

Fault-Diagnosing DCN-SLAM for 3D Change Object Detection: A Method based on Masking Input Images

Tanaka Kanji

Abstract—Although image change detection (ICD) methods provide good detection accuracy for many scenarios, most of the existing methods rely on place-specific background modeling. The time/space cost for such place-specific models becomes prohibitive for large-scale scenarios, such as long-term robotic visual simultaneous localization and mapping (SLAM). Therefore, we propose a novel ICD framework that is specifically tailored for long-term SLAM. This study is inspired by the multi-map-based SLAM framework, where N multiple localizers are capable of mutual diagnosis, thus not requiring any explicit background modeling/model. We extend this multi-map diagnosis approach toward a more generic single-map-based object-level diagnosis framework (i.e., ICD), where state-of-the-art self-localization systems can be used in their original form, which is as the change object indicator. The available single localizer is extended to different N localizers by introducing different N masked input images. Further, we also consider map diagnosis on a state-of-the-art deep-visual-SLAM system (rather than on conventional bag-of-words or landmark-based systems) in which the blackbox nature of the deep convolutional neural network (DCN) complicates the diagnosis problem. We also consider a 3D point cloud (PC)-based SLAM, and for the first time (to the best of our knowledge) adopt the state-of-the-art scan context PC descriptor for the purpose of map diagnosis.

I. INTRODUCTION

Background modeling plays a crucial role in image change detection (ICD) methods. A significant amount of work has been conducted on background modeling, which has led to the development of several interesting and effective algorithms [1]–[9]. However, most of the existing methods rely on place-specific background models such as place-specific background images [10] or place-specific background classifiers [11]. The time/space cost for place-specific modeling/models grows linearly with the environment size and becomes prohibitive for large-scale applications, such as long-term robotic visual simultaneous localization and mapping (SLAM).

Therefore, we consider a new ICD problem that is specifically tailored to long-term robotic visual SLAM. Many recent SLAM scenarios (e.g., office robots [12], car robots [13]) require the ability to detect changed objects in a live image with respect to background images or a pre-trained background model (i.e., ICD) to maintain an environment map that can change over time. Long-term SLAM, in particular, is self-supervised scenario [14]. It needs to maintain the map database along with frequently updating the background model by reflecting each new detected change [15].

We address this issue by adopting the multi-map-based SLAM [14], which does not require an explicit background model or modeling. In the seminal work of [14] and other

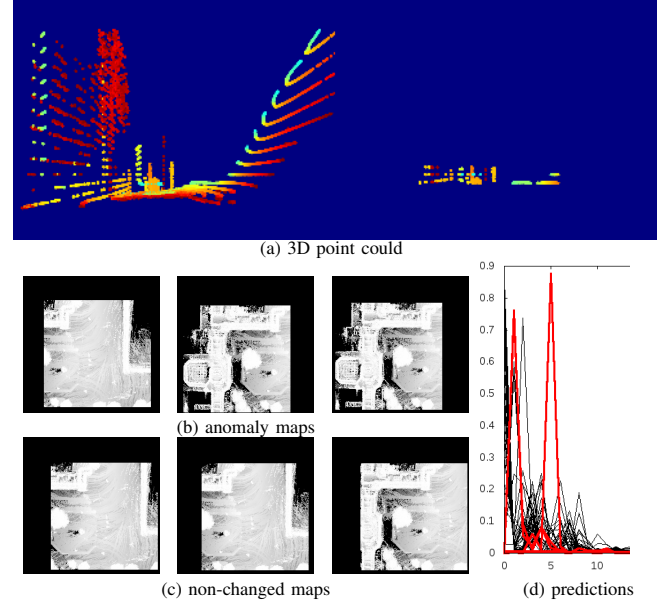


Fig. 1. A new object-level image change detection (ICD) framework that does not require the maintenance of any external background modeling/model, but exploits the available single-map-based SLAM system as the sole background model. In our approach, the input 3D map is truncated (a) and described using a scan context descriptor (SCD). Further, the SCD is used to generate multiple synthetic maps (b,c). Further, individual synthetic maps are fed to the SLAM (i.e., DCN) to obtain the posterior probability distribution over place class (d). Finally, synthetic maps are classified into anomaly (b) or non-change (c) depending on whether they are consistent with other synthetic maps.

recent studies [16]–[18], a multi-map-based SLAM framework was demonstrated. This involves multiple localizer engines trained for different experiences (i.e., maps) based on different domains (such as seasons, time of day, and weather conditions) in the same target environment. During self-localization stage, the maps mutually perform diagnosis on each other to identify erroneous maps that are inconsistent with the other maps and, hence, require modification (or replacement). In particular, this map diagnosis mechanism does not require an explicit background model but only requires the available SLAM to serve as the sole background model. Furthermore, it can directly detect the degradation in map quality in terms of self-localization performance.

We extend the map diagnosis approach onto the ICD application (Fig. 1¹). The key differences between our method and the multi-map-based approaches are as follows: (1) Instead

¹In the panel (d), the horizontal and vertical axes are place ID and the DCN's output probability, respectively. For the sake of visualization, the place ID is sorted in the descending order of $\sum_{i=1}^N p_i$, where N is the number of synthetic maps and p_i is the posterior probability of the class ID i .

of aiming the detection at map-level, we focus the detection on more detailed, object-level map errors (i.e., ICD), and we require just a single map (not multiple maps). (2) We consider a deep convolutional neural network (DCN) -based (rather than the conventional landmark-based [19] or bag-of-words-based [20]) maps, where the black box nature of the DCN complicates the map diagnosis problem. (3) We implement our idea on a 3D point cloud (PC) -based SLAM (rather than a typical monocular color image -based one), and for the first time (to the best of our knowledge) adopt the state-of-the-art scan context PC descriptor in [21] for map diagnosis.

Our main contributions through this study are summarized as follows. First, this paper extends the multi-map-based map-level diagnosis framework onto the general single-map-based object-level diagnosis (i.e., ICD) framework. Second, our map diagnosis framework is the first of its kind to address the black box nature of the state-of-the-art DCN-based SLAM and the scan context 3D PC descriptor. Third, our extensive experiments using the publicly available north campus long term (NCLT) cross-season dataset [22] (Fig. 2) have shown that the proposed framework is well suited for challenging cross-domain long-term SLAM scenarios [23].

II. APPROACH

A multi-map-based SLAM system can be treated as a mutual diagnosis system. Prior to the self-localization stage, multiple N visual experiences (i.e., maps) $M_1, \dots, M_N$ of the target environment are collected in different N domains (e.g., seasons, time of day, weathers) and multiple N localizer engines $L_1, \dots, L_N$ are trained on these individual maps. During the self-localization stage, a query live map Q (i.e., local map) is independently fed to the multiple localizer engines, and various N responses from the multiple localizers are aggregated into a single final prediction. As a byproduct, an erroneous map is identified when responses of its corresponding localizer is highly inconsistent with those of the other localizers.

For our study, let us assume an environment or the robot workspace as a set of C pre-defined place classes. We define a localizer as a ranking function that undergoes permutation and assigns ranks $1, \dots, C$ to each place class according to its relevance to a localizer with a query live map. A map-localizer pair (M_i, L_i) is said to be consistent if its ground-truth place class in M_i is top-ranked by the localizer L_i . In other words, such an inconsistency of a pair (M_i, L_i) can be viewed as an indicator for potential failure mode of a map M_i . Similarly, it is likely that a map M_i is in the failure mode if it is inconsistent for several other pairs.

We extend this idea to address a more general case of single-map-based object-level map diagnosis. The idea is to obtain different N virtual localizers $L_1, \dots, L_N$ from the available pre-trained localizer L . In short, each i -th localizer L_i is approximated by a virtual localizer L'_i that takes a masked query image as input. The masking operation erases sub-region of a given query map Q with a binary region mask (Fig. 3). The size of the resulting N synthetic maps is same

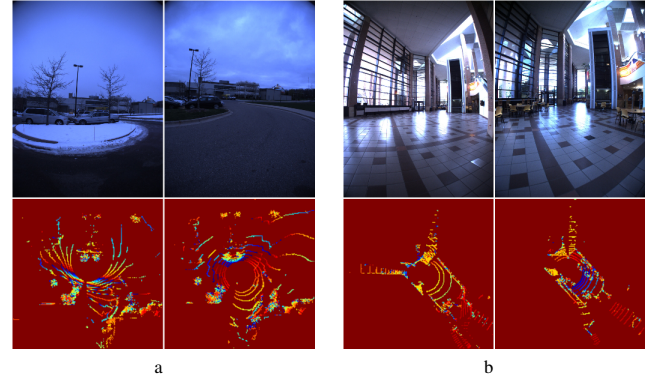


Fig. 2. Examples of scenes and point clouds (PCs). (a),(b): The left and right panels show map and live scenes/PCs, respectively.

as that of the original map Q , and its discriminative power is usually weaker and depends strongly on the change ratio of the non-masked region. Through the experiments, we test with several different mask sizes, 50, 100, 120, 150, 200, 220, and 250 as shown in Fig. 3.

We independently sample these N region masks. Therefore, a masked region may partially overlap with other masked regions. In such cases, we fuse these multiple responses at the overlapping region to determine a single final prediction for that region (Section II-D).

In the remainder of this section, we will provide a more detailed description of map representation, localizer, masking, and consistency evaluation.

A. Map Representation

We adopt the scan context descriptor (SCD) [21] as our map representation. For simplicity, we assume that the roll and pitch rotations of the 3D PC sensor with respect to the world coordinate system are negligible, which frequently occurs for car robot scenarios such as the NCLT dataset. The SCD extraction procedure involves three steps. Firstly, we eliminate useless scan points that might have originated from the ground plane and that are too far from the sensor viewpoints. For this, a raw PC is truncated in height $[0.5, 1.5]$ m and represented using a 2D horizontal elevation grid map in an ego-centric overhead image coordinate, with size 256×256 and resolution 0.3m. Further, the grid map is represented using an SCD image whose pixel value is the height at the cell.

We observe that the SCD-based DCN localizer has an attractive feature as a change indicator. In a preliminary experiment, we tested the DCN localizer under various scenes with different ratios of changes, and determined that the SCD-based localizer has the following two distinctive behaviors: (1) the localization is frequently successful when the ratio of changed regions (RoC) over the mapped region is small, and (2) the localization is frequently not successful when the RoC is significantly large. Intuitively, this clear difference and dependency of the behavior acts as a good indicator of changes in our approach.

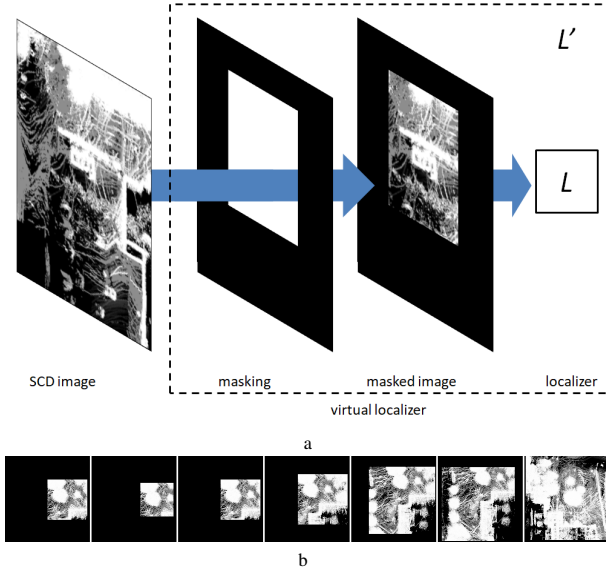


Fig. 3. Masked maps. (a) A virtual localizer L'_i is approximated by the combination of the available pre-trained localizer L and an i -th masking operation, without re-training the localizer L . (b) From left to right, each panel corresponds to the mask size 50, 100, 120, 150, 200, 220, and 250.

B. Localizer

We formulate self-localization as a visual place classification task. This formulation requires to partition the robot workspace into place classes. The concept of place partitioning is an important topic of on-going research in the field of robot self-localization (e.g., [24], [25]), as was demonstrated in our previous study [26]. In this study, we loosely follow the relatively simple equal-spaced partitioning method [27], and partition the workspace into $40\text{m} \times 40\text{m}$ square regions. As a result, each place class is assigned a unique place ID and a subset of the training map images that belong to the grid cell.

We train a DCN as a place classifier. The place-specific training images for each place class are converted to the SCD format. A VGG16 classifier network [28] is adopted as the classifier network. The learning rate is set 1×10^{-4} and the momentum is 0.9. The SGD optimizer is used. We believe that our approach could be generalized to deeper powerful networks such as ResNet [29], or more specialized localizer network such as NetVLAD [30].

C. Region Masks

The field-of-view (FoV) of typical 3D PC sensors (such as Lidar) is significantly larger when compared to the sizes of typical change objects (such as cars and furniture). When using such a raw scan, the SCD-based localizer is frequently successful and might not fail even when change objects exist in the scene. This could be problematic for collecting failure experiences (the training set) to train our failure detector (FD-based ICD).

To address this issue, we introduce a synthetic map with a narrower FoV (Fig. 3), called weak feature map (WFM). A WFM is generated by erasing a subregion of an SCD image,

by masking the subregion to have zero pixel value (such as height). Thus, it will have weaker discriminative power than the original SCD. We use WFM in place of the original SCD in the DCN localizer. The WFM-based DCN localizer is expected to be successful owing to the inherent robustness of DCN against occlusions (i.e., erasing) [31] as long as the non-masked region is dominated by change objects. In other words, the non-masked region is likely to be dominated by change objects if the WFM-based localizer is non-successful.

Preliminary experiments on different combinations of region masks including fixed/flexible masks and random/pre-defined masks showed that flexible random region masks often have the best performing strategy. This strategy implements the extraction of WFM as a logical conjunction operation. It should be noted that the erasing task can be addressed using techniques other than masking, such as adversarial erasing [32] or scene completion [33]. The testing of such alternative erasing techniques in our map diagnosis approach will be addressed in our future study.

D. Consistency Evaluation

This section discusses the treating mechanism of overlapping masked regions. For multiple overlapping masked regions, the rank values are fused at that region. We follow our study in [20] and adopt the unsupervised rank fusion strategy derived from multi-media information retrieval [34]. A set of rank values $r_1, \dots, r_k$ for the overlapping region is fused in the form: $\bar{r} = 1 / \sum_{i=1}^k s_i$ where $s_i = (1/r_i)$. We further adopt the entropy based novel scene detection method [27] to our object-level map diagnosis purpose. An input WFM is categorized as a novel scene part (i.e., change object) if the entropy value $E = -\sum_{c=1}^C p_c \log p_c$ computed from a DCN output $\{p_c\}_{c=1}^C$ is higher than an empirical threshold value of 1.0. A high entropy value is determined to serve as a reliable indicator of change, leading to the modification of the value $s_i \rightarrow s_i + 1$.

III. EXPERIMENTS

Various experiments were conducted to evaluate the proposed FD-based map diagnosis. We performed sensitivity analysis with respect to its hyper-parameters. For the ICD task, we employed mean average precision (mAP) as the performance metrics.

We evaluated the method under a challenging scenario of cross-season long-term SLAM, where the training and test sets contained images from different seasons (Fig. 4). We followed the work discussed in [35] and used four datasets “2012/01/22 (WI)”, “2012/03/31 (SP)”, “2012/08/04 (SU)”, and “2012/11/17 (AU)”. Fig. 2 shows examples of the scenes and point cloud data. We use three successive season pairs, which are “WI-SP”, “SP-SU”, and “SU-AU”, as the pairings of training and test sets. We used the publicly available VGG16 architecture [36] as the CNN classifier. We fixed the grid map size to 256×256 using a spatial resolution of 0.3 m. Sufficiently large local maps are created from the train set and such fixed-size grid maps are randomly sampled from the local maps. We pre-trained the CNN weights on

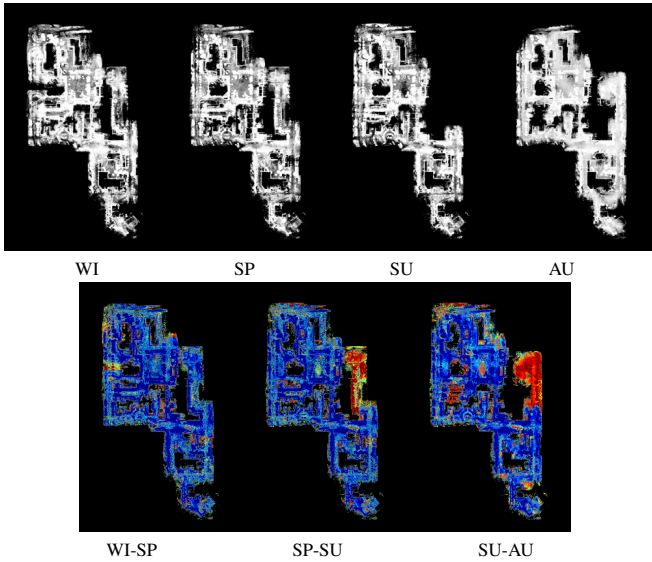


Fig. 4. Global environment maps and ground-truth changes for four seasons. Each cell is colored using a heatmap, and ranges from 0 (blue) to 1.5 (red).

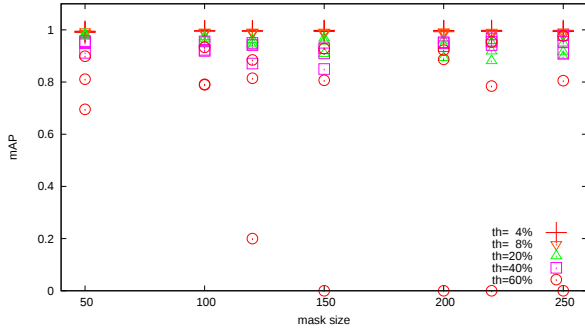


Fig. 5. Performance results in terms of mAP versus mask size (CR=30[%]).

ImageNet ILSVRC dataset [28], and then fine-tuned them with individual training sets. The training set consists of 40-60 classes and around 100 training images per class.

The sensitivity analysis of the WFM-based DCN localizer was performed with respect to its three hyper-parameters: mask size (MS), change threshold (TH), and change ratio (CR). MS defines the size of mask, which influences the size of change objects that can be detected by the proposed method. TH defines the threshold [%] on the difference in cell value normalized by its maximum value, which splits change and no-change cells. CR defines the threshold [%] on the ratio of changed cells, which splits change and no-change maps. For our experiments, we set MS=100[pixels], TH=20[%], and CR=30[%] as the default values.

We computed the ground-truth change objects using a fine-grained background subtraction technique. Formally, we implemented a background subtraction method consisting of two distinct steps: registration and differencing [37]. It should be noted that the first registration step is essentially an ill-posed problem, which requires accurate pixel-level

viewpoint localization. During implementation, we assume a perfect registration and approximate it by using the GPS information available in the NCLT dataset, followed by fine-grained registration. We emphasize that the proposed method does not rely on such fine-grained viewpoint information. Instead, it only requires more coarse viewpoint information (such as the place class ID), provided by an external localization engine (as found in [27]).

In Fig. 5, we compare seven different settings of the mask size (MS): 50, 100, 120, 150, 200, 220, and 250. For each setting, three mAP values corresponding to the three season pairs “WI-SP”, “SP-SU”, “SU-AU” are plotted in the graph. The best performance is achieved when MS=50, 100, 120, and 150. The primary reason for this is the failure of self-localization with such a small mask, which acts as a good indicator of existence of a significant change, as discussed in II-A.

In Fig. 5, we also compare the influence of the threshold TH to mAP performance. We find that the proposed method is successful for the settings TH=4, 8, and 20 [%]. On the other hand, the proposed method failed for the settings TH=40 and 60 [%]. A main reason is that failure mode of the self-localization system tends to occur even with much weaker changes (i.e., $TH \leq 20$).

IV. CONCLUSIONS

In this paper, we have proposed a map diagnosis framework for long-term robotic visual SLAM. Our framework extends the multi-map-based map-level diagnosis and explicitly considers single-map-based object-level map diagnosis. This enables the use of a state-of-the-art self-localization system in its original form as a change object indicator. Unlike the existing ICD approaches, the proposed FD-based ICD approach does not require fine-grained environment maps or fine-grained viewpoint information. Our observation indicated that we observed that the proposed ICD method performs quite well compared to previous approaches. The idea behind the proposed approach is the generation of multiple synthetic object-level maps from an available single map and performing mutual diagnosis on these synthetic maps to find inconsistent objects (change objects). We have demonstrated the effectiveness of the proposed framework on challenging cross-season NCLT dataset using the state-of-the-art self-localization system, which is based on scan context descriptor and deep convolutional neural network. In the future, our framework will be applied to more general self-localization frameworks such as multi-hypothesis tracking [38] and Monte Carlo localization [39].

REFERENCES

- [1] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, “Real-time foreground-background segmentation using codebook model,” *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.
- [2] M. Shah, J. D. Deng, and B. J. Woodford, “A self-adaptive codebook (sacb) model for real-time background subtraction,” *Image and Vision Computing*, vol. 38, pp. 52–64, 2015.
- [3] M. Wu and X. Peng, “Spatio-temporal context for codebook-based dynamic background subtraction,” *AEU-International Journal of Electronics and Communications*, vol. 64, no. 8, pp. 739–747, 2010.

- [4] A. Taneja, L. Ballan, and M. Pollefeys, "Geometric change detection in urban environments using images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2193–2206, 2015.
- [5] E. Palazzolo and C. Stachniss, "Fast Image-Based Geometric Change Detection Given a 3D Model," in *Proc. the IEEE Int. Conf. Robotics and Automation (ICRA)*, 2018.
- [6] L. Luft, A. Schaefer, T. Schubert, and W. Burgard, "Detecting changes in the environment based on full posterior distributions over real-valued grid maps," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1299–1305, 2018.
- [7] D. Fox, W. Burgard, and S. Thrun, "Markov localization for mobile robots in dynamic environments," *Journal of artificial intelligence research*, vol. 11, pp. 391–427, 1999.
- [8] J. Andrade-Cetto and A. Sanfeliu, "Concurrent map building and localization on indoor dynamic environments," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 03, pp. 361–374, 2002.
- [9] L. Gueguen and R. Hamid, "Large-scale damage detection using satellite imagery," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1321–1328.
- [10] M. Babaei, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.
- [11] P. Christiansen, L. N. Nielsen, K. A. Steen, R. N. Jørgensen, and H. Karstoft, "Deepanomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field," *Sensors*, vol. 16, no. 11, p. 1904, 2016.
- [12] L. J. Manso, P. Núñez, S. d. Silva, and P. Drews-Jr, "A novel robust scene change detection algorithm for autonomous robots using mixtures of gaussians," *International Journal of Advanced Robotic Systems*, vol. 11, no. 2, p. 18, 2014.
- [13] M. Fehr, F. Furrer, I. Dryanovski, J. Sturm, I. Gilitschenski, R. Siegwart, and C. Cadena, "Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery," in *Robotics and Automation (ICRA), 2017 IEEE Int. Conf. IEEE*, 2017, pp. 5237–5244.
- [14] W. Churchill and P. M. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *IEEE Int. Conf. Robotics and Automation, ICRA 2012, 14-18 May, 2012, St. Paul, Minnesota, USA*, 2012, pp. 4525–4532.
- [15] M. Bürki, M. Dymczyk, I. Gilitschenski, C. Cadena, R. Siegwart, and J. I. Nieto, "Map management for efficient long-term visual localization in outdoor environments," in *2018 IEEE Intelligent Vehicles Symposium*, 2018, pp. 682–688.
- [16] P. Mühlhoffner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, and P. Furgale, "Summary maps for lifelong visual localization," *Journal of Field Robotics*, vol. 33, no. 5, pp. 561–590, 2016.
- [17] M. Paton, K. MacTavish, M. Warren, and T. D. Barfoot, "Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat," in *2016 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1918–1925.
- [18] M. Bürki, M. Dymczyk, I. Gilitschenski, C. Cadena, R. Siegwart, and J. I. Nieto, "Map management for efficient long-term visual localization in outdoor environments," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 682–688.
- [19] T. Sugimoto, K. Yamaguchi, and K. Tanaka, "Fault-diagnosing SLAM for varying scale change detection," *CoRR*, vol. abs/1909.09592, 2019.
- [20] K. Tanaka, "Detection-by-localization: Maintenance-free change object detector," in *Int. Conf. Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, 2019, pp. 4348–4355.
- [21] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *2018 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4802–4809.
- [22] 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*, pp. 1023–1035, 2015.
- [23] —, "University of Michigan North Campus long-term vision and lidar dataset," *International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2015.
- [24] R. Paul, D. Feldman, D. Rus, and P. Newman, "Visual precis generation using coresets," in *2014 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2014, pp. 1304–1311.
- [25] Ö. Erkent and I. Bozma, "Place representation in topological maps based on bubble space," in *2012 IEEE Int. Conf. Robotics and Automation*, 2012, pp. 3497–3502.
- [26] K. Tanaka, "Mining minimal map-segments for visual place classifiers," *CoRR*, vol. abs/1909.09594, 2019.
- [27] 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.
- [28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
- [29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [30] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," pp. 5297–5307, 2016.
- [31] E. Osherov and M. Lindenbaum, "Increasing cnn robustness to occlusions by reducing filter support," in *Proc. the IEEE Int. Conf. Computer Vision*, 2017, pp. 550–561.
- [32] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1568–1576.
- [33] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graphics (ToG)*, vol. 36, no. 4, p. 107, 2017.
- [34] M. Imhof and M. Brschler, "A study of untrained models for multimodal information retrieval," *Inf. Retr. Journal*, vol. 21, no. 1, pp. 81–106, 2018.
- [35] Y. Kojima, K. Tanaka, Y. Naiming, Y. Hirota, and K. Yamaguchi, "From comparison to retrieval: Scalable change retrieval from discriminatively learned deep three-dimensional neural codes," in *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, 2019, pp. 789–795.
- [36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
- [37] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE transactions on image processing*, vol. 14, no. 3, pp. 294–307, 2005.
- [38] P. Jensfelt and S. Kristensen, "Active global localization for a mobile robot using multiple hypothesis tracking," *IEEE Trans. Robotics and Automation*, vol. 17, no. 5, pp. 748–760, 2001.
- [39] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte carlo localization for mobile robots," in *ICRA*, vol. 2, 1999, pp. 1322–1328.