Accurate matching between BIM-rendered and real-world images

Houhao Liang[1] and Justin K.W.Yeoh[1]

Abstract— **As the digital representation of the built environment, BIM has been used to assist robot localization. Real-world images captured by the robot camera can be compared with BIM-rendered images to estimate the pose. However, there is a perception gap between the BIM environment and reality; image styles are typically too different to be matched. Hence, this study investigates an advanced image feature detection technique, D2-Net, to identify key points and descriptors on BIM-rendered and real-world images. These key features are further matched via K Nearest Neighbor Search and RANSAC. The ability to bridge the perception gap can be evaluated by the image matching performance, which is the Euclidean distance between the projected key points and the number of inliers. SIFT, as the traditional feature detection technique, was compared in this study. Results show that the average projection error of D2-Net is only 16.55 pixels, while the error of SIFT is 187.46 pixels. It demonstrates that the advanced D2-Net can be utilized to detect representative features on BIM-rendered and real-world images. The matched image pairs can be further utilized to estimate the robot pose in BIM. Overall, it aims to enhance the BIM-assisted localization and improve the robot's reliability as a decision-making tool on-site.**

I. INTRODUCTION

In recent decades, Building Information Modeling (BIM) has been widely used to enable digital transformation in Architecture, Engineering, and Construction(AEC) sector. It is able to store infrastructure information digitally and has the potential to reshape the management in an automated way. For example, as-built status collected in a format of images [1] or point cloud [2] can be automatically compared against as-designed BIM to analyze the progress deviation. Instead of sending inspectors on-site to collect data, the feasibility of leveraging robot has been studied in [3]–[5] to replace human work. Light Detection and Ranging (LiDAR) and camera can be mounted and programmed on robot to automate the data acquisition task. It can be further developed as a decision-making tool by analyzing the acquired as-built data.

In order to deploy the robot on-site, waypoints are usually designed based on reference models such as BIM, and corresponding tasks are automatically conducted at these waypoints. Localization is an essential module to ensure the robot reaches its designated waypoints, as it determines the robot position within the environment. However, accurate indoor localization is still challenging to be achieved due to the inaccuracy and incompleteness of the robot's sensors and effectors [7]. Consequently, the robot might not be able to reach the position exactly. The perspective retrieved in BIM at the designated waypoint is prone to be misaligned with the robot perspective. Analysis such as progress inference and productivity calculation tends to be falsely made when referring to incorrect BIM elements. Hence, it is necessary to enhance the pose estimation in BIM and improve the perspective alignment to strengthen the reliability of the decisions made by the robot.

Since BIM can represent the built environment digitally, recent studies have sought to use it to assist localization. BIM-rendered and real-world images are compared to estimate the pose. However, there is a perception gap between the two domains. Considering this, BIM-PoseNet [8] was developed as a regression model to roughly estimate the pose using BIM-rendered images as the training dataset. Improvement was proposed in [9] by using an advanced recurrent CNN model. Also, the style of BIM-rendered images were transferred into a realistic one by Generative Neural Network (GNN) model, and these generated images were then further compared against real-world images to estimate the pose [10], [11]. Nevertheless, deep learning-based methods still need a massive amount of data to develop the model, and the generalization abilities need to be further studied.

To address the aforementioned limitations, this study investigates an advanced image feature detection technique to bridge the cross-domain perception gap. Detected key features on BIM-rendered and real-world images are further matched via a neighboring search. The Euclidean distance error between projected key points, and the number of matched pixel pairs are reported to demonstrate the matching performance. These matched pixel pairs can be further developed to estimate the robot pose.

II. METHOD

The appearances of BIM-rendered images are significantly different from real-world images due to changes in depiction style, texture, and illumination difference, making cross-domain image matching challenging [10], [12], [13]. Traditional key points detection methods such as SIFT, SURF, ORB, and DoG perform poorly as they only consider small image regions at a low level [14]. The detection results are significantly affected by changes in pixel intensities. Therefore, images with significant appearance differences are difficult to be matched based on these key features.

Considering the style difference, this study exploits a pre-trained D2-Net model [14] to detect key features on BIM-rendered and real-world images. Instead of detecting key points and then describing features given the image patches extracted around the key points, D2-Net detects and describes

[1] Authors are with Department of Civil and Environmental Engineering, College of Design and Engineering, National University of Singapore, 117576, Singapore, {houhaol@u.nus.edu, justinyeoh@nus.edu.sg}

TABLE I

THE IMAGE MATCHING EVALUATION USING SIFT AND D2-NET ON BIM-RENDERED AND REAL-WORLD IMAGES

Methods	Image pair 1		Image pair 2		Image pair 3		Overall	
	Projection error	Inliers	Projection error	Inliers	Projection error	Inliers	Mean projection error	Mean inliers
SIFT	196.76	21	124.17	5	241.46	7	187.46	11
D2-Net	14.56	63	19.19	62	15.89	40	16.55	55

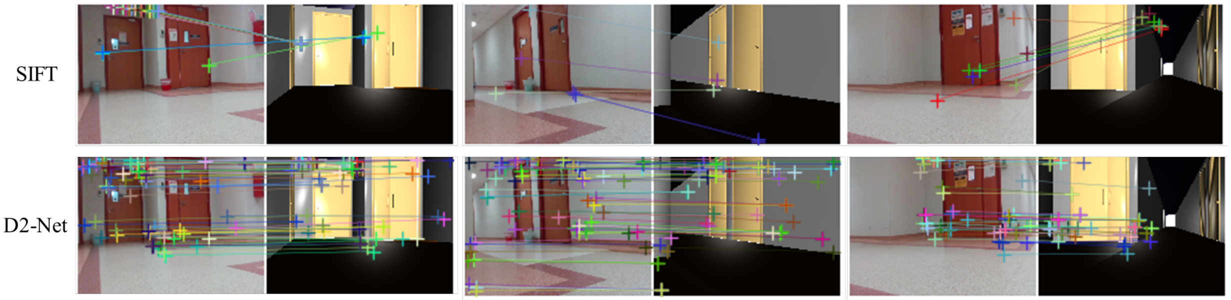

Image pair1

Image pair 2

Image pair 3

Fig. 1. Matched key features on three image pairs

them simultaneously. It uses VGG16 architecture, pre-trained on ImageNet, to extract the feature maps on images. At a high dimensional space, the key point is the local maxima of VGG derived feature maps by applying Non-Maximum Suppression (NMS). At the same time, the descriptor is computed by chunking VGG derived feature maps for each key point. Finally, upon obtaining the detected key features, K Nearest Neighbour (KNN) search and a ratio test are applied to select the good matches [15]. Random sample consensus (RANSAC) is further adopted to remove outliers.

III. RESULTS AND DISCUSSIONS

To demonstrate the matching performance, a traditional and well-known image feature extraction technique, SIFT, was used as a comparison in this study. Key points and descriptors were detected by SIFT and D2-Net, followed by a KNN search and RANSAC to match the descriptors. The “good” matched pairs are named as inliers. A homography matrix was manually computed by selecting corresponding key points on each set of BIM and real-world images. The matching performance can be measured as the projection error of key points. Specifically, key points detected on BIM-rendered images were projected to the position on real-world images using the homography matrix. The Euclidean distance between projected BIM key points and matched real-world key points was measured in pixels. In this study, three pairs of BIM-rendered and real-world images were used to demonstrate the matching performance of using D2-Net against using SIFT.

Table I reported the average projection error and the number of inliers. It can be seen the overall projection error of SIFT reaches 187.46 pixels, implying that the SIFT failed to find correctly matched key points between both BIM-rendered and real-world images. However, the D2-Net is robust to the cross-domain perception issue as the projection error is only 16.55 pixels. In terms of the amount of inliers, D2-Net outperformed the traditional SIFT, by 55 to 11. Fig.1 shows matched key points pairs on BIM-rendered and real-world images. Overall, the quantitative and qualitative results show that D2-Net is able to find the point pairs despite there being differences in appearances within the images. Besides, it is necessary to mention that BIM in this study have not been rendered to produce more detailed and realistic images. With these better-rendered images, it may be posited that the detection and matching could be significantly improved by using a pre-trained D2-net model. However this implies greater computational effort to obtain these rendered images.

IV. CONCLUSIONS

This study investigates the performance of finding matched key features on BIM-rendered and real-world images using a pre-trained D2-Net model. Without the efforts to prepare a large building-oriented dataset to train new deep learning model, this study shows a pre-trained D2-Net model is potentially able to associate BIM environment and reality to a certain extent. One potential research work that can be conducted in the future is to estimate the robot pose by using the matched 2D key features. It also can be developed to rectify the robot pose to have a better-aligned perspective between BIM and the real-world.

REFERENCES

[1] Kevin K. Han and Mani Golparvar-Fard. Appearance-based material classification for monitoring of operation-level construction progress using 4d bim and site photologs. *Automation in Construction*, 53:44–57, 2015.

[2] Frédéric Bosché, Mahmoud Ahmed, Yelda Turkan, Carl T Haas, and Ralph Haas. The value of integrating scan-to-bim and scan-vs-bim techniques for construction monitoring using laser scanning and bim: The case of cylindrical mep components. *Automation in Construction*, 49:201–213, 2015.

[3] Pileun Kim, Jingdao Chen, and Yong K Cho. Slam-driven robotic mapping and registration of 3d point clouds. *Automation in Construction*, 89:38–48, 2018.

[4] Khashayar Asadi, Hariharan Ramshankar, Harish Pullagurla, Aishwarya Bhandare, Suraj Shanbhag, Pooja Mehta, Spondon Kundu, Kevin Han, Edgar Lobaton, and Tianfu Wu. Vision-based integrated mobile robotic system for real-time applications in construction. *Automation in Construction*, 96:470–482, 2018.

[5] A Adán, B Quintana, SA Prieto, and F Bosché. An autonomous robotic platform for automatic extraction of detailed semantic models of buildings. *Automation in Construction*, 109:102963, 2020.

[6] Amir Ibrahim, Ali Sabet, and M. Golparvar-Fard. Bim-driven mission planning and navigation for automatic indoor construction progress detection using robotic ground platform.

[7] Prabin Kumar Panigrahi and Sukant Kishoro Bisoy. Localization strategies for autonomous mobile robots: A review. *Journal of King Saud University - Computer and Information Sciences*, 2021.

[8] Debaditya Acharya, Kourosh Khoshelham, and Stephan Winter. Bimposenet: Indoor camera localisation using a 3d indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:245–258, 2019.

[9] Debaditya Acharya, Sesa Singha Roy, Kourosh Khoshelham, and Stephan Winter. A recurrent deep network for estimating the pose of real indoor images from synthetic image sequences. *Sensors*, 20(19), 2020.

[10] Junjie Chen, Shuai Li, Donghai Liu, and Weisheng Lu. Indoor camera pose estimation via style-transfer 3d models. *Computer-Aided Civil and Infrastructure Engineering*, 37(3):335–353, 2022.

[11] Junjie Chen, Shuai Li, and Weisheng Lu. Align to locate: Registering photogrammetric point clouds to bim for robust indoor localization. *Building and Environment*, 209:108675, 2022.

[12] Hao Zhou, Torsten Sattler, and David W. Jacobs. Evaluating local features for day-night matching. In *ECCV Workshops*.

[13] Torsten Sattler, William P. Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, M. Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomás Pajdla. Benchmarking 6dof outdoor visual localization in changing conditions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018.

[14] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *ArXiv*, abs/1905.03561, 2019.

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