# Fault-Diagnosing SLAM for Varying Scale Change Detection

Sugimoto Takuma Yamaguchi Kousuke Tanaka Kanji

***Abstract*— In this paper, we present a new fault diagnosis (FD) -based approach for detection of imagery changes that can detect significant changes as inconsistencies between different sub-modules (e.g., self-localization) of visual SLAM. Unlike classical change detection approaches such as pairwise image comparison (PC) and anomaly detection (AD), neither the memorization of each map image nor the maintenance of up-to-date place-specific anomaly detectors are required in this FD approach. A significant challenge that is encountered when incorporating different SLAM sub-modules into FD involves dealing with the varying scales of objects that have changed (e.g., the appearance of small dangerous obstacles on the floor). To address this issue, we reconsider the bag-of-words (BoW) image representation, by exploiting its recent advances in terms of self-localization and change detection. As a key advantage, BoW image representation can be reorganized into any different scaling by simply cropping the original BoW image. Furthermore, we propose to combine different self-localization modules with strong and weak BoW features with different discriminativity, and to treat inconsistency between strong and weak self-localization as an indicator of change. The efficacy of the proposed approach of FD with/without combining AD and/or PC was experimentally validated.**

## I. Introduction

For long-term map maintenance in dynamic environments, a robotic visual SLAM system must detect changed objects (e.g., furniture movement, and building construction) in a live image with respect to the map, while ignoring nuisance changes (e.g., sensor noises, registration errors, and occlusions) during long-term multi-session navigation. One approach is to formulate the problem as a pair-wise image comparison (PC), to compare each live-map-image-pair using image differencing techniques [1]. However, this requires that a robot memorizes every map image; hence, scaling to large-size environments is difficult. An alternative approach is to formulate the problem as anomaly detection (AD), and predict anomalies (changes) in a live image with respect to a pre-trained normal model (map) [2]. However, this requires a robot to re-train the normal model frequently to keep it up-to-date every time the map is updated.

An alternative is to formulate the problem as a fault diagnosis (FD) to treat inconsistency (changes) between the responses of different sub-modules of SLAM (e.g., map-relative self-localization vs. pose-tracking [3], self-localization vs. mapping [4]) as an indicator of the likelihood-of-changes (LoC). The first solution to a multi-experience based mapping system was pioneered by [5], which maintains a collection of mapping sub-modules using data from differing environmental conditions (i.e., visual experiences). If self-localization with respect to such a map performs sufficiently well, it covers the encountered conditions and there is no significant change. However, if localization performance is poor, the map does not cover the conditions and environmental changes can occur with high probability. Such a new FD-based change detection framework has two main advantages:

- No additional storage or detector engine (but only existing map database and localization engine) is required;
- Degradation of map quality (i.e., need for map update) in terms of localization performance can be measured.

A significant challenge when incorporating different SLAM sub-modules into FD involves addressing varying scales of changed objects. This is an important issue because a robot is often required to inspect not only image-level but also sub-image-level changes, such as the appearance of small dangerous obstacles on the floor [6]. Typical feature extractors used by self-localization (e.g., ConvNet [7], autoencoders [8]) utilize live images as the input at the same scaling as the map images. If they are tested with varying scale images, they tend to fail, even when there is no change. As such, it is difficult to apply them to sub-image-level change detection.

In this paper, we reconsider the bag-of-words (BoW) image representation based on recent advances in self-localization techniques (Fig. 1). We consider that BoW has several desirable properties. Firstly, a recently developed BoW-based method has developed into the state-of-the-art in self-localization [9]. Secondly, BoW is a compressed (and yet discriminative) image representation that assists in the suppression of the map maintenance cost. Thirdly, BoW has a good affinity to codebook-based image representation [10], which has attracted increasing attention in the field of change detection [11]. Most importantly, the full-image-level BoW representation can be flexibly reorganized into sub-image-level BoW representation by simply cropping the BoW image with a smaller ROI.

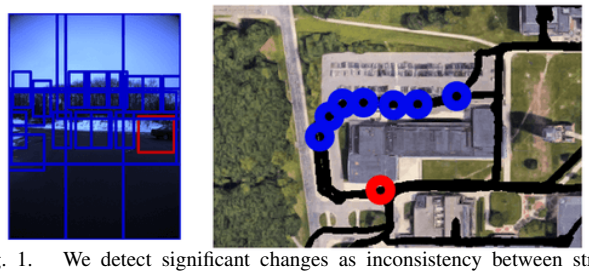


Fig. 1. We detect significant changes as inconsistency between strong and weak self-localization modules in visual SLAM. In the figure, blue and red colors indicate self-localization results for the strong and weak BoW features.

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 26330297, and (C) 17K00361.

The authors are with Graduate School of Engineering, University of Fukui. tnkknj@u-fukui.ac.jp

In this contribution, we present a new single-view FD-based change detection approach that can detect sub-image-level changes while simultaneously localizing the robot-self. The idea is to crop the original BoW image with different ROIs to reorganize it into different sub-image-level BoW features. Specifically, we propose two different types of sub-image-level BoW features: strong and weak, with different levels of discriminativity. The former are discriminative features with large ROIs that are useful for reliable self-localization. It should be noted that such strong features on their own are *not* sufficient for the FD-based change detection because strong features merely cause a self-localization fault, which is required by the FD-based change detection. Therefore, we introduce weak features with small object-level ROIs to trigger a self-localization fault intentionally. Intuitively, changes are expected to be detected as inconsistencies between such strong and weak self-localization modules. Experiments on challenging cross-season change detection using publicly available NCLT dataset [12] validate the efficacy of the proposed approach of FD with/without combining AD and/or PC.

## II. Approach

The main focus of this paper is to introduce a novel fault-diagnosis -based approach to improve object-level change detection. As mentioned in the Section I, existing techniques can be categorized into three groups, PC (II-A), AD (II-B), and FD (II-C). It should be noted that PC and AD are not always available, depending on the availability of map images and the update frequency of place-specific anomaly detectors. Therefore, there are four possible combinations of available change detection modules, including FD, FD+AD, FD+PC, and FD+AD+PC (Fig. 2). Each of these modules will be investigated in the experimental section. It should be noted that in all change detection approaches, knowledge of the current viewpoint is required, to pair a live query image with the appropriate map image or the place-specific models. To this end, a robust viewpoint localization scheme is also introduced in II-C.

### A. Pairwise Image Comparison (PC)

For PC, we introduce a SIFT feature-based PC approach. Based on the literature (e.g., [1]), the LoC of a query live feature is measured according to its dissimilarity to the most similar normal feature. Firstly, every live/map image is represented as a collection of SIFT features with Harris-Laplace keypoints [13]. The LoC at each keypoint in the query image is measured using the L2 distance between a live SIFT descriptor at that keypoint and its nearest-neighbor map SIFT in the 128-dim SIFT feature space.

### B. Anomaly Detection (AD)

In contrast, the AD approach formulates the problem as a one-class classification [14], in which the goal is to classify (not a live-map-image-pair but) a query live image as a "change" or a "no-change" [10], [11], [15]–[22], with respect to an offline pretrained normal model. Unlike PC approaches,

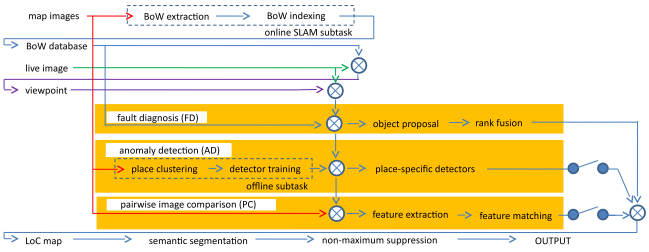


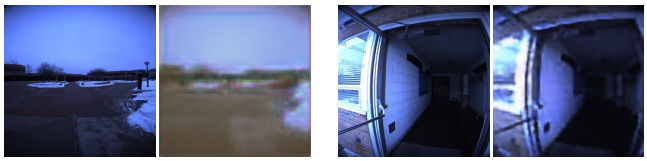

Fig. 2. System architecture.

this formulation facilitates the utilization of compressed normal models such as bag-of-visual-features (BoVFs) [10], [11], [15], 3D/landmark/grid maps [16]–[20], compact manifold learning [21]–[23], and autoencoders (AEs) [2]. Early studies employed one-class support vector machines [14], or support vector data description [24]. However, their computational scaling was poor because of the construction and manipulation of the kernel matrix. Subspace-based methods [25]–[31] are effective means of finding anomalous objects in relevant subspaces that are not anomalous in the full-dimensional space. However, most existing methods use shallow methods that typically require substantial feature engineering. Recently, deep AEs [2] have developed into a predominant approach for learning-based AD. In this context, AEs are used differently in two approaches: (1) mixed approaches [22], in which the learned embeddings are plugged into classical AD methods, and (2) fully deep approaches, in which the reconstruction error (RE) is directly utilized as an anomaly score [32]. Our approach belongs to the latter [32]–[35].

The basic idea is to reconstruct a query live image $I$ by using its counterpart (or linked) normal model (i.e., AE) $c_j$, as shown in Fig. 3. The AE is designed to extract the common factors of variation from normal samples and to reconstruct them accurately. As such, anomalous samples do not contain these common factors of variation; hence, they cannot be accurately reconstructed. Therefore, the pixel-level LoC for a given image region $P$ can be evaluated by the RE at each pixel: $V_{RE}(P) = \sum_{p \in P} |I(p) - I'(p)|$, in which the images $I$ and $I'$ are the input image and the image is reconstructed by the AE that is linked to the corresponding map image, respectively, and $|\cdot|$ is an absolute value operator. If $V_{RE}$ exceeds a pre-defined threshold $V^*_{RE}$, the region of interest $P$ is determined to be an anomalous object.

A place-specific anomaly detector is required to define a-priori what constitutes a place. Recently, we have explored several place definition approaches to partition the map image set into $k$ place-classes for AE-based change detection [36]. In the current study, we employ the k-means approach, in which map images are grouped into $k$ place classes by a k-means clustering algorithm. Each of the $k$ trained AEs is used as the normal model of the map images that belong to the cluster.

A notable design issue in training place-specific AEs is the determination of the threshold $V^*_{RE}$ on $V_{RE}$ for different place-specific AEs. The RE outputs by different AEs are



(a) RE = 1.2907 (b) RE = -0.0335

Fig. 3. Examples of input and reconstructed image pairs and the normalized reconstruction errors.

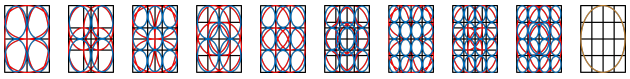



Fig. 4. Library of OBB templates. A grid with cells (black) is imposed on the image region and OBBs with different combinations of the cells (red/blue circles) are defined.

not comparable because individual AEs are trained using different training sets. To address this issue, we normalize each RE value by the AE-specific normalizer constant. In the normalization process, we approximate the probability distribution function (PDF) of the REs using a Gaussian distribution, and normalize the RE value by subtracting the mean value $\mu$ and dividing by its standard deviation (SD) $\sigma$ and by a normalizer coefficient $c$ with a value of $c = 0.8$ as the default. This normalization allows outputs from different AEs to be directly compared.

### C. Fault Diagnosis (FD)

The basic idea of the FD-based approach is to introduce strong and weak self-localization modules based on different levels of feature discriminativity. Inconsistencies between responses from the strong and weak self-localization are then treated as a change indicator. Formally, such a difference in discriminativity can be realized by varying the number of used BoW sub-images between the strong and weak self-localization modules. Detailed explanations of sub-image extraction, weak self-localization, strong self-localization, fault diagnosis and change detection is presented in the following.

For sub-image extraction, we employ both supervised and unsupervised object proposal approaches that are expected to act as strong and weak features, respectively. The supervised YOLO method from [37] is employed to extract 1-11 OBBs per image (Fig. 1). The unsupervised BING proposal method from [38] is employed to extract 42-50 class-agnostic OBBs per image (Fig. 1). In addition, we introduce a various combinations of non-adaptive fixed OBBs as shown in Fig. 4. A natural design choice is to use the weak and strong features only for the weak and strong self-localization modules, respectively. However, we propose to use every feature-type for both self-localization modules, which works better in practice.

For weak self-localization, we use a single sub-image-level BoW (without enhancing it using the available contextual information) to trigger a self-localization fault intentionally, when and only when there is a significant change in the sub-image region. Specifically, we use the recently developed state-of-the-art BoW framework in [9]. In a previous work, we studied this framework in a different context of simultaneous mapping and localization (i.e., SLAM) [39]. In the current study, we extend this framework to sub-image-level BoW and implement mapping and localization as two separate (i.e., offline and online) processes rather than a single SLAM process. Formally, in the offline map-building process, a collection of ORB features [40] is extracted from each map sub-image and then indexed to the inverted file. Simultaneously, we update the BoW vocabulary by incrementally incorporating newly arrived visual words. In the online localization process, the inverted file is retrieved using each word in each live sub-image as a query, and each retrieval result is further refined by the TF-IDF scoring scheme [41], the ratio test [42], RANSAC geometric verification, and island clustering [43].

For strong self-localization, we extend the weak self-localization framework to aggregate multiple BoW sub-images to increase robustness and discriminativity. It should be noted that a weak self-localization module outputs a ranked list of all the map images in the order of relevance. For strong self-localization, these weak self-localization modules are ensembled and the ranked lists are aggregated into a single strong rank list using the robust rank aggregation technique in [44]. As such, the rank values $r_1, \cdots, r_N$ of a given map image from $N$ different self-localization modules are aggregated into a relevance score by rank fusion: $\sum_{i=1}^{N} r_i^{-1}$. In experiments in III, different combinations of the OBB templates from Fig. 4 will be tested.

For FD, we evaluate inconsistency between the different ranked lists output using weak and strong self-localization modules. Given a relevant map-image top-ranked by strong self-localization, inconsistency can be evaluated by the rank of that relevant map image in the weak rank list. As such, a larger rank value represents greater inconsistency. To address the inherent retrieval noise in the strong self-localization, we consider multiple top-$Y$ ranked map images ($Y = 10$) in the strong rank list, and then the corresponding $Y$ rank values in the weak rank list are aggregated into the final decision. For the aggregation, we propose to use pixel-wise min pooling. This is performed because the retrieval noise influences and can spuriously increase the rank values. This increases the difficulty associated with implementing the other typical pooling techniques such as max or average pooling.

For change detection, we aggregate the sub-image-level rank values into image-level LoC maps by incorporating individual OBBs. The rank aggregation problem was explored in the context of part-based self-localization in our previous study [44]. The method used in the current study is based on our previous method with a few key modifications: Firstly, our previous study emphasized at image-level ranking, whereas this study aims to obtain pixel-level rank values. Secondly, the previous method utilized non-overlapping query live sub-images (from color-based segmentation) as

inputs, whereas the current method utilizes overlapping query sub-images (unsupervised/supervised OBBs) with *variable* amounts of overlap per pixel. To address this issue, we must perform a new task of pixel-wise rank fusion. Formally, we adopt the recently presented extension of *variable* length rank lists for multi-media retrieval [45], and fuse per-pixel ranking results as follows [46]:

$$r[p] = |J[p]| \left( \sum_{j \in J[p]} r_j[p]^{-1} \right)^{-1}. \quad (1)$$

$J[p]$ is the set of identifiers of OBBs to which pixel $p$ belongs to. $r_j$ is a rank value of the $j$-th OBB in the map image.

## III. Experiments

We evaluated the proposed change detection framework for a challenging cross-season scenario.

### A. Settings

We used a large-scale long-term autonomy dataset, North Campus long-term (NCLT) dataset, which is publicly available in [12]. The data used in this study includes view image sequences along vehicle trajectories acquired using the front facing camera (i.e., Ladybug3) of a Segway vehicle platform (Fig. 5). The image size is 1232×1616. This dataset includes various types of changing images such as cars, pedestrians, building construction, construction machines, posters, tables and whiteboards with wheels, from seamless indoor and outdoor navigations. Additionally, it has recently gained significant popularity as a benchmark in the SLAM community [47].

In this study, we used four datasets labeled "2012/1/22 (WI)", "2012/3/31 (SP)", "2012/8/4 (SU)", and "2012/11/17 (AU)" that were collected from four different seasons. For all the possible 12 live-map-season-pairs, we annotated 986 different changing objects with bounding boxes. Additionally, we prepared a collection of 1,973 random destructor images that were independent of the 986 annotated images. These images did not include changing objects. We then merged the 1,973 destructor images and 986 annotated images to obtain a map database containing 2,959 images. Fig. 5 presents examples of changing objects in the dataset.

### B. Performance Results

The performance for the change detection task was evaluated in terms of top-$X$,$Y$ accuracy. Pixel values in the LoC image are initialized by a function $\varepsilon$ that generates a tiny random number. Firstly, we estimated an LoC image using a change detection algorithm on the top-$Y$ self-localization hypotheses ($Y = 10$). We then imposed a 2D grid with $10 \times 10$ pixel sized cells on the query image and estimate an LoC for each cell by max-pooling the pixel-wise LoC values from all pixels that belong to that cell. To fuse results from multiple change detection algorithms (e.g., FD+AD+PC), the rank fusion algorithm in II-C is reused. Next, all the cells from all the map images are sorted in descending order of LoC, and the accuracies of the top-$X$ items in the list were evaluated. For a specific $X$ threshold, a successful detection is defined as a changed object with an annotated bounding box that is sufficiently covered (intersection-over-union $\geq$ 50%) by the top-$X$ percent of the cells.

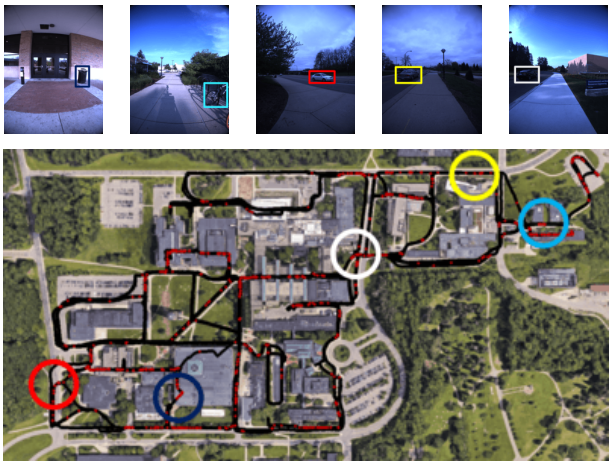


Fig. 5. Experimental environments and robot trajectories. Colored rectangles and circles respectively indicate the bounding boxes and the GPS locations of the ground-truth change objects. Different colors represent different change objects.

TABLE I

Effectiveness of BoW on change detection.

| $X$ [%] | 5 | 10 | 15 | 20 |
|---|---|---|---|---|
| BoW-based FD | 9.3 | 18.6 | 32.5 | 42.8 |
| global FD | 6.2 | 11.0 | 28.8 | 46.6 |

In self-localization, the combination of J+B+BING+YOLO outperforms the other settings. This combination is used for viewpoint localization in the following change detection experiments.

The number of place-specific AEs is set as $k = 10$ based on the preliminary experiments in [36].

For performance evaluation, we further introduce post-processing on the LoC map, which can stably boost the proposed and all the change detection comparison techniques. The idea is to use recent deep learning based semantic segmenation to detect non-interesting regions with "sky" or "ground" labels, where no interesting change is expected. LoC values for these non-interesting regions are reset by $\varepsilon$. Formally, we employed a DeepLabV3+ model in [48] that combines the advantages of spatial pyramid pooling and encoder-decoder structure. Input images are resized to 512×512, and the model was trained on the Cityscapes dataset [49].

In the first experiment, we compared the performance of the proposed BoW-based FD approach to that of an alternative non-BoW-based FD approach, termed global FD. The only difference between the global FD and the proposed FD approach is that a global image feature is used in place of the local feature BoW. Specifically, we introduced a state-of-the-art autoencoder as the global feature similar to [8], motivated by the recent success of AEs in self-localization applications [8]. Our AE architecture uses a layer structure of 64-32-16-16-32-64 nodes, and is trained on a collection of map images. Each image is resized to 64×64 pixels prior to input into the AE. For the evaluation, we considered a relatively small size $N = 100$ map database that consisted of

TABLE II

CHANGE DETECTION PERFORMANCE.

| $X$ [%] | 5 | 10 | 15 | 20 | solo-leader | co-leader |
|---|---|---|---|---|---|---|
| FD | 7.8 | 15.8 | 27.2 | 39.6 | 21.7 | 27.7 |
| PC | 0.0 | 2.1 | 7.2 | 19.6 | 13.0 | 20.7 |
| AD | 0.4 | 2.5 | 10.6 | 25.2 | 15.7 | 22.3 |
| FD+AD+PC | 2.6 | 20.1 | 34.1 | 42.1 | 13.2 | 29.4 |
| FD+AD | 2.6 | 19.7 | 33.8 | 41.9 | 6.4 | 23.0 |
| FD+PC | 3.1 | 11.0 | 20.6 | 34.9 | 9.7 | 13.2 |

one ground-truth map image and $(N-1)$ non-ground-truth map images consisting of random samples from the 2,959 map images. As shown in Table I, the proposed BoW -based FD outperforms the global FD. It should be noted that the AE features fail badly when matched with different scaling features with very different OBB sizes.

We performed a joint viewpoint-change prediction using each full season dataset as the map images, to compare the proposed method to other comparing methods including AD, PC, FD+AD, FD+PC, and FD+AD+PC. It should be noted that the full season dataset is much more challenging than the one considered in Tab. I, due to the increased difficulty in viewpoint localization, and confusing map images. The OBB templates J+B+BING+YOLO from Fig. 4 are used as the default setting. Table II summarizes the comparing result. In the table, "solo-leader" and "co-leader" represent the ratio of query images in which the method outperforms the other methods in terms of the minimum $X$ threshold (i.e., an annotated bounding box is sufficiently covered by the top-$X$ percent of the cells). It is evident that the proposed FD method combined with AD and PC (FD+AD+PC) frequently outperformed the other combinations of methods. Importantly, the FD method outperforms the AD and PC methods.

## IV. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed a novel method for fault diagnosis-based change detection based on 2D on-board imagery in a 3D real-world environment. The method is substantially different from existing methods and uses a BoW image representation to reorganize BoW flexibly into varying scales. It is shown experimentally that the proposed method boosts the accuracy of the image change detection, while suppressing the map maintenance cost. Furthermore, the proposed method accounts for joint viewpoint-change prediction by introducing strong and weak BoW features with different levels of discriminativity.

Future work should address a general framework of change detection from image sequence rather than a single image. Currently, in our experimental implementation, only a single-view joint viewpoint-change prediction is considered. Future work can leverage richer information from multi-view sequential query images [50]. Furthermore, performance gains can be expected when change detection results from different approaches (i.e., PC, AD, FD) are combined not only in late fusion (i.e., at the level of output LoC maps) but also in early fusion (e.g., at the level of input features) [51].

## REFERENCES


[1] J. Košecka, "Detecting changes in images of street scenes," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 590–601.

[2] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[3] A. Jacobson, Z. Chen, and M. Milford, "Online place recognition calibration for out-of-the-box SLAM," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*, 2015, pp. 1357–1364.

[4] M. Burki, M. Dymczyk, I. Gilitschenski, C. Cesar, R. Siegwart, and J. Nieto, "Map management for efficient long-term visual localization in outdoor environments," in *Intelligent Vehicle Symposium, IEEE*. IEEE, 2018.

[5] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1645–1661, 2013.

[6] M. Nava, J. Guzzi, R. O. Chavez-Garcia, L. M. Gambardella, and A. Giusti, "Learning long-range perception using self-supervision from short-range sensors and odometry," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1279–1286, 2019.

[7] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*, 2015, pp. 4297–4304.

[8] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018.

[9] E. Garcia-Fidalgo and A. Ortiz, "ibow-lcd: An appearance-based loop-closure detection approach using incremental bags of binary words," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3051–3057, Oct 2018.

[10] 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.

[11] 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.

[12] 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.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[14] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class svm for learning in image retrieval," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1. IEEE, 2001, pp. 34–37.

[15] 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.

[16] 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.

[17] E. Palazzolo and C. Stachniss, "Fast Image-Based Geometric Change Detection Given a 3D Model," in *Proceedings of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018.

[18] 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.

[19] 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.

[20] 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.

[21] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.

[22] 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.

[23] L. Gueguen and R. Hamid, "Large-scale damage detection using satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1321–1328.

[24] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[25] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, no. 2. ACM, 2001, pp. 37–46.

[26] J. Zhang, M. Lou, T. W. Ling, and H. Wang, "Hos-miner: a system for detecting outlyting subspaces of high-dimensional data," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 1265–1268.

[27] I. Assent, R. Krieger, E. Müller, and T. Seidl, "Inscy: Indexing subspace clusters with in-process-removal of redundancy," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 719–724.

[28] H. V. Nguyen, V. Gopalkrishnan, and I. Assent, "An unbiased distance-based outlier detection approach for high-dimensional data," in *International Conference on Database Systems for Advanced Applications*. Springer, 2011, pp. 138–152.

[29] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2009, pp. 831–838.

[30] E. Müller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE, 2011, pp. 434–445.

[31] F. Keller, E. Muller, and K. Bohm, "Hics: high contrast subspaces for density-based outlier ranking," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012, pp. 1037–1048.

[32] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Fully convolutional neural network for fast anomaly detection in crowded scenes," *CoRR*, vol. abs/1609.00866, 2016.

[33] A. Makhzani and B. Frey, "K-sparse autoencoders," *arXiv preprint arXiv:1312.5663*, 2013.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[35] A. Makhzani and B. J. Frey, "Winner-take-all autoencoders," in *Advances in Neural Information Processing Systems*, 2015, pp. 2791–2799.

[36] K. Yamaguchi, K. Tanaka, T. Sugimoto, R. Ide, and T. Koji, "Recursive background modeling with place-specific autoencoders for large-scale image change detection," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[37] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *CoRR*, vol. abs/1612.08242, 2016. [Online]. Available: http://arxiv.org/abs/1612.08242

[38] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3286–3293.

[39] R. Yamamoto, K. Tanaka, and K. Takeda, "Invariant spatial information for loop-closure detection," in *16th International Conference on Machine Vision Applications, MVA 2019, Tokyo, Japan, May 27-31, 2019*, 2019, pp. 1–6.

[40] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision, ICCV*, 2011, pp. 2564–2571.

[41] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[42] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[43] D. Galvez-Lopez and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012.

[44] K. Tanaka, "Unsupervised part-based scene modeling for visual robot localization," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 6359–6365.

[45] A. Mourão, F. Martins, and J. Magalhães, "Multimodal medical information retrieval with unsupervised rank fusion," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 35–45, 2015.

[46] K. Tanaka, "Detection-by-localization: Maintenance-free change object detector," in *Robotics and Automation (ICRA), 2019 IEEE International Conference on*. IEEE, 2019.

[47] J. G. Mangelson, D. Dominic, R. M. Eustice, and R. Vasudevan, "Pairwise consistent measurement set maximization for robust multi-robot map merging," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 1–8.

[48] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, 2018, pp. 833–851. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2_49

[49] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[50] M. Fehr, M. Dymczyk, S. Lynen, and R. Siegwart, "Reshaping our model of the world over time," in *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, 2016, pp. 2449–2455. [Online]. Available: https://doi.org/10.1109/ICRA.2016.7487397

[51] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.