Diagnosing Deep Self-localization Network: Accelerated Model-free Subimage-level Diagnosis via Image Synthesis and Relevance Feedback

Tanaka Kanji

***Abstract*— Deep convolutional neural network (DCN) has become a common approach in visual robot self-localization. In a typical self-localization system, a DCN is trained as a visual place classifier from past visual experiences in the target environment. However, its classification performance can be deteriorated when it is tested in a different domain (e.g., times of day, weathers, seasons), due to domain shifts. Therefore, an efficient domain-adaptation (DA) approach to suppress per-domain DA cost would be desired.**

In this study, we address this issue with a novel "domain-shift localization (DSL)" technique that diagnosis the DCN classifier with the goal of localizing which region of the robot workspace is significantly affected by domain-shifts. In our approach, the DSL task is formulated as a fault-diagnosis (FD) problem, in which the deterioration of DCN-based self-localization for a given query image is viewed as an indicator of domain-shifts at the imaged region. In our contributions, we address the following non-trivial issues:

(1) We address a subimage-level fine-grained DSL task given a typical coarse image-level DCN classifier, in which the target DCN system is queried with a region-of-interest (RoI) masked synthesized query image to diagnosis the RoI region;

(2) We extend the DSL task to a relevance feedback (RF) framework, to perform a further query and return improved diagnosis results; and

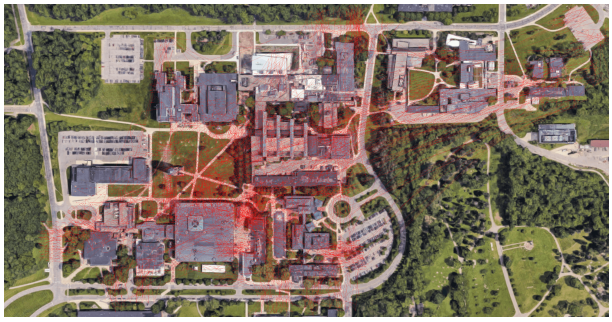
(3) We implement the proposed framework on 3D point cloud imagery -based self-localization and experimentally demonstrate the effectiveness of the proposed algorithm.

I. Introduction

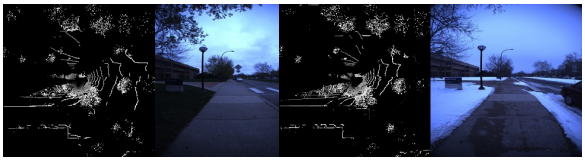
Deep convolutional neural network (DCN) has become a common approach in visual robot self-localization. In a typical self-localization system, a DCN is trained as a visual place classifier from past visual experiences in the target environment [1]. However, its classification performance can be deteriorated when it is tested in a different domain (e.g., times of day, weathers, seasons), due to environmental and optical effects, such as occlusions, dynamic objects, and confusing features. A possible method to maintain the self-localization performance is introducing a domain adaptation (DA) module that collects a new visual experience over the entire workspace and trains a new DCN classifier. However, such a naive DA method requires significant per-domain DA cost for training data collection and retraining, which increases in proportion to the environment size. Thus, an alternative efficient DA approach would be desired.

We address this issue with a novel "domain-shift localization (DSL)" technique that diagnosis the DCN classifier with the goal of localizing which region of the robot workspace is significantly affected by domain-shifts (Fig. 1). Such a DSL technique enables a DA module to focus the available resources on the domain-shifted region, rather than the entire workspace, which leads to significant reduction in the per-domain DA cost for data collection and retraining. In our approach, the DSL task is formulated as a fault-diagnosis (FD) problem as in our previous study [2], in which the deterioration of the DCN classifier for a given query image is viewed as an indicator of domain-shifts at the imaged region. Figure 1 shows an overview of the DSL task for a typical self-localization system based on 3D point cloud imagery.

In our contributions, we tackle several non-trivial issues: (1) We address the subimage-level fine-grained DSL. This is a non-trivial task because a typical DCN place classi-

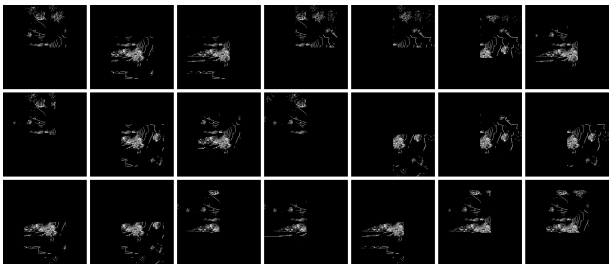


(a) robot workspace and viewpoints



query image/scene mapped image/scene

(b) raw images and corresponding scenes



(c) chain of synthetic queries

Fig. 1. Given a visual experience along a viewpoint trajectory (a), we aim to diagnosis a deep self-localization network with the goal of localizing which region of the robot workspace is significantly affected by domain shifts. Given raw images (b), effective synthetic images are generated and queried for subimage-level domain-shift localization (c).

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 17K00361, and (C) 20K12008.

K. Tanaka is with Faculty of Engineering, University of Fukui, Japan. tnkknj@u-fukui.ac.jp

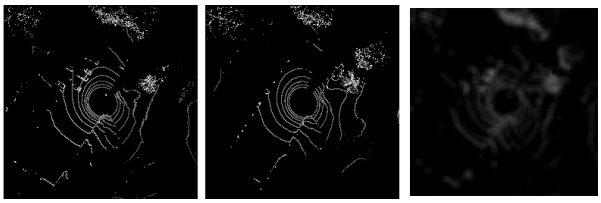


Fig. 2. DSL example of 3D point cloud imagery. Left: live image. Middle: map image. Right: localized domain-shifts.

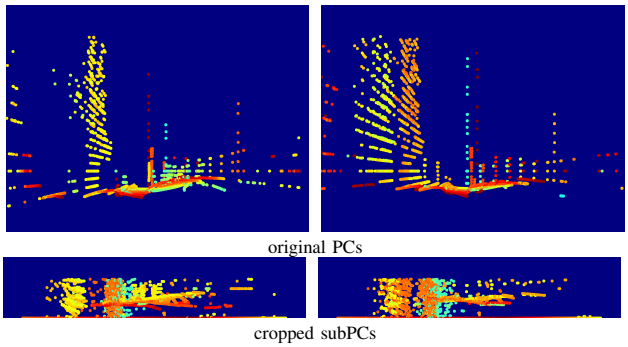


Fig. 3. PC preprocessing

fier provides only coarse image-level localization cues. To address this, we introduce a data augmentation method to generate synthesized query images that are effective for subimage-level DSL. (2) We further consider an efficient relevance feedback (RF) framework, derived from the field of content-based image retrieval (CBIR) [3], to reduce the total number of queries and accelerate the diagnosis process. In the RF framework, each query result is used as a cue to perform a further query and return improved diagnosis results. (3) We implement the proposed framework on a 3D point cloud imagery -based self-localization (Fig. 2) and experimentally demonstrate the effectiveness of the proposed algorithm using publicly available NCLT dataset.

II. Approach

This section presents the FD-based DSL approach. We first introduce our experimental system of the deep self-localization network to clarify our DSL task (Section II-A), and formulate the DSL problem (Section II-B). Then, we describe the proposed DSL method for subimage-level diagnosis (Section II-C). Further, we extend it to a relevance feedback framework to accelerate the diagnosis process (Section II-D). Figure 4 illustrates the overview of the DSL systems.

A. Experimental System

Although our framework is sufficiently general and applicable to a broad class of DCN systems, our experiments will focus on 3D point cloud imagery -based DCN place classifier. The experimental system uses an on-board LIDAR as the sole sensor system (Fig. 3), which is available in the NCLT dataset [4]. Similar with the recently developed scan context technique [1], a 2D grid (spatial resolution: 0.05m$\times$0.05m) is imposed on the horizontal plane and datapoints that belong to each grid cell are represented by their maximum normalized truncated height (MNTH). That is, height values of the datapoints are truncated in range [0.5, 1.5], normalized to the range [0, 1], and then max pooled. However, instead of the polar-coordinate, the bird's eye view image coordinate is used as the coordinate for the per-cell MHNT data, to make the image-to-subimage translation (Section II-B) more straightforward. Figure 3 shows an example of original and preprocessed PCs.

The DCN place classifier aims to classify a given query image into one of predefined place classes. It is trained via unsupervised and supervised learning stages. First, the unsupervised learning stage aims to partition the robot workspace into place classes. Although the problem of place partitioning (PP) is significantly ill-posed and an on-going research topic as addressed in our recent paper [5], in this work, we use a simple standard PP method that partitions the workspace into grid cells (resolution: 40m$\times$40m) and defines each grid cell as a place class. Next, the supervised learning stage aims to fine-tune a pretrained DCN with a place-specific train image set. In implementation, we use one specific season in the NCLT dataset [4] as a train set, and as a result, we obtain around 100 place classes and 258 train images per class on average. For the train set, each image is annotated with its corresponding place label, by using GPS information that is also available in the NCLT dataset. For the fine-tuning, SGD optimizer (learning rate: 10^{-4}, momentum: 0.9) is used.

B. Problem Formulation

The DSL task is formulated as follows. Suppose a DCN is pre-trained in the procedure described in II-A, as a place classifier $C[\cdot]$ that maps a given query image x to a class-specific probability density vector (PDV) $C[x]$. Suppose we are given a sequence of live images $(x^1,\cdots,x^L)$ along the viewpoint trajectory. Suppose we are given the ground-truth place class $t[x^j]$ for each live image x^j $(j \in [1,L])$, from an external positioning system such as structure-from-motion [6] or SLAM [7]. Then, a passive DSL task can be formulated as follows: Given a live image x with its ground-truth place class $t[x]$, generate an effective set of synthesized query images

$$X = f(x) \tag{1}$$

from x, and diagnosis the output PDV $C[x']$ for each query input x' $(x' \in X)$ to estimate the likelihood $Y[x'](p)$ of each pixel p in the imaged region being affected by domain-shifts:

$$Y[x'] = g(x', t[x'], C[x']). \tag{2}$$

Then, an active DSL task can be formulated as follows: Given a sequence of previous queries and diagnosis results $q[x'_1],...,q[x'_{i-1}]$, where $q[x'] = (x', t[x'], C[x'], Y[x'])$, select an effective next i-th query

$$x'_i = h(q[x'_1],...,q[x'_{i-1}]), \quad x'_i \in f(x^1) \cup \cdots \cup f(x^L) \tag{3}$$

among all the possible synthesized images to perform a further query and return improved results. To design the above generator function f, diagnosis function g, and selection

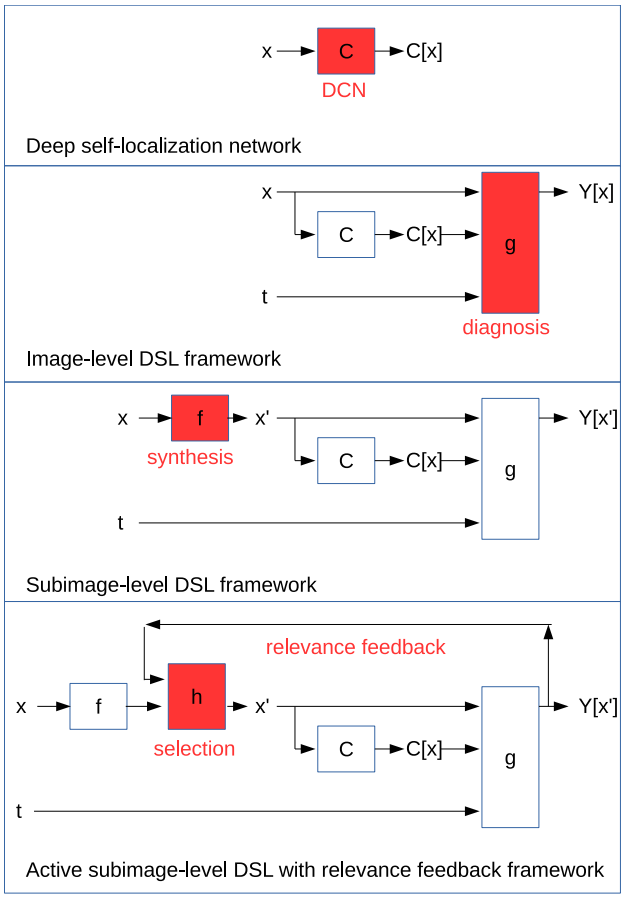


Fig. 4. System integration of deep self-localization network, diagnosis, generation, and selection functions.

function h is not a trivial issue and is the main issue of this study.

In this work, the PDV $C[x]$ is used in two ways. First, it is translated to an inverse rank (IR) list of place classes in the descending order of place-specific probability, which can be viewed as a main output of the self-localization module. Second, the entropy $E[C[x]]$ is computed from the PDV, and then its inverse is used as an evaluation of confidence of the estimate. This confidence value is used for causal analysis of the self-localization performance deterioration. That is, when the confidence value is lower than a pre-set threshold 1.0, our diagnosis system simply ignores and does not reflect the PDV to the diagnosis result.

C. Subimage-level Diagnosis

Realizing a fine-grained subimage-level DSL from a coarse image-level DCN is not a trivial issue. In our approach, we propose to use a masked image as a synthesized image. This strategy is inspired by image masking approach [8] in the field of image saliency, where the target DCN system is queried with an RoI masked query image to analyze saliency of the RoI region. In our approach, we create a synthesized image by replacing the height value with zero at every pixel outside the RoI. Although simple, this strategy

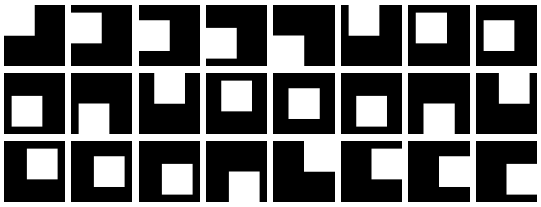


Fig. 5. The 24 masks used for image synthesis.

frequently provides a realistic synthesized image (e.g., Fig. 1(c)), as the original input MNTH image is already truncated and normalized to the range of [0, 1]. A possible direction of future research is to replace the simple binary mask with a more advanced fuzzy mask, such as adversarial erasing, to improve the quality of the subimage-level DSL.

In a subimage-level diagnosis framework, one pixel may belong to multiple overlapping RoI regions, and thus receive multiple different IR values from such overlapping RoIs. Here, we fuse such multiple IRs into a single IR value, with a similar information fusion method as in our previous work in [2], which is theoretically supported by multi-modal information retrieval in [9]. That is, we start with a set of multiple diagnosis results in the form of multiple PDF vectors. Then, we average the IR r_c^{-1} for each class c over all the related diagnosis results, and output the averaged IR $(\sum_c r_c^{-1})^{-1}$ as the final pixel-level diagnosis result.

D. Relevance Feedback

One possible way to improve the diagnosis performance is to increase the number of synthesized images queried. However, this naive strategy comes with high computational cost, which is proportional to the cost for DCN evaluation and the number of synthesized images queried. To accelerate the diagnosis process, we here consider an active diagnosis task with the RF framework, derived from the field of CBIR [3]. However, a key difference of our RF task is that the original train image set is no longer accessible but only the trained DCN is available, which makes our RF task more challenging.

The active diagnosis framework is formulated as follows. For simplicity, we discretize the mask space and obtain $M = 24$ RoI mask candidates, which are described in Fig. 5. Suppose we are given a set of N live images. Thus, we have $M \times N$ RoI candidates in total. Our goal is to intelligently query a small portion (e.g., 10 percent) of effective candidates, instead of exhaustively querying for all possible RoI candidates.

We implemented three different strategies for such an RoI sampling task. One is "random" strategy that samples a new RoI randomly and independently of the previous samples. The second strategy is "nearest-neighbor (NN)" strategy that is basically the same as the "random" strategy but if the previous query is "successful", it samples a new RoI from the same image as the previous query, rather than from all the input images. Here, the condition for "successful" query is that its rank value should be equal or lower than a pre-

learned threshold (1.0). The third strategy is "spatial NN" strategy that is basically same as the NN strategy, but if the previous query is "successful", it prioritizes the non-visited RoI candidates by spatial nearness $|p^{new} - p^{prev}|^{-1}$ of each RoI's center p^{new} to the previous RoI's center p^{prev}, and samples a non-visited RoI with the nearest spatial location $p^{new} = \arg\min_{p^{new}} |p^{new} - p^{prev}|^{-1}$.

III. Experimental Evaluation

We evaluated the proposed methods using NCLT cross-season dataset [4]. The NCLT dataset is a large-scale, long-term autonomy dataset for robotics research collected at the University of Michigan's North Campus by a Segway vehicle robotic platform for 15 months, 27 sessions, in both indoor and outdoor. During the navigation, the vehicle encounters several different environment changes, including car parking, falling leaves, building construction, and furniture movement. These environmental changes can be main causes of self-localization performance deterioration. In the current experiments, we used three pairings of query (i.e., test) and map (i.e., train) seasons from the NCLT dataset, (2012/03/31, 2012/01/22), (2012/08/04, 2012/03/31), and (2012/11/17, 2012/08/04), and for each dataset, a size 1,500 image set is randomly sampled. Given a wide-range 70m radius view provided by the 3D point cloud imagery, we crop the image into 12.8m$\times$ 12.8m image region with spatial resolution of 0.05m, which is centered at the robot viewpoint, and whose x-axis is aligned with the heading direction of the robot.

Figure 6 shows example diagnosis results. As shown in Fig. 1(c), the proposed active diagnosis framework generates an effective chain of queries based on the previous query responses. The locations of each synthesized image in the sequence with respect to the bird's eye view coordinate is visualized in Fig. 1(a). Effectiveness of such an active diagnosis technique will be shown in the following quantitative evaluation.

For quantitative performance evaluation, we prepare a ground truth diagnosis result for each test image. The ground truth is in the same data structure as the diagnosis result image, but the pixel value is determined by using the fine-grained GPS and RANSAC based point cloud registration and precise pixel-wise pairwise 3D point cloud comparison, based on RANSAC scoring using subPCs truncated in height [-2, 2]. Suppose we are given a pre-defined hyper-parameters Th and $Height$. First, each grid cell (spatial resolution: 0.05m$\times$0.05m) in the ground-truth data is labeled as "domain-shifted" if height difference between the live and map images at that grid cell is greater than the preset $Height$ value. Then, an image region in the ground-truth data is labeled as "domain-shifted" if the ratio of grid cells with "domain-shifted" labels within the region exceeds the threshold Th.

For the performance index, we used pixel-wise precision-recall. This index is computed as follows. First, every query image is discretized into a regular grid with the spatial resolution of 0.3m. Then, all the cells in the image are sorted in the descending order of the likelihood of domain-shifts. Then, 51-point interpolated mean-Average-Precision (mAP) is computed over the top-50% recall points. Figure 7 shows the performance results for a specific (default) setting of the hyper-parameters, $Height$ and Th.

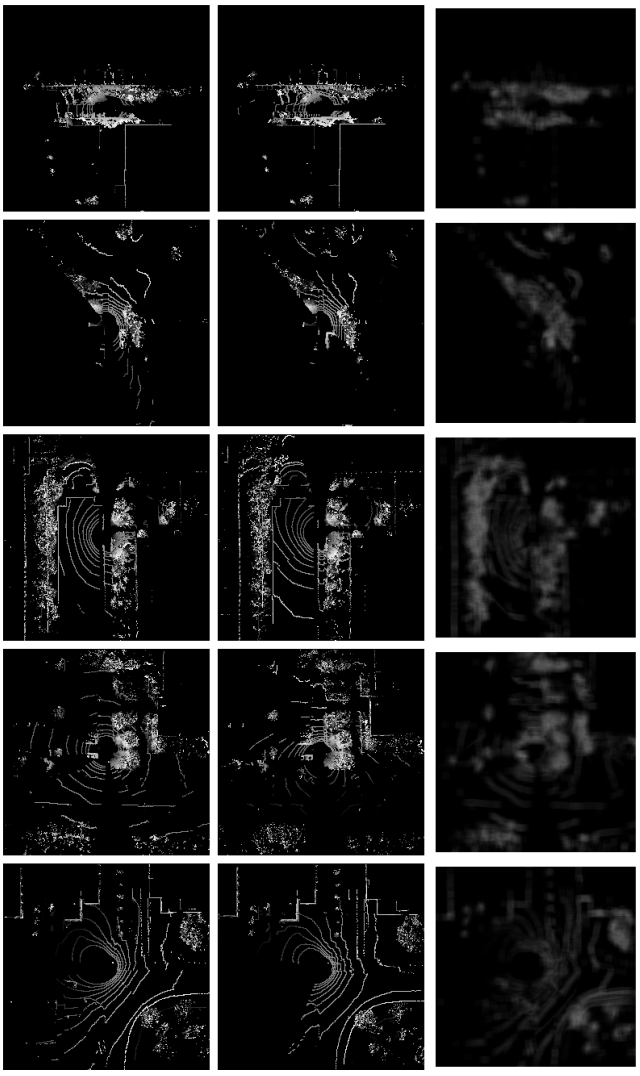


Fig. 6. Examples of DSL results (left: query, middle: map, right: pixel-level diagnosis result).

For performance comparison, we implemented and tested six different strategies: "FULL", "RANDOM", "RANK>1", "RANK>AVG", "RANK>1 CHAIN", and "RANK>1 CHAIN SPATIAL". The "FULL" method is the subimage-level DSL method that utilizes all the available RoI candidates in exhaustive manner. The "RANDOM" method randomly samples and uses 10% of RoI candidates. The other four methods also samples and uses 10% of RoI candidates, but in an intelligent manner within the relevance feedback framework explained in Section II-D. Among them, the "RANK>1" method samples the next RoI from the same image if the current query is successful, otherwise randomly. The "RANK>AVG" method is similar with "RANK>1", but

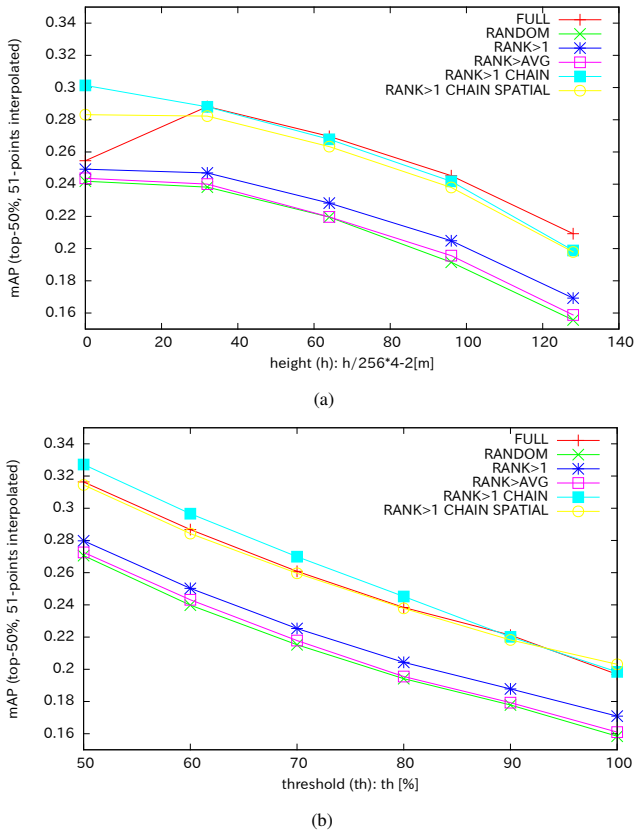


Fig. 7. Performance results for the default setting (*Th*: 80, *Height*: 64).

the definition of "successful" is different, that is, average of the rank values of all the queries is used as the threshold instead of the default definition (1). The "RANK>1 CHAIN" method is similar with "RANK>1", but the next query is sampled from the same image as long as the query response is "successful". The "RANK>1 CHAIN SPATIAL" is similar with "RANK>1 CHAIN", but the NN method is used to prefer the spatial nearest queries, as explained in II-D.

As can be seen, the proposed active diagnosis method "RANK>1 CHAIN SPATIAL" or "RANK>1 CHAIN" outperforms almost all the other methods, except for the exhaustive "FULL" strategy, and it is competitive to the exhaustive strategy despite the fact that the proposed method is ten times more efficient in terms of the number of synthesized images queried. On the other hand, the "RANDOM" strategy often fails to generate effective samples, which led to the worst performance result. On the other hand, the proposed method was successful in generating an effective chain of queries that yield much higher performance.

IV. Conclusions

In this study, we addressed the novel problem of domain-shift localization by fault-diagnosing a deep self-localization network. The key idea is that the deterioration of DCN-based self-localization for a given query image is viewed as an indicator of domain-shifts at the imaged region. In our contributions, several non-trivial issues were addressed: (1) We addressed a subimage-level fine-grained DSL task given only an typical coarse image-level DCN classifier. (2) We extended the DSL task to a relevance feedback framework, in which each query response is used as a cue to make the next query more effective. (3) We implemented the proposed framework on 3D point cloud imagery -based self-localization and experimentally demonstrated the effectiveness of the proposed algorithm.

References

[1] G. Kim, B. Park, and A. Kim, "1-day learning, 1-year localization: Long-term lidar localization using scan context image," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1948–1955, 2019.

[2] T. Kanji, "Detection-by-localization: Maintenance-free change object detector," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4348–4355.

[3] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on circuits and systems for video technology*, vol. 8, no. 5, pp. 644–655, 1998.

[4] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[5] K. Tanaka, "Self-supervised map-segmentation by mining minimal-map-segments," in *2020 IEEE Intelligent Vehicles Symposium*. IEEE, 2020.

[6] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[7] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1168–1174.

[8] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[9] M. Imhof and M. Braschler, "A study of untrained models for multimodal information retrieval," *Information Retrieval Journal*, vol. 21, no. 1, pp. 81–106, 2018.