arXiv:submit/5122002 [cs.RO] 19 Sep 2023

Vision-based Situational Graphs Generating Optimizable 3D Scene Representations

Ali Tourani[1], Hriday Bavle[1], Jose Luis Sanchez-Lopez[1], Deniz Işınsu Avşar[2],
Rafael Muñoz Salinas[3], and Holger Voos[1]

***Abstract*— 3D scene graphs offer a more efficient representation of the environment by hierarchically organizing diverse semantic entities and the topological relationships among them. Fiducial markers, on the other hand, offer a valuable mechanism for encoding comprehensive information pertaining to environments and the objects within them. In the context of Visual SLAM (VSLAM), especially when the reconstructed maps are enriched with practical semantic information, these markers have the potential to enhance the map by augmenting valuable semantic information and fostering meaningful connections among the semantic objects. In this regard, this paper exploits the potential of fiducial markers to incorporate a VSLAM framework with hierarchical representations that generates optimizable multi-layered vision-based situational graphs. The framework comprises a conventional VSLAM system with low-level feature tracking and mapping capabilities bolstered by the incorporation of a fiducial marker map. The fiducial markers aid in identifying walls and doors in the environment, subsequently establishing meaningful associations with high-level entities, including corridors and rooms. Experimental results are conducted on a real-world dataset collected using various legged robots and benchmarked against a Light Detection And Ranging (LiDAR)-based framework (*S-Graphs*) as the ground truth. Consequently, our framework not only excels in crafting a richer, multi-layered hierarchical map of the environment but also shows enhancement in robot pose accuracy when contrasted with state-of-the-art methodologies.**

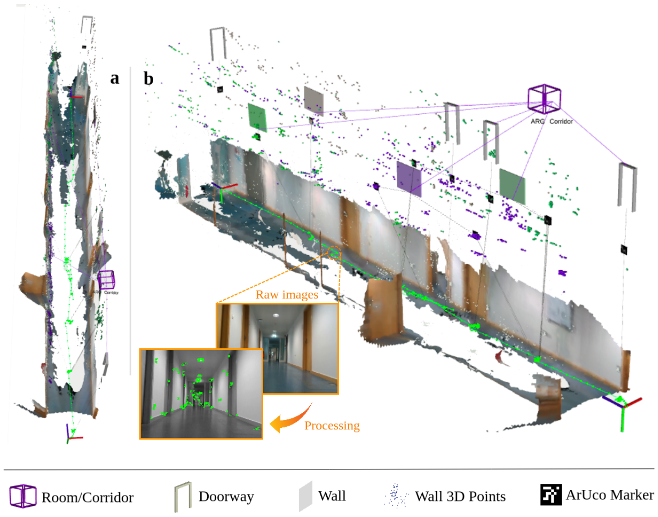

Fig. 1: A reconstructed map with its hierarchical representation generated by our framework, containing various detected semantic entities and the connections among them: a) the top view of the reconstructed map represented in 2D, b) the generated 3D view.

I. Introduction

Employing vision sensors for Simultaneous Localization and Mapping (SLAM) applications can bring about several merits, including the ability to achieve rich visual information using a low-cost hardware setup, making them attractive approaches compared to Light Detection And Ranging (LiDAR)-based tools [1]. This variant of SLAM systems is known as Visual SLAM (VSLAM), as they employ visual data for map reconstruction. When the goal is to incorporate semantic data, it becomes possible to enrich VSLAM with high-level information about the environment [2], [3] Nonetheless, many of these approaches do not integrate valuable relational information among pertinent semantic entities. 3D scene graph methodologies such as [4]–[6] exhibit promising results generating meaningful 3D scene graphs from underlying SLAM. In this regard, Situational Graphs (*S-Graphs*) [7], [8] tightly couples SLAM graphs with 3D scene graphs to provide meaningful maps with state-of-the-art accuracy, but only works with LiDAR sensors.

On the other hand, fiducial markers such as AprilTag [9] and ArUco [10], have the capability to easily extract semantic information from the environment for creating meaningful 3D scene graphs. However, current marker-based approaches such as UcoSLAM [11] and TagSLAM purely create geometric map representations. In this context, the previous work introduced by the authors of this paper [12] capitalized on fiducial markers to craft a visual situational graph of the environment. However, this prior work was constrained by its exclusive reliance on monocular cameras and exhibited limited scalability in large-scale environments.

This paper in hand presents an improved version of

[1]Authors are with the Automation and Robotics Research Group, Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg. Holger Voos is also associated with the Faculty of Science, Technology, and Medicine, University of Luxembourg, Luxembourg. {ali.tourani, hriday.bavle, joseluis.sanchezlopez, holger.voos}@uni.lu

[2]Author is with the Department of Physics & Materials Science, University of Luxembourg, Luxembourg. {deniz.avsar}@uni.lu

[3]Author is with the Department of Computer Science and Numerical Analysis, Rabanales Campus, University of Córdoba, Spain. rmsalinas@uco.es

*This work was partially funded by the Institute of Advanced Studies (IAS) of the University of Luxembourg (project TRANSCEND), and the Fonds National de la Recherche of Luxembourg (FNR) (project C22/IS/17387634/DEUS)

*For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

marker-based visual SLAM capable of generating a three-layered situational graph of the environment using RGB-D cameras. It tightly couples the robot poses with fiducial markers and semantic entities of walls, doorways, rooms, and corridors in a multi-layered hierarchical optimizable graph. Fig. 1 presents the generated three-layered graphs for an indoor environment, interconnecting the robot poses with diverse semantic entities. The principal contributions of the paper are summarized below:

- A novel marker-based three-layered optimizable framework supporting RGB-D visual sensor,
- A new map reconstruction and hierarchical representation process with the extraction of semantic entities, including markers, walls, doorways, corridors, and rooms,
- Addition of new semantic and geometric constraints for improving the quality of the reconstructed map and reducing localization errors,
- And revealing the concept and potential of "*imperceptible fiducial markers*" for situational awareness applications.

II. Related Works

A. SLAM and 3D Scene Graphs

Over the years, VSLAM approaches have reached a level of maturity, and various innovative solutions have pushed the boundaries of this domain. The authors of this paper have surveyed state-of-the-art approaches in VSLAM and discussed its current trends and possible directions in [12]. In accordance with the mentioned study, semantic VSLAM has evolved with methods such as [2], [13], [14] estimating a geometric map and adding semantic objects in the environment for jointly optimizing the robot pose and semantic object landmarks. Sun *et al.* [15] proposed a deep learning-based method that extracts semantic information from the scene and performs multi-object tracking based on them. DS-SLAM [16] utilizes semantic information achieved by SegNet [17] for semantic mapping. Yang *et al.* [18] proposed another approach with dynamic object removal for generating static semantic maps. Although all the above methods outperform their geometric VSLAM counterparts and are able to classify and map different semantic elements in the environment, they can still suffer errors due to misidentification and the semantic elements' pose estimation errors. Thus, adding structural/topological constraints among various semantic elements can further increase the robustness of the environmental understanding.

To explore the limitations of semantic SLAM techniques, 3D scene graph approaches such as [4], [5], [19] generate hierarchical representations of the environment interconnecting different semantic entities with suitable relations. While the above techniques consider SLAM and 3D scene graphs as two different optimization problems, methods like [6]–[8] tightly couple SLAM graphs and 3D scene graphs for improving accuracy while generating meaningful, multi-layered hierarchical environment maps.

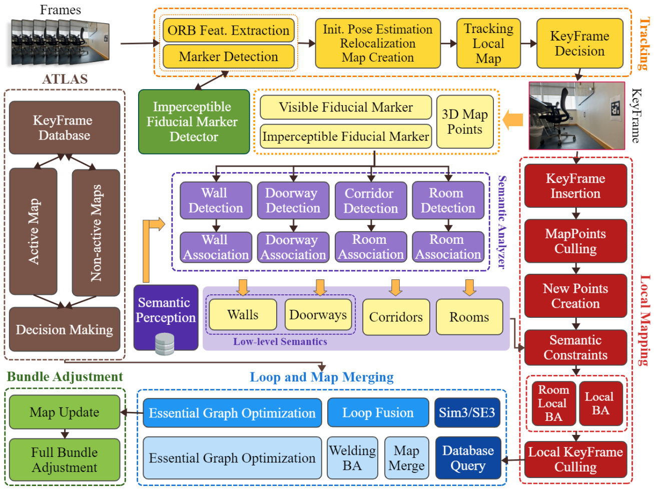

Fig. 2: The primary system components and pipeline of the proposed approach.

B. Marker-Based SLAM

While several existing works focus on semantic data extraction, it is worth noting that fiducial markers can also serve as valuable tools for achieving enhanced scene understanding and mapping. In this regard, UcoSLAM [11] is a marker-based VSLAM work that utilizes visual features obtained from natural landmarks and ArUco markers. It works in keypoint-only, marker-only, and mixed modes and is equipped with a marker-based loop closure detector, which requires placing distinct fiducial markers in the environment. TagSLAM [20] is another VSLAM approach that uses AprilTags to perform SLAM tasks. However, the system runs in marker-only mode and needs to see the markers constantly for localization and tracking stages. In another marker-based work, Romero-Ramirez *et al.* introduced a lightweight VSLAM work titled sSLAM [21], which utilizes the potential of various customized markers to facilitate tracking. However, all of the mentioned approaches are focused on creating geometric maps, and they do not employ the potential of fiducial markers for decoding semantic information.

Thus, and in order to exploit the potentials of fiducial markers in SLAM while simultaneously generating 3D scene graphs, the previous work of the authors of this paper [22] proposed a modified version of UcoSLAM detecting semantic entities from the surroundings with the aid of ArUco markers and interconnecting them with topological relations in a single optimizable graph. However, it was limited to monocular cameras and lacked scalability in larger environments. Accordingly, this work utilizes fiducial markers to generate a three-layered optimizable hierarchical graph incorporating semantic objects with appropriate relational constraints.

III. Proposed Method

The presented framework is built upon ORB-SLAM 3.0 [23] and aims to provide more comprehensive reconstructed maps and their multi-level topological graph representations with semantic information derived from fiducial markers. It also seeks to employ visual data to represent the robot's pose within an optimizable graph with an approach comparable to *S-Graphs* [7] and *S-Graphs+* [8]. In this regard, our framework computes semantic elements' spatial positioning, leveraging fiducial markers affixed to them instead of relying on LiDAR-based odometry readings and planar surface extractions. It reconstructs a semantic map with hierarchical representations in the presence of ArUco markers and a database of high-level semantic information about the affiliations of markers to objects. The current version of the framework supports the RGB-D sensor.

Fig. 2 depicts the pipeline of the proposed methodology and its constituent components. The framework benefits from a multi-thread architecture for processing data, including *tracking*, *local mapping*, *loop and map merging*, and *marker detector*. The operation commences by processing the frames captured by an RGB-D camera and conveying them to the *tracking* module where Oriented FAST and Rotated BRIEF (ORB) features and ArUco markers are extracted. The outcome of this module contains KeyFrame candidates with pose information, 3D map points, and possibly fiducial markers. If the *tracking* module decides whether the current frame should be a KeyFrame, the *local mapping* module triggers to add the KeyFrame and points to the map, alongside refining the map structure. Concurrently, the identification of semantic entities defined in the framework, including walls, doorways, corridors, and rooms, is being done within the framework. The mentioned process takes place by leveraging pose information obtained from fiducial markers and a pre-defined semantic perception database housing real-world data about the environment. Extracted semantic information serves as a resource for enhancing local bundle adjustment and KeyFrame culling the *local mapping*. The system constantly cooperates with an enhanced version of *Atlas* for establishing connections among disparate maps, representing the currently existing map (*i.e.,* active) and previously generated maps (*i.e.,* non-active). Loops and shared regions are detected within the active map and maps archived in *Atlas*, in the *loop and map merging* module. Finally, and in the case of loop detection and correction, the *global bundle adjustment* module is invoked to refine the constructed map further.

A. Fundamentals

The proposed method introduces four main coordinate systems at time t: the odometry frame of reference O, the camera coordinate system C_t, the marker coordinate system M_t, and the global coordinate system G_t. The vision sensor captures a set of frames $\mathbf{F} = \{\mathbf{f}\}$, with each frame $\mathbf{f} = \{t, \mathbf{T}, \delta\}$ containing the camera pose $\mathbf{T} \in SE(3)$ acquired through the transformation of C_t to G_t, as well as intrinsic camera parameters δ. Each frame f undergoes sub-sampling into an image pyramid and is processed by ORB feature extractor to obtain a set of key points for selecting keyframes, denoted as $\mathbf{K} = \{k\} \subset \mathbf{F}$. These keyframes contain feature points $\mathbf{P} = \{\mathbf{p}\}$ and fiducial markers $\mathbf{M} = \{\mathbf{m}\}$ used for detecting semantic entities. A feature point $\mathbf{p} = \{\mathbf{x}, \mathbf{v}, \hat{\mathbf{d}}\}$ is characterized by its corresponding 3D position $\mathbf{x} \in \mathbb{R}^3$, viewing direction $\mathbf{v} \in \mathbb{R}^3$, and descriptor $\hat{\mathbf{d}}$. On the other hand, each marker $\mathbf{m} = \{id, t, s, \mathbf{p}\}$ holds unique ArUco marker identifier $m_i \in \mathbb{N}$, length $s \in \mathbb{R}$, and pose $\mathbf{p} \in SE(3)$ derived from the transformation of M_t to G_t. Consequently, the final representation of the reconstructed map of the environment $\mathbf{E}$ is defined as follows:

$$\mathbf{E} = \{\mathbf{K}, \mathbf{P}, \mathbf{M}, \mathbf{W}, \mathbf{D}, \mathbf{R}\} \tag{1}$$

where $\mathbf{W} = \{\mathbf{w}\}$ encompasses the detected walls within the environment, in which each wall $\mathbf{w} = \{t, \mathbf{q}, \mathbf{m_w}\}$ holds the wall equation $\mathbf{q} \in \mathbb{R}^4$ and the list of attached markers $\mathbf{m_w} \subset \mathbf{M}$. $\mathbf{D} = \{\mathbf{d}\}$ comprises the doorways found in the environment, where each doorway $\mathbf{d} = \{t, m_d, \mathbf{p}\}$ contains attached marker $\mathbf{m_d} \in \mathbf{M}$ and pose $\mathbf{p} \in SE(3)$ computed from M_t to G_t transformation. Finally, $\mathbf{R} = \{\mathbf{r}\}$ represents the set of rooms/corridors present in the environment, each $\mathbf{r} = \{t, \mathbf{r_c}, \mathbf{r_w}\}$ with a center point $\mathbf{r_c} \in \mathbb{R}^3$ and a list of walls $\mathbf{r_w} \subset \mathbf{W} = (w_1...w_n)|w_i \in \mathbb{N}$ that constitute the boundaries of the room or corridor.

B. Semantic Entities

Reconstructing a rich semantic map of the environment incorporating the mentioned entities entails employing diverse methodologies, which will be discussed in this section. In the proposed approach, fiducial markers play a vital role, and their synergy with the system, in conjunction with the *semantic perception* module, aids in recognizing the semantic entities of interest. It is worth emphasizing that the dictionary encodes exclusively the *marker-ids* corresponding to rooms and doorways, obviating the need for any supplementary pose information to be incorporated.

Markers. Fiducial markers serve as crucial reference points in our framework, enabling the system to interpret and contextualize semantic data in the environment. Owing to their distinctive textures and the capacity to compute pose information ($\mathbf{p} \in SE(3)$), fiducial markers are primary sources of information in the proposed work for identifying and labeling targeted semantic entities, *i.e.,* walls and doorways. Each marker $\boldsymbol{m}_i$ in G_t is constrained by the keyframe K_i observing it, which can be formulated as:

$$c_{m_i}({}^{G}\boldsymbol{K}_i, {}^{G}\boldsymbol{m}_i) = \|{}^{L}\boldsymbol{m}_i \boxplus {}^{G}\boldsymbol{K}_i \boxminus {}^{G}\boldsymbol{m}_{i_1}\|^2_{\boldsymbol{\Lambda}_{\tilde{\boldsymbol{m}}_i}} \tag{2}$$

where ${}^{L}\boldsymbol{m}_i$ represents the locally observed fiducial marker's pose, $\boxplus$ and $\boxminus$ refer to the composition and inverse composition, $\|\ldots\|$ is the Mahalanobis distance, and $\boldsymbol{\Lambda}_{\tilde{\boldsymbol{m}}_i}$ is information matrix associated to $\tilde{m}_i$.

Walls. The procedure employed to identify wall surfaces, which serve as the planar substrates on which ArUco markers and feature points features are situated, relies on pose information derived from detected markers. Wall detection takes place whenever a fiducial marker is visited within the current keyframe. Hence, by accessing the real-world environment data obtained from the *semantic perception* module, in cases where the marker's identifier is not found within the list of markers associated with doorways, the planar equation of the wall is calculated using the poses of the attached marker and the surrounding map points. It should be noted that this work operates under the assumption that all fiducial markers are affixed directly onto the walls in an environment. As a consequence, the equations characterizing these walls are obtained based on the poses of the markers attached to them.

Each wall $\boldsymbol{w}_i$ in G_t is represented by ${}^G\boldsymbol{w}_i = \begin{bmatrix} {}^G\boldsymbol{n}_i & {}^G d \end{bmatrix}$, where ${}^G\boldsymbol{n}_i = \begin{bmatrix} n_x & n_y & n_z \end{bmatrix}^T$ denotes the normal vector of the wall. The vertex node of the wall within the graph is denoted as ${}^G\boldsymbol{w_i} = [{}^G\phi, {}^G\theta, {}^G d]$, where ${}^G\phi$ and ${}^G\theta$ represent the azimuth and elevation angles of the wall in the global coordinate G_t, respectively. Consequently, the cost function associated with each marker ${}^G\boldsymbol{m}_i$ affixed to the wall ${}^G\boldsymbol{w}_i$ can be calculated as follows:

$$c_{w_i}({}^G\boldsymbol{w}_i, {}^G\boldsymbol{m}_i) = \|[{}^M\delta\phi_{w_{i_{m_i}}}, {}^M\delta\theta_{w_{i_{m_i}}}, {}^M d_{w_i}]^T\|^2_{\boldsymbol{\Lambda}_{\tilde{\boldsymbol{w}}_i}} \tag{3}$$

where ${}^M\delta\phi_{w_{i_{m_i}}}$ represents the disparity between the azimuth angle of wall w_i and its attached marker m_i in M_t, ${}^M\delta\theta_{w_{i_{m_i}}}$ denotes the difference in elevation angles between the two, while ${}^M d_{w_i}$ signifies the perpendicular distance separating the wall from the marker. This distance should ideally be zero for given marker-wall pairings.

Corridors (Two-Wall Rooms). Our framework leverages an adapted version of the ”room segmentation” methodology originally presented in *S-Graphs+* [8], wherein markers are employed for detecting walls belonging to rooms. In this context, a corridor is defined as a room in the environment in which two parallel walls are labeled with ArUco markers. Due to the complexities associated with detecting rooms with diverse layouts, this work extends its definition to include rooms with inaccessible walls as corridors.

A corridor ${}^G\mathbf{r}_x = [{}^G\mathbf{w}_{x_{a_1}}, {}^G\mathbf{w}_{x_{b_1}}]$ encompasses wall planes aligned with the x-axis. To calculate the center point of a corridor ${}^G\mathbf{r}_{x_i}$, the two equations representing the x-wall planes are employed in conjunction with the center point ${}^G\mathbf{c}_i$ of the marker $\mathbf{m}_i$ in the following manner:

$$\begin{aligned} {}^G\mathbf{k}_{x_i} &= \tfrac{1}{2}\left[|{}^G d_{x_{a_1}}| \cdot {}^G\mathbf{n}_{x_{a_1}} - |{}^G d_{x_{b_1}}| \cdot {}^G\mathbf{n}_{x_{b_1}}\right] + |{}^G d_{x_{b_1}}| \cdot {}^G\mathbf{n}_{x_{b_1}} \\ {}^G\boldsymbol{\eta}_{x_i} &= {}^G\hat{\mathbf{k}}_{x_i} + \left[{}^G\mathbf{c}_i - [\,{}^G\mathbf{c}_i \cdot {}^G\hat{\mathbf{k}}_{x_i}]\ \cdot {}^G\hat{\mathbf{k}}_{x_i}\right] \end{aligned} \tag{4}$$

where ${}^G\boldsymbol{\eta}_{x_i}$ represents the center point of the corridor ${}^G\mathbf{r}_{x_i}$ and ${}^G\hat{\mathbf{k}}_{x_i}$ is derived from ${}^G\hat{\mathbf{k}}_{x_i} = {}^G\mathbf{k}_{x_i}/\|{}^G\mathbf{k}_{x_i}\|$. The center point ${}^G\mathbf{c}_i$ of the marker is determined based on the marker's pose within the G frame. It is noteworthy that the computation of the center for a two-wall room in the y direction follows a similar procedure. The cost function to minimize the corridor's vertex node and its corresponding wall planes is defined as follows:

$$\begin{aligned} &c_{\boldsymbol{r}_{x_i}}({}^G\boldsymbol{r}_{x_i}, [{}^G\mathbf{w}_{x_{a_1}}, {}^G\mathbf{w}_{x_{b_1}}, {}^G\mathbf{c}_i]) \\ &\quad = \sum_{t=1,i=1}^{T,K} \|{}^G\hat{\boldsymbol{\eta}}_{x_i} - f({}^G\tilde{\mathbf{w}}_{x_{a_1}}, {}^G\tilde{\mathbf{w}}_{x_{b_1}}, {}^G\mathbf{c}_i)\|^2_{\boldsymbol{\Lambda}_{\tilde{\boldsymbol{r}}_{i,t}}} \end{aligned} \tag{5}$$

where $f({}^G\tilde{\mathbf{w}}_{x_{a_1}}, {}^G\tilde{\mathbf{w}}_{x_{b_1}}, {}^G\mathbf{c}_i)$ is a mapping function that associates the wall planes with the corridor's center point.

Rooms (Four-Wall Rooms). In the scenario where a room in the environment consists of four walls, each labeled with ArUco markers (*i.e.,* two pairs of perpendicular labeled walls), the room is represented as ${}^G\mathbf{r}_i = [{}^G\mathbf{w}_{x_{a_1}} {}^G\mathbf{w}_{x_{b_1}} {}^G\mathbf{w}_{y_{a_1}} {}^G\mathbf{w}_{y_{b_1}}]$. The center point ${}^G\boldsymbol{\rho}_i$ of the room ${}^G\mathbf{r}_i$ is computed using the following equation:

$$\begin{aligned} {}^G\mathbf{q}_{x_i} &= \tfrac{1}{2}\left[|{}^G d_{x_{a_1}}| \cdot {}^G\mathbf{n}_{x_{a_1}} - |{}^G d_{x_{b_1}}| \cdot {}^G\mathbf{n}_{x_{b_1}}\right] + |{}^G d_{x_{b_1}}| \cdot {}^G\mathbf{n}_{x_{b_1}} \\ {}^G\mathbf{q}_{y_i} &= \tfrac{1}{2}\left[|{}^G d_{y_{a_1}}| \cdot {}^G\mathbf{n}_{y_{a_1}} - |{}^G d_{y_{b_1}}| \cdot {}^G\mathbf{n}_{y_{b_1}}\right] + |{}^G d_{y_{b_1}}| \cdot {}^G\mathbf{n}_{y_{b_1}} \\ {}^G\boldsymbol{\rho}_i &= {}^G\boldsymbol{q}_{x_i} + {}^G\boldsymbol{q}_{y_i} \end{aligned} \tag{6}$$

where the equation is positive if $|d_{x_1}| > |d_{x_2}|$. The cost function to minimize the room's vertex node and its corresponding wall planes is defined as follows:

$$\begin{aligned} &c_{\boldsymbol{\rho}}({}^G\boldsymbol{\rho}, [{}^G\mathbf{w}_{x_{a_i}}, {}^G\mathbf{w}_{x_{b_i}}, {}^G\mathbf{w}_{y_{a_i}}, {}^G\mathbf{w}_{y_{b_i}}]) \\ &= \sum_{t=1,i=1}^{T,S} \|{}^G\hat{\boldsymbol{\rho}}_i - f({}^G\tilde{\mathbf{w}}_{x_{a_i}}, {}^G\tilde{\mathbf{w}}_{x_{b_i}}, {}^G\tilde{\mathbf{w}}_{y_{a_i}}, {}^G\tilde{\mathbf{w}}_{y_{b_i}})\|^2_{\boldsymbol{\Lambda}_{\tilde{\boldsymbol{\rho}}_{i,t}}} \end{aligned} \tag{7}$$

where $f({}^G\tilde{\mathbf{w}}_{x_{a_i}}, {}^G\tilde{\mathbf{w}}_{x_{b_i}}, {}^G\tilde{\mathbf{w}}_{y_{a_i}}, {}^G\tilde{\mathbf{w}}_{y_{b_i}})$ is a mapping function that associates wall planes with the room's center point.

Doorways. Detecting doorways is comparatively less challenging than the wall detection process. In this case, the pose information of ArUco markers placed on a door frame is employed for doorway definition in the map. The procedure involves visiting fiducial markers present in the current keyframe and subsequently verifying their association with doorways through the utilization of the *semantic perception* module. Once confirmed, the pose of the visited marker is designated for the doorway.

Accordingly, the cost function for each doorway $\boldsymbol{d}_i$ in G_t and its corresponding room (or corridor) ${}^G\mathbf{r}_i$ is computed as follows [24]:

$$c_{d_i}({}^G\boldsymbol{d}_i, {}^G\boldsymbol{r}_i) = \|{}^G\hat{\boldsymbol{\delta}}_{d_i,r_j} - f({}^G\boldsymbol{d}_i, {}^G\boldsymbol{r}_i)\|^2_{\boldsymbol{\Lambda}_{\tilde{\boldsymbol{d}}_{i,t}}} \tag{8}$$

where ${}^G\hat{\boldsymbol{\delta}}_{d_i,r_j}$ represents the relative distance between the door and the room and $f({}^G\boldsymbol{d}_i, {}^G\boldsymbol{r}_j)$ maps the relative distance among their nodes.

C. Final Graph

The structure of the final semantic graph generated by our framework is depicted in Fig. 3. Accordingly, the primary sources of tracking information are located in the keyframes, which contain geometric data and pose information of objects, including 3D points and visited ArUco markers. The constraints among the mentioned objects guarantee proper computation of the odometry and loop closure detection. Lower-level semantic entities, including walls and doorways, are linked to constraints associated with the mapped fiducial markers and establish next-level constraints with higher-level entities, including rooms and corridors.

D. Imperceptible Fiducial Markers

Inconspicuous markers are a new generation of fiducial markers that can be imperceptible to humans but visible

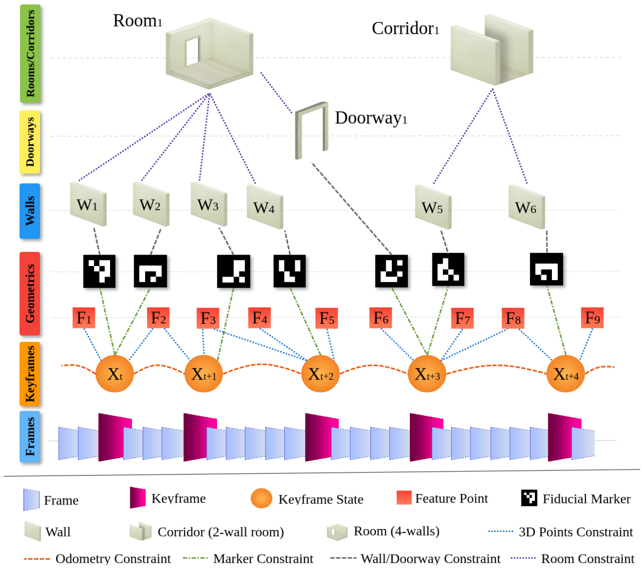

Fig. 3: The graph-based hierarchical representation of our proposed work incorporates semantic constraints, including walls, doorways, corridors, and rooms. The pre-existing geometric constraints have been complemented with semantic constraints acquired with the aid of fiducial markers.

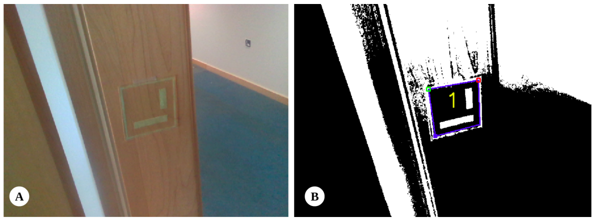

Fig. 4: An imperceptible marker introduced by the authors: a) a marker placed on a door frame, b) the recognized marker.

under certain lighting conditions with specific camera setup [25]. The mentioned markers can be fabricated in diverse wavelength ranges of light (*i.e.,* Infrared, visible, and Ultraviolet) to avoid visual clutter of the environment, rendering them entirely invisible to humans in the invisible ranges of the light spectrum. The detection of such markers involves leveraging particular optical components and computer vision methods, with detailed technical explanations found in [25]. As conventional fiducial markers may cause problems with visual clutter, they can be seamlessly replaced with invisible markers in our framework. Accordingly, an instance of such markers is shown in Fig. 4. The performance and potential of such markers will be evaluated in detail in Section IV.

IV. Evaluation

To evaluate the performance and robustness of the proposed method, labeled as *vS-Graphs*, compared to other existing frameworks, various experiments have been done using the proposed approach, UcoSLAM [11], Semantic UcoSLAM [22], and ORB-SLAM 3.0 [23] as the baseline. Due to less noisy outputs of LiDAR sensors compared to vision-based sensors, *S-Graphs+* [7], [8] as a LiDAR-based framework has been used to produce ground truth data. A computer equipped with an *11th Gen. Intel® Core™ i9 @2.60GHz* processor and 32 GigaBytes of memory was used for conducting the mentioned evaluation.

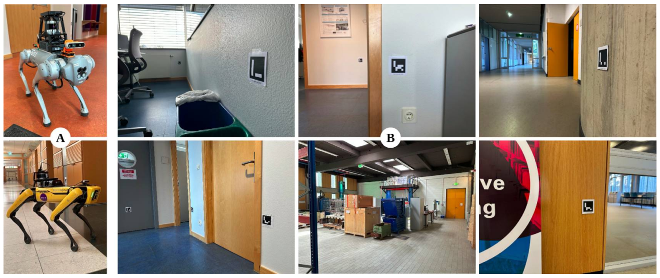

Fig. 5: Dataset collected to evaluate the proposed method: a) the legged robots used for data collection, b) some instances of the dataset.

TABLE I: The characteristics of the collected indoors dataset.

*Sequence**	*Duration*	*Description*
Seq-01	06m 27s	Two rooms connected via a door
Seq-02	07m 55s	A corridor connected to a room and another corridor
Seq-03	12m 32s	Five rooms connected to a corridor
Seq-04	07m 34s	Two corridors connected via a landing area
Seq-05	16m 42s	Four corridors connected to a room, forming a loop
Seq-06	01m 44s	A single room connected to a corridor

*data were stored as packages of *rosbag* files.

A. Evaluation Setup

For evaluating the performance of the proposed method in real-world conditions, a 3D LiDAR sensor and an *Intel® RealSense™ Depth Camera D435i* were mounted on legged robots, including *Boston Dynamics Spot®* and *Unitree Go1* (shown in Fig. 5). The robots collected data provided by sensors while traversing various indoor environments with different room and corridor configurations. Each wall and door of the rooms and corridors were labeled with printed $8cm \times 8cm$ ArUco markers and the unique identifiers of the markers were stored in a database to feed the proposed method (*i.e.,* the *semantic perception* module in Fig. 2). Accordingly, the characteristics of the collected dataset are presented in Table I.

B. Experimental Results

Accuracy. To validate the accuracy of the proposed method in comparison to its baseline and ground truth, ATE measurements have been employed in this paper. Regarding the evaluation results presented in Table II, it becomes evident that *vS-Graphs* outperforms its baseline counterpart and other frameworks across various scenarios. This enhancement can be attributed to the ability of *vS-Graphs* to introduce new constraints to the map through the association of semantic entities. It is noteworthy that the improvements are particularly pronounced in comparison to the two marker-based methodologies. While on *Seq-02* and *Seq-05* ORB-SLAM 3.0 exhibits slightly better performance than *vS-Graphs*, the difference is negligible (*i.e.,* $< 3cm$). This

TABLE II: Evaluation results on the collected dataset using Root Mean Square Deviation (RMSE) error in *meters* and Standard Deviation (STD). The best results are boldfaced and the second best are underlined. Our method outperforms the state-of-the-art in most of the sequences.

	RMSE						STD					
	Seq-01	*Seq-02*	*Seq-03*	*Seq-04*	*Seq-05*	*Seq-06*	*Seq-01*	*Seq-02*	*Seq-03*	*Seq-04*	*Seq-05*	*Seq-06*
vS-Graphs (ours)	**0.5127**	0.6662	**2.3555**	**0.4479**	2.1794	**0.2189**	**0.2454**	**0.3332**	**0.7441**	**0.2422**	**0.7107**	**0.0796**
UcoSLAM [11]	5.7996	3.0521	3.3034	2.1573	15.0184	1.5601	3.1814	1.3999	1.2332	1.2284	6.1595	0.8055
ORB-SLAM 3.0 [23]	0.5351	**0.6484**	2.5011	0.4895	**2.1404**	0.2479	0.2572	0.3334	0.8602	0.2653	0.7366	0.0815
Semantic UcoSLAM [22]	4.9437	2.8363	2.5154	1.9154	4.6672	1.5552	2.7065	1.3191	0.8582	1.1547	2.3891	0.8014

discrepancy could be due to the noisy detection of the fiducial markers as the primary source of semantic information in real-world scenarios with changing light conditions. However, it is important to emphasize that the proposed approach, in addition to improving ATE in most cases, also has the capacity to generate a three-layered situational graph of the environment. Accordingly, Fig. 6 depicts some qualitative results alongside the accuracy of the proposed approach against the LiDAR-based benchmark.

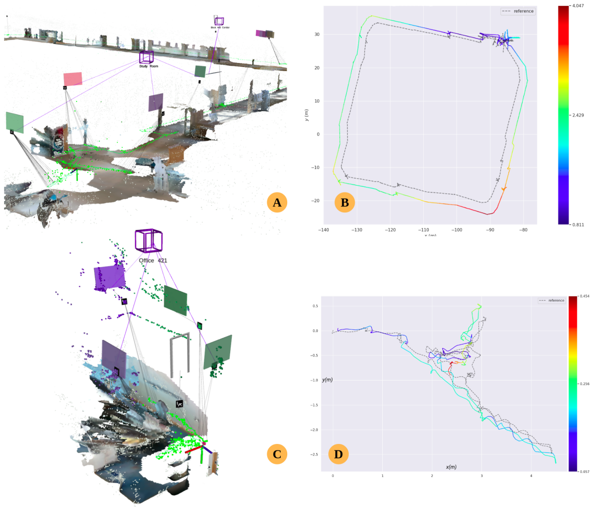

Fig. 6: The qualitative results and Absolute Trajectory Error (ATE) of the proposed approach w.r.t. translation in *meters* on *Seq-01* (A and B) and *Seq-06* (C and D). The dotted lines in the charts are LiDAR ground truth values.

Imperceptible Markers' Performance. To verify the applicability and potential of such markers for the first time, a similar experiment on *Seq-06* was done, where the door frame was labeled with an imperceptible marker. The reconstructed map using the mentioned markers is depicted in Fig. 6. Given that the sole distinction between imperceptible and ordinary fiducial markers lies in the detection step, the resulting semantic map remained unaltered. The authors believe that this will provide ample room for future investigation on the utilization of such markers in their future works.

V. Conclusions

This paper introduced a VSLAM framework that effectively harnesses the outputs provided by RGB-D cameras to achieve highly accurate map reconstruction. The proposed approach employs pose and topological information derived from strategically positioned ArUco markers within indoor environments for detecting semantic objects, including walls, doorways, corridors, and rooms, and utilizing the added semantic entities and the topological constraints among them for an elevated reconstructed map quality. Considering the evaluations performed on a real-world dataset collected by legged robots and benchmarked against a LiDAR-based framework as the ground truth, the proposed method showed an accuracy and performance improvement compared to state-of-the-art works.

As the proposed framework is part of a broader research project, in future works, the authors intend to determine transparent objects (*e.g.,* windows, mirrors, and glass doors) using the imperceptible markers due to their challenging recognition using computer vision algorithms and potential difficulties for robots performing SLAM tasks. Moreover, supporting more visual (*i.e.,* mono and stereo cameras) and inertial (*i.e.,* Inertial Measurement Unit (IMU)) sensors along with the efficient implementation of modules is another target of future works.

Acknowledgment

The authors would like to thank Jan P.F. Lagerwall from the Department of Physics & Materials Science (DPHYMS) of the University of Luxembourg for his efforts and valuable remarks on employing invisible fiducial markers as a futuristic scenario.

References

[1] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, "A comprehensive survey of visual slam algorithms," *Robotics*, vol. 11, no. 1, p. 24, 2022.

[2] H. Bavle, P. De La Puente, J. P. How, and P. Campoy, "Vps-slam: Visual planar semantic slam for aerial robotic systems," *IEEE Access*, vol. 8, pp. 60 704–60 718, 2020.

[3] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," 2020.

[4] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5664–5673.

[5] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans," 2020.

[6] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," 2022.

[7] H. Bavle, J. L. Sanchez-Lopez, M. Shaheer, J. Civera, and H. Voos, "Situational graphs for robot navigation in structured indoor environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9107–9114, 2022.

[8] ——, "S-graphs+: Real-time localization and mapping leveraging hierarchical representations," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4927–4934, 2023.

[9] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 3400–3407.

[10] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[11] R. Muñoz-Salinas and R. Medina-Carnicer, "Ucoslam: Simultaneous localization and mapping by fusion of keypoints and squared planar markers," *Pattern Recognition*, vol. 101, p. 107193, 2020.

[12] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, and H. Voos, "Visual slam: What are the current trends and what to expect?" *Sensors*, vol. 22, no. 23, p. 9297, 2022.

[13] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1722–1729.

[14] K. Doherty, D. Baxter, E. Schneeweiss, and J. Leonard, "Probabilistic data association via mixture models for robust semantic slam," 2019.

[15] Y. Sun, J. Hu, J. Yun, Y. Liu, D. Bai, X. Liu, G. Zhao, G. Jiang, J. Kong, and B. Chen, "Multi-objective location and mapping based on deep learning and visual slam," *Sensors*, vol. 22, no. 19, p. 7576, 2022.

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

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[18] S. Yang, C. Zhao, Z. Wu, Y. Wang, G. Wang, and D. Li, "Visual slam based on semantic segmentation and geometric constraints for dynamic indoor environments," *IEEE Access*, vol. 10, pp. 69 636–69 649, 2022.

[19] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," 2021.

[20] B. Pfrommer and K. Daniilidis, "Tagslam: Robust slam with fiducial markers," 2019. [Online]. Available: https://arxiv.org/abs/1910.00679

[21] F. J. Romero-Ramirez, R. Muñoz-Salinas, M. J. Marín-Jiménez, M. Cazorla, and R. Medina-Carnicer, "sslam: Speeded-up visual slam mixing artificial markers and temporary keypoints," *Sensors*, vol. 23, no. 4, p. 2210, 2023.

[22] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, R. M. Salinas, and H. Voos, "Marker-based visual slam leveraging hierarchical representations," *arXiv preprint arXiv:2303.01155*, 2023.

[23] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[24] M. Shaheer, H. Bavle, J. L. Sanchez-Lopez, and H. Voos, "Robot localization using situational graphs and building architectural plans," *arXiv preprint arXiv:2209.11575*, 2022.

[25] H. Agha, Y. Geng, X. Ma, D. I. Avşar, R. Kizhakidathazhath, Y.-S. Zhang, A. Tourani, H. Bavle, J.-L. Sanchez-Lopez, H. Voos *et al.*, "Unclonable human-invisible machine vision markers leveraging the omnidirectional chiral bragg diffraction of cholesteric spherical reflectors," *Light: Science & Applications*, vol. 11, no. 1, pp. 1–19, 2022.