DOI: 10.1109/LRA.2023.3242872

Towards Long-Term Retrieval-based Visual Localization in Indoor Environments with Changes

Julia Kabalar*[1], Shun-Cheng Wu*[1], Johanna Wald[1,4], Keisuke Tateno[1,3], Nassir Navab[1,2], Federico Tombari[1,3]

Visual localization is a challenging task due to the presence of illumination changes, occlusion, and perception from novel viewpoints. Re-localizing the camera pose in long-term setups raises difficulties caused by changes in scene appearance and geometry introduced by human or natural deterioration. Many existing methods use static scene assumptions and fail in dynamic indoor scenes. Only a few works handle scene changes by introducing outlier awareness with pure learning methods. Other recent approaches use semantics to robustify camera localization in changing setups. However, to the best of our knowledge, no method has yet used scene graphs in feature-based approaches to introduce change awareness. In this work, we propose a novel feature-based camera re-localization method that leverages scene graphs within retrieval and feature detection and matching. Semantic scene graphs are used to estimate scene changes by matching instances and relationship triplets. The knowledge of scene changes is then used for our change-aware image retrieval and feature correspondence verification. We show the potential of integrating higher-level knowledge about the scene within a retrieval-based localization pipeline. Our method is evaluated on the *RIO10* benchmark with comprehensive evaluations on different levels of scene changes.

***Index Terms*—Localization.**

I. Introduction

VISUAL localization aims to estimate the camera pose from a given image with respect to a reference scene [1], [2]. This is challenging and also a fundamental requirement for many computer vision applications, such as augmented, mixed, and virtual reality (AR/MR/VR) [3]–[5], robotics, and autonomous driving [6]–[8]. Despite the difficulty of this task, it is a rather mature research field where recent methods [9]–[11] reported high pose accuracy on popular indoor benchmarks such as *7-scenes* [12] and *12-scenes* [13]. Structure-based methods [11], [14], [15] are one of the most famous approaches. They rely on local and global features for finding correspondences across images and then use them to estimate accurate camera poses. An image retrieval step [16], [17] is often used to reduce memory usage and run-time.

Although those methods show excellent performance on the benchmarks with static scenes, recent studies [11], [18] have shown that they fail to perform under scene changes, while only a few methods targeted in solving and studying this task [11], [18]. Dong *et al.* [10] proposes to handle

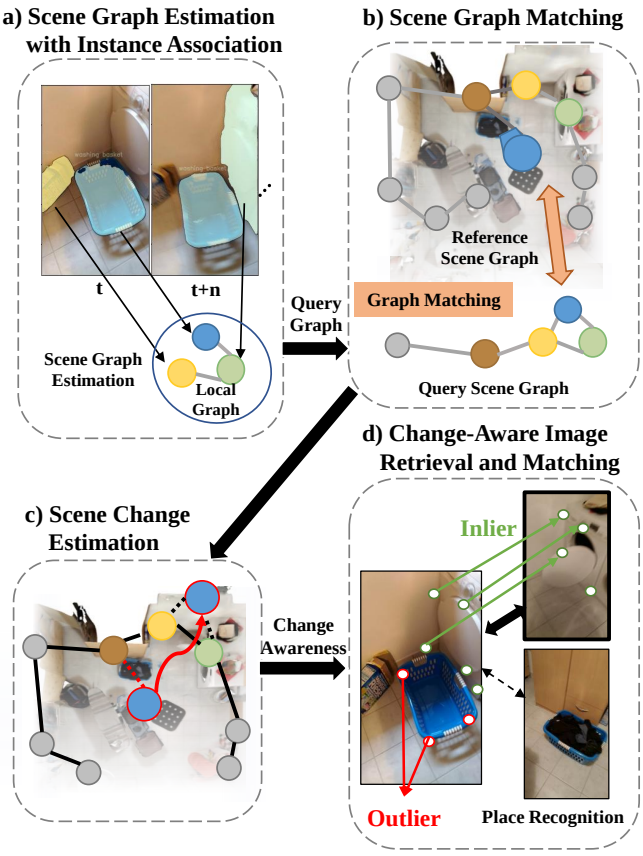

Fig. 1. Method overview: a) We use image sequences to generate a per-frame scene graph and merge them into a global representation; b) Given a reference and a query graph, we match two graphs with graph matching; c) The matching results are used to find static and changed objects; d) The knowledge of scene changes is integrated into image retrieval, by only considering inlier objects (green), and feature matching, by detecting outlier keypoints (red).

Manuscript received: September, 27, 2022; Revised December, 19, 2022; Accepted February, 2, 2023.

This paper was recommended for publication by Editor Sven Behnke upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Technical University of Munich.

*The authors contributed equally to this paper.

[1]The Laboratory for Computer Aided Medical Procedures, Technical University of Munich (TUM), Boltzmannstr. 3, 85748 Garching, Germany

[2]The Laboratory for Computer Aided Medical Procedures, Johns Hopkins University, 3400 N. Charles Street, Baltimore, USA

[3]Google LLC, Brandschenkestrasse 110, 8002 Zürich, Switzerland

[4]Everyday Robots, Erika-Mann-Straße 33, München, Germany.

Kabalar, J. and Wald, J. were at TUM when the work was conducted.

Digital Object Identifier (DOI): see top of this page.

DOI: 10.1109/LRA.2023.3242872 ©2023 IEEE

dynamic indoor scenes by introducing outlier awareness and rejecting points during the hierarchical routing process in a decision tree, outperforming other methods in the *RIO10* benchmark [18]. However, compared to other baselines, the fraction of the correctly localized frames is greatly affected by the changes in scene geometry. Inspired by their work, we observe that the awareness of moved objects is the key to localizing cameras in changed scenes. Existing structure-based methods [11], [15] do not consider scene changes in the step of image retrieval and correspondence finding.

To this end, we propose a novel feature-based visual localization pipeline that aims at improving visual localization in changed scenes by identifying scene changes using semantic scene graphs. The overview of our method is shown in Fig. 1. Our method estimates a reference and a query graph by associating the local scene graphs from all their respective input images. We detect scene changes on a scene graph level by comparing the spatial relationship within relationship triplets (subject, predicate, object), *e.g*. the basket is on the left of a washing machine, between the scene graphs of a reference and revisited scene. With the awareness of these scene changes, we propose two major changes in a retrieval-based visual localization pipeline [11]. First, we propose a retrieval method that retrieves image candidates based on the similarity of static object sets. Second, we modify the descriptor matching step by eliminating potential outliers on the moved objects.

We evaluate the performance of our method on the *RIO10* dataset [18], which was designed to evaluate long-term visual camera re-localization. We provide a comprehensive evaluation of our method under different levels of scene changes. The evaluation results show that compared to baseline methods, our method has slightly better results given low-level scene changes and increasing performance along with the scene change level. In summary, we contribute: (1) A novel pipeline aims to solve visual localization for dynamic indoor scenes. (2) A scene change estimation method using scene graph matching.

II. Related Work

A. Camera Re-Localization

Camera localization can be roughly classified into four categories: image retrieval, direct pose regression, structure-based, and scene coordinate regression. *Image retrieval* methods find the pose of a query image by retrieving the pose of a reference image within a pre-built image database. Due to the reliance on a reference image database, these methods only work when the query poses are identical or very similar to the reference set [19]. Strategies have been proposed to solve this issue, *e.g*. using view interpolation [20], [21], and scene graphs [22]. *Direct pose regression methods* estimate the pose of a given image without preamble. It is usually based on learning-based approaches [23]–[25]. Despite their end-to-end fashion in estimating camera pose, such methods usually underperform the state-of-the-art RGB-D and structure-based methods. *Scene coordinate regression methods* estimate dense scene coordinates on each pixel of the query image [9], [26], [27]. However, since those methods typically target regression scene coordinates with static image sequences, the quality of the pose estimation is susceptible to scene changes. Only a few methods [10] try to handle changes by classifying them as outliers. *Structure-based methods* rely on the feature matches of 2D features and the 3D points in a scene. Matches are then used to estimate camera poses within a RANSAC-based method. The quality of feature description and matching directly affect the quality of the estimated pose. Many methods have been proposed to extract features with learning-based approaches [14], [28], [29] others proposed to improve feature matching with hierarchical methods [11], [15]. Semantics have successfully been used to improve long-term camera re-localization in outdoor scenes [30]–[37]. However, [31] argues the applicability of semantics in indoor scenes. Some methods use objects, instead of keypoints, as landmarks for visual localization. However, they still rely on static scene assumptions that struggle in localizing camera poses when scene content changes.

B. Scene changes

Handing scene changes in visual localization is either done by estimating overall similarity [30] or filtering outliers [10]. Only a few works focus on estimating the scene changes. Most scene change detection networks are siamese-like architectures [38], [39] that learn changes under the assumption that the test trajectory is similar to the reference trajectory. On the other hand, while scene graphs are a great representation to encode the high-level content of the scene, they were only recently proposed for scene change detection [40]. A direction to use scene graphs to learn scene changes [41] is costly, requires similar views, and is labeling-work intense, thus relatively unexplored. However, the idea of using a graph representation to link objects is used in various works such as object-based SLAM systems [42] and semantic graphs [43], [44]. A line of work [45]–[47] use graph representation in visual localization with the use of graph topology and random walk descriptors. However, they focus on handling the changes in viewpoint and visual appearance instead of the changes in semantics or geometry. To the best of our knowledge, no previous work uses scene graphs for change detection in a visual localization pipeline.

III. Method

In this work, we focus on improving feature-based methods by introducing awareness of scene changes. Our method is built on top of the state-of-the-art retrieval-based visual localization framework, Kapture [11], which combines image retrieval and structure-based methods for visual localization, by integrating the scene change awareness estimated with scene graph matching. The overview of our method is shown in Fig. 2. We will briefly discuss the original Kapture plus the proposed changes, and then detail the proposed method in the following method sections.

Reconstruction pipeline: Given a reference sequence, the reconstruction pipeline reconstructs a sparse point map through three steps: image retrieval, descriptor matching, and point

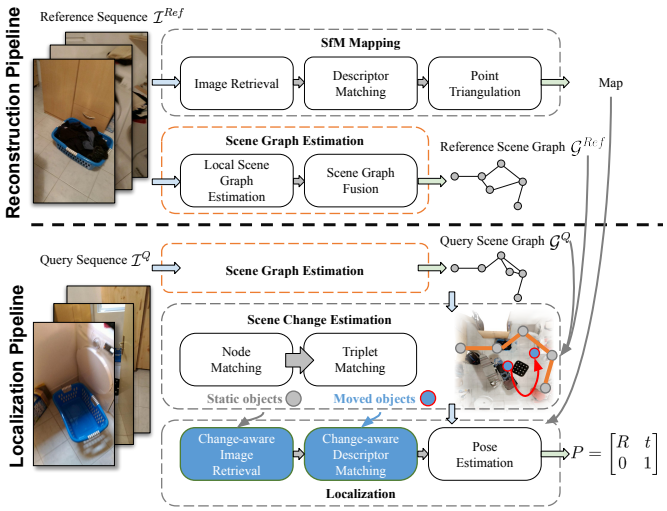

Fig. 2. Overview of the proposed change-aware localization pipeline. On top of the typical Structure from Motion (SfM) Mapping and Localization, we integrate scene change awareness into image retrieval and descriptor matching using scene graphs.

triangulation. The image retrieval step reduces the search space in an image database to allow localization at scale, the descriptor matching step finds feature correspondences between images, and the point triangulation estimates 3D sparse points, which are associated with 2D points to be used to build a global map of the environment. In our method, we additionally estimate a reference scene graph $\mathcal{G}^{Ref}$ using input reference images (Sec. III-A). The reference scene graph is used for scene change estimation (Sec. III-B).

Localization pipeline: Given a query sequence, the localization pipeline estimates the 6-degree-of-freedom (DoF) camera poses for the input images with respect to the reference scene. An image retrieval step is applied to reduce the search space to the top-k most similar images based on the visual appearance, followed by a descriptor matching step that finds 2D-3D correspondence by matching local image features to 3D points using the associated 2D points from the database images. Then, camera poses are estimated with those matches with perspective-n-point (PnP) [48] and RANSAC [49]. In this work, we modify image retrieval and feature matching with the knowledge of scene changes, estimated by comparing the query and reference scene graphs (Sec. III-B). The identified static and moved objects are used to improve image retrieval (Sec. III-C) and feature matching (Sec. III-D), respectively.

A. Scene Graph Estimation

In this work, we define a scene graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are a set of object nodes and directed edges represent the relationships between two nodes. Each node $v_i \in \mathcal{V}$ is an object instance that has the attributes of an instance label l_i, an object class label $c_i^o \in \mathcal{C}^o$, a set of 2D bounding boxes $\mathcal{B}_i$, and a set of appearance descriptors $\mathcal{D}_i$. $\mathcal{C}^o$ is a node category set, which is the 40 object classes in NYUv2 [50] in our implementation. Each edge $e_{i\to j} \in \mathcal{E} \mid v_i, v_j \in \mathcal{V}, i \neq j$ is defined by a predicate class label $c^p_{e_{i\to j}} \in \mathcal{C}^p$, where $\mathcal{C}^p$ is a predicate category set consisting of *left*, *right* and *none*.

Given an image set, we estimate a global scene graph by merging per-frame local scene graph estimation. Sec. III-A1 describes how local scene graphs are estimated, and Sec. III-A2 explains the merging step for global scene graph generation through node association and predicate fusion.

1) Local Scene Graph Estimation: Given an image with an index t, its scene graph $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$ is estimated by first detecting nodes, then classifying predicates between them.

a) Node Estimation: The nodes are estimated by combining the instance and semantic masks estimated by an instance and a semantic segmentation network. Each detected instance mask corresponds to a node $v_i^t \in \mathcal{V}^t$. The instance label l_i^t is inherited from the instance label of this mask, the semantic class label is assigned with the dominant semantic label from the semantic mask that corresponds to this mask, and the bounding box $b_i^t \in \mathbb{R}^4$ is the maximum and minimum pixel locations of this mask region. Moreover, the appearance descriptor d_i^t is computed with an image encoder given the input of the RoI aligned with the bounding box. The b_i^t and d_i^t are used to initialize $\mathcal{B}_i^t$ and $\mathcal{D}_i^t$ respectively.

b) Edge Estimation: We estimate a predicate class between all node pairs in the image. For an edge $e_{i\to j}^t \in \mathcal{E}^t$, the predicate is determined with

$$e_{i\to j}^t = \begin{cases} left & \text{if } b^t_{i,\text{left}} < b^t_{j,\text{left}} \wedge b^t_{i,\text{right}} < b^t_{j,\text{right}} \\ right & \text{if } b^t_{j,\text{left}} < b^t_{i,\text{left}} \wedge b^t_{j,\text{right}} < b^t_{i,\text{right}} \\ none & \text{else} \end{cases}, \quad (1)$$

where $b^t_{i,\text{left}}$ and $b^t_{i,\text{right}}$ are the left and right boundary of b_i^t of node v_i^t.

2) Global Scene Graph Estimation: Given a set of local scene graph estimates, the global scene graph $\mathcal{G}$ is estimated by first finding the global nodes $\mathcal{V}$ through objects association across frames, then the global predicates $\mathcal{E}$ are estimated via predicate fusion.

a) Node Association: We use the intersection-over-union (IoU) tracker from [51] with some modifications to associate nodes across frames. It associates objects in consecutive frames if objects have sufficient IoU from the previous frame. We modify their approach by adding semantic consistency, *i.e.* objects should have the same semantic class, and we use the image retrieval step as in [11] to associate objects across frames when consecutive frames are unavailable. The association results in a unique instance label for each associated node, and the instance label will be used to associate its node and its properties.

b) Predicate Fusion: After the node association step, we determine the global relationship $e_{i\to j} \in \mathcal{E}$ between nodes v_i and v_j by taking the dominant predicate that appears in all local scene graphs of those two nodes. In addition, to prevent ambiguity in predicate estimation due to the change of viewpoint, *e.g.* two nodes placed in the center with a camera observing the objects in a circular trajectory, we override the relationship estimation to *none* if the percentage of the dominant predicate estimation over all predicate estimations are less than a threshold f_{min}.

B. Scene Change Estimation

Given a reference and query image sequence, we use the scene graphs estimated from the respective image sequences to identify scene changes. Scene changes are identified by solving a graph-matching problem. We propose to solve this problem by relaxing it to the node matching problem (Sec. III-B1) and using the matched nodes for triplet matching (Sec. III-B2).

After the matching step, for all registered nodes with valid triplet matches, we select nodes that have semantic classes with a low likelihood of changes (Sec. IV-A) as a set of static nodes $\mathcal{V}_{\text{static}}$, and for all registered nodes that do not have valid triplet matches, they are marked as changed nodes $\mathcal{V}_{\text{change}}$.

1) Node Matching: Given two sets of nodes $\mathcal{V}^Q$ and $\mathcal{V}^{Ref}$ from $\mathcal{G}^Q$ and $\mathcal{G}^{Ref}$ respectively, we find node matches for every node $v_i \in \mathcal{V}^Q$ to a node $v_j \in \mathcal{V}^{Ref}$, where $i \neq j$, by using the consensus of all appearance descriptors matching. Given a query node v_i, its node label l_i, semantic label c_i^o, and a set of appearance descriptors $\mathcal{D}_i$, we use every descriptor $d_i \in \mathcal{D}_i$ to find the most similar descriptor for all descriptors in the set of reference nodes that have the same semantic class. As a result, a node-set is obtained, which is used to find the final node match by selecting the node with the highest statistical frequency.

2) Triplet Matching: For every triplet $\{v_i, e_{i\rightarrow j}, v_j\}$ in the query graph $\mathcal{G}^Q$, a match exists if a corresponding edge $e_{i\rightarrow j} \in \mathcal{E}^{Ref}$ exists between the two nodes matching $(v_i, v_j) \in \mathcal{V}^{Ref}$. Since indoor objects are potentially repetitive and might have similar appearances, object ambiguity is considered in the triplet matching process. For all nodes in $\mathcal{V}^{Ref}$ that do not have a valid triplet match, a node ambiguity set is built, including the nodes with the same semantic label. Then we construct triplets using all ambiguous nodes to include all potential matches. We experimentally found that this improves the final pose estimation result.

C. Change-aware Image Retrieval

In contrast to classical image retrieval, where a global image representation is used to find matches, we propose to estimate the similarity of two given images using the objects that have not been exposed to scene changes. Given a set of query nodes $\mathcal{V}^Q$ and reference nodes $\mathcal{V}^{Ref}$, we estimate a set of static nodes $\mathcal{V}_{\text{static}}$ from previous step (Sec. III-B2) and use it to obtain the node sets of all static objects $\bar{\mathcal{V}}^Q$ and $\bar{\mathcal{V}}^{Ref}$, with $\bar{\mathcal{V}}^Q = \mathcal{V}^Q \cap \mathcal{V}_{\text{static}}$ and $\bar{\mathcal{V}}^{Ref} = \mathcal{V}^{Ref} \cap \mathcal{V}_{\text{static}}$. Then estimate the instance similarity between those two frames with Jaccard index as

$$J(\bar{\mathcal{V}}^Q, \bar{\mathcal{V}}^{Ref}) = \frac{|\bar{\mathcal{V}}^Q \cap \bar{\mathcal{V}}^{Ref}|}{|\bar{\mathcal{V}}^Q \cup \bar{\mathcal{V}}^{Ref}|}. \tag{2}$$

For every query node, we check all the reference nodes from all reference images until the top-k similar matches are found. To prevent trivial matches where the Jaccard index is too low, we use a minimum similarity threshold j_{min} to select valid candidates. When the number of candidates does not reach the given top-k, the best matches from the original image retrieval using global image representation are used to populate the missing image candidates.

D. Feature Matching with Correspondences Verification

Instead of matching all features, we propose to reject feature correspondences on moving objects. Feature correspondences are computed with the closeness of the local feature descriptors of a set of selected keypoints. Two images I^Q, I^{Ref} are considered to have scene overlap if a sufficient number of feature correspondences are found. The output of this stage is a set of potentially overlapping image pairs

$$\mathcal{Z} = \{(I^Q, I^{Ref}, \mathcal{M}) \mid I^Q \in \mathcal{I}^Q, I^{Ref} \in R(I^Q, \mathcal{I}^{Ref})\}, \tag{3}$$

where $\mathcal{M} \subset \mathcal{F}^Q \times \mathcal{F}^{Ref}$ are feature matches [52], $\mathcal{I}^Q$ is a set of query images, $\mathcal{I}^{Ref}$ is a set of reference images, $R(I^Q, \mathcal{I}^{Ref})$ is an image retrieval function that returns a set of retrieved images given a query image I^Q and reference images $\mathcal{I}^{Ref}$.

Those feature correspondences $\mathcal{M}$ may include changed objects resulting in wrong camera pose estimation using PnP. We apply correspondence filtering on $\mathcal{M}$ to filter out moving objects. We filter out a correspondence pair $(p^Q, p^{Ref}) \in \mathcal{M}$, with p^Q, p^{Ref} feature points in I^Q, I^{Ref}, if a query point p^Q is located in the pixel occupation $\mathcal{O}_i^Q$ of node v_i^Q and whether this is part of $\mathcal{V}_{\text{change}}$. The final feature correspondences are then used to estimate camera pose using PnP [48] solver in a RANSAC loop [49], [53].

IV. Evaluation

A. Implementation Details

In Sec. III-A1, the instance masks are estimated with Entity Segmentation Network [54] trained model on the MS COCO [55] dataset, and the semantic masks are estimated from FuseNet [56] trained on ScanNetv2 [57] fine-tuned with 3RScan [58] with the class mapping of [59] for 40 dominant object classes in the NYUv2 [50] labeling scheme.

In Sec. III-B, the likelihood of objects being changed is determined from the statistics computed using a subset of the *3RScan* dataset [58]. *3RScan* contains multiple scans of the same scenes with potential scene changes. We compute the likelihood of an object class being changed statistically. Precisely, *human, mirror, pillow, box, towel, shower curtain, chair, bag*, and *otherprop* are identified as objects with a high likelihood of changes. We filter them out when selecting a set of static nodes.

The global and ROI image features are computed with a pre-trained AP-GeM network [60], and the local features are computed using the l2-normalized output layer of the R2D2 [29] network architecture for local feature extraction.

For all experiments, we use the *config2* COLMAP parameter configurations from Table 1 in [11] with 20,000 local features extracted on the top-20 image pairs.

B. Dataset

RIO10 [18], a subset of *3RScan* [58], was published for long-term camera re-localization in indoor scenes. Unlike typical benchmarks for static scene localization [13], [61], it provides multiple scans of the same scenes with changes. In addition, compared to other benchmarks containing scene

changes [62], [63], *RIO10* also provides an evaluation scheme to analyze the impact of scene change on the performance of a method, which allows us to evaluate our method holistically against different levels of changes.

C. Metrics

a) Pose Difference: We use the absolute translation error in meters computed as euclidean distance Δt, and the absolute angular error $\Delta\theta$ measured in degrees. We follow other methods by reporting the metrics with a fraction of frames localized within a given error threshold $\varepsilon_a(\epsilon_t, \epsilon_\theta)$, where the conditions $\Delta t < \epsilon_t$ and $\Delta\theta < \epsilon_\theta$ holds true.

b) Image Retrieval: In image retrieval, we use top-20 precision (p@20) to present the fraction of relevant images from 20 retrieved images. An image is considered relevant if it was taken spatially close to the ground truth position from where the query image was taken. The spatial closeness is determined by the absolute translation and angular errors given the poses of the query and the retrieved images.

c) Dense Correspondence Re-Projection Error (DCRE): This metric is proposed by *RIO10* benchmark [18], computed as the dense correspondence error of the 2D flow of dense 3D points rendered from an underlying 3D model normalized by the image diagonal. This metric reflects the alignment in visual perception without considering perceptual aliasing. We use Cumulative Dense Correspondence Re-Projection Error $\varepsilon_f(\epsilon_f)$, with ϵ_f a given error threshold, to measure a given sequence. In addition, we use $\overline{\varepsilon}_f(\epsilon_f) = 1 - \varepsilon_f(\epsilon_f)$ that reports the number of outliers, N/A reports the fraction of frames failed re-localizations, and SCORE $= 1 + \varepsilon_f(0.05) - \overline{\varepsilon}_f(0.5)$ reports the overall performance.

d) Change Measures: As in [18], we evaluate the result regarding semantic ζ_s and geometric ζ_g change measure, which describe the percentage of per-pixel differences in the 2D instance segmentation images and depth renderings, respectively. The other measures used are the normalized correlation coefficient ρ_v and the normalized sum of squared differences ζ_v, which capture visual change in appearance.

D. Results

We report the evaluation result of our method against two state-of-the-art feature-based methods, *i.e.* Active Search [64] and Kapture with the R2D2 method [11], an RGBD method, *i.e.* D2-Net [14], and a scene coordinate regression method, *i.e.* Grove v2 [9]. Since we implement our method based on the Kapture with the R2D2 method [11], we depict it as *Baseline*, and our method as *Ours*. In addition, since our method relies on instance and semantic results, we also report ours with ground truth instance and semantic segmentation to show the upper bound of our method, which denotes as *Ours**.

a) Image Retrieval: We report the performance of *Ours* to *Baseline* in the task of image retrieval in the *RIO10* validation set using the metrics of p@20 and ε_a.

The result is shown in Tbl. I. There are clear differences between the performance in our method that shows an increase of p@20 and ε_a. We consider an image relevant according to p@20 if it is within a tolerance of $(\epsilon_t, \epsilon_\theta) <$ (3m, 50°). We also report the recall based on the absolute pose error ε_a with mid-precision $(\epsilon_t, \epsilon_\theta)$ = (5m, 10°) and low-precision thresholds $(\epsilon_t, \epsilon_\theta)$ = (25m, 40°) over all sequences. In addition, we report the percentage change of the frames being retrieved of our change-aware retrieval process over the baseline method in the last column (+/-). It can be seen that our method brings around 5% increase in retrieved images and has the potential up to 16%. This is important since the percentage change positively correlates to all other metrics.

TABLE I
COMPARISON OF OUR IMAGE RETRIEVAL METHOD AGAINST *Baseline*

	p@20	ε_a(5m, 10°)	ε_a(25m, 40°)	+/-
Baseline	0.544	0.108	0.612	0%
Ours	0.557	0.110	0.624	4.7%
*Ours**	0.593	0.113	0.662	16.3%

b) Visual Localization: The following results are obtained using the *RIO10* evaluation framework[1], which allows quantifying the performance under different types of changes, *e.g.* visual, geometric and semantic, on the *RIO10* validation sequences. We visualize the overall results of our method compared to the baselines in Fig. 3. The results show that our method can localize a similar fraction of frames as D2-Net [14] in the high-precision $\varepsilon_f(0.05)$ and outperform all methods in the mid-precision $\varepsilon_f(0.15)$ zone.

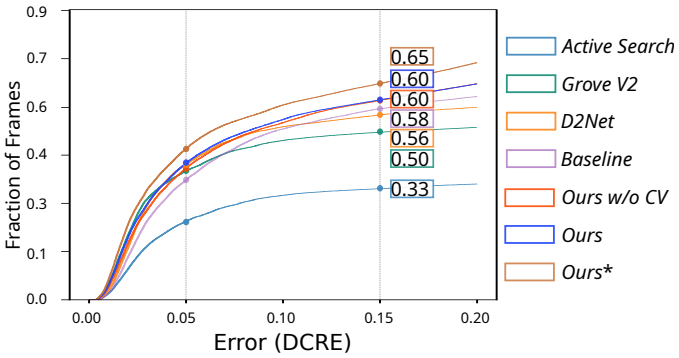

Fig. 3. Cumulative plots of DCRE for all camera relocalization methods on the RIO10 validation dataset [18].

The visual localization statistics in Table II show that our method achieves the highest SCORE according to the visual

[1] github.com/WaldJohannaU/RIO10

TABLE II
COMPARISON OF VISUAL LOCALIZATION STATISTICS OF ALL METHODS ON RIO10 VALIDATION DATASET [18]

	SCORE	$\varepsilon_f(0.05)$	$\varepsilon_f(0.15)$	ε_a(0.05m, 5°)	$\overline{\varepsilon}_f(0.5)$	N/A
Active Search [64]	1.22	0.236	0.335	0.081	**0.016**	0.6050
Grove v2 [9]	0.99	0.388	0.505	**0.202**	0.395	0.0
D2-Net [14]	1.30	0.408	0.557	0.135	0.107	0.0335
Baseline [11]	1.23	0.360	0.576	0.115	0.126	0.0017
Baseline 30 [11]	1.27	0.374	0.595	0.119	0.109	0.0027
Ours w/o CV	1.32	0.396	0.601	0.115	0.080	0.0066
Ours	**1.34**	**0.410**	**0.602**	0.116	0.074	0.0068
*Ours**	1.41	0.454	0.652	0.119	0.043	0.0034

localization performance measure from *RIO10* public benchmark. By comparing *Ours* and ours without correspondence verification (*Ours w/o CV*), the improvement validates the integration of our filtering method in local feature matches. We also show the baseline method with top-30 retrieved images (*Baseline 30*), which is claimed to achieve significantly better results than the methods using top-20 images [1]. Due to its extensive feature matching, we can outperform *Baseline 30*, which takes around 40% higher runtimes compared to the baseline. In terms of outliers, *Active Search* [64] provides less outlier pose estimates $\bar{\varepsilon}_f(0.5)$. However, *Active Search* generally does not localize sufficient good poses seen from the DCRE inlier statistics, which can be verified by checking the SCORE metric as it considers both inlier and outlier results and thus provides a real indicator for the performance of a method. Compared to learning methods, D2-Net [14] and Grove v2 [9] achieve the best performance in the threshold-based metrics ϵ_a, but their DCRE performance drop increasingly as the fraction of frames increases (see Fig. 3) which indicates that they tend to output inaccurate poses [18].

To better understand how different scene change factors affect the localization result, we plot the overall fraction of localized frames of all methods with increasing ζ_s, ζ_g, ρ_v, and ζ_v values and for a fixed DCRE error set to $\varepsilon_f(0.15)$ in Fig. 4. In the upper left sub-figure, *Ours* and *Ours** show increased localization accuracy of 4% and 10%, respectively comparing to the *Baseline* for large semantic differences ζ_s. Similarly, the results in the right upper corner visualize that we can achieve at least 5% better under increasing depth differences ζ_g. In the second row, the evaluation regarding visual appearance changes demonstrates that our method based on a robust local feature extraction allows outperforming other methods.

c) Case Study: We provide results on the SCORE metric in Tbl. III and can compare our method to the baseline method on each scene. We investigate the results of Tbl. III based on scene 6, which shows the lowest improvement in terms of the SCORE metric for our method. The left sub-figure in Fig. 5 shows the visual localization performance in this scene. We investigate the cause by checking the statistics in semantic and geometric changes (Fig. 5 right) and found that validation sequence 6 exhibits the lowest semantic and geometric change statistics. Our method only improves slightly over the baseline method under such a scenario. However, it achieves a better performance under more extensive changes as our method explicitly focuses on working under scene changes.

d) RIO10 Visual Localization Benchmark: We show the performance of our method on the *RIO10* test sequences in Tbl. IV compared to other approaches evaluated on the *RIO10* benchmark, which includes 54 unseen test sequences. Our method outperforms all others in $\varepsilon_f(0.15)$, while having a slightly worse SCORE metric compared to *D2-Net*. We suspect a high domain gap between the validation and test sequences in terms of semantics and the level of scene changes since our method outperforms *D2-Net* in most of the validation sequences. However, we cannot do further evaluation since the ground truth annotations for the test sequences are unavailable.

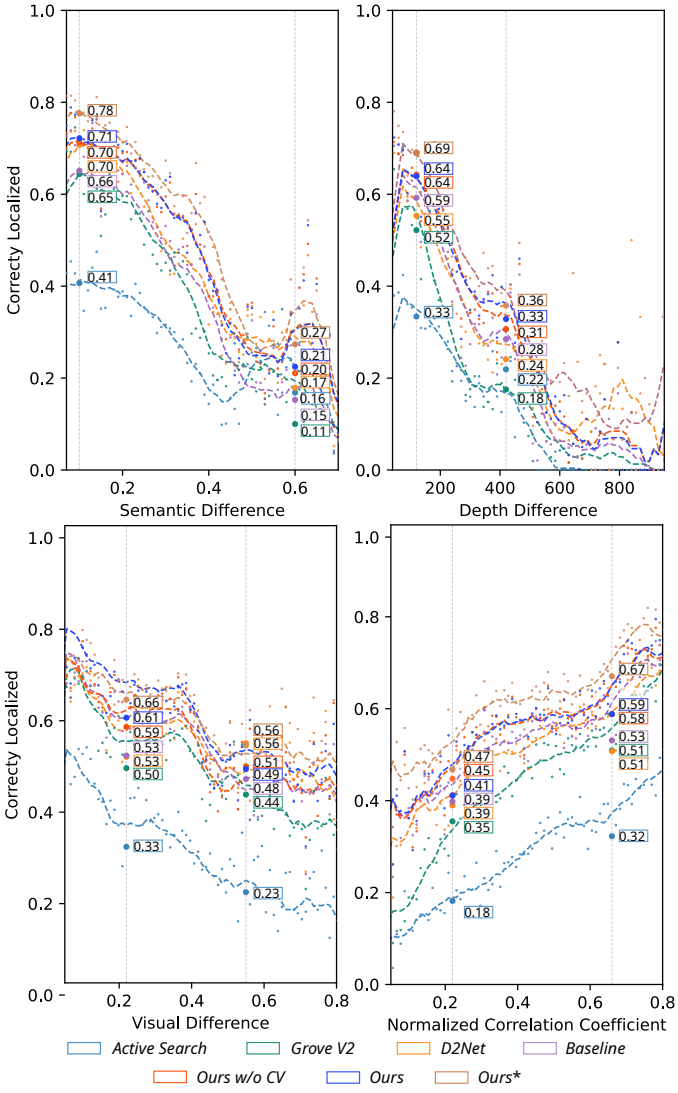

Fig. 4. Evaluation of fraction of frames localized according to semantic, geometric and visual appearance change measures. In comparison, our method is able to localize a higher fraction of frames under high semantic and geometric change.

V. Conclusion

In this work, we propose a novel pipeline to solve long-term visual localization in dynamic indoor scenes by adding awareness of scene change. We use scene graphs to estimate scene changes and use them for our robust change-aware image retrieval and correspondence verification. The results from the evaluation suggest that our method can generate accurate localization results and obtain a significantly higher fraction of reasonable pose estimates under high geometric and semantic scene changes. Furthermore, the result using ground truth instances and semantics shows the potential of our method with more robust instance and semantic estimation networks. To the best of our knowledge, this is the first work that integrates scene graphs for scene change detection and object re-identification in image retrieval. Possible directions for future works could be building a more detailed scene representation, such as scene graphs in 3D, and using semantic edge predicates for scene graph matching.

TABLE III
PER-SCENE COMPARISON OF *Baseline*, *D2-Net* AND *Ours* IN SCORE METRIC IN THE RIO10 DATASET [18].

SCORE	*Baseline* [11]	*D2-Net* [14]	*Ours*
Scene 01	1.37	1.47	**1.48**
Scene 02	1.53	1.53	**1.63**
Scene 03	1.26	**1.42**	1.33
Scene 04	0.93	1.00	**1.17**
Scene 05	1.40	**1.49**	**1.49**
Scene 06	1.66	**1.75**	1.68
Scene 07	1.54	1.57	**1.60**
Scene 08	1.06	1.06	**1.16**
Scene 09	1.11	1.12	**1.20**
Scene 10	1.00	1.09	**1.15**

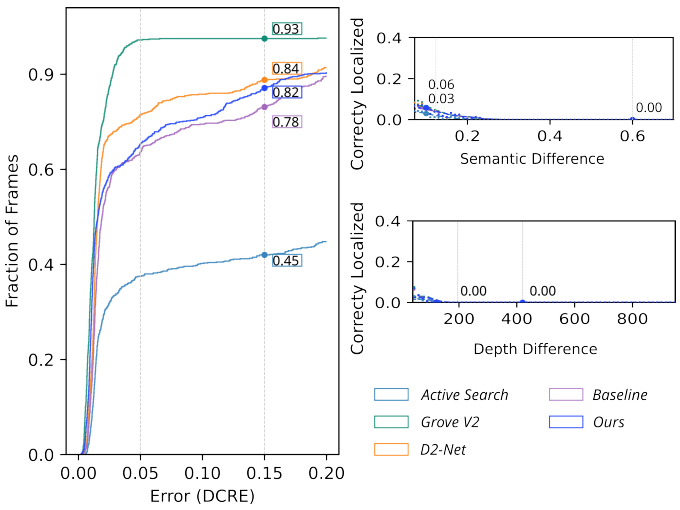

Fig. 5. Case study of Scene 6, which has the lowest scene change. When a scene has nearly no scene changes, our method outperforms slightly against the *Baseline* while failing to compare against other state-of-the-art methods.

REFERENCES

[1] N. Pion, M. Humenberger, G. Csurka, Y. Cabon, and T. Sattler, "Benchmarking image retrieval for visual localization," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 483–494.

[2] Y. Wu, F. Tang, and H. Li, "Image-based camera localization: an overview," *Visual Computing for Industry, Biomedicine, and Art*, vol. 1, no. 1, pp. 1–13, 2018.

[3] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular slam on a mobile phone," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 4, pp. 531–539, 2014.

[4] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg, "Wide area localization on mobile phones," in *2009 8th ieee international symposium on mixed and augmented reality*. IEEE, 2009, pp. 73–82.

[5] R. Castle, G. Klein, and D. W. Murray, "Video-rate localization in multiple maps for wearable augmented reality," in *2008 12th IEEE International Symposium on Wearable Computers*. IEEE, 2008, pp. 15–22.

[6] H. Lim, S. N. Sinha, M. F. Cohen, M. Uyttendaele, and H. J. Kim, "Real-time monocular image-based 6-dof localization," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 476–492, 2015.

TABLE IV
EVALUATION OF VISUAL LOCALIZATION APPROACHES ON THE RIO10 BENCHMARK [18].

Visual Localization Benchmark Statistics						
	SCORE	$\varepsilon_f(0.05)$	$\varepsilon_f(0.15)$	ε_a(0.05m, 5°)	$\bar{\varepsilon}_f(0.5)$	N/A
Active Search [64]	1.166	0.185	0.250	0.070	**0.019**	0.690
Grove v2 [9]	1.162	**0.416**	0.488	**0.274**	0.254	0.162
D2-Net [14]	**1.247**	0.392	0.521	0.155	0.144	0.014
Baseline [11]	1.238	0.382	0.558	0.138	0.144	0.000
Ours	1.241	0.384	**0.559**	0.140	0.143	0.000

[7] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger *et al.*, "Project autovision: Localization and 3d scene perception for an autonomous vehicle with a multi-camera system," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4695–4702.

[8] Y. Zhou, G. Wan, S. Hou, L. Yu, G. Wang, X. Rui, and S. Song, "Da4ad: End-to-end deep attention-based visual localization for autonomous driving," in *European Conference on Computer Vision*. Springer, 2020, pp. 271–289.

[9] T. Cavallari, S. Golodetz, N. A. Lord, J. Valentin, V. A. Prisacariu, L. Di Stefano, and P. H. Torr, "Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2465–2477, 2019.

[10] S. Dong, Q. Fan, H. Wang, J. Shi, L. Yi, T. Funkhouser, B. Chen, and L. J. Guibas, "Robust neural routing through space partitions for camera relocalization in dynamic indoor environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8544–8554.

[11] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, V. Leroy, J. Revaud, P. Rerole, N. Pion, C. de Souza, and G. Csurka, "Robust image retrieval-based visual localization using kapture," *arXiv preprint arXiv:2007.13867*, 2020.

[12] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time rgb-d camera relocalization," in *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2013, pp. 173–179.

[13] J. Valentin, A. Dai, M. Nießner, P. Kohli, P. Torr, S. Izadi, and C. Keskin, "Learning to navigate the energy landscape," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 323–332.

[14] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8092–8101.

[15] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *CVPR*, 2019.

[16] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1808–1817.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[18] J. Wald, T. Sattler, S. Golodetz, T. Cavallari, and F. Tombari, "Beyond controlled environments: 3d camera re-localization in changing indoor scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 467–487.

[19] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8601–8610.

[20] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 929–938.

[21] A. Torii, J. Sivic, and T. Pajdla, "Visual localization by linear combination of image descriptors," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 102–109.

[22] S. Yoon, W. Y. Kang, S. Jeon, S. Lee, C. Han, J. Park, and E.-S. Kim, "Image-to-image retrieval by learning similarity between scene graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 718–10 726.

[23] D. Acharya, K. Khoshelham, and S. Winter, "Bim-posenet: Indoor camera localisation using a 3d indoor model and deep learning from synthetic images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 245–258, 2019.

[24] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5974–5983.

[25] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 879–886.

[26] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6684–6692.

[27] T. Cavallari, L. Bertinetto, J. Mukhoti, P. Torr, and S. Golodetz, "Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 564–573.

[28] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[29] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[30] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 383–399.

[31] H. Taira, I. Rocco, J. Sedlar, M. Okutomi, J. Sivic, T. Pajdla, T. Sattler, and A. Torii, "Is this the right place? geometric-semantic pose verification for indoor visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4373–4383.

[32] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6896–6906.

[33] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Localization from semantic observations via the matrix permanent," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 73–99, 2016.

[34] J. Li, D. Meger, and G. Dudek, "Semantic mapping for view-invariant relocalization," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7108–7115.

[35] P. Weinzaepfel, G. Csurka, Y. Cabon, and M. Humenberger, "Visual localization by learning objects-of-interest dense match regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5634–5643.

[36] T. Jenicek and O. Chum, "No fear of the dark: Image retrieval under varying illumination conditions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9696–9704.

[37] H. Fan, Y. Zhou, A. Li, S. Gao, J. Li, and Y. Guo, "Visual localization using semantic segmentation and depth prediction," *arXiv preprint arXiv:2005.11922*, 2020.

[38] K. Sakurada, M. Shibuya, and W. Wang, "Weakly supervised silhouette-based semantic scene change detection," in *2020 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 6861–6867.

[39] J.-M. Park, J.-H. Jang, S.-M. Yoo, S.-K. Lee, U.-H. Kim, and J.-H. Kim, "Changesim: Towards end-to-end online scene change detection in industrial indoor environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8578–8585.

[40] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3961–3970.

[41] S. Kim, K.-n. Joo, and C.-H. Youn, "Graph neural network based scene change detection using scene graph embedding with hybrid classification loss," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2021, pp. 190–195.

[42] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[43] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7515–7525.

[44] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," in *Robotics: Science and Systems (RSS)*, 2022.

[45] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, 2018.

[46] Y. Liu, Y. Petillot, D. Lane, and S. Wang, "Global localization with object-level semantics and topology," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4909–4915.

[47] S. Lin, J. Wang, M. Xu, H. Zhao, and Z. Chen, "Topology aware object-level semantic mapping towards more robust loop closure," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7041–7048, 2021.

[48] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *CVPR 2011*. IEEE, 2011, pp. 2969–2976.

[49] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[50] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[51] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[52] J. L. Schönberger, "Robust methods for accurate and efficient 3d modeling from unstructured imagery," Ph.D. dissertation, ETH Zurich, 2018.

[53] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized ransac–full experimental evaluation," in *British machine vision conference*, vol. 2. Citeseer, 2012.

[54] L. Qi, J. Kuen, Y. Wang, J. Gu, H. Zhao, P. Torr, Z. Lin, and J. Jia, "Open world entity segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[56] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*. Springer, 2016, pp. 213–228.

[57] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[58] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Nießner, "Rio: 3d object instance re-localization in changing indoor environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7658–7667.

[59] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 564–571.

[60] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5107–5116.

[61] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2930–2937.

[62] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[63] 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.

[64] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1744–1756, 2016.