this regard, this study is the first to explore the use of the deep learning-based rendering on terrestrial laser scan data to generate synthetic images with known poses. After the initial coarse localization, the improved localization result can be estimated by solving the PnP problem based on the automatically extracted correspondences between images and point cloud data. The performance of the proposed method is compared with other state-of-the-art visual localization methods to test our hypothesis.

By designing the above hierarchical structure, the proposed algorithm can improve the localization accuracy and reduce failure rate. The sophisticated 2D-3D

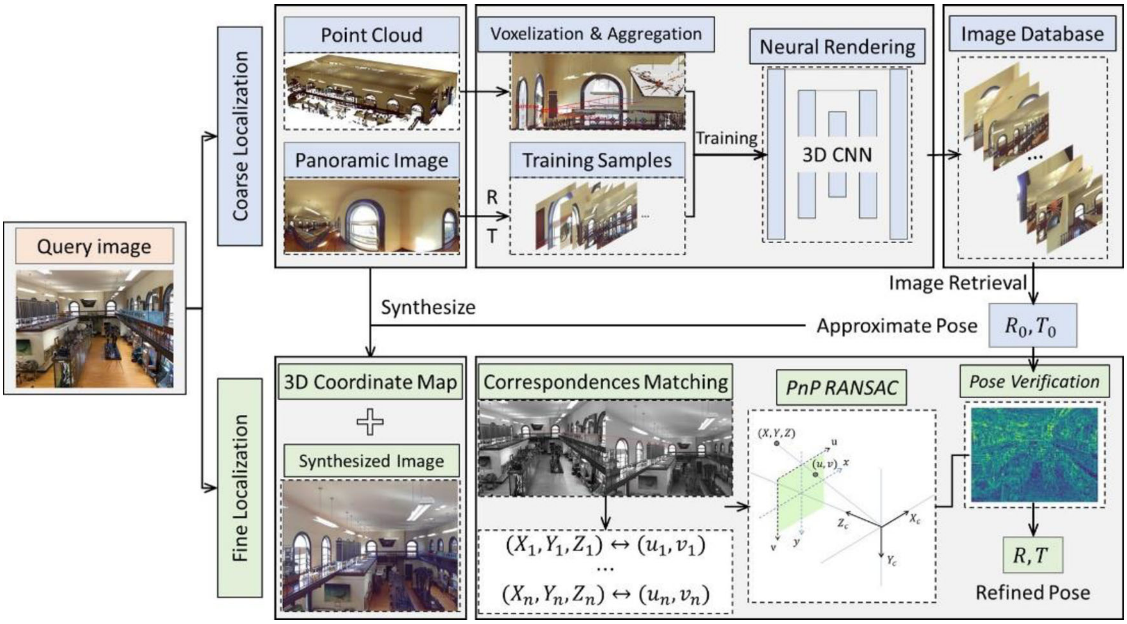

FIGURE 1 Framework of the proposed method. The method can be mainly divided into coarse and fine localizations. In the coarse localization stage, training images with known reference poses are generated from panoramic images, and point cloud is voxelized and aggregated based on the corresponding poses. Then, the neural point cloud rendering model is trained to build the image database, and approximate poses are estimated through the image retrieval method. Given the approximate pose obtained in the coarse stage, point cloud is projected into a 3D coordinate map, and synthesized images are generated from panoramic images. The new poses are estimated by solving the PnP problem based on matched 2D-3D correspondences. Finally, the refined poses are verified by comparing the query images with views synthesized from the rendering model

matching problem is cleverly transformed into 2D-2D matching and can be solved more efficiently. In addition to the computational novelties, this paper also makes the following contributions: (1) it established a method to automatically generate images with highly accurate poses from laser scan data. The generated images can be used to train indoor localization models; (2) it developed a facility scan data based indoor localization method, which can achieve state-of-the-art performances.

2 | METHOD

The overall workflow of the proposed method consists of two hierarchical modules: coarse and fine localizations (Figure 1). The point cloud data and panoramic images of the study area are collected simultaneously by a FARO Focus 3D scanner. In the coarse localization stage, approximate poses are obtained through the image retrieval method. The image database is established by a neural point cloud rendering model that can imitate a camera to synthesize image from arbitrary location and orientation. The model is trained based on the point cloud representation and images generated from panoramic images. The poses will be refined in the following module. With an approximate position and orientation, a new view and its 3D coordination map are generated from corresponding panoramic image and point cloud, respectively. The query image is matched with the synthesized image followed by building 2D-3D correspondences between the query image and the point cloud space. Refined poses are calculated by solving the PnP problem with these corresponding points, and they are verified through comparing the view synthesized from the estimated pose with the query image.

2.1 | Coarse localization

2.1.1 | Point cloud rendering

Rendering techniques can be applied to many graphics and vision-related fields. Image-based rendering has been widely studied in previous research, but their generalization is limited (Gortler et al., 1996; Levoy & Hanrahan, 1996). In this research, we use a rendering method based on point cloud data (Dai et al., 2020) without the need to construct the geometry of the environment.

Given an arbitrary viewpoint, the neural rendering model can generate photo-realistic images from point cloud data without data gaps. The training images with reference poses are synthesized from the panoramic images produced by the FARO Focus 3D scanner. As illustrated in Figure 2, the panoramic image is first converted into a panoramic sphere. Then training samples are generated

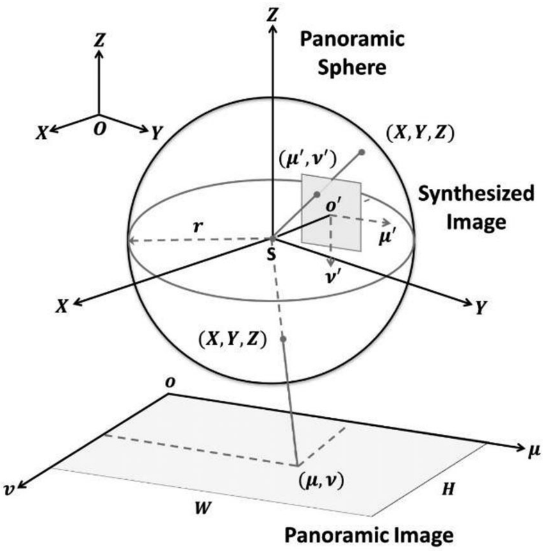

FIGURE 2 Coordinate systems in generating images from pre-defined poses. This figure presents how to generate images from a panoramic image based on given camera viewpoints. *0–XYZ* is the coordinate system of point cloud. *S–XYZ* is a panoramic sphere coordinate system, and it is converted from *0–XYZ*. *r* is the radius of the panoramic sphere. (*X, Y, Z*) is the panoramic sphere coordinate. *o–μv* is coordinate system of the panoramic image, and (*μ, v*) is the coordinate of the panoramic image pixel. H and W is the height and width of the panoramic image. *o′-μ′v′* is the coordinate system of the synthesized image. (*μ′, v′*) is the coordinate of the synthesized image pixel. The axis *So′* is perpendicular to the synthesized image plane

based on the pinhole camera model in3D sphere space. The details are described as follows:

(1) The conversion from a panoramic image to the panoramic sphere. The panoramic sphere is located at the camera center of the FARO Focus scanner. The coordinate of a panoramic image (μ, v) can be first expressed in the form of polar coordinates (θ, φ). As Figure 2 shows, the [0, W] and [0, H] of the panoramic image will be projected into the horizontal 360° and vertical 180° view, respectively (Cui et al., 2017). Thus, the relationship between (μ, v) and (θ, φ) can be represented by Equation (1):

$$\begin{cases} \theta &= (2\mu - W)\cdot \frac{\pi}{W} \\ \varphi &= \left(1 - \frac{2v}{H}\right)\cdot \frac{\pi}{2} \end{cases} \tag{1}$$

The polar coordinates can be further transformed into Cartesian coordinates. Assuming that r is the radius of the panoramic sphere, the Cartesian coordinate can be calculated by Equation (2):

$$\begin{cases} X &= r\cdot\cos\varphi\cdot\sin\theta \\ Y &= r\cdot\cos\varphi\cdot\cos\theta \\ Z &= r\cdot\sin\varphi \end{cases} \tag{2}$$

where (X, Y, Z) is the Cartesian coordinate on the panoramic sphere.

(2) The generation of images from given poses. The parameters of a virtual camera are set as those found in the back camera equipped on the IPhone 6s: focal length −4.5 mm and pixel size −1.22 μm, and the virtual camera is located at the center of the panoramic spheres S. As illustrated in Figure 2, the relationship between the sphere Cartesian coordinate *(X, Y, Z)* and the synthesized image coordinate *(*μ', v') can be expressed by the pinhole camera model:

$$s\begin{bmatrix}\mu' \\ v' \\ 1\end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\cdot\begin{bmatrix} R\ T \end{bmatrix}\cdot\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3}$$

where s is the scale factor. f_x and f_y are the focal length along μ' and v' axis expressed in pixel units. (C_x, C_y) is the coordinate of the principal point o' related to the upper left corner of the synthesized image. In our framework, [RT] is the joint rotation-translation matrix (i.e., the matrix of extrinsic parameters). R and T can be calculated from the pre-defined viewpoint of the virtual camera. Due to the alignment between the virtual camera and the panoramic sphere, T is a zero vector. R can be calculated using Equation (4), where R_x, R_y, R_z are the rotation angle of the coordinate system along X, Y, Z axis. Due to $T = [000]^T$, we can know the specific value of panoramic sphere radius r has no effect on the final projection result based on Equations (2) and (3):

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} \cos R_z \cos R_y & \cos R_z \sin R_y \sin R_x - \sin R_z \sin R_x & \cos R_z \sin R_y \cos R_x + \sin R_z \sin R_x \\ \sin R_z \sin R_y & \sin R_z \sin R_y \sin R_x + \cos R_z \sin R_x & \sin R_z \sin R_y \cos R_x - \cos R_z \sin R_x \\ -\sin R_y & \cos R_y \sin R_x & \cos R_y \cos R_x \end{bmatrix} \tag{4}$$

Same as what was used in Dai et al. (2020), a 32-layer voxel schema is established in this study along the virtual camera frustum between the nearest and farthest points. The features of every frustum voxel are aggregated from the 3D points that are projected on the voxel on the same layer. The extracted features from the point cloud are served as the input of the U-net-like 3D convolutional neural network (CNN) network (Ronneberger et al., 2015), and the model is trained by minimizing the differences between synthesized images with reference poses and network outputs.

2.1.2 | Creating the image database

The accuracy of image retrieval-based coarse localization is highly related to the size of the image database. In general, a lack of a sufficient number of rendered images in 3D space can result in bad image retrieval performance. Therefore, the poses of the rendered images need to be defined to capture the whole study area with dense coverage. Accordingly, rendered images in this study are generated within every 2-feet 3D block using the trained neural point cloud rendering model. The virtual camera is located at the center of each block, and 72 images are generated with a sampling stride of 5° in horizontal direction. In summary, around 7000 and 16,000 synthesized images are generated for lab and museum scenes.

2.1.3 | Image retrieval

The content-based image retrieval (CBIR) has been extensively studied by researchers in various fields for decades, and its performance depends critically on the feature representations and metrics for similarity measurements (Wan et al., 2014). In this research, we adopt a simple yet effective CBIR method consisting of three phases: (1) Feature extractions of the images in the pre-built database: Due to the development of DNN, many outstanding networks have emerged, and they can be used in different computer vision tasks, such as image classification, object detection, semantic segmentation, and place recognition. In this research, we use the NetVLAD as the feature extractor. NetVLAD was proposed by Arandjelović et al. (2016) and has been widely adopted as the backbone in many visual localization tasks based on image retrieval. The whole NetVLAD architecture consists of two steps: (1) a multi-layers CNN is cropped at the last convolutional layer and function as a dense descriptor extractor; (2) NetVLAD layer with learnable parameters is used to pool previous extracted dense descriptors into a fixed image representation. The NetVLAD layer is developed by mimicking the vector of locally aggregated descriptors (VLAD) in a CNN framework and making the VLAD pooling differentiable (Jégou et al., 2010). In our case, an image can be described as a 1 × 4096 feature vector through the NetVLAD (Cieslewski et al., 2018). Of particular note is that this phase is an off-line stage, and the feature vectors of these images are calculated before the localization process; (2) feature extraction of the query image: The feature of the query image is calculated using the same procedure as phase (1), and it is an on-line stage where the feature extraction is followed with image searching; (3) image searching: We take advantage of the CSI to select the image from the database, which is the most similar to the query image, and the index is calculated as the following Equation (5):

$$CSI = \frac{\vec{A} \cdot \vec{B}}{\left\|\vec{A}\right\| \cdot \left\|\vec{B}\right\|} = \frac{\sum_{n=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5}$$

where $\vec{A}$ and $\vec{B}$ are the feature vector, n is the length of the feature vector, in this research, n is 4096. The larger CSI indicates that the two images are more similar. The position and orientation of the selected image are used to represent the approximate pose of the query image.

2.2 | Fine localization

2.2.1 | Data preparation

Based on the position estimated in the coarse localization stage, the nearest panoramic image taken by the FARO Focus scanner is selected to generate five images around the approximate viewpoint using the method described in the previous module. Meanwhile, the 3D coordinates of the point cloud collected by the same scanning can also be projected onto the synthesized image. As the FARO scanner collects both point cloud and panoramic images, these two types of data can provide supplementary information. Point cloud data can provide accurate 3D information, and optical images can capture high-resolution textures (Pu & Vosselman, 2009). An important step to fuse these datasets is to bring them under the same coordinate system. For coordination transformation, the first step is to move the origin of the point cloud coordinate system to where the scanner stands and then project 3D points onto the unit sphere using Equation (6):

$$\begin{cases} X = \dfrac{X^O - X_S}{\sqrt{\left(X^O - X_S\right)^2 + \left(Y^0 - Y_S\right)^2 + \left(Z^0 - Z_S\right)^2}} \\ Y = \dfrac{Y^0 - Y_S}{\sqrt{\left(X^0 - X_S\right)^2 + \left(Y^0 - Y_S\right)^2 + \left(Z^O - Z_S\right)^2}} \\ Z = \dfrac{Z^O - Z_S}{\sqrt{\left(X^O - X_S\right)^2 + \left(Y^0 - Y_S\right)^2 + \left(Z^o - Z_S\right)^2}} \end{cases} \tag{6}$$

where X^0, Y^0, Z^0 are the Cartesian coordinates of the point cloud, X_s, Y_s, Z_s are the coordinates of the scanner under the point cloud coordinate system, and X, Y, Z is the projected coordinates on the unit sphere. Then the point (X, Y, Z) can be projected onto the panoramic image based on the inverse transformation of Equations (1) and (2). The panoramic image coordinates can be further projected to the synthesized image coordinates. In this way, the 3D point cloud can be connected with the 2D synthesized image.

2.2.2 | Image matching

The 2D-3D correspondences matching between the query image and the point cloud is achieved in an indirect way. Given the fact that the pose of the synthesized image is defined (i.e., the joint rotation-translation matrix can be calculated), 3D point cloud can be projected onto the 2D synthesized view based on the pinhole camera model (i.e., formula 3). The query image is then matched with the synthesized image, and the 3D coordinate of the matched key points can be estimated using the projection of the point cloud. A reliable set of 2D-3Dcorrespondences plays a vital role in recovering the position and orientation of the query image. In this research, we have selected two types of features for image matching: (1) traditional handcrafted features and (2) learning-based features. The former includes SIFT (Lowe, 2004), RootSIFT (Arandjelović & Zisserman, 2012), SURF (Bay et al., 2008), BRIEF (Calonder et al., 2010), and ORB (Rublee et al., 2011). For the handcrafted features, the ratio test proposed by Lowe (2004) is used to obtain consistent matching results among extracted feature vectors. We select SuperPoint (DeTone et al., 2018) as the learned feature and use SuperGlue network, which is a neural network designed to find correspondences and reject non-matchable points among two sets of features through optimizing a transport problem with a graph neural network, to perform matching on the extracted learned features of two images (Sarlin et al., 2020). This self-supervised architecture is appealing in terms of efficiency, and its time consumption is independent of the number of detected key points. We also evaluate the influence of the fusion of different features on the final localization results.

2.2.3 | Pose estimation

Assuming n pairs of 2D-3D correspondences are obtained from the previous step, we can have the following terms from the formula (3):

$$\begin{cases} \mu' = f_\mu \left(\frac{X}{Z}\right) \\ \upsilon' = f_\upsilon \left(\frac{Y}{Z}\right) \end{cases} \tag{7}$$

where f_μ and f_v are the function of the image pose, that is, six position and orientation values $\chi = [X_0, Y_0, Z_0, R_x, R_y, R_z]^T$. The above equation can be linearized as Equation (8) with the Taylor series:

$$\begin{cases} \mu' = \mu_0' + \frac{\partial f_\mu}{\partial X_0}\Delta X_0 + \frac{\partial f_\mu}{\partial Y_0}\Delta Y_0 + \cdots + \frac{\partial f_\mu}{\partial R_Z}\Delta R_Z \\ \upsilon' = \upsilon_0' + \frac{\partial f_\upsilon}{\partial X_0}\Delta X_0 + \frac{\partial f_\upsilon}{\partial Y_0}\Delta Y_0 + \cdots + \frac{\partial f_\upsilon}{\partial R_Z}\Delta R_Z \end{cases} \tag{8}$$

With sufficient 2D-3D correspondences, $\Delta = [\Delta X_0, \Delta Y_{0,}$ $\Delta Z_{0,}$ $\Delta R_{x,}$ $\Delta R_{y,}$ $\Delta R_Z]^T$ can be estimated with the least-squares regression. The pose parameters can be updated through $\chi = \chi + \Delta$, and they will be calculated iteratively until the Δ value reaches the pre-defined threshold. Considering that outliers might exist in the matched 2D-3D coordinates, random sample consensus (RANSAC) is used in the process of estimating pose (Fischler & Bolles, 1981). RANSAC is a widely used paradigm that can interpret the data containing a large percentage of gross errors.

2.3 | Pose verification

The estimated position and orientation in the fine localization stage can deviate from the true pose in cases that sufficient or accurate 2D-3D correspondences cannot be found. Under this circumstance, the accuracy of the calculated pose can fail to outperform the coarse pose. Pose verification is used to integrate the localization result of the coarse and fine stages. New virtual views are rendered from the dense and accurate point clouds based on the estimated poses in two stages. The synthesized images are compared with the corresponding query images through calculating local descriptors in a patch-wise manner, that is, DenseSIFT (C. Liu et al., 2008). The final similarities between the rendered views and the query images are computed as the median of the descriptor distances across the entire images. The final pose is selected from the coarse or fine result whose rendered view is more similar to the query image.

2.4 | Pose evaluation measures

The accuracy is evaluated by comparing the differences between the estimated and corresponding reference poses, and two indicators, including position error and orientation error, are calculated in this paper. The position error (E_{pos}) is measured with Equation (9):

$$E_{pos} = \sqrt{\left(X - X_{ref}\right)^2 + \left(Y - Y_{ref}\right)^2 + \left(Z - Z_{ref}\right)^2} \quad (9)$$

X, Y, Z are the estimated position coordinates, and X_{ref}, Y_{ref}, Z_{ref} are the reference position coordinates. The orientation error (E_{ori}) is measured as an angle in degrees, and it can be calculated with estimated and reference camera rotation matrix as illustrated in the work of Hartley et al. (2013) like the following Equation (10):

$$2\cos\left(E_{\mathrm{ori}}\right) = \mathrm{trace}\left(R_{\mathrm{ref}}^{-1} \cdot R\right) - 1 \quad (10)$$

R_{ref} and R are the reference camera rotation matrix and estimated camera rotation matrix, respectively. Besides, we follow the standard practice to calculate the percent of query images within three pose accuracy intervals, and the defined thresholds are similar with Sattler et al. (2018) and Taira et al. (2018): high-precision (0.25 m, 2°), medium-precision (0.50 m, 5°) and coarse-precision (1.00 m, 10°).

3 | EXPERIMENT MATERIALS

3.1 | Lab and Museum datasets

Two indoor datasets (as illustrated in Figure 3), that is, Lab and Museum datasets—both were captured with a FARO Focus scanner, are utilized in this research. The first dataset covers an indoor laboratory environment with a rough size of 40 m^2 area. The Lab dataset is built upon six registered scans, and the final point cloud contains around 3.6 million points (1.28 GB). The lab scene mainly consists of computers, office desks, chairs, and lab equipment. The other dataset represents a more complicated scene at the Rutgers University Geology Museum (RUGM). The RUGM dataset covers two floors, where 11 and 17 scans were captured and registered for each floor. The final registered point cloud has a size of 6.39 GB with around 14 million points. There are many display windows and fossil specimens in the museum. Due to its larger area (around 300 m^2) and clutterness, the localization problem is more challenging in the museum case.

3.2 | Images with ground truth poses

Obtaining images with accurate poses as the ground truth is always challenging and often requires a lot of manual effort. In many previous studies, SfM has been widely applied to estimate reference poses. However, the accuracy of image poses estimated by SfM is limited because the method depends on the local feature, and it is prone to fail if large differences between images such as view angles exist. Two kinds of ground truth images, that is, synthesized images and real-taken photos, are used in this study to evaluate the performance of the proposed algorithm. In order to generate virtual images with known poses as the ground truth dataset, another 10 and 12 scans are collected in the laboratory and museum, and 100 and 120 reference images are synthesized with a sampling stride of 36° along the vertical axis. The other kind of reference images are taken by an iPhone 6s, and the poses of these photos are estimated through solving the PnP problem with manually annotated 2D-3D correspondences. For every query photo, we select six to 10 corresponding points between the image and 3D point cloud to calculate pose as the reference. Then the point cloud is projected into a 2D image based on the estimated pose, and visual comparison between the iPhone photos and the projected images ensures the correctness of the reference poses. We would reselect correspondences and calculate the reference poses if inconsistency exists among the photos and the projected images. The process of manual annotation is labor-intensive and time-consuming, so the method is difficult to be scaled up. Twenty-one photos are taken randomly but evenly distributed at the laboratory room and the RUGM. Of particular note is that the ground truth images are generated from scans collected from different time at the facility; therefore, it serves the purpose to validate if the proposed method is effective for different illumination conditions and changing environments.

4 | RESULTS AND DISCUSSIONS

4.1 | Qualitative evaluation of the point cloud rendering results

According to Dai et al. (2020), the density of a point cloud has a great influence on the final rendering result. But the choice of point cloud resolution has to be balanced with the training time as denser point cloud inevitably requires significantly more processing time. In our case, using one NVIDIA Quadro RTX 5000 Graphic Processing Unit (GPU), it takes about 2 and 4 weeks to train

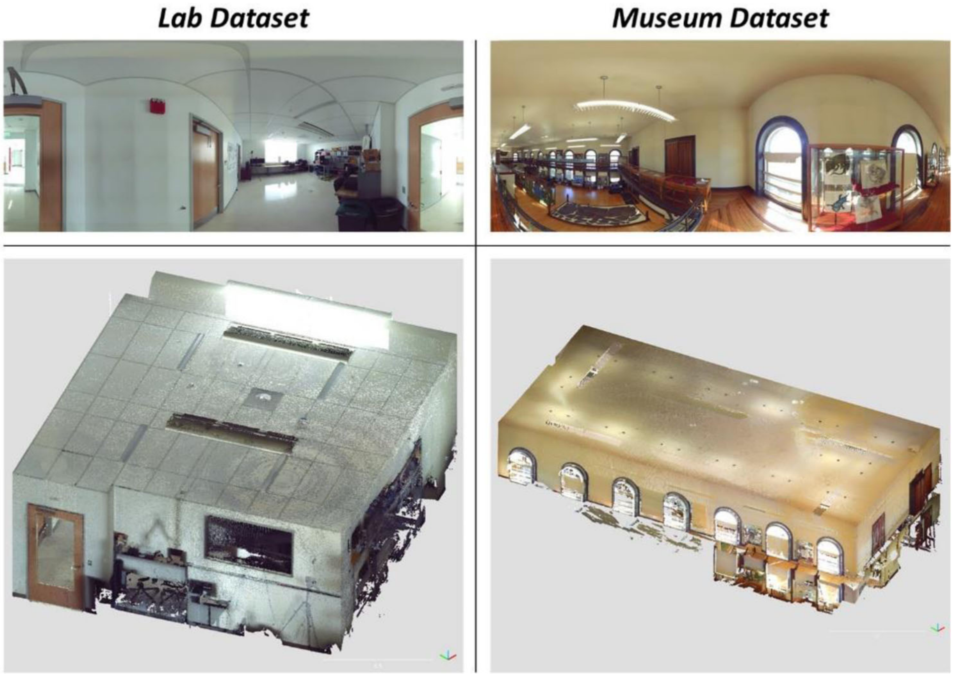

FIGURE 3 Illustration of the Lab and Museum datasets. Both the two datasets consist panoramic images and point cloud, which is registered from different scans. The first row is the example of panoramic images, and the second row is the registered point cloud

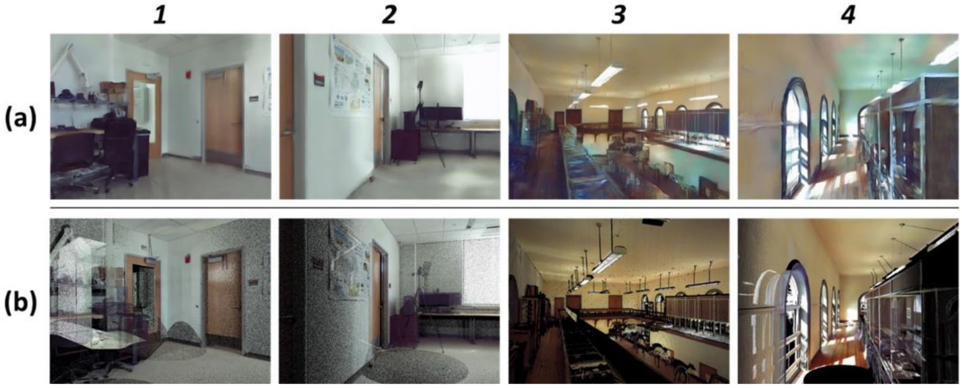

FIGURE 4 Illustration of the synthesized images from point cloud based on the deep neural rendering model. The first row (a) is the synthesized images, and the second row (b) is the corresponding projection of point cloud. Columns 1 and 2 are from Lab dataset, and Columns 3 and 4 are from Museum dataset

the deep neural point cloud rendering model for the Lab (432 training images) and Museum (2016 training images) scene, respectively. Figure 4 shows the visual comparison between the synthesized images and the corresponding direct projection of the point cloud. Two main problems exist in using direct projection of point cloud (Row b): (1) gaps in scan data result in noisy and incomplete projected images; (2) obscured point cloud would also be projected onto the images, and this will make the projection looks unrealistic. These drawbacks are unavoidable as the point cloud is not continuously distributed in 3D space. When a 3D point cloud is projected onto 2D images, certain part of the projection ray can go across the front points to reach the invisible background or nothing, and many important details are likely to be lost in the process. In contrast to the direct projection of point cloud, the deep neural rendering model can fill the gaps (i.e., blank holes) and important details can be recovered. The multi-plane design of the model can reduce the depth noise of the point cloud and prevent the occluded background point cloud from affecting the synthesized images. Although some places of the rendered images can be blurry (e.g., the poster in first row, the floors), the photo-realistic views are sufficient for the purpose of coarse localization.

TABLE 1 Accuracy of the estimated pose in the coarse localization phase

Dataset	Position error (m)	Orientation error (°)	High precision (0.25 m, 2°)	Medium precision (0.50 m, 5°)	Coarse precision (1.00 m, 10°)
Lab	0.51	6.27	1.65%	25.62%	67.77%
Museum	0.83	5.03	0.00%	16.53%	48.76%

Note: Position and orientation errors are the median values. The three intervals for pose accuracy represent the percentage (%) of correctly estimated poses evaluated on real photos and synthesized images.

4.2 | Evaluation of the coarse localization results

As shown in Table 1, the performance of the coarse localization based on image retrieval (the evaluation of image retrieval using NetVLAD can be found in Appendix B) is evaluated on the synthesized query images and real iPhone photos. For the Lab dataset, the median position and orientation error are 0.51 m and 6.27° on the used query images. A relatively large position error, 0.83 m, occurred on the Museum dataset in comparison with what was achieved in the Lab dataset, partially due to the large spatial coverage of the Museum data. But its orientation error is smaller, compared to that of the Lab dataset. In both cases, the image retrieval method alone can hardly position a camera with high precision. The percentage of position and orientation errors within 0.25 m and 2° is nearly zero for both the Lab and Museum cases. Statistically, the position and orientation based on the retrieved images have the following characteristics: The probability of position error being smaller than 1.00 m and orientation error being smaller than 10° is 67.77% and 48.76% on Lab and Museum datasets, respectively. We attribute these accuracy differences to two possible reasons: (1) the spatial size of the studied area and (2) the different spatial distribution of the query images on the two datasets. Similar photos can result from more different perspectives in a larger environment, and it poses a huge challenge for the image retrieval-based localization (L. Liu et al., 2017; Sarlin et al., 2019). The synthesized query images of the Museum are mainly distributed along the corridor close to the wall. Some images along this path face the wall and have very little semantic information that can adversely affect the CBIR (Arandjelović et al., 2016). The overall performance of the coarse localization depends mainly on the quality of the pre-built image database and the image retrieval algorithms. In this research, the database consisting of images with known poses is generated with the deep neural point cloud rendering network. The quality of the rendered images can affect the following image retrieval process. In general, we conduct a visual evaluation to confirm the quality of these generated images. A larger database can improve the accuracy of the retrieved poses but at the cost of processing time. In our case, it takes nearly 2 weeks to generate around 23k rendered images. The image retrieval algorithm used in this study also plays a critical role in the coarse localization stage. Figure 5 gives a few examples of the three most similar retrieved images through calculating the CSI based on the feature vectors extracted by the pre-trained NetVLAD. The simple yet effective image retrieval approach can search for similar images from the pre-built database even if the query images are taken under different viewpoints and illumination. In the coarse localization stage, the pose of the most similar image was used to represent the position and orientation of the query image. Although the accuracy of the coarse localization is not very high, it can provide good initial poses for the followed fine localization. In this way, it can reduce the overall localization time, and in some cases serve as the backstop if the fine localization fails.

4.3 | Evaluation of the fine localization results

The final localization accuracy achieved by coarse localization followed by fine localization is evaluated in Table 2. The performance of the selected hand-crafted and learned features show significant differences. For the Lab dataset, the learned feature (i.e., SuperPoint and SuperGlue) achieves the highest accuracy, and its position and orientation error is 7 cm and 1.21°, respectively. The performance of BRIEF and ORB is disappointing, and they do not greatly improve the localization accuracy, compared to the coarse localization. SIFT, RootSIFT, and SURF achieve similar accuracy, that is, around 0.1 m position error and 2° orientation error. The percent of correctly estimated poses with high precision has been significantly increased for all the features. The learned feature improves the rate from 1.65% in the coarse localization stage to 67.77%, and BRIEF and ORB also increase the rate by more than 20%. The percentage of query images with coarse localization precision is also increased by these features, especially with the combination of SuperPoint and SuperGlue, and it increases the percentage by 23.97%.

For the Museum dataset, the learned feature also performs well. It gets the suboptimal localization accuracy,

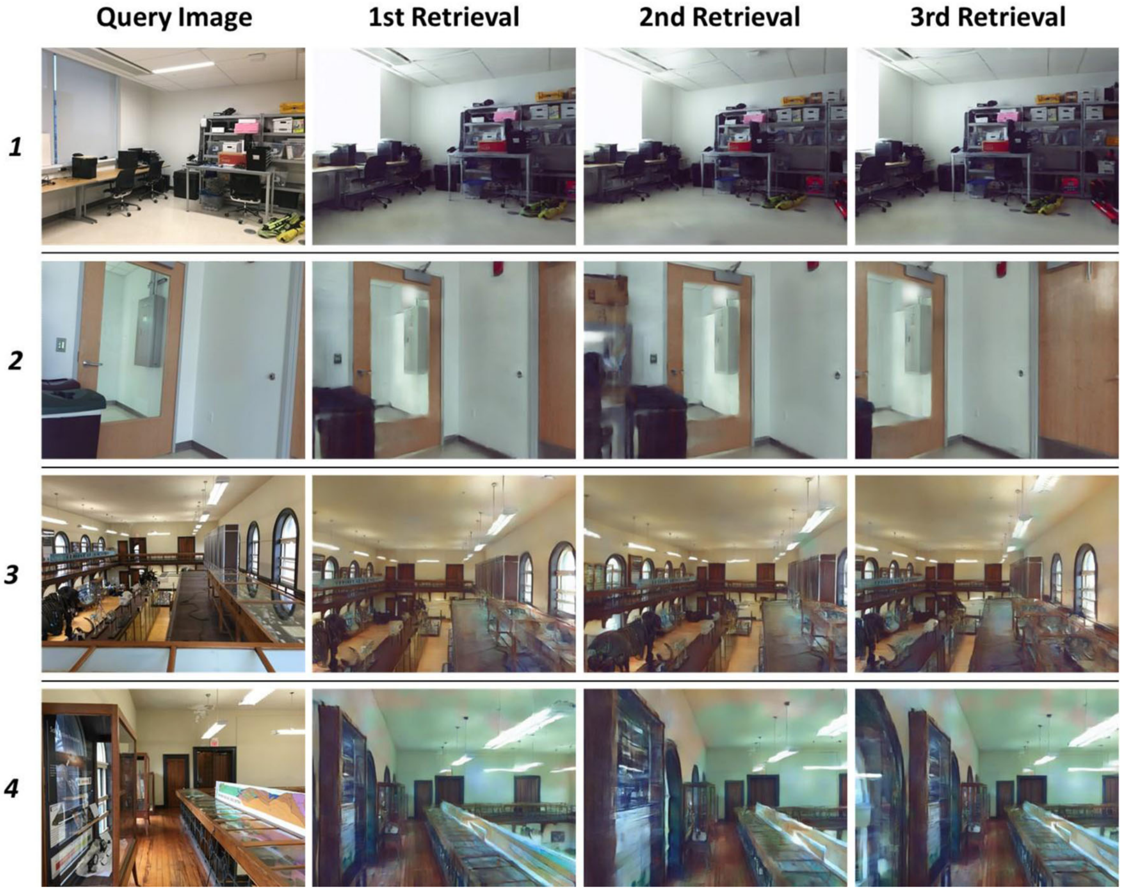

FIGURE 5 Example image retrieval results: The query images in Rows 1 and 3 are taken by an iPhone; the query images in Rows 2 and 4 are synthesized images

TABLE 2 Accuracy evaluation of the estimated poses after both coarse and fine localizations

Dataset	Feature	Position error (m)	Orientation error (°)	High precision (0.25 m, 2°)	Medium precision (0.50 m, 5°)	Coarse precision (1.00 m, 10°)
Lab	SIFT	0.09	1.72	50.41%	66.94%	78.51%
	RootSIFT	0.09	1.55	52.07%	67.77%	80.17%
	BRIEF	0.45	5.48	26.47%	40.50%	62.81%
	ORB	0.38	4.51	28.10%	47.93%	74.38%
	SURF	0.12	2.07	48.76%	61.16%	74.38%
	SP+SG	**0.07**	**1.21**	**67.77%**	**85.12%**	**90.08%**
Museum	SIFT	0.07	1.18	66.11%	76.86%	87.60%
	RootSIFT	**0.05**	1.16	69.42%	80.17%	87.60%
	BRIEF	0.09	1.31	58.68%	68.60%	75.21%
	ORB	0.55	3.12	35.54%	46.28%	57.85%
	SURF	0.06	**1.12**	68.60%	76.03%	82.64%
	SP+SG	0.07	1.22	**72.73%**	**83.47%**	**91.74%**

Note: Position and orientation errors represent the median value. The three pose accuracy intervals indicate the percentage (%) of correctly estimated poses. The poses are evaluated on the real photos and synthesized images.

Abbreviations: BRIEF, binary robust independent elementary feature; ORB, oriented fast and rotated BRIEF; SIFT, scale-invariant feature transform; SP+SG, the SuperPoint and SuperGlue feature matching; SURF, speeded up robust features.

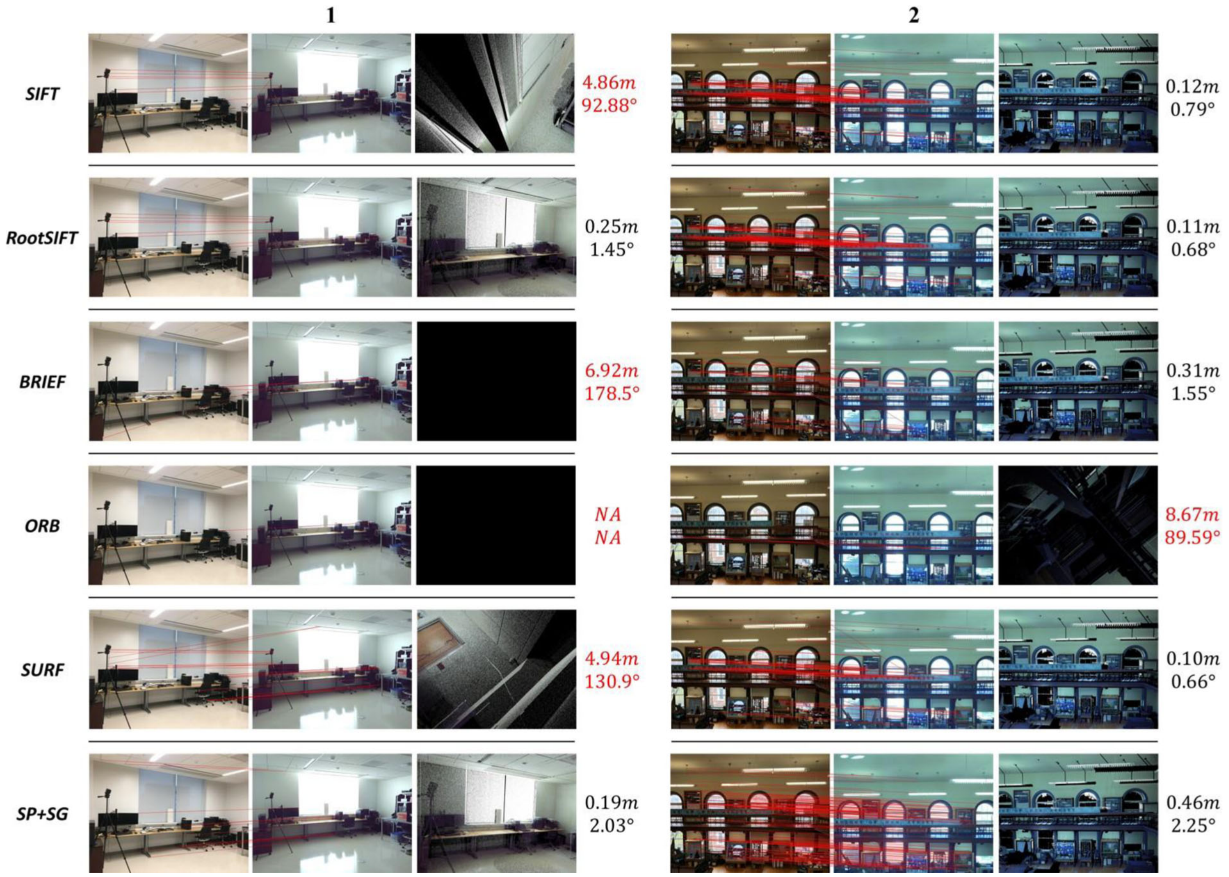

FIGURE 6 Qualitative comparison of different features on extracting correspondences between images. Examples 1 and 2 are from Lab and Museum scenes, respectively. For every selected example, query image, synthesized image from the panoramic image, direct projection of point cloud, and localization error (NA represents cannot estimate pose because of lack of sufficient correspondences.) are shown from left to right

and its localization result is robust considering that the percentage of correctly estimated poses within the three intervals are all higher than that of the handcrafted features. SIFT, Root-SIFT, and SURF achieved similar localization accuracy and fractions of correctly localized queries with high, medium, and coarse accuracy. As opposed to the Lab dataset, BRIEF achieves higher localization accuracy on the Museum dataset. In general, most features perform better on the Museum dataset, compared to the Lab dataset. This is likely due to the geometric complexity of the study area (Karami et al., 2017; Ma et al., 2020). The laboratory has a simple layout with a homogeneous spatial configuration of floor, wall, and ceiling, and does not contain rich texture information. Therefore, it is challenging for these features to extract reliable keypoints. Taking the simple query from the laboratory as an example (Figure 6), it is difficult for the utilized features to extract sufficient and reliable correspondences, and many features (except RootSIFT and learned feature) fail to estimate the pose correctly under this circumstance. Regarding the scene containing rich texture information, more correspondences can be extracted by these features and more accurate poses can be estimated based on these matched key points. Overall, the coarse localization stage can obtain approximate position and orientation for the query images, and the fine localization stage can improve the pose accuracy. However, the unreliable matched correspondences resulting from the different viewpoints and illumination conditions can lead to incorrect pose estimations (Figure 6). The accuracy evaluation of the directly estimated poses is recorded in Table B1. Due to the failures that exist in the poses estimated from solving PnP problems based on 2D-3D correspondences, the localization accuracy or the percentage of the correctly estimated poses is limited, especially the performance of the BRIEF and ORB on Lab dataset. For the sake of improving the directly estimated poses, we compare them with the coarse poses through pose verification and replace the incorrect estimated poses with the coarse localization result. Fusing coarse and fine localization results can avoid the localization failure of our proposed algorithm when lacking enough reliable extracted correspondences. Both the localization accuracy and percentage of correctly estimated poses have been

TABLE B1 Accuracy evaluation of the poses directly estimated from solving PnP problems based on 2D-3D correspondences

Dataset	Feature	Position error (m)	Orientation error (°)	High precision (0.25 m, 2°)	Medium precision (0.50 m, 5°)	Coarse precision (1.00 m, 10°)
Lab	SIFT	0.09	1.95	50.41%	60.33%	63.64%
	RootSIFT	0.10	1.80	51.24%	60.33%	61.98%
	BRIEF	3.12	108.80	25.62%	30.58%	31.40%
	ORB	4.23	153.74	27.27%	32.23%	34.71%
	SURF	0.12	2.07	49.59%	57.02%	57.85%
	SP+SG	**0.07**	**1.21**	**67.77%**	**83.47%**	**84.30%**
Museum	SIFT	0.07	1.18	66.12%	76.03%	80.17%
	RootSIFT	**0.05**	1.19	69.42%	77.69%	80.99%
	BRIEF	0.09	1.31	58.68%	66.94%	69.42%
	ORB	3.83	21.86	35.54%	40.50%	42.15%
	SURF	0.06	**1.17**	68.60%	74.38%	76.03%
	SP+SG	0.07	1.26	**72.73%**	**81.82%**	**86.78%**

Note: Position and orientation errors represent the median value. The three pose accuracy intervals indicate the percentage (%) of correctly estimated poses. The poses are evaluated on the real photos and synthesized images. SP+SG represents the SuperPoint and SuperGlue feature matching.
Abbreviations: BRIEF, binary robust independent elementary feature; ORB, oriented fast and rotated BRIEF; SIFT, scale-invariant feature transform; SP+SG, the SuperPoint and SuperGlue feature matching; SURF, speeded up robust features.

improved through the fusion, for example, the rate of estimated poses with coarse precision of learned feature is increased from 86.78% to 91.74% on the Museum dataset.

As illustrated in Figure 6, the performance of the selected features varies on the same query image. We explore if the feature fusion can reduce the probability of localization failure. Considering the bad performance of BRIEF and ORB on the Lab dataset, only SIFT, RootSIFT, SURF, and the learned feature are selected for the fusion study. The results indicate that fusion of appropriate features can improve the localization accuracy (Figure 7). The combination of SURF and the learned feature makes the most progress on the Lab dataset, and its position and orientation errors are reduced to 4 cm and 0.96°. Most of the feature combinations improve the localization performance to a certain degree, and it has resulted from more reliable extracted correspondences. Nevertheless, we also find that the percentage of query images that cannot be correctly localized cannot be significantly reduced with feature fusion. The curve peak value of the feature combination is similar to that of a separate feature, and the rate of correctly estimated poses is reduced in some cases, for example, the combination of SIFT and RootSIFT reduces the number by around 8% on the Museum dataset. The performance of feature fusion can be attributed to the correctly extracted correspondences can enhance the localization result, while the outliers can cause wrong pose estimates.

To our best knowledge, this study is the first effort on building image-based indoor localization solutions based on terrestrial laser scanner data. In the following, we can compare our method with previous work. Taira et al. (2018) proposed their indoor localization method based on Red, Green, Blue, and Depth (RGBD) images (i.e., InLoc dataset), but their rate of correctly localized queries within 1 m and 10° only reaches about 70%, which is far lower than our result. Sarlin et al. (2020) improved the percentage on the same InLoc dataset to 82.4% and won the indoor localization challenge at CVPR 2020. Although the structure-based method (i.e., activate search) achieve more accurate localization result on the 7-scenes dataset, for example, the position error reaches 2 cm on the heads scene, it is highly related to that the seven scenes have small spatial coverage and simple geometric features (Sattler et al., 2016). To summarize, the proposed method can estimate the pose of query images with comparable, if not better, accuracy.

4.4 | Application of synthesized images with reference poses

Images with high-quality reference poses are fundamental for evaluating and improving existing visual localization approaches (Zhang et al., 2020). How to obtain sufficient images with accurate poses is always a challenging task considering it is nearly impossible for humans to directly acquire poses from images. Reference pose of images in visual localization usually depends on other algorithms, for example, SfM (Schonberger & Frahm, 2016; Snavely et al., 2008) and PnP (Fischler & Bolles, 1981; Haralick et al., 1994), which may produce inaccurate pose results or require labor-intensive annotation. These drawbacks limit the availability of large and reliable benchmark datasets for visual localization.

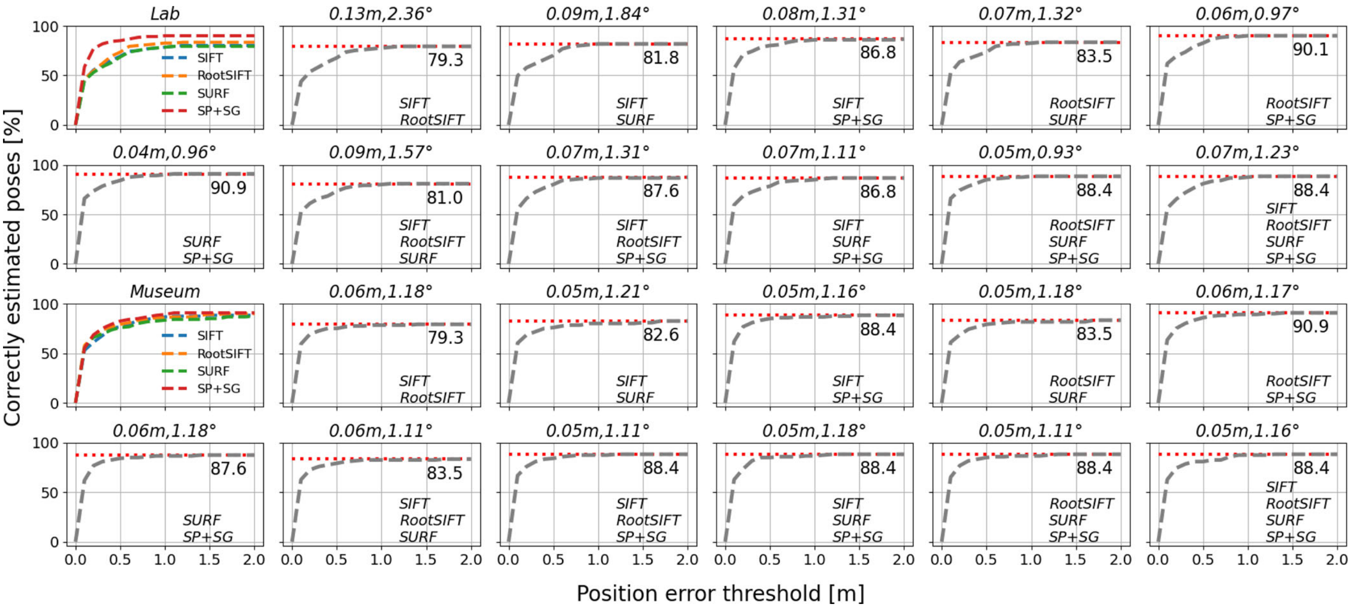

FIGURE 7 Comparison between different feature combinations. The graphs show the impacts of different feature fusion combinations on the final localization result, and plots indicated the percentage of correctly localized query images within a distance threshold whose orientation error is at most 10°. The number above the figure is the median position and orientation error of the corresponding feature fusion

TABLE 3 Comparison of PoseNet localization accuracy on different datasets

Scene	Spatial Extent (m)	Position error (m)	Orientation error (°)
Chess	3 × 2 × 1	0.32	4.06
Fire	2.5 × 1 × 1	0.47	7.33
Heads	2 × 0.5 × 1	0.29	6.00
Office	2.5 × 2 × 1.5	0.48	3.84
Pumpkin	2.5 × 2 × 1	0.47	4.21
Red kitchen	4 × 3 × 1.5	0.59	4.32
Stairs	2.5 × 2 × 1.5	0.47	6.93
Lab	**6 × 6 × 2.7**	**1.18**	**3.99**

Note: Numbers indicate the median position and orientation errors.

As described in Section 2, we propose to generate images with accurate poses from terrestrial laser scan data. In order to prove the synthesized images can be applied to complicated visual localization tasks, an end-to-end deep pose regression model, that is, PoseNet (Kendall et al., 2015), is built for the scene of the laboratory room. As a pioneer in pose regression based on DNN, PoseNet inspired many other learning-based pose estimation methods (Shavit & Ferens, 2019). Adequate images with accurate poses are the foundation of training PoseNet. Therefore, approximately 30k images are synthesized from 67 scans, and 10% of the images are randomly chosen as the test set. As illustrated in Table 3, the PoseNet on the laboratory room achieves a position error of 1.18 m and an orientation error of 3.99°. Although the position error is larger than the result of the original PoseNet paper (Kendall et al., 2015), it is related to the spatial extent of the study area. The position errors of the PoseNet trained by Kendall et al. (2015) on larger outdoor scenes also increase, and the laboratory room is significantly larger than the seven indoor scenes. In general, the PoseNet trained on the synthesized images obtains similar accuracy as the original PoseNet.

5 | CONCLUSION

This study proposes a hierarchical computational algorithm that is capable of estimating the pose of an image with high accuracy, and it is built on the foundation of laser scan data, which are commonly collected for facility modeling purposes. The algorithm consists of two stages: coarse localization and fine localization. In the first stage, the approximate pose of an image is obtained through searching from a pre-built database in which the photo-realistic images with known poses are rendered from facility scan data through a deep neural point cloud rendering model. Our results indicate the image retrieval-based coarse localization can achieve sub-meter accuracy as shown on the Lab and Museum scenarios (the position error is 0.51 and 0.83 m respectively). The coarse poses can be further improved to the centimeter accuracy in the followed fine localization stage. Our studied features including SIFT, RootSIFT, SURF, and learned features (SuperPoint and Super-Glue) in the fine localization stage achieved good performance on the two datasets. Especially for the learned feature, its performance is robust (around 7-cm position error and 1° orientation error), and the percentage of estimated queries with coarse accuracy (i.e., 1.00 m and 10°) reaches nearly 90%. The effect of feature fusion on the final localization result indicates that localization error can be reduced by integrating appropriate features, but the percentage of correctly localized images can be hardly increased. In addition to the contribution of developing the localization method, this study has found a new way of generating training samples with high-quality pose information. High-quality images with accurate poses are the foundation of many visual localization studies. It is time-consuming and labor-intensive to calculate the poses of images from manual annotation. In the proposed method, we directly generate images with accurate reference poses from the panoramic images taken by commercial static laser scanners. The synthesized reference images have been successfully used to train the PoseNet, and its accuracy, which is similar to the result of the original paper, proves that the generated images can be further used to benchmark and improve image-based localization methods. Some of the future research activities planned for this study include: (1) optimizing the image search in terms of speed by using GPUs; (2) using multiple GPUs to accelerate the neural point cloud rendering process; (3) looking into more powerful machine learning algorithms for fast learning.

ACKNOWLEDGMENTS

This research project was performed by Rutgers, The State University of New Jersey, and was sponsored by US National Science Foundation under awards 1827505 and 1737533. The funding and support from US National Science Foundation are appreciated.

REFERENCES

Alarifi, A., Al-Salman, A., Alsaleh, M., Alnafessah, A., Al-Hadhrami, S, Al-Ammar., M A., & Al-Khalifa, H. S. (2016). Ultra wideband indoor positioning technologies: Analysis and recent advances. *Sensors*, *16*(5), 707.

Arabi, S., Haghighat, A., & Sharma, A. (2020). A deep learning-based computer vision solution for construction vehicle detection. *Computer-Aided Civil and Infrastructure Engineering*, *35*(7), 753–767.

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV (pp. 5297–5307).

Arandjelović, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI (pp. 2911–2918).

Asimakopoulou, E., & Bessis, N. (2011). Buildings and crowds: Forming smart cities for more effective disaster management. *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, Seoul, South Korea (pp. 229–234).

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer Networks*, *54*(15), 2787–2805.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Borrion, H., Mitchener-Nissen, T., Taylor, J., & Lai, K.-M. (2012). Countering bioterrorism: Why smart buildings should have a code of ethics. *2012 European Intelligence and Security Informatics Conference*, Odense, Denmark (pp. 68–75).

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). Brief: Binary robust independent elementary features. *European Conference on Computer Vision*, Crete, Greece (pp. 778–792).

Chen, Z., Zou, H., Jiang, H., Zhu, Q., Soh, Y. C., & Xie, L. (2015). Fusion of WiFi, smartphone sensors and landmarks using the Kalman filter for indoor localization. *Sensors*, *15*(1), 715–732.

Cieslewski, T., Choudhary, S., & Scaramuzza, D. (2018). Data-efficient decentralized visual slam. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia (pp. 2466–2473).

Cui, T., Ji, S., Shan, J., Gong, J., & Liu, K. (2017). Line-based registration of panoramic images and LIDAR point clouds for mobile mapping. *Sensors*, *17*(1), 70.

Dai, P., Zhang, Y., Li, Z., Liu, S., & Zeng, B. (2020). Neural point cloud rendering via multi-plane projection. In T. Boult, G. Medioni, R. Zabih, E. Mortensen, & M. Masson (Eds.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7830–7839). IEEE Computer Society, Conference Publishing Services.

Davidson, P., & Piché, R. (2016). A survey of selected indoor positioning methods for smartphones. *IEEE Communications Surveys & Tutorials*, *19*(2), 1347–1370.

Davison, A. J., Reid, I. D., Molton, N. D., & Stasse, O. (2007). Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1052–1067.

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT (pp. 224–236).

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., & Sattler, T. (2019). D2-net: A trainable CNN for joint description and detection of local features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA (pp. 8092–8101).

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.

Gálvez-López, D., & Tardos, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, *28*(5), 1188–1197.

Gortler, S. J., Grzeszczuk, R., Szeliski, R., & Cohen, M. F. (1996). The lumigraph. *Proceedings of the 23rd Annual Conference on*

Computer Graphics and Interactive Techniques, New Orleans, LA (pp. 43–54).

Haralick, B. M., Lee, C. -N., Ottenberg, K., & Nölle, M. (1994). Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal of Computer Vision*, *13*(3), 331–356.

Hartley, R., Trumpf, J., Dai, Y., & Li, H. (2013). Rotation averaging. *International Journal of Computer Vision*, *103*(3), 267–305.

Hazas, M., & Hopper, A. (2006). Broadband ultrasonic location systems for improved indoor positioning. *IEEE Transactions on Mobile Computing*, *5*(5), 536–547.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA (pp. 3304–3311).

Karami, E., Prasad, S. & Shehata, M. (2017). Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. arXiv:1710.02726.

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

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., Behar, J. V., Hern, S. C., & Engelmann, W. H. (2001). The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemiology*, *11*(3), 231–252.

Kriz, P., Maly, F., & Kozel, T. (2016). Improving indoor localization using Bluetooth low energy beacons. *Mobile Information Systems*, *2016*, 2083094.

Levoy, M., & Hanrahan, P. (1996). Light field rendering. *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, New Orleans, LA (pp. 31–42).

Liu, C., Yuen, J., Torralba, A., Sivic, J., & Freeman, W. T. (2008). SIFT flow: Dense correspondence across different scenes. *European Conference on Computer Vision*, Marseille, France (pp. 28–42).

Liu, L., Li, H., & Dai, Y. (2017). Efficient global 2D-3D matching for camera localization in a large-scale 3D map. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy (pp. 2372–2381).

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

Lu, X., Chen, Y., & Li, X. (2017). Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features. *IEEE Transactions on Image Processing*, *27*(1), 106–120.

Luo, X., Li, H., Yu, Y., Zhou, C., & Cao, D. (2020). Combining deep features and activity context to improve recognition of activities of workers in groups. *Computer-Aided Civil and Infrastructure Engineering*, *35*(9), 965–978.

Ma, J., Jiang, X., Fan, A., Jiang, J., & Yan, J. (2020). Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, *129*, 23–79.

Ni, F., Zhang, J., & Noori, M. N. (2020). Deep learning for data anomaly detection and data compression of a long-span suspension bridge. *Computer-Aided Civil and Infrastructure Engineering*, *35*(7), 685–700.

Niu, J., Wang, B., Shu, L., Duong, T. Q., & Chen, Y. (2015). ZIL: An energy-efficient indoor localization system using zigbee radio to detect WiFi fingerprints. *IEEE Journal on Selected Areas in Communications*, *33*(7), 1431–1442.

Nouwakpo, S. K., Weltz, M. A., & McGwire, K. (2016). Assessing the performance of structure-from-motion photogrammetry and terrestrial LIDAR for reconstructing soil surface microtopography of naturally vegetated plots. *Earth Surface Processes and Landforms*, *41*(3), 308–322.

Oh, B. K., Kim, K. J., Kim, Y., Park, H. S., & Adeli, H. (2017). Evolutionary learning based sustainable strain sensing model for structural health monitoring of high-rise buildings. *Applied Soft Computing*, *58*, 576–585.

Park, H. S., Lee, H., Adeli, H., & Lee, I. (2007). A new approach for health monitoring of structures: terrestrial laser scanning. *Computer-Aided Civil and Infrastructure Engineering*, *22*(1), 19–30.

Park, S., Park, H. S., Kim, J., & Adeli, H. (2015). 3D displacement measurement model for health monitoring of structures using a motion capture system. *Measurement*, *59*, 352–362.

Pittaluga, F., Koppal, S. J., Kang, S. B., & Sinha, S. N. (2019). Revealing scenes by inverting structure from motion reconstructions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico (pp. 145–154).

Pu, S., & Vosselman, G. (2009). Building facade reconstruction by fusing terrestrial laser points and images. *Sensors*, *9*(6), 4525–4542.

Rafiei, M. H., & Adeli, H. (2016). A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, *142*(2), 04015066.

Rafiei, M. H., & Adeli, H. (2018). Novel machine learning model for construction cost estimation taking into account economic variables and indices. *Journal of Construction Engineering and Management*, *144*(12), 04018106.

Rodriguez Lera, F. J., Rico, F. M., & Olivera, V. M. (2019). Neural networks for recognizing human activities in homelike environments. *Integrated Computer-Aided Engineering*, *26*(1), 37–47.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: C onvolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany (pp. 234–241).

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision*, Barcelona, Spain (pp. 2564–2571).

Sarlin, P. -E., Cadena, C., Siegwart, R., & Dymczyk, M. (2019). From coarse to fine: Robust hierarchical localization at large scale. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA (pp. 12716–12725).

Sarlin, P. -E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In T. Boult, G. Medioni, R. Zabih, E. Mortensen, & M. Masson (Eds.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4938–4947). IEEE Computer Society, Conference Publishing Services.

Sattler, T., Leibe, B., & Kobbelt, L. (2016). Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(9), 1744–1756.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., & Stenborg, E., Safari, D., Okutomi, M., Pollefeys,M., Sivic, J., Kahl, F., & Pajdla, T. (2018). Benchmarking 6DOF outdoor visual localization in changing conditions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT (pp. 8601–8610).

Schonberger, J. L., & Frahm, J. -M. (2016). Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI (pp. 4104–4113).

Shavit, Y. & Ferens, R. (2019). Introduction to camera pose estimation with deep learning. arXiv preprint arXiv:1907.05272.

Smith, M. J., Paron, P., & Griffiths, J. S. (2011). *Geomorphological mapping: Methods and applications* (Vol. 15). Elsevier.

Snavely, N., Seitz, S. M., & Szeliski, R. (2008). Modeling the world from internet photo collections. *International Journal of Computer Vision*, *80*(2), 189–210.

Snoonian, D. (2003). Smart buildings. *IEEE Spectrum*, *40*(8), 18–23.

Stella, M., Russo, M., & Begušić, D. (2012). RF localization in indoor environment. *Radioengineering*, *21*(2), 557–567.

Stewart, A. D., & Newman, P. (2012). Laps-localisation using appearance of prior structure: 6-DOF monocular camera localisation using prior pointclouds. *2012 IEEE International Conference on Robotics and Automation*, Guangzhou, China (pp. 2625–2632).

Subbu, K. P., Gozick, B., & Dantu, R. (2013). Locateme: Magnetic-fields-based indoor localization using smartphones. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(4), 1–27.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., & Torii, A. (2018). InLoc: Indoor visual localization with dense matching and view synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT (pp. 7199–7209).

Taleb, T., & Kunz, A. (2012). Machine type communications in 3GPP networks: potential, challenges, and solutions. *IEEE Communications Magazine*, *50*(3), 178–184.

Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., & Pajdla, T. (2015). 24/7 place recognition by view synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Waikoloa Beach, HI (pp. 1808–1817).

Vera-Olmos, F. J., Pardo, E., Melero, H., & Malpica, N. (2019). DeepEye: Deep convolutional network for pupil detection in real environments. *Integrated Computer- Aided Engineering*, *26*(1), 85–95.

Wallace, L., Lucieer, A., Malenovský, Z., Turner, D., & Vopěnka, P. (2016). Assessment of forest structure using two uav techniques: A comparison of airborne laser scanning and structure from motion (SfM) point clouds. *Forests*, *7*(3), 62.

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., & Li, J. (2014). Deep learning for content-based image retrieval: A comprehensive study. *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL (pp. 157–166).

Weyand, T., Kostrikov, I., & Philbin, J. (2016). Planet photo geolocation with convolutional neural networks. *European Conference on Computer Vision*, Amsterdam, The Netherlands (pp. 37–55).

Wolcott, R. W., & Eustice, R. M. (2014). Visual localization within LIDAR maps for automated urban driving. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Chicagi, IL (pp. 176–183).

Wu, Y., Tang, F., & Li, H. (2018). Image-based camera localization: an overview. *Visual Computing for Industry, Biomedicine, and Art*, *1*(1), 8.

Yang, T., Cappelle, C., Ruichek, Y., & El Bagdouri, M. (2019). Multi-object tracking with discriminant correlation filter based deep learning tracker. *Integrated Computer-Aided Engineering*, *26*(3), 273–284.

Zafari, F., Gkelias, A., & Leung, K. K. (2019). A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials*, *21*(3), 2568–2599.

Zelenkauskaite, A., Bessis, N., Sotiriadis, S., & Asimakopoulou, E. (2012). Interconnectedness of complex systems of internet of things through social network analysis for disaster management. *2012 Fourth International Conference on Intelligent Networking and Collaborative Systems*, Bucharest, Romania (pp. 503–508).

Zhang, Z., Sattler, T. & Scaramuzza, D. (2020). Reference pose generation for visual localization via learned features and view synthesis. Int J Comput Vis (2020); doi:10.1007/s11263-020-01399-8

How to cite this article: Xia, J., & Gong, J. (2022). Precise indoor localization with 3D facility scan data. *Comput Aided Civ Inf*, *37,* 1243–1259. https://doi.org/10.1111/mice.12795

APPENDIX

A. EVALUATION OF IMAGE RETRIEVAL USING NetVLAD

To evaluate how well the NetVLAD extracts image features, a test dataset with 42 image pairs was generated through manual labeling from Museum and Lab scenes. Features extracted by the pre-trained VGG-16, which is also the backbone network used in the NetVLAD method, were applied to the image retrieval. The evaluation of the image retrieval result based on features extracted by NetVLAD and VGG-16 is illustrated in Figure A1. Generally, feature representations extracted by NetVLAD outperform VGG-16 by a large margin on the test dataset. For example, features extracted by NetVLAD achieve 54.6% for recall@1 in comparison to 42.9% obtained by VGG-16 features. When considering recall@5, NetVLAD features can nearly retrieve all the correct images (97.6%). However, it is undeniable that improper NetVLAD features can also result in inaccurate image retrieval results (Figure A2),

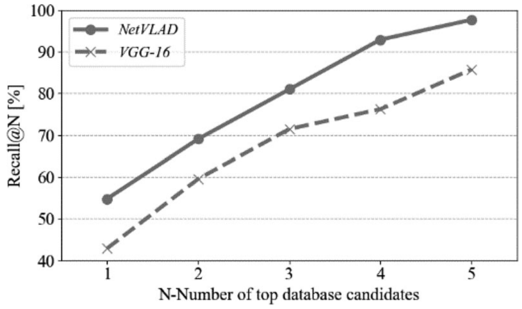

FIGURE A1 Evaluation of image retrieval based on features extracted by NetVLAD

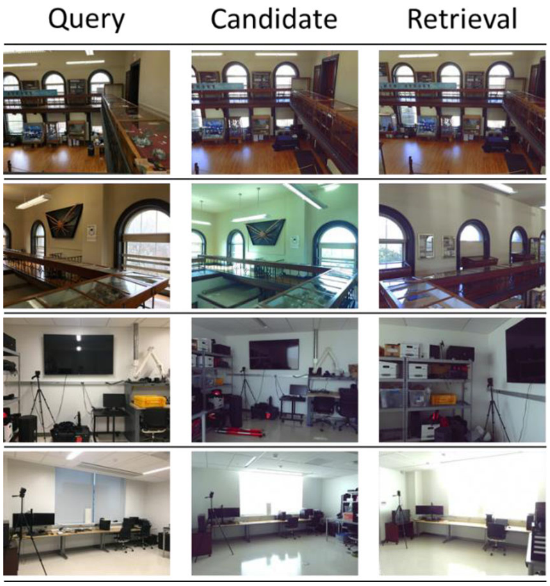

FIGURE A2 Examples of inaccurate image retrieval resulted from improperly extracted NetVLAD features. The candidate column is the manually selected most similar image from the database

and this will further affect the following correspondences matching process.

B. EVALUATION OF DIRECTLY ESTIMATED POSES

The accuracy evaluation of the directly estimated poses in the fine localization stage is recorded in Table B1. For Lab dataset, without fusion with the coarse poses, the position and orientation error of BRIEF and ORB cannot even meet the requirement of coarse precision localization (1.00 m, 10°). This indicates these two features cannot retrieve sufficient reliable correspondences on Lab dataset for most query images and result in wrong estimated poses. The performance of ORB is also disappointing on the Museum dataset, and its position and orientation error reach 3.83 m and 21.86°. However, the poses accuracy of the BRIEF has been improved a lot, and the position and orientation errors are decreased to 0.09 m and 1.31°. The performance of the other four features is robust, and they obtain approximately the same localization accuracy on the Lab and Museum dataset. The learned feature (i.e., the combination of SuperPoint and SuperGlue) achieved the best performance in terms of correctly estimated poses within the three defined intervals. For example, the percentage of the correctly estimated pose with high precision (i.e., 0.25 m, 2°) of learned feature is 67.77%, which is much higher than the other feature on the Lab dataset. The localization accuracy of the learned feature is also very high, and the position and orientation errors are 0.07 m and 1.21° on the Lab dataset and 0.07 m and 1.26° on the Museum dataset.

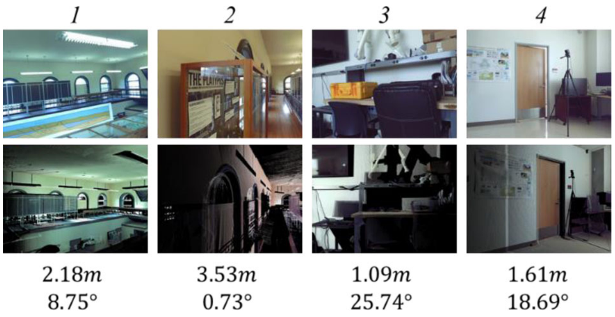

FIGURE C3 Examples of localization error based on SuperPoint feature and SuperGlue feature matching. The first row is query images, and the images of the second row is projected from point cloud using the estimated pose. The numbers represent the position and orientation error

C. EXAMPLES OF LOCALIZATION ERROR

Examples of localization error based on SuperPoint and SuperGlue feature matching are illustrated in Figure C3. The errors mainly arise from the incorrect or insufficient 2D-3D correspondences. To be more specific, abnormal illumination conditions, as example 1 shows, can result in inaccurate correspondences matching. If the field of view of the query image (shown in Example 2) is limited to a small environment, it is also not easy to match corresponding points. The change of environment (the yellow box on the desk in Example 3) could generate inaccurate image retrieval results, and further leads to imprecise 2D-3D correspondences. In another condition, the estimated sufficient correspondences mainly located on the same plane, the pose calculated through solving the PnP problem based on the 2D-3D points could not reach high accuracy.