Robotic Vision Scene Understanding Challenge 2023
MCSLab Report

Antyanta Bangunharcana   Kyung-Soo Kim
Mechatronics, Systems, and Control Laboratory
Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
{antabangun, kyungsoo}@kaist.ac.kr

1. Introduction

This report presents our approach to the 2023 Robotic Vision Scene Understanding Challenge [6], a major research area for tasks such as navigation, object recognition, and interaction in robotics. Unlike the 2022 challenge, this year's challenge primarily focuses on active and noisy pose estimation scenarios, further complicating the task at hand.

Building on our success with the challenge of last year [1], we have expanded and adapted our methodology to address these challenging situations. The backbone of our solution remains our proven SLAM (simultaneous localization and mapping) and scene change detection semantic techniques, refined and extended to meet the unique challenges of this year.

The following sections describe methodological aspects and predict future research directions.

2. Methodology

The challenge is executed within a BenchBot simulator [7], posing unique complexities that our methodology is designed to tackle. Our strategy comprises two principal components: 3D object/scene change map creation and localization within noisy dead reckoning robotic systems. The overall pipeline is shown in Fig. 1.

2.1. 3D Object and Scene Change Map

Building on our previous solution, our strategy for the current challenge relies on our established 3D semantic mapping and scene change detection mechanisms. Instead of overhauling the entire mapping pipeline, we focused on updating the instance segmentation and 3D object detection models, which are central to our solution.

Similarly to last year, we use MMDetection [3]. We attempted to transition from QueryInst [5] to Mask2Former [4] model but did not observe noticeable improvement. On the contrary, the upgrade led to a degradation in the final OMQ metrics of the submitted results. Furthermore, there was a noticeable increase in false positives in the predictions.

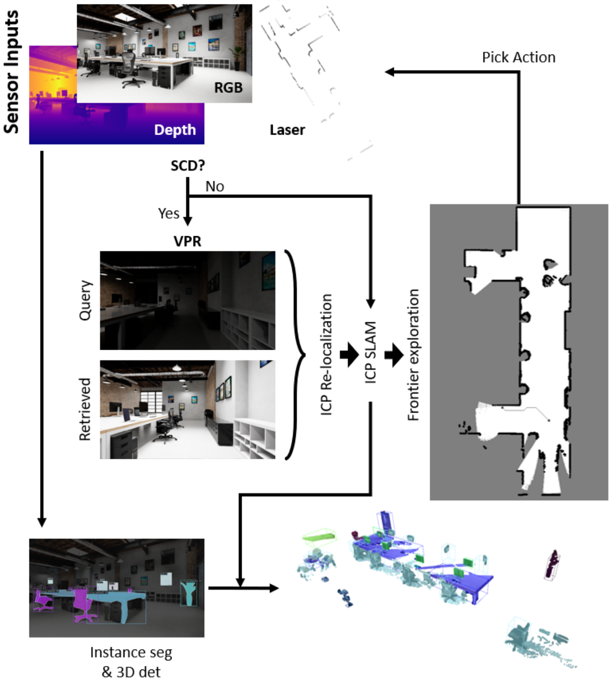
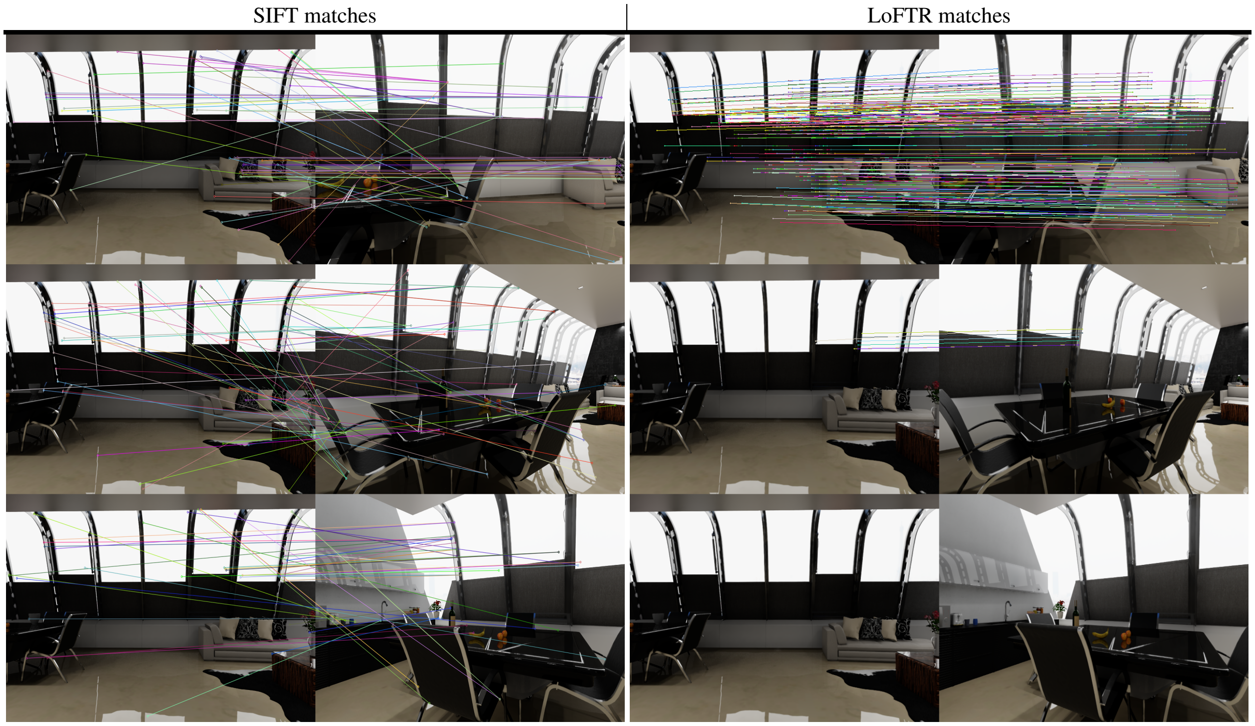


Figure 1. Overall pipeline.

We observed minimal improvements when we updated the FCAF3D [11] detection model to the newer TR3D [12] model. Despite demonstrating superior metrics in public benchmarks, these recent state-of-the-art models do not seem to transfer effectively to the BenchBot simulation system.

2.2. Localization

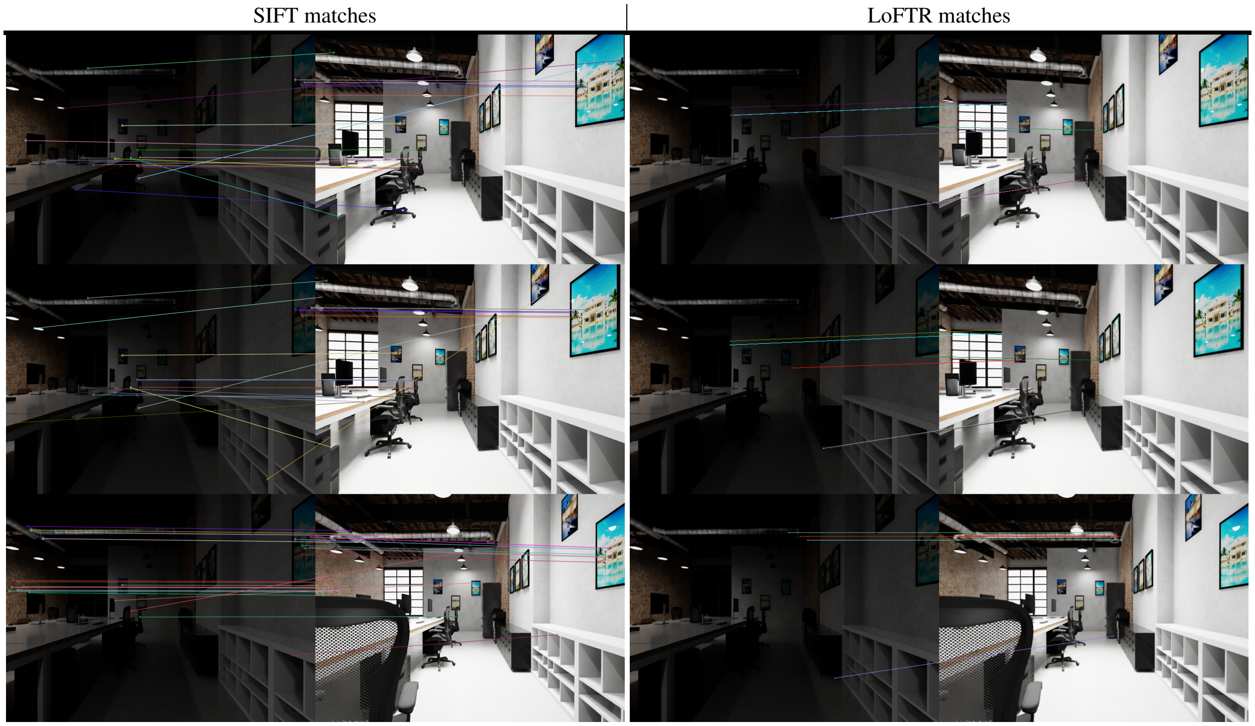
The main focus of this year's challenge is to deal with scenarios where the pose observation obtained through dead-reckoning is noisy and imperfect, necessitating the computation of a more accurate localization via SLAM.



SIFT matches

LoFTR matches

Figure 2. Feature matching using SIFT and LoFTR in the **apartment_1_4** scene. A case where LoFTR is a better option than SIFT.



SIFT matches

LoFTR matches

Figure 3. Feature matching using SIFT and LoFTR in the **office_1_5** scene. A case where LoFTR might not have been better than SIFT.

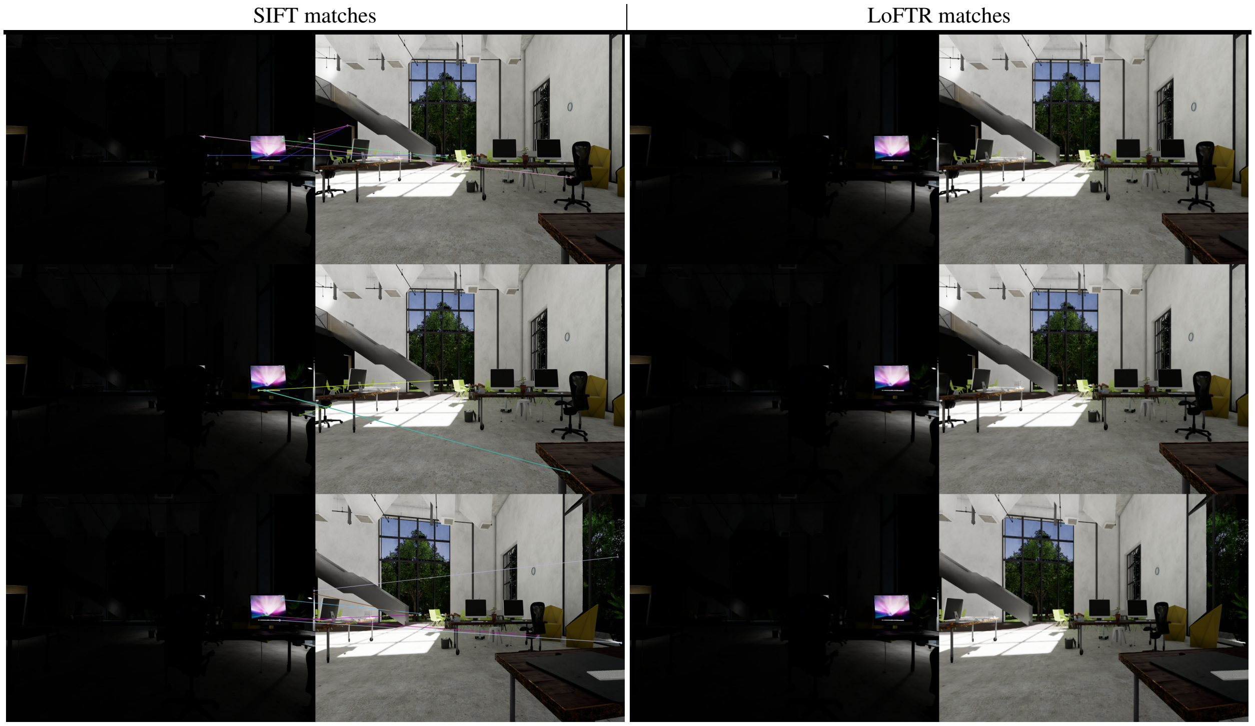


Figure 4. Feature matching using SIFT and LoFTR in the **company_4_5** scene. A case where both LoFTR and SIFT fail.

2.2.1 SLAM

Given that the simulation environment provides accurate camera depths and 2D laser depth measurements around the robot, we had two options to choose from: vision-based SLAM or 2D point cloud-based SLAM. We opted for the latter for several reasons.

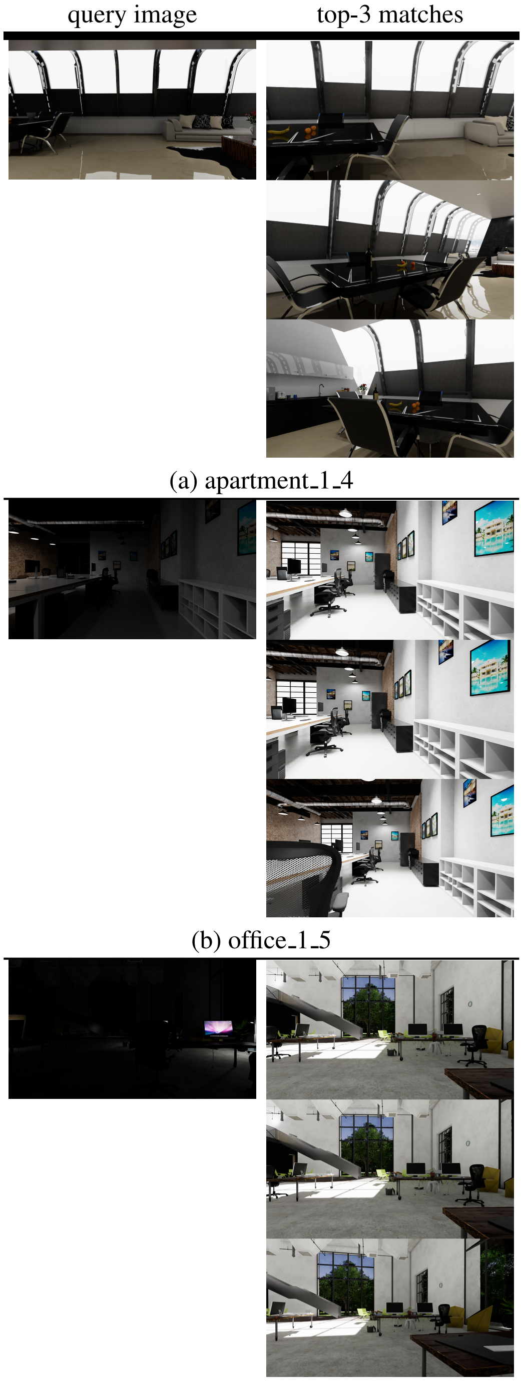
Firstly, the challenge requires scene change detection in an indoor environment over time. Lighting conditions could shift significantly (e.g., from bright daylight to a dark office room). When revisiting the scene, obtaining matches between the two environmental conditions is crucial. We found that traditional keypoint-based matching using SIFT [9] struggled to find matches between scenes and often completely failed. The deep learning-based LoFTR method [14] sometimes offers better matching but may fail in more difficult cases. We also tried using adaptive histogram equalization [10] as an attempt to improve matching [13] but did not observe any improvement. Moreover, based on what we observed, the LoFTR matching sometimes does not appear to provide ”pixel-perfect” matches. These observations are illustrated in Figs. 2 to 4. These matching are performed between image matches extracted via visual place recognition, as will be discussed in the next section. In contrast, the 2D points from the laser scans did not have the visual sensor's variations and were more straightforward to localize towards the previously constructed map.

Secondly, our active exploration method from the previous year's solution was based on frontier exploration, which finds frontiers based on a 2D map constructed using laser scans. Therefore, constructing a SLAM method that works in conjunction with the frontier map would be more efficient.

2.2.2 Long-term Visual Localization

In the scene change detection task, relocalizing our robot towards the previously constructed map (akin to a kidnapped robot situation) is necessary when the scene changes. We explored the scan context [8], commonly used in LiDAR point-cloud loop detection. However, it did not perform well, likely because of the sparsity of the 2D laser scans.

As an alternative, we then investigated the visual place recognition method and used CosPlace [2] to find the best image match. We chose Cosplace because of its ease of use and availability in the Pytorch hub. We used CosPlace to compute the image visual descriptor and rank the top matches from the previously observed scenes. We found that CosPlace often provided the correct scene match within the top-k closest matches (Fig. 5). Using the poses of the closest matches as the initial pose estimates, we use ICP to align the current scan and compute the point-cloud distances with the previously constructed map. We then choose the pose that yields the smallest average point distances. However, sometimes even the closest matches extracted by CosPlace are not entirely correct, leading to a situation where the minimum distances of the aligned scans remain large. In



(a) apartment_1_4

(b) office_1_5

(b) company_4_5

Figure 5. Visual place recognition using CosPlace. The left image is the query image, and the right images are the top 3 matches extracted by CosPlace.

these cases, we extract the next-closest images ranked by CosPlace until we find a pose with the aligned point distances below our predetermined threshold. Of course, we could also estimate the poses using vision-based feature matching. But, as observed in the previous section, it is not robust even with a state-of-the-art method like LoFTR.

After relocalizing our pose in the new scene, we can continue and use ICP to localize our robot within the previously constructed point cloud map. This efficient yet effective approach allows us to navigate through noisy pose estimation scenarios effectively.

3. Discussion

3.1. Current Limitations

Several limitations must be acknowledged in our RVSU challenge submission. Our solution is heavily based on the effectiveness of instance segmentation and 3D object detection models, which are the key components of our approach. Our attempts to upgrade these models to the latest state-of-the-art versions did not yield the expected improvements, leading us to surmise that these models may not transfer well to the BenchBot simulation environment.

Our localization system's performance relies on accurate 2D laser scans as well as visual place recognition to extract correct scene matches successfully. But the complexity of the scenario, especially in larger scenes, may sometimes result in suboptimal matches.

In addition, the components of the methodology, *i.e*. semantic map, localization, and active planning, all act independently. Integrating all these methods into a single framework could potentially result in a more optimal solution. These observations present areas for potential refinement.

References

[1] Antyanta Bangunharcana, Soohyun Kim, and Kyung-Soo Kim. Robotic vision scene understanding challenge: Msclab report. 1

[2] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 3

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[4] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1

[5] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6910–6919, 2021. 1

[6] David Hall, Ben Talbot, Suman Raj Bista, Haoyang Zhang, Rohan Smith, Feras Dayoub, and Niko Sünderhauf. The

robotic vision scene understanding challenge. *arXiv preprint arXiv:2009.05246*, 2020. 1

[7] David Hall, Ben Talbot, Suman Raj Bista, Haoyang Zhang, Rohan Smith, Feras Dayoub, and Niko Sünderhauf. Benchbot environments for active robotics (bear): Simulated data for active scene understanding research. *The International Journal of Robotics Research*, 41(3):259–269, 2022. 1

[8] Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4802–4809. IEEE, 2018. 3

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3

[10] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987. 3

[11] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. *arXiv preprint arXiv:2112.00322*, 2021. 1

[12] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Tr3d: Towards real-time indoor 3d object detection. *arXiv preprint arXiv:2302.02858*, 2023. 1

[13] Pranjay Shyam, Antyanta Bangunharcana, and Kyung-Soo Kim. Retaining image feature matching performance under low light conditions. In *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, pages 1079–1085. IEEE, 2020. 3

[14] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 3