Indoor Visual Re-Localization for Long-Term Autonomous Robots Based on Object-Level Features and Semantic Relationships

Yuanyan Xie, *Member, IEEE*, Yu Guo, *Member, IEEE*, Zhenqiang Mi, *Member, IEEE*, Xiaokun Wang, Yang Yang, and Mohammad S. Obaidat, *Life Fellow, IEEE*

***Abstract*—Visual re-localization has become one of the key technologies for long-term autonomous robots. Existing methods, mostly focusing on addressing day-night, weather, and seasonal changes, are not applicable in indoor scenarios. At the same time, the layouts of objects in indoor scenes are highly dynamic over time due to human interactions with the environment, which makes indoor re-localization challenging. This letter presents a novel indoor visual re-localization method for long-term autonomous robots. First, a scene graph model is proposed, incorporating object-level features and semantic relationships, which overcomes the influence of dynamic objects by understanding the interactions among objects. Then, a visual re-localization method is developed based on the proposed scene graph model. It adopts graph matching technologies to incorporate pairwise object interactions as important features for re-localization, and designs a feature reweighting strategy to further reduce the impact of outliers in dynamic scenes. The proposed re-localization method has been verified in both photorealistic simulation environments and real-world scenarios. The results show that our approach exhibits higher robustness to diverse object changes and performs comparably to the state-of-the-art methods when illumination changes occur.**

***Index Terms*—Localization, long-term autonomy, mobile robots, RGB-D perception, semantic scene understanding.**

I. Introduction

VISUAL re-localization is a key ability for long-term autonomous robots and can determine where the robot is with a known map, which can help the robot maintain a consistent map when it repeatedly operates in a specific scenario. The core problem that needs to be solved is how to seek invariant and distinctive scene representation to achieve robust scene matching across diverse scene changes.

Due to the boom of self-driving vehicles, outdoor re-localization [1], [2], [3], [4] in large-scale urban environments has received much attention and focuses on addressing scene matching across day-night, weather, and seasonal changes. In contrast, indoor re-localization [5], [6], [7] has rarely been investigated in recent decades. At the same time, indoor re-localization is a more challenging problem than outdoor re-localization. 1) The backgrounds of indoor scenes are mostly similar as well as textureless and consist of walls, floors, ceilings, and windows. Therefore, indoor re-localization mainly depends on the features of objects in the foreground. 2) However, the layouts of objects in indoor scenes are highly dynamic over time caused by human interaction with the environment. This kind of object changes are more diverse, and harder to predict than that in outdoor scenes. 3) The appearance of indoor scenes also changes significantly due to complex changes in illumination, including both natural and artificial lighting. The above difficulties make indoor re-localization far from being solved [6].

Existing approaches can be categorized into three types: feature-based methods [8], [9], [10], learning-based methods [2], [11], [12], [13], and object-based methods [1], [4], [14], [15]. Feature-based methods extract handcrafted local features and establish matches between query images and known maps to achieve re-localization [8]. They assume that all local features in a scene are a whole and that there exists a rigid transformation between matched scenes, resulting in their sensitivity to dynamic points. Moreover, handcrafted features have limited illumination invariance and are not robust to changes in lighting conditions [10]. Learning-based methods utilize the powerful

Manuscript received 6 September 2022; accepted 20 December 2022. Date of publication 30 December 2022; date of current version 6 January 2023. This letter was recommended for publication by Associate Editor N. Radwan and Editor S. Behnke upon evaluation of the reviewers' comments. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grants FRF-TP-20-063A1Z and FRF-IDRY-20-018, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515110463, in part by EU Horizon 2020-Marie Skåodowska Curie Action-Individual Fellowships under Grant 895941, in part by the National Natural Science Foundation of China under Grant 61772068, and in part by the Ministry of Education of the People's Republic of China Overseas Distinguished Professorship Grant given to Prof. Obaidat under Grant MS2017BJKJ003. *(Corresponding author: Yu Guo.)*

Yuanyan Xie, Zhenqiang Mi, and Yang Yang are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: yyxie@xs.ustb.edu.cn; mizq@ustb.edu.cn; yyang@ustb.edu.cn).

Yu Guo is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, and also with the Shunde Innovation School, University of Science and Technology Beijing, Foshan 100083, China (e-mail: guoyu@ustb.edu.cn).

Xiaokun Wang is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, and also with the National Center for Computer Animation, Faculty of Media and Communication, Bournemouth University, Poole BH12 5BB, U.K. (e-mail: xwang1@bournemouth.ac.uk).

Mohammad S. Obaidat is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, also with the Computer science department and the Cybersecurity Center, University of Texas-Permian Basin, Odessa, TX 79762 USA, also with the King Abdullah II School of Information Technology, University of Jordan, Amman 11152, Jordan, and also with Amity University, Noida 201313, India (e-mail: m.s.obaidat@ieee.org).

Digital Object Identifier 10.1109/LRA.2022.3233235

representation capability of convolutional neural networks (CNNs) to learn local features that are robust to illumination and viewpoint changes [11] and to learn to aggregate local features [12], [13], reducing the influence of dynamic and distracting features on re-localization. However, they can only address simple and cyclic object changes in outdoor scenes, such as moving cars or pedestrians, and the growth of vegetation and cannot be applied to indoor scenes where object changes are more diverse and harder to predict [6]. In object-based methods, object instances are generally regarded as elementary units for scene matching [15]. Geometrical relationships among objects is also incorporated as important features [1], [14], and graph matching technologies are used to solve the re-localization problem. This makes object-based methods robust to dynamic objects to a certain extent. Unfortunately, object changes in indoor scenes are very drastic, and geometrical relationships among objects mostly featured by distance also change significantly. Existing object-based methods are unable to handle this kind of object changes.

This letter follows the line of object-based methods, but introduces semantic relationships to augment the scene representation instead of geometrical relationships. The semantic relationships are generated according to the prior semantic knowledge and the current scene layout, which are more robust against dynamic objects than geometrical relationships. Moreover, the letter adopts graph matching technologies to incorporate pairwise object interactions as important features for re-localization, and designs a feature reweighting strategy to further reduce the impact of outliers in dynamic scenes. The main contributions of this letter are summarized as follows:

1) A novel scene graph model is proposed that understands the scene at the level of object instances and can reason the semantic relationships among objects.
2) Based on the proposed scene graph model, a robust indoor visual re-localization method for long-term autonomous robots is proposed based on graph matching and feature reweighting.
3) The experimental results show that our approach exhibits higher robustness with diverse object changes and has performance comparable to that of state-of-the-art methods when illumination changes occur.

The remainder of this letter is organized as follows. Section II briefly reviews previous studies. Section III explains the proposed scene graph model. An indoor visual re-localization method is proposed for long-term autonomous robots based on the scene graph model in Section IV. Section V introduces the experimental setup, reports the experimental results, and analyzes the performance of the proposed approach. Finally, Section VI concludes the letter and discusses our future work.

II. Related Work

The visual re-localization problem has been traditionally solved by extracting local features from query images and establishing 2D-2D or 2D-3D correspondences between query images and the known maps [8]; thus, the camera poses can be estimated by multiple view geometry theory [16]. Obtaining the invariant and distinctive local features and avoiding incorrect data associations, however, remain extremely challenging due to the complex and time-varying environments in practical applications [6].

Classical approaches [8], [9], [10] have devised a series of state-of-the-art handcrafted shallow features, which are usually invariant to image scale and rotation and even provide robust matching with changes in viewpoints and illumination. Lowe et al. [9] computed scale-space extrema in the difference-of-Gaussians as keypoint locations to achieve scale invariance and adopted image gradient directions to obtain orientation-invariance descriptors. Rublee et al. [10] augmented FAST keypoints with pyramid schemes to obtain the scale-invariant features and measured the keypoint orientation using the intensity centroid. To achieve illumination invariance, existing methods utilize the relative values of intensities or gradients to acquire robust descriptions [9], [10], but they cannot work when drastic illumination changes occur, such as day-night or weather changes [8].

With the boost of CNNs, many learning-based approaches [2], [11], [12], [13] have emerged for robust localization in changing environments. DeTone et al. [11] applied deep learning to both interest point detection and description to acquire sparse as well as rich features, which shows good invariance to changes in illumination and viewpoints. Arandjelovic et al. [12] developed a training procedure based on the imagery of the same places over time to learn the means of aggregating local features. It was robust to viewpoint and illumination changes and could deal with object changes in outdoor scenarios, including moving cars or pedestrians, and changes in vegetation. Jin et al. [13] designed a contextual reweighting network (CRN) to predict the importance of different regions in a feature map, thus reducing the influence of distracting and non-discriminative features on place recognition. Sarlin et al. [2] combined SuperPoint [11] and NetVLAD [12] to realize robust hierarchical localization in large-scale scenarios.

In recent years, object-based localization methods have become promising solutions for achieving robust localization, as object-level features have excellent invariance to object appearance changes. Yang et al. [15] optimized camera poses by combining object-level features with handcraft local features, which focuses on addressing perception aliasing and incorrect data associations. Wang et al. [17] replaced handcraft local features with CNN features to augment the scene representation, which are more robust to illumination and viewpoint changes. But unfortunately, object-level features cannot always be a type of invariant scene representation because human interactions with environments may significantly change the layout of objects in the same place. Rosinol et al. [18] constructed 3D dynamic scene graphs for robotic motion planning, but they did not provide the localization methods based on scene graphs. Liu et al. [14] and Lin et al. [4] realized the scene-graph-based localization, with objects as nodes and geometrical relationships among objects as edges. These methods adopted graph matching technologies, which make them robust to dynamic objects to a certain extent. However, they fail to handle drastic object changes, because geometrical relationships featured by distance will change significantly when the layout of objects is changed drastically.

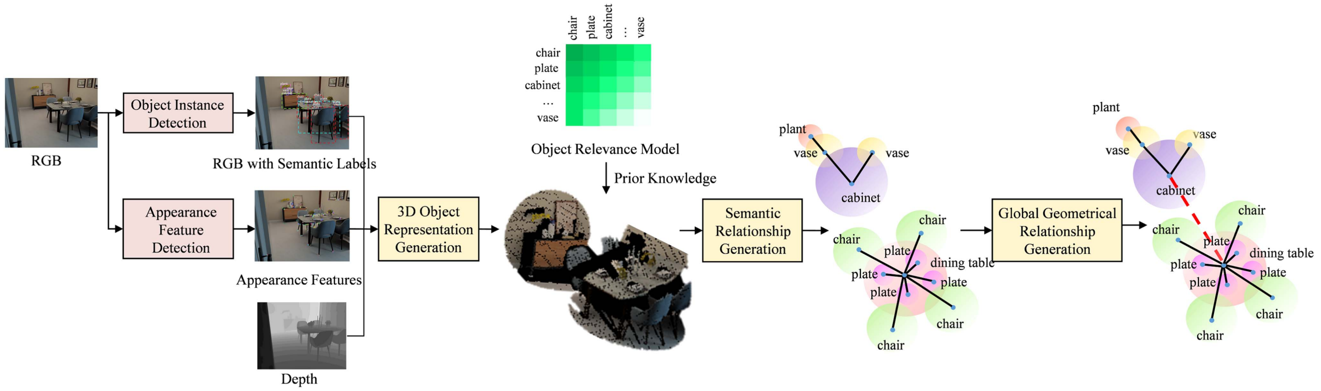


Fig. 1. Process of scene graph model construction, where black lines represent the semantic relationships and red lines represent the global geometrical relationships.

III. Scene Graph Model With Object-Level Features and Semantic Relationships

This section presents a novel scene graph model that understands the scene at the level of object instances and can reason the semantic relationships among objects, showing its strong robustness against dynamic environments, especially when objects are moved or replaced. More specifically, a scene is formulated as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the graph nodes $\nu \in \mathcal{V}$ denote objects observed in the scene, and the edges of graph $e = (\nu_i, \nu_j) \in \mathcal{E}$ represent pairwise object interactions. As shown in Fig. 1, the construction of the scene graph model consists of two steps, i.e., object representation and semantic relationship generation. The details are introduced below.

A. Object Representation

Due to the superior invariance to object appearance changes, object-level features have been adopted in recent works [1], [4], [14] to deal with dynamic environments. However, the object-level features obtained by modern object detectors [19] contain only the semantic label information of objects, so they are not sufficiently discriminative for place recognition. This letter adopts the joint representation of the semantic label c and appearance s information for object description. To describe the object appearance while retaining invariance with illumination changes, this letter applies SuperPoint [11], which is a state-of-the-art learned feature and has shown good robustness to changes in illumination.

In addition to the semantic label and appearance information, the size of the object is also introduced to assess the matches among object-level features. Previous studies have used cubes [15], ellipsoids [20], and other geometries to represent object landmarks. These methods need multiple parameters to describe the object size, but it is difficult to precisely measure these parameters due to the limitation of observation viewpoints. Therefore, this letter utilizes a sphere to describe the size of an object with only one parameter, spherical radius r. It can be calculated by (1).

$$r = \frac{z\sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}}{2f_x} \tag{1}$$

where (u_1, v_1) and (u_2, v_2) are the pixel coordinates of the top-left and bottom-right corners of the object bounding box, respectively, f_x is the intrinsic parameter of the camera, and z is the depth of the center of the object bounding box.

The position of an object is also important information for object association. The 2D positions on image planes cannot reflect the spatial relationships among objects [21]. In contrast, this letter uses the 3D coordinate of spherical center $X = [x, y, z]$, which can be calculated by

$$\begin{bmatrix} x & y & z \end{bmatrix} = z \begin{bmatrix} u_c & v_c & 1 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \end{bmatrix}^{\mathrm{T}} \tag{2}$$

where (u_c, v_c) is the center of the object bounding box, and f_x, f_y, c_x, and c_y are all the intrinsic parameters of the camera.

Consequently, an object is represented as $\nu = (c, s, r, X)$, where the semantic label c, appearance s, size r, and position X of the object are considered.

B. Semantic Relationship Generation

However, object-level features are still not sufficient for place recognition, as many objects may change their positions due to human activities interacting with environments, especially in indoor scenarios [6]. To solve this problem, this letter introduces the semantic relationships among objects to augment the scene representation, which can better understand the object interactions and overcome the influence of dynamic objects.

The semantic relationships that this letter aims to generate are actually the most stable relationships among objects and have lower probabilities to be changed in dynamic environments. To achieve this goal, this letter adopts the indoor object relevance model proposed in our previous work [22], and the semantic realtionships are established among the most relevant objects, taking the scene spatial layout into consideration. The object relevance refers to the co-occurrence probability of objects. The higher the probability is, the higher relevance they have. In [22], the object relevance model was obtained by extracting the object co-occurrence information and the spatial distance among objects from large indoor scenario datasets and training these data with the skip-gram model.

In this letter, we suppose that the relationships among the most relevant objects are more stable when scenes are changed, as the probabilities of these relationships being changed are lower than those of other relationships with lower relevance. Moreover, semantic relationships do not change when objects are moved within an acceptable distance and are more robust than geometrical relationships. Note that there is no semantic relationship when the distance between two objects exceeds a certain range, even though they are highly relevant in the object relevance model.

To clarify our strategy, the definition of the distance among objects is first given. For objects ν_i^P and ν_j^P in scene P, the distance between them can be defined as given below:

$$D(\nu_i^P, \nu_j^P) = \delta(|X_i - X_j| - (r_i + r_j)) \tag{3}$$

where $\delta(x)$ is an active function. $\delta(x) = x$, if $x > 0$; $\delta(x) = 0$, if $x \leq 0$. X_i and X_j are the center positions of objects. r_i and r_j represent the radiuses of objects.

The local area of an object ν_i^P can be defined as (4), where X represents any 3D point. The local area can be regarded as a new object with the center of X_i and the radius of $2r_i$. The distance definition in (3) can also apply to local areas.

$$N(\nu_i^P, 2r_i) = \left\{X \middle| \, |X - X_i| \leq 2r_i\right\} \tag{4}$$

Thus, the neighbors of an object ν_i^P can be defined as the objects that have a distance of zero from the local area of ν_i^P, as shown in (5), where $\mathcal{V}^P$ represents all objects in scene P.

$$A_i = \{\nu_j^P | D(\nu_j^P, N(\nu_i^P, 2r_i)) = 0, \forall \nu_j^P \in \mathcal{V}^P - \{\nu_i^P\}\} \tag{5}$$

The semantic relationship e_l can be established between object ν_i^P and the most relevant object ν^P within its neighbors A_i, as shown in (6).

$$\nu^P = \max_{\nu_j^P \in A_i} \{\psi_{ij}\} \tag{6}$$

where ψ_{ij} is the relevance between objects ν_i^P and ν_j^P. Please refer to [22] for the details about the relevance calculation.

After finishing the abovementioned semantic relationship generation, the constructed scene graph may consist of multiple connected branches. The objects in different connected branches cannot establish the semantic relationships because they are far from each other. To describe the relationships among connected branches, we establish the global geometrical relationships e_g among different connected branches. More specifically, the node with the maximum node degree in each connected branch is regarded as the center of the connected branch, and the global geometrical relationships are established among these center nodes in the manner of full connection.

IV. Indoor Visual Re-Localization Method Based on Graph Matching and Feature Reweighting

In this section, a novel visual re-localization method for long-term autonomous robots is proposed based on graph matching technology and feature reweighting, which can incorporate pairwise object interactions as the important features to establish correspondences between scenes and be robust to the deformation and outliers in dynamic scenes. An overview of the proposed visual re-localization method is shown in Fig. 2.

A. Graph Matching Theory

With the scene graph model presented in Section III, the re-localization problem can be reduced to a graph matching problem. Usually, the graph matching problem [23] can be interpreted as finding out the node correspondences between two graphs to maximize the sum of node and edge affinities, as shown in (7). $\mathcal{V}^P$ and $\mathcal{V}^Q$ represent the sets of nodes in graph P and graph Q, respectively. An assignment matrix $\mathcal{X} \in \{0, 1\}^{|\mathcal{V}^P| \times |\mathcal{V}^Q|}$ is defined to represent the correspondence. $\mathcal{X}_{ia} = 1$ indicates that node correspondence between $\nu_i^P \in \mathcal{V}^P$ and $\nu_a^Q \in \mathcal{V}^Q$ exists. $K^{|\mathcal{V}^P||\mathcal{V}^Q| \times |\mathcal{V}^P||\mathcal{V}^Q|}$ represents the affinity matrix, which is defined to measure the confidence of node and edge correspondences between two graphs. A diagonal element $k_{ia;ia}$ represents the unary affinity of the node correspondence (ν_i^P, ν_a^Q), and a non-diagonal term $k_{ia;jb}$ denotes the pairwise affinity between node correspondences (ν_i^P, ν_a^Q) and (ν_j^P, ν_b^Q).

$$\begin{aligned} &x^* = \arg\max(vec(\mathcal{X})^T K vec(\mathcal{X})) \\ &s.t. \quad \forall i \textstyle\sum_{a=1}^{|\mathcal{V}_Q|} \mathcal{X}_{ia} \leq 1, \forall a \textstyle\sum_{i=1}^{|\mathcal{V}_P|} \mathcal{X}_{ia} \leq 1 \end{aligned} \tag{7}$$

B. Object Association With Random Walks

Since the graph matching problem formulated in (7) is NP-hard, approximate solutions are required. This letter adopts Markov random walk statistics [23], which has shown excellent performance on a wide range of applications, such as object recognition and image matching.

First, an association graph, also known as a direct product graph [24], is used to demonstrate the candidate correspondences between two graphs. For two graphs $\mathcal{G}^P = (\mathcal{V}^P, \mathcal{E}^P)$ and $\mathcal{G}^Q = (\mathcal{V}^Q, \mathcal{E}^Q)$, the association graph is defined as $\mathcal{G}^\times = \mathcal{G}^P \times \mathcal{G}^Q = (\mathcal{V}^\times, \mathcal{E}^\times)$, where $\mathcal{V}^\times = \{(\nu^P, \nu^Q) | \nu^P \in \mathcal{V}^P \wedge \nu^Q \in \mathcal{V}^Q\}$ and $\mathcal{E}^\times = \{\{(\nu_i^P, \nu_a^Q), (\nu_j^P, \nu_b^Q)\} | (\nu_i^P, \nu_j^P) \in \mathcal{E}^P \wedge (\nu_a^Q, \nu_b^Q) \in \mathcal{E}^Q\}$. In this letter, the size of the association graph can be reduced by some distinctive node or edge attributes. Only when two nodes ν^P and ν^Q have the same object semantic label and their object size difference is within a preset threshold, does the node correspondence exist, i.e., $(\nu^P, \nu^Q) \in \mathcal{V}^\times$. For two edges e^P and e^Q, edge correspondence exists when the edge types are consistent. When e^P and e^Q are both global geometrical edges, the lengths of edges e^P and e^Q, represented as d^P and d^Q, should satisfy $|d^P - d^Q| / \max\{d^P, d^Q\} < \varepsilon$.

An affinity matrix K can be defined on the association graph. The affinities of node correspondences (i.e., diagonal elements $k_{ia:ia}$ mentioned in IV.A) are calculated by comparing the appearance similarities of objects, and the affinities of edge correspondences (i.e., non-diagonal elements $k_{ia:jb}$ mentioned in IV.A) are the product of the affinities of pairwise node correspondences.

Then, a random walk is defined on the association graph; it takes off from an arbitrary node and then successively visits

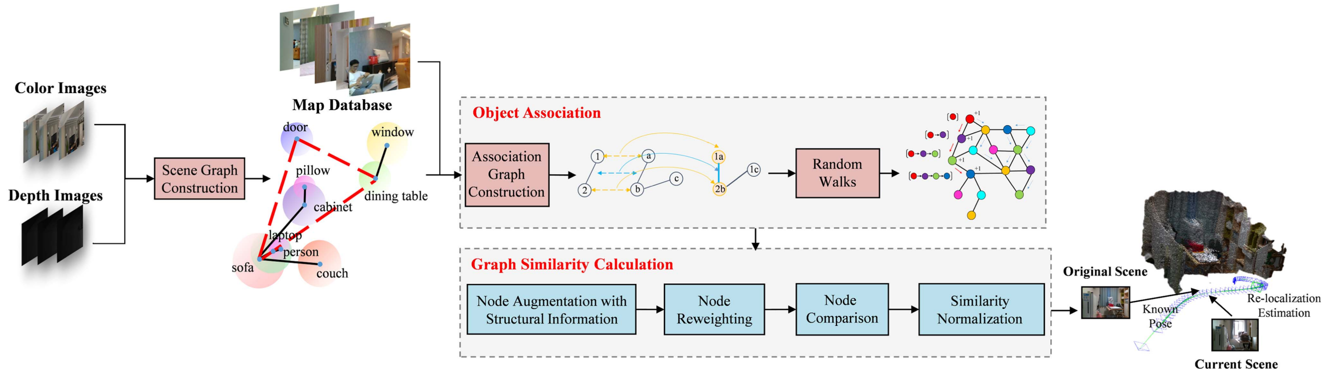


Fig. 2. Overview of the proposed visual re-localization method.

adjacent nodes according to Markov transition probabilities. Nodes in the association graph $\mathcal{G}^{\times}$ (i.e., the node correspondences between $\mathcal{G}^P$ and $\mathcal{G}^Q$) are ranked by the frequency of random walkers passing by. In this way, the graph matching problem can be transformed into node ranking and selection on the association graph.

The transition matrix can be derived from the affinity matrix K. The probabilities that nodes can be selected as the start of a walk are calculated by

$$P_{ia} = \frac{K_{ia;ia}}{\sum_{\left(\nu_i^P,\nu_a^Q\right)\in\mathcal{V}^{\times}} K_{ia;ia}} \tag{8}$$

And the transition probabilities with the current node of (ν_i^P, ν_a^Q) are calculated by

$$P_{ia;jb} = \frac{K_{ia;jb}}{\sum_{\left(\nu_j^P,\nu_b^Q\right)\in\mathcal{V}^{\times}} K_{ia;jb} - K_{ia;ia}} \tag{9}$$

In the above process, the matching constraints shown in (7) are temporarily ignored. In this letter, the final node correspondences are determined using a post-processing discretization step such as that in [25] to check the matching constraints. Specifically, node correspondences are checked in descending order of their weights. For node correspondence (ν_i^P, ν_a^Q), object ν_i^P in graph $\mathcal{G}^P$ and object ν_a^Q in graph $\mathcal{G}^Q$ are successfully matched if both ν_i^P and ν_a^Q have not been matched with other object instances in earlier rounds.

C. Similarity Metric Calculation With Feature Reweighting

With the random walk weighting, node matches between two graphs can be obtained. However, unlike [23], the objective of this letter is not to establish the exact correspondences between visual features from two scenes, but to find the most similar scene from a map database with respect to visual features. Therefore, a similarity metric is defined to measure the consisency of any two graphs, according to the obtained node matches.

Graph similarity is generally calculated using the high-order interactions among nodes [24], [26], which scales with at least $O(n^3)$, where n represents the number of nodes. To handle large-scale scenarios, researchers attempt to augment the representation of nodes with structural information, and compute the graph similarity by directly comparing the unary node information [27]. This letter adopts an approach similar to [27] based on the node weights computed in Section IV.B, which jointly maps the unary affinity of node correspondences and pairwise compatibilities among node correspondences to node weights. Note that node weights are only associated with the relative affinity values, and thus, graph similarity should be calculated by multiplying node weights with the absolute affinity values:

$$\kappa(\mathcal{G}^P,\mathcal{G}^Q) = \sum_{(\nu_i^P,\nu_a^Q)\in\mathcal{V}_*^{\times}} \frac{\omega_{ia}}{nl} K_{ia;ia} \tag{10}$$

where $\mathcal{V}_*^{\times}$ denotes the successfully matched nodes in association graph $\mathcal{G}^{\times}$, ω_{ia} is the weight value of node (ν_i^P, ν_a^Q), n and l represent the number and length of random walks, respectively, and $\frac{\omega_{ia}}{nl}$ refers to the average increase in the weight of node (ν_i^P, ν_a^Q) when random walkers move one step on the association graph.

The graph similarity defined in (10) is highly related to the number of successfully matched nodes. Any moved, replaced, or deformable object in the scene imposes a significant impact on the scene graph similarity. To alleviate the influence of dynamic objects, this letter reweights the object-level features by the term frequency-inverse document frequency (TF-IDF), which has been used to weight the visual geometrical features in BoW [8]. Therefore, graph similarity can be defined as

$$\kappa(\mathcal{G}^P,\mathcal{G}^Q) = \sum_{(\nu_i^P,\nu_a^Q)\in\mathcal{V}_*^{\times}} \eta_i^P \frac{\omega_{ia}}{nl} K_{ia;ia} \tag{11}$$

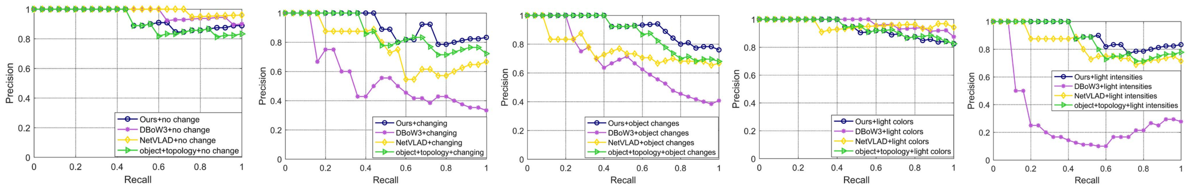
where η_i^P represents the importance of object ν_i^P in graph $\mathcal{G}^P$ and can be calculated by

$$\eta_i^P = \frac{|\{\nu_j^P \in \mathcal{V}^P | c(\nu_j^P) = c(\nu_i^P)\}|}{|\mathcal{V}^P|} \ln \frac{|\Omega|}{|\{\mathcal{G}^P \in \Omega | c(\nu^P) = c(\nu_i^P), \exists \nu^P \in \mathcal{V}^P\}|} \tag{12}$$

where Ω represents all scenes in the map database.

Finally, we normalize the graph similarity by the self-similarity of the query graph:

$$\kappa'(\mathcal{G}^P,\mathcal{G}^Q) = \frac{\kappa(\mathcal{G}^P,\mathcal{G}^Q)}{\kappa(\mathcal{G}^P,\mathcal{G}^P)} \tag{13}$$



(a) (b) (c) (d) (e)

Fig. 3. Precision-recall curves of different methods with different configurations of scene dynamic factors: (a) static environments (b) the environments with complex scene changes; (c) the environments with object changes; (d) the environments with light color changes; and (e) the environments with light intensity changes.

TABLE I
RECALL@5 OF DIFFERENT RE-LOCALIZATION METHODS IN THE ENVIRONMENTS WITH DIFFERENT CONFIGURATIONS OF SCENE DYNAMIC FACTORS

Recall@5	**Ours**	**DBoW3**	**NetVLAD**	**object +topology**
IndoorVR-OC+Object	**0.76**	0.4	0.68	0.68
IndoorVR-OC+NoChange	0.84	0.88	**0.92**	0.84
InteriorNet+LightColors	0.825	0.875	**0.95**	0.825
InteriorNet+NoChange	0.85	0.9	**0.95**	0.85
InteriorNet+LightIntensities	**0.825**	0.275	0.725	0.775
InteriorNet+NoChange	0.85	0.9	**0.95**	0.85
OpenLORIS+Changing	**0.84**	0.32	0.68	0.72
OpenLORIS+NoChange	0.88	0.92	**0.96**	0.84

TABLE II
THE RE-LOCALIZATION PERFORMANCE OF OUR APPROACHES WITH DIFFERENT COMPONENT CHOICES

Methods	**Recall@3**	**Recall@5**	**Recall@10**
Ours_w/o_SGM	0.659	0.723	0.826
Ours_w/o_FR	0.712	0.835	0.867
Ours	**0.745**	**0.846**	**0.889**

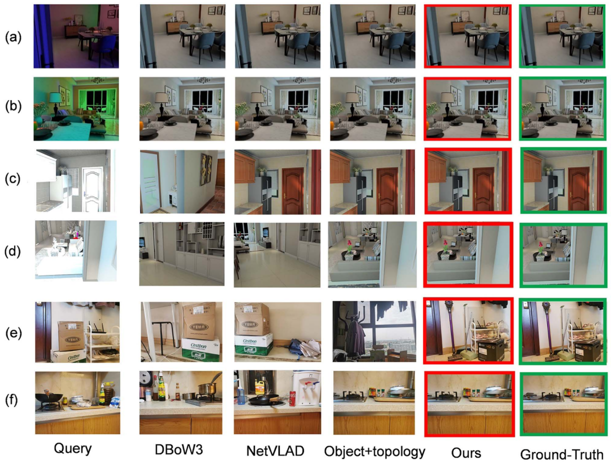


Fig. 4. Examples of retrieved images with a single scene dynamic factor changed. In lines (a) and (b), light colors are changed. In lines (c) and (d), light intensities are changed. In lines (e) and (f), objects are changed.

V. EXPERIMENTS

A. Experimental Settings and Evaluation Metrics

Evaluation Datasets: The OpenLORIS-Scene benchmark dataset[1] was used to evaluate our method, which consists of more than 22 data sequences captured from real-life indoor scenarios with RGB-D cameras. It records the imagery of scenes under various lighting conditions and different layouts. In our experiments, corridor scenarios in this dataset were excluded, as featureless regions occurring in corridors are not considered in this letter.

Our experiments were also conducted on the InteriorNet benchmark dataset.[2] It consists of 22 million synthetic indoor scenes and can flexibly configure different scene settings, including lighting, illumination, and object layouts. In addition, we collected our own dataset, named IndoorVR-OC, using TurtleBot3 equipped with Kinect V2 in indoor scenarios. It records the imagery of scenes with objects moved, replaced, or deformed under the same lighting conditions. Object changes in scenes were deliberately made by authors; however, all changes are very likely to occur in daily life.

Evaluation Metrics: Similar to [12], this letter adopts the recall $Recall@N$ given the top N candidates to measure the performance of our approach. It refers to the probability of correctly matched scenes being in the top N candidates. Precision-recall curves were also used to evaluate our method. Recall here refers to the proportion of correctly re-localized query images to all query images that have matched scenes in known maps. Precision refers to the proportion of correctly re-localized query images to all successfully re-localized images. In our experiments, a re-localized pose is regarded as the correct pose when the translation error and orientation error are within 0.25˜m and 5°, respectively.

Contrastive Methods: To perform a contrastive analysis with existing methods, this letter compared our method with the classical feature-based approach DBoW3 [28], the state-of-the-art CNN-based method NetVLAD [12], and the representative object-based localization method [14] (referred to as object+topology in the following part of this letter).

B. Re-Localization Performance

First, the re-localization performance of our approach was evaluated in the environments with single scene dynamic factor changes, such as the changes in light colors, light intensities, and scene layouts. Then, our approach's performance was further evaluated in daily changing indoor environments, where multiple dynamic factors may be changed.

[1] [Online]. Available: https://lifelong-robotic-vision.github.io/dataset/scene

[2] [Online]. Available: https://interiornet.org/

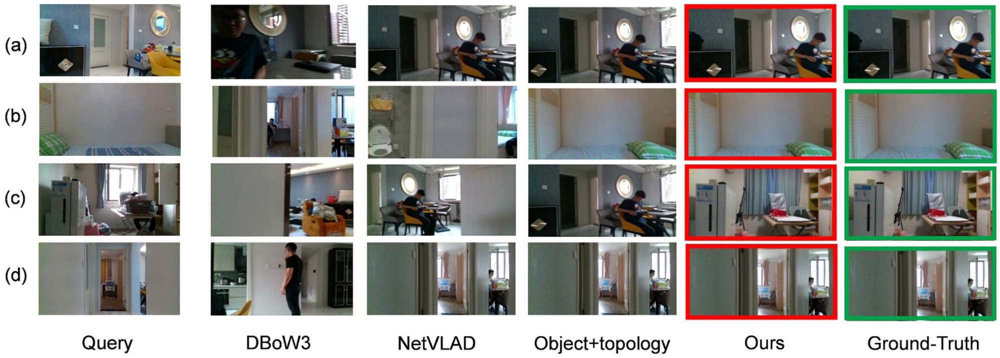


Fig. 5. Examples of retrieved images with complex scene changes.

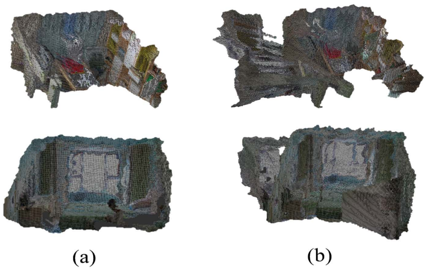


Fig. 6. Maps constructed with different methods: (a) our approach and (b) ORB_SLAM3.

Table I presents the $Recall@5$ values of different methods in the environments with different configurations of scene dynamic factors. As shown in Table I, the recalls of our method, DBoW3, NetVLAD, and object+topology have slight declines when only the light colors of scenes are changed. When objects or light intensities are changed in scenes, the recalls of all four methods decline more obviously. DBoW3 has the lowest recalls among the four methods. The recalls of NetVLAD and object+topology with objects changed are lower than those with light intensity changed. Our approach has the highest recalls with either objects or light intensity changed.

Table I also reports the $Recall@5$ values of different methods in daily changing environments. The results show that DBoW3 and NetVLAD can re-localize query images with high recalls in static environments, but suffer drastic performance degradation in changing environments. In contrast, the recalls of our approach and object+topology change more slightly when moving from static environments to changing environments, and our approach surpasses object+topology.

Fig. 3 shows the precision-recall curves of different methods in different environments. Fig. 3(a) reports the experimental results in static environments. It can be found that the precision of DBoW3 and NetVLAD is higher than that of ours and object+topology with the same recalls. The superiority of DBoW3 and NetVLAD can be maintained when only light colors are changed in the environments, as shown in Fig. 3(d). But when light intensities (Fig. 3(e)) or objects (Fig. 3(c)) are changed in the environments, DBoW3 and NetVLAD suffers an obvious decline in precision. The precision decline of DBoW3 is more severe than NetVLAD, and DBoW3 shows higher precision with object changes than with light intensity changes. In contrast, our method can maintain higher precision when light intensities or objects are changed. Fig. 3(b) further demonstrates the precision-recall curves of different methods in complex dynamic environments. The results show that the precision of our method is highest among four methods with the same recalls.

From the above results, it can be found that our approach exhibits higher robustness than state-of-the-art methods in complex indoor dynamic environments and shows superiority especially when diverse object changes occur.

Besides, to verify the effectiveness of scene graph model and feature reweighting, we also compare the re-localization performance of our approach with the method that replaces the proposed scene graph model with geometry-based graph model (i.e., Ours_w/o_SGM) and the method without the feature reweighting strategy (Ours_w/o_FR). As shown in Table II, our approach exhibits the highest recall values. This is because our scene graph model can establish more stable relationships among objects against scene changes and the feature reweighting strategy can further alleviate the influence of dynamic objects.

C. Qualitative Evaluation

Finally, to intuitively demonstrate the characteristics of different methods, Figs. 4 and 5 show some examples of our results, which display the top retrieved images in known maps for various queries using different methods. As can be seen, our approach retrieves correct images despite the presence of both dynamic objects (Fig. 4(e)–(f) and Fig. 5(a)–(c)) and changed illumination (Fig. 4(a)–(d) and Fig. 5(a), (c)–(d)). NetVLAD shows good robustness to illumination changes (Fig. 4(a)–(c) and Fig. 5(a), (d)) and simple object changes (Fig. 5(a)), but fails to retrieve correct images with diverse object changes (Fig. 4(e)–(f) and Fig. 5(c)). The object+topology approach performs well in the presence of illumination changes (Fig. 4(a)–(d) and Fig. 5(a), (d)), but shows poor robustness to object changes (Fig. 4(e)–(f) and Fig. 5(c)). DBoW3 has limited illumination invariance (Fig. 4(a)–(b)) and is not robust to object changes (Fig. 4(e)–(f) and Fig. 5(a)–(d)).

Moreover, we integrated the proposed re-localization module into the mapping system and compared the mapping performance with that of ORB_SLAM3, which adopts DBoW3 as the re-localization module. Fig. 6 demonstrates an example of the maps constructed with our re-localization approach and ORB_SLAM3. The results show that our method (Fig. 6(a)) successfully localizes the robot in the known map and constructs a consistent map. ORB_SLAM3 (Fig. 6(b)) gives the wrong re-localization in the known map and results in map aliasing.

VI. Conclusion

This letter presented a novel indoor visual re-localization method for long-term autonomous robots. A scene graph model with object-level features and semantic relationships was proposed. The semantic relationships are generated according to the prior semantic knowledge and the current scene layout, showing strong robustness against dynamic environments, especially those with diverse object changes. Based on the proposed scene graph model, a visual re-localization method was proposed based on graph matching and feature reweighting, which incorporates pairwise object interactions as the important features to establish correspondences between scenes and is robust to the deformation and outliers in dynamic scenes. The results show that our approach exhibits higher robustness to diverse object changes and has performance comparable to that of state-of-the-art methods for environments with illumination changes.

In the future, we will focus on the optimization of our re-localization method and speed up our approaches to satisfy the real-time requirements for more rigorous robotic applications based on our previous work on cloud robotics [29].

References

[1] X. Guo, J. Hu, J. Chen, F. Deng, and T. L. Lam, “Semantic histogram based graph matching for real-time multi-robot global localization in large scale environment,” *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8349–8356, Oct. 2021.

[2] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, “From coarse to fine: Robust hierarchical localization at large scale,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12716–12725.

[3] T. Sattler et al., “Benchmarking 6DOF outdoor visual localization in changing conditions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8601–8610.

[4] S. Lin, J. Wang, M. Xu, H. Zhao, and Z. Chen, “Topology aware object-level semantic mapping towards more robust loop closure,” *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 7041–7048, Oct. 2021.

[5] H. Taira et al., “InLoc: Indoor visual localization with dense matching and view synthesis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1293–1307, Apr. 2021.

[6] J. Wald, T. Sattler, S. Golodetz, T. Cavallari, and F. Tombari, “Beyond controlled environments: 3D camera re-localization in changing indoor scenes,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 467–487.

[7] H. Taira et al., “Is this the right place? Geometric-semantic pose verification for indoor visual localization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4373–4383.

[8] D. Gálvez-López and J. D. Tardos, “Bags of binary words for fast place recognition in image sequences,” *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[9] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to SIFT or SURF,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN architecture for weakly supervised place recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[13] H. Jin Kim, E. Dunn, and J.-M. Frahm, “Learned contextual feature reweighting for image geo-localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2136–2145.

[14] Y. Liu, Y. Petillot, D. Lane, and S. Wang, “Global localization with object-level semantics and topology,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 4909–4915.

[15] S. Yang and S. Scherer, “CubeSLAM: Monocular 3-D object slam,” *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 925–938, Aug. 2019.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[17] Y. Wang, Y. Qiu, P. Cheng, and X. Duan, “Robust loop closure detection integrating visual–spatial-semantic information via topological graphs and cnn features,” *Remote Sens.*, vol. 12, no. 23, 2020, Art. no. 3890.

[18] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, “3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans,” in *Proc. Robot.: Sci. Syst.*, 2020, pp. 1–11.

[19] H. Liu, R. A. R. Soto, F. Xiao, and Y. J. Lee, “YolactEdge: Real-time instance segmentation on the edge,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 9579–9585.

[20] L. Nicholson, M. Milford, and N. Sünderhauf, “QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented slam,” *IEEE Robot. Automat. Lett.*, vol. 4, no. 1, pp. 1–8, Jan. 2019.

[21] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, “X-View: Graph-based semantic multi-view localization,” *IEEE Robot. Automat. Lett.*, vol. 3, no. 3, pp. 1687–1694, Jul. 2018.

[22] Y. Guo, Y. Xie, Y. Chen, X. Ban, B. Sadoun, and M. S. Obaidat, “An efficient object navigation strategy for mobile robots based on semantic information,” *Electronics*, vol. 11, no. 7, 2022, Art. no. 1136.

[23] M. Cho, J. Lee, and K. M. Lee, “Reweighted random walks for graph matching,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 492–505.

[24] A. Feng, C. You, S. Wang, and L. Tassiulas, “KerGNNs: Interpretable graph neural networks with graph kernels,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 6, pp. 6614–6622.

[25] M. Leordeanu and M. Hebert, “A spectral technique for correspondence problems using pairwise constraints,” in *Proc. IEEE 10th Int. Conf. Comput. Vis.*, 2005, pp. 1482–1489.

[26] T. Gärtner, P. Flach, and S. Wrobel, “On graph kernels: Hardness results and efficient alternatives,” in *Proc. Learn. Theory Kernel Mach.*, 2003, pp. 129–143.

[27] E. Stumm, C. Mei, S. Lacroix, J. Nieto, M. Hutter, and R. Siegwart, “Robust visual place recognition with graph kernels,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4535–4544.

[28] C. Campos, R. Elvira, J. J. G. Rodrlguez, J. M. Montiel, and J. D. Tardós, “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam,” *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[29] Y. Xie, Y. Guo, Z. Mi, Y. Yang, and M. S. Obaidat, “Edge-assisted real-time instance segmentation for resource-limited iot devices,” *IEEE Internet Things J.*, vol. 10, no. 1, pp. 473–485, Jan. 2023, doi: 10.1109/JIOT.2022.3199921.