

3D Active Metric-Semantic SLAM

Yuezhan Tao*, Xu Liu*, Igor Spasojevic, Saurav Agarwal and Vijay Kumar

Abstract—In this letter, we address the problem of exploration and metric-semantic mapping of multi-floor GPS-denied indoor environments using Size Weight and Power (SWaP) constrained aerial robots. Most previous work in exploration assumes that robot localization is solved. However, neglecting the state uncertainty of the agent can ultimately lead to cascading errors both in the resulting map and in the state of the agent itself. Furthermore, actions that reduce localization errors may be at direct odds with the exploration task. We develop a framework that balances the efficiency of exploration with actions that reduce the state uncertainty of the agent. In particular, our algorithmic approach for active metric-semantic SLAM is built upon sparse information abstracted from raw problem data, to make it suitable for SWaP-constrained robots. Furthermore, we integrate this framework within a fully autonomous aerial robotic system that achieves autonomous exploration in cluttered, 3D environments. From extensive real-world experiments, we showed that by including Semantic Loop Closure (SLC), we can reduce the robot pose estimation errors by over 90% in translation and approximately 75% in yaw, and the uncertainties in pose estimates and semantic maps by over 70% and 65%, respectively. Although discussed in the context of indoor multi-floor exploration, our system can be used for various other applications, such as infrastructure inspection and precision agriculture where reliable GPS data may not be available.

Index Terms—Aerial Systems; Perception and Autonomy; Mapping; Perception-Action Coupling

I. INTRODUCTION

MANY real-world applications require the construction of accurate metric-semantic maps of *a priori* unknown 3D environments. Unlike traditional maps that are concerned only with geometric information in the environment, metric-semantic maps encode both geometric and semantic information. Semantic objects provide a sparse but informative representation of the environment. In addition to benefiting robot navigation, they also provide actionable information for humans, e.g. they aid estimation of yield in agriculture or inventory in factories.

Due to the remarkable progress in deep learning during the past decade, extracting semantic information from the environment, such as object detection or scene classification, can be achieved with off-the-shelf pre-trained neural network models. As a result, we have seen many significant advances in metric-semantic SLAM [1]–[6].

Manuscript received: September 13, 2023; Revised December 8, 2023; Accepted January 24, 2024.

This paper was recommended for publication by Editor Tetsuya Ogata upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by TILOS under NSF Grant CCR-2112665, IoT4Ag ERC under NSF Grant EEC-1941529, the ARL DCIST CRA W911NF-17-2-0181, and ONR Grant N00014-20-1-2822.

*Equal Contribution. All authors are with GRASP Laboratory, University of Pennsylvania {yztao, liuxu, igorspas, sauravag, kumar}@seas.upenn.edu.

Copyright ©2024 IEEE

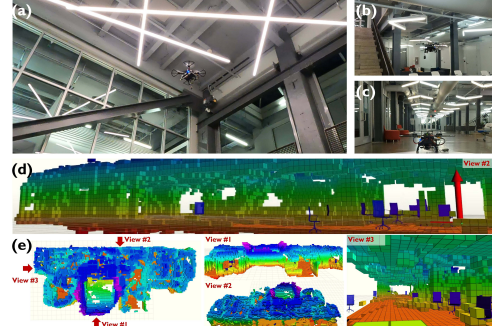


Figure 1: Falcon 250 UAV exploring a multi-floor environment. The robot explores the first (b-c) and second (a) floors, while constructing a metric-semantic map (d-e) in real time. Our framework enables efficient 3D exploration and accurate metric-semantic mapping.

Autonomous exploration has been widely studied and various approaches and systems have been proposed [7]–[10]. With the increase in computing power and the emergence of UAVs, recent work has been focused on expanding the planning space into 3D domains [11]–[14].

However, very few of prior works considered the problem of exploration in metric-semantic maps, or active metric-semantic mapping. Even those that do consider active metric-semantic mapping [15]–[17], they decouple the active mapping problem and the localization problem. This is suboptimal, especially when robots have noisy vision-based sensing. While the robot navigates in the environment, the Visual-Inertial Odometry (VIO) system inevitably accumulates drift. Such errors will eventually lead the robot to deviate from the desired path, resulting in erroneous mapping results and unsafe behaviors.

Motivated by this gap, in this paper, we present a unified framework that addresses the challenge of concurrent exploration, localization, and metric-semantic mapping. The contributions of the paper consists of:

- 1) An active Semantic Loop Closure (SLC) module and an SLC algorithm. The active SLC module generates and evaluates SLC candidates with a sparse but semantically meaningful representation of the environment. The SLC algorithm builds upon this representation to detect loop closures and estimate relative pose transformations.
- 2) A framework that trades off exploration and exploitation. The former is modeled as the Correlated Orienteering Problem (COP), and the latter is achieved using SLC-enabled active uncertainty reduction planning.
- 3) A 3D exploration and navigation stack for a fully autonomous UAV with real-time metric-semantic localization and mapping. Extensive real-world experiments in multi-floor indoor environments demonstrate the performance of the proposed system and its core modules.

To our knowledge, we are the first to develop and demonstrate a framework that enables SWaP-constrained UAVs to actively balance 3D exploration and uncertainty reduction using

metric-semantic maps, while operating in multi-floor environments without using any infrastructure. A demo video can be found at: <https://www.youtube.com/watch?v=Kb3s3IJ-wNg>.

II. RELATED WORK

A. 3D Autonomous Exploration

Several methods have been proposed for path planning for autonomous exploration in 2D domain [7]–[10]. With the emergence of UAVs, especially multirotor micro UAVs, recent works focus on developing algorithms and systems that could effectively plan and explore the full 3D space. In [11], 3D frontiers are detected through a stochastic equation-based method. In [12], Next Best View (NBV) is sampled in 3D space to maximize Information Gain (IG).

Most works select exploration waypoint greedily or within a finite horizon, while the Travelling Salesman Problem (TSP) has been applied to generate non-myopic plans. In [13, 14], 3D viewpoints around frontiers are sampled, and a global TSP tour is planned throughout the exploration process. Since TSP requires the robot to visit *all* viewpoints in the graph, it does not consider the information provided at each viewpoint. In this paper, we use the COP [18], which has three main attributes: (1) the vertices have rewards associated with them, (2) there is a correlation of rewards between vertices, and (3) a budget constraint limits the number of vertices that can be visited. The COP, which is a generalization of the Orienteering Problem (OP), maximizes the total reward while exploiting the correlation between vertices. Correlations capture the fact that visiting a vertex may provide information about other nearby vertices. The TSP, on the other hand, has no notion of rewards, correlations, or budget constraints. Thus, from a theoretical perspective, along with the existing qualitative and quantitative results [18, 19], the COP models the environment more accurately than the OP and TSP. Hence, the COP is our choice for path planning for 3D autonomous exploration.

B. Active Semantic SLAM

Prior work has investigated the problem of semantic SLAM or metric-semantic SLAM. Metric-semantic SLAM differs from traditional SLAM in that it not only utilizes traditional geometric features, such as points, lines or planes, but also leverages semantic features, such as object classes.

The benefits of utilizing semantic features in a SLAM framework are twofold: First, it helps robot localization because object-level features are more informative, storage efficient, and robust to viewpoint changes [2]–[4]. This is especially beneficial when integrated in real time with autonomous navigation in GPS-denied, unstructured environments [20]. Second, it offers robots a high-level understanding of the environment. Such advanced perception capabilities allow the robot to perform tasks with semantically meaningful mission specifications, such as actively gathering information on objects of interest [1, 16] or collaboratively surveying the environment to discover objects [21].

In light of these benefits, we utilize sparse semantic landmarks in the environment to reduce the uncertainties in robot state estimation during exploration. We achieve this by actively

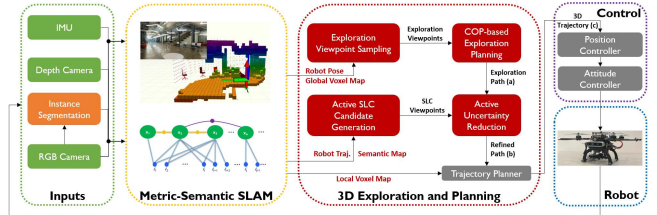


Figure 2: **System Diagram.** Our system takes in data from an RGB-D camera and the pose estimates from the VOXL VIO module. Instance segmentation is performed on RGB images with a pre-trained deep neural network (YOLO-V8 [25]) model. The *metric-semantic SLAM* module then takes in these inputs and estimates (1) a global voxel map for sampling exploration viewpoints, (2) a local voxel map for trajectory planning, (3) optimized robot pose estimates, and (4) a semantic map comprising object landmarks to generate SLC candidates. Next, a COP-based exploration planning algorithm takes in the exploration viewpoints and plans a long-horizon exploration path (a) consisting of a sequence of viewpoints, which seeks to maximize the Information Gain (IG) given the travel budget. This exploration path is then refined by inserting SLC viewpoints so that the robot can trade off exploration with uncertainty reduction. The refined path (b) is used to generate goals for the low-level trajectory planning algorithm, which constantly replans dynamically feasible 3D trajectories (c) in the local voxel map.

establishing Semantic Loop Closure (SLC). Specifically, the objective is to revisit a viewpoint for SLC, at which a cluster of semantic objects has been discovered to reduce the uncertainties in SLAM. Existing approaches can be found in utilizing semantic maps for passive loop closures [22, 23], or using geometric observations for opportunistic loop closure [24]. We propose to use the semantic maps for active SLC, which allows the robot to keep track of the environment at a much larger scale, efficiently detect and estimate relative transformations upon loop closures, and optimize pose estimation and semantic mapping accuracy simultaneously.

III. PROBLEM SPECIFICATION

Our goal is to maximize the accuracy of the metric-semantic map of a given region in 3D space within a given exploration budget. This requires the robot to (1) efficiently explore the environment, and (2) actively reduce uncertainties in its state estimates and the metric-semantic map.

In the following sections, we detail the proposed system that enables (1) and (2). From a high level, for (1), we model the problem as COP, the solution of which provides a long-horizon exploration path. For (2), we refine the path by actively establishing SLC, trading off exploration and uncertainty reduction in metric-semantic SLAM.

IV. SYSTEM OVERVIEW

We utilize the Falcon 250 platform in this work. This platform, as shown in Fig. 1, carries an Intel Realsense D435i camera, where the RGB images are used for instance segmentation, while the depth images are used for mapping and obstacle avoidance. A Pixhawk 4 Mini flight controller is responsible for low-level attitude control.

On top of the platform first introduced in [26], we added a VOXL VIO module [27], which outputs six degrees-of-freedom poses at 30 Hz. This, together with the IMU data, is fed into an Unscented Kalman Filter (UKF) to obtain 150 Hz pose estimates. The platform carries an Intel NUC onboard computer with an i7-10710U processor. The full software stack, including instance segmentation, metric-semantic

SLAM, exploration, planning, and control, runs in real time on board. A system diagram and its brief overview are provided and explained in Fig. 2. In the following sections, we provide detailed explanations of each module of the system.

V. METRIC-SEMANTIC SLAM

A. Hierarchical Volumetric Mapping

The hierarchical volumetric mapping module maintains two maps with different resolutions: a low-resolution (f_{gr}) global map and a high-resolution (f_{lr}) ego-centric local map. The former is used for frontier detection, viewpoint sampling, and COP-based exploration planning, with a size of no less than the experiment region. The latter is used to plan safe local trajectories and has a smaller size ($f_{lx} \times f_{ly} \times f_{lz}$).

When the robot receives pose information together with the depth images, ray-casting is conducted to project the readings from the depth images into 3D space, followed by a log-odds-based update on the probabilities of occupancy for all voxels traversed. Map updates are conducted asynchronously for the global map and the local map. The global map is updated with the optimized pose from the metric-semantic SLAM module, while the local map is updated with the estimated pose from the VIO algorithm. After each map update, a bounding box enclosing the updated region in the global map is recorded and used by the subsequent frontier detection (see Sec. VI-A).

B. Semantic SLAM

We use a factor graph-based semantic SLAM algorithm. Fig. 3 shows a close-up view of our factor graph diagram. Our semantic SLAM algorithm supports different types of objects. It encodes robot pose to object model constraints via customized factors in the GTSAM backend [28, 29]. We refer the readers to our previous work [15] for details. In this work, we utilize a centroid-based model for the semantic objects in our environment.

The semantic SLAM module takes into account the estimated relative transformation of VIO between two consecutive key poses, i.e., $\mathbf{x}_t^{vio} \ominus \mathbf{x}_{t-1}^{vio}$ as the odometry factor, and the estimated centroid locations of the detected objects as the range and bearing factor. Due to the sparsity of the semantic map, our factor graph keeps track of historical measurements over the entire mission of the robot, and optimizes the robot poses and object landmarks in a globally consistent manner.

C. Semantic Loop Closure

The input of the SLC module consists of the map of the portion of the environment the robot has explored thus far, and the “local map” within its current sensing range. The module determines whether the robot is currently located at a position it has been in the past, and if so, where. It is critical to minimize the drift in odometry and errors in map estimates. The main challenge of detecting loop closures involves aligning two maps with only partial overlap. If the two maps are encoded in the form of dense point clouds, the problem can be computationally challenging. Popular existing approaches

either iterate between estimating data association and performing point cloud alignment, or solving convex relaxations of the problem. Nevertheless, they are either susceptible to being stuck in suboptimal local minima, or involve substantial computational resources. In addition, they may fail to find the match due to viewpoint changes. We overcome these difficulties using semantic information to reduce the size of aligned maps, only focusing on aligning *objects* detected in the two maps. This approach (1) significantly reduces the size of the alignment problem and (2) is robust to viewpoint changes, allowing us to use an exact exhaustive search algorithm.

To illustrate SLC in more detail, we are given two sets of points, $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^3$ of cardinalities $|\mathcal{A}| = n$ and $|\mathcal{B}| = m$, respectively. Points in \mathcal{A} represent centroids of semantic objects in our global map, whereas those in \mathcal{B} represent centroids of semantic objects currently within field of view. Furthermore, there are subsets $\mathcal{A}_1 \subseteq \mathcal{A}$ and $\mathcal{B}_1 \subseteq \mathcal{B}$ with $|\mathcal{A}_1| = |\mathcal{B}_1| = k$, together with a translation parameter $\mathbf{t} \in \mathbb{R}^3$ and a yaw angle $\psi \in (-\pi, \pi]$, such that

$$\mathbf{p}_{\sigma_B(i)} = \mathbf{R}(\psi)^T (\mathbf{p}_{\sigma_A(i)} - \mathbf{t}), \quad 1 \leq i \leq k \quad (1)$$

for some permutations $\sigma_A, \sigma_B \in Sym(k)$. Vector \mathbf{t} encodes the position of the UAV w.r.t. the world frame. We assume the roll and pitch angles of the robot can be accurately obtained from its IMU. Therefore, matrix $\mathbf{R}(\psi) = \exp([\psi \mathbf{e}_3]_{\times})$, with ψ denoting the yaw angle of the UAV, represents its roll-and-pitch-adjusted orientation w.r.t. the world frame. Subsets \mathcal{A}_1 and \mathcal{B}_1 encode the intersection $\mathcal{A} \cap \mathcal{B}$, while σ_A, σ_B encode data association. Ultimately, $\mathcal{A}_1, \mathcal{B}_1, k, \mathbf{t}, \psi, \sigma_A, \sigma_B$ are all *unknown*, and the task of our SLC module is to compute them. Loosely speaking, it is a search-based procedure that works by iterating through the Cartesian product of variations of \mathcal{A} and \mathcal{B} in decreasing order of cardinality, and stores the pair for which the quality of the match is as high as possible. We measure the quality of the match between sequences of points $(\mathbf{p}_{\sigma_A(i)})_{i=1}^k$ and $(\mathbf{p}_{\sigma_B(i)})_{i=1}^k$ via the residual function \mathcal{R} defined as

$$\begin{aligned} \mathcal{R}((\mathbf{p}_{\sigma_A(i)})_{i=1}^k, (\mathbf{p}_{\sigma_B(i)})_{i=1}^k) = \\ \min_{\mathbf{t} \in \mathbb{R}^3, \psi \in (-\pi, \pi]} \frac{1}{k} \sum_{i=1}^k \|\mathbf{R}(\psi)^T (\mathbf{p}_{\sigma_A(i)} - \mathbf{t}) - \mathbf{p}_{\sigma_B(i)}\|_2^2. \end{aligned} \quad (2)$$

Smaller residuals correspond to “better” matches. The residual can be computed analytically by noting that for a fixed yaw angle ψ , the optimal translation is given by

$$\mathbf{t}^*(\psi) = \frac{1}{k} \sum_{i=1}^k \underbrace{\mathbf{p}_{\sigma_A(i)}}_{=: \bar{\mathbf{p}}_{\sigma_A}} - \mathbf{R}(\psi) \frac{1}{k} \sum_{i=1}^k \underbrace{\mathbf{p}_{\sigma_B(i)}}_{=: \bar{\mathbf{p}}_{\sigma_B}}. \quad (3)$$

Defining $\Delta \mathbf{p}_{\sigma_A(i)} = \mathbf{p}_{\sigma_A(i)} - \bar{\mathbf{p}}_{\sigma_A}$, and $\Delta \mathbf{p}_{\sigma_B(i)} = \mathbf{p}_{\sigma_B(i)} - \bar{\mathbf{p}}_{\sigma_B}$, the optimal ψ can be recovered by minimizing the expression $\sum_{i=1}^k \|\Delta \mathbf{p}_{\sigma_A(i)} - \mathbf{R}(\psi) \Delta \mathbf{p}_{\sigma_B(i)}\|_2^2$ over ψ , which in turn, is equivalent to maximizing $\text{tr}(\mathbf{R}(\psi) \sum_{i=1}^k \Delta \mathbf{p}_{\sigma_B(i)} (\Delta \mathbf{p}_{\sigma_A(i)})^T)$. Defining the matrix $M := \sum_{i=1}^k \Delta \mathbf{p}_{\sigma_B(i)} (\Delta \mathbf{p}_{\sigma_A(i)})^T \in \mathbb{R}^{3 \times 3}$, we have that the trace under question equals

$$\cos(\psi)(M_{11} + M_{22}) + \sin(\psi)(M_{12} - M_{21}) + M_{33}, \quad (4)$$

and its maximum value

$$\sqrt{(M_{11} + M_{22})^2 + (M_{12} - M_{21})^2} + M_{33} \quad (5)$$

is attained for

$$\psi^* = \arctan 2(M_{12} - M_{21}, M_{11} + M_{22}). \quad (6)$$

The running time of this module is $\mathcal{O}((\min(m, n) + 1)! 2^{\max(m, n)})$. Even though the the algorithm is exponential, given that sets \mathcal{A} and \mathcal{B} comprise of a small number of semantic objects instead of dense point clouds, we consider the worst case computational burden to be acceptable in practice. Finally, it is worth noting that by disregarding pairs of variations and combinations of \mathcal{A} and \mathcal{B} that match objects of different classes, the complexity bound above can be reduced further. Nevertheless, the speed-up involved depends on the distribution of objects across the different classes, and this is something which is unknown a priori.

VI. EXPLORATION WITH ACTIVE SLC

In this section, we introduce the exploration with active SLC module, which utilizes the metric-semantic maps to generate paths that balance exploration and uncertainty reduction.

A. Frontier Detection and Exploration Viewpoint Sampling

We employ the incremental frontier detection and viewpoint sampling module presented in our previous work [26]. All existing frontiers within the bounding box from the map update will be re-evaluated and removed if observed. New frontiers are detected and clustered. Large clusters are broken down into small ones recursively if they are greater than the desired size, so that the robot can cover the cluster with the limited sensing range and field of view. 3D viewpoints are sampled for each frontier cluster following a two-step process. In the first step, candidate positions are uniformly sampled around the cluster centroid. For the second step, multiple yaw angles are uniformly sampled at each candidate position. Different from [26], cell-counting-based IG is estimated for all sampled yaw angles without information prediction. The candidate yaw angle with maximum estimated IG is selected as the sampled yaw angle and associated with the candidate position. We take the sampled pose with the highest estimated IG as the viewpoint for the frontier cluster.

B. COP-based Exploration Planning

The COP operates on a given complete graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. A vertex $v \in V$ has a reward $r_v \geq 0$ associated with it, and an edge $(i, j) \in E$ has a travel cost $c_{ij} \geq 0$. The edge costs are symmetric, i.e., $c_{ij} = c_{ji}$. For exploration planning, the vertices represent the sampled viewpoints, and the edges represent optimal paths between viewpoints. The reward r_v of a vertex v is the estimated IG at the viewpoint v , while the edge costs are computed using the A^* path cost between the viewpoints. Additionally, a correlation function $w(u, v) \in [0, 1]$ is defined for each pair of vertices $u, v \in V$, which measures the correlation between the rewards of the two

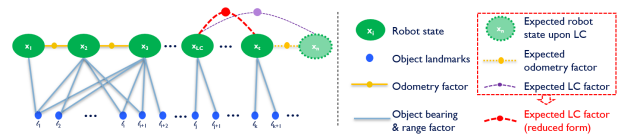


Figure 3: **Active Metric-Semantic SLAM.** The proposed semantic factor graph consists of nodes for both robot poses and object landmarks, and edges that represent odometry constraints, robot-to-object constraints, and semantic loop closure constraints. This graph also illustrates how the virtual factors and nodes are added in the active semantic loop closure step (see Sec. VI-C).

vertices. We compute the correlation between two vertices as the percentage of the overlap of the two viewpoints, assuming they are occlusion-free. The correlation function is symmetric in this case, i.e., $w(u, v) = w(v, u)$.

The goal of COP is to find a tour (or path) π that visits a subset of the vertices to maximize the total reward collected while respecting a given budget B on the total travel cost. Let $x_v \in \{0, 1\}$ denote whether a vertex v has been visited and let $y_{ij} \in \{0, 1\}$ denote whether an edge (i, j) has been traversed from vertex i to vertex j by a tour π . The COP maximizes the total reward:

$$R(\pi) = \sum_{v \in V} r_v (x_v + \omega_v (1 - x_v)), \quad (7)$$

subject to the following constraints:

$$\omega_v - \sum_{u \in V \setminus \{v\}} w(u, v) x_u \leq 0 \quad (8)$$

$$\sum_{(i, j) \in E} (c_{ij} y_{ij} + c_{ji} y_{ji}) \leq B \quad (9)$$

Tour constraints for π [18, 19]

$$y_{ij}, y_{ji} \in \{0, 1\}, \quad \forall (i, j) \in E$$

$$\omega_v \in [0, 1], x_v \in \{0, 1\} \quad \forall v \in V,$$

The variable ω_v models the portion of the reward r_v that is collected by vertices other than the vertex v (8). Similar to [19], we permit the sum of correlations, the second term in (8), to be greater than one, unlike the original, more restrictive COP formulation [18]. Note that the variable ω_v will always be either one or the sum of correlations, as it is in the objective function of a maximization problem. Although our edge costs and correlation functions are symmetric, the COP formulation allows them to be asymmetric. Constraint (9) is the budget constraint, which limits the total travel cost of the tour. We heuristically set the budget by estimating the cost from the robot's position to several nearby frontiers and scale it with a constant factor to limit the resulting tour length and reduce computational complexities. Tour constraints ensure that the tour has no disconnected subtours and at least one edge connected to a visited vertex is traversed. We refer readers to [18, 19] for details on the tour constraints.

The COP is NP-hard, and the MIQP formulation [18] is not suitable for online computation in the exploration problem. Hence, we use a simplified version of the greedy constructive heuristic algorithm from [19]. The algorithm starts with an empty tour, greedily selects a vertex to be added, and computes an efficient tour with the selected vertices. These steps are iteratively executed until the budget constraint is violated. The greedy criterion is based on the value of a vertex computed as:

value(v) = $r_v + \sum_{u \in V \setminus S} r_u w(v, u) - \sum_{u \in S} r_u w(u, v)$, where S is the subset of vertices already selected in the previous iterations. The complexity of the algorithm is $\mathcal{O}(|V|^3)$ [19]. In practice, by selecting a proper frontier cluster size (f_{sz}), we can bound the number of viewpoints in the environment to be less than 10, making it possible to compute an optimal tour for the selected vertices with the Bellman-Held-Karp Algorithm in real time on board. The exploration tour is re-planned if either of the following conditions is met: (1) a fraction (f_{r1}) of frontier changes in the current environment, (2) the percentage of the refined exploration tour (detailed in Sec. VI-D) that has been executed exceeds a given threshold (f_{r2}). The second condition is also known as receding-horizon planning. The COP-based exploration planning module is asynchronous with the rest of the software stack; the robot continues executing the refined tour until a new refined tour is received.

C. Active Semantic Loop Closure

While the robot is constructing the metric-semantic map, it needs to generate candidate SLC submap and viewpoint pairs. This is done by first finding the submaps by clustering the semantic landmarks, and then selecting the corresponding viewpoints. In the first step, we use the DBSCAN algorithm [30] to cluster the centroids of the semantic landmarks in the Euclidean space, as illustrated in the purple box of Fig. 4. We obtain valid submaps by choosing clusters with no less than a specific number (f_{cs}) of landmarks. Note that since we have range and bearing measurements from each landmark, with *a priori* unknown data association, we need at least three landmarks to uniquely determine the position and yaw of the robot upon loop closure, as explained in detail in our previous work [31]. In the second step, for each submap, we need to generate a viewpoint that is reachable, detectable, and informative. To make it reachable, we choose the viewpoint from the set of key poses (which the robot has reached before) in the factor graph. Next, we limit the choice of the key pose so that any of the landmarks in the submap is within the sensing range (f_{sr}). Third, to maximize the possible information gain brought about by the SLC, the oldest key pose (the pose first added to the factor graph) among all key poses that satisfy the previously mentioned conditions is selected as the viewpoint. The robot establishes a loop closure by taking a panorama (by yawing in place) at such an SLC viewpoint. One example SLC viewpoint is shown by the red arrow in Fig. 4.

Once the loop closure viewpoint-submap pairs are sampled, they will be used in the active uncertainty reduction planning module of our system, as illustrated in Fig. 2. This module seeks to insert loop closure viewpoints along the COP exploration path, such that the combined IG is maximized while respecting the travel budget constraint. Our pseudo-code in Algorithm 1 further explains this procedure.

An important step is to predict the IG brought about by each of the candidate SLC viewpoint-submap pairs. We achieve this by adding a virtual factor to the semantic factor graph as illustrated in Fig. 3. Conceptually, in this step, we added two factors, an expected odometry factor and an expected loop closure factor. The former brings the robot to the loop

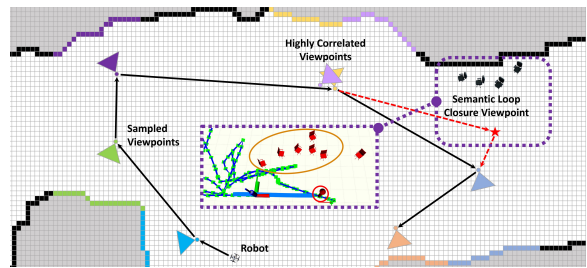


Figure 4: **An illustration of exploration with active SLC.** The solid black arrows show the nominal COP-based exploration path. The dashed red arrows highlight the difference between the refined (red) and nominal (black) paths. The active uncertainty reduction planning balances exploration and uncertainty reduction. A pair of active SLC landmark cluster and viewpoint is highlighted in the purple box, in which the orange-circled chairs belong to the cluster, and the red-colored arrow is the SLC viewpoint (i.e., \mathbf{x}_{lc} in Fig. 3).

closure viewpoint (\mathbf{x}_n) to establish an SLC with an existing key pose (\mathbf{x}_{lc}) in the graph, and the latter connects \mathbf{x}_n and \mathbf{x}_{lc} . Since our loop closure viewpoint is sampled from one of the existing key poses in the graph, \mathbf{x}_n and \mathbf{x}_{lc} are the same nodes. Therefore, the procedure reduces to adding the expected odometry factor between \mathbf{x}_t and \mathbf{x}_{lc} , with a motion noise scaled by the expected travel distance. For a long-horizon path, we can sequentially perform such operations to evaluate the IG for a sequence of actions with multiple SLCs.

This simulates the effect of the robot directly navigating to establish SLC with its noisy odometry measurements. This virtual factor leaves the estimates intact, but alters the covariance matrix of the factor graph. We calculate the reduction in the trace of the covariance matrix before and after adding this virtual factor as the IG measure: $IG = tr(\Sigma_t) - tr(\Sigma_{t+1})$, where tr denotes the trace of the covariance matrix. Once we evaluate the IG along a given path, we remove such virtual factors from the factor graph, so that these virtual factors do not alter the factor graph's estimates.

D. Active Uncertainty Reduction Planning

1) *Algorithm:* This algorithm converts the planned exploration path from the COP module into a refined path by balancing the IG from exploration and uncertainty reduction from SLC. Pseudo-code for this module is provided in Algorithm 1. It iterates through the sequence of viewpoints comprising the COP-based path, at each point evaluating if the robot should actively seek an SLC before resuming exploration. At every iteration, the non-negative remaining IG of the subsequent exploration viewpoint is evaluated by subtracting the correlated information gathered by the viewpoints already present in the refined path from its IG. Then, the cost of every SLC candidate is evaluated via A^* and its IG is calculated using the method detailed in Sec. VI-C. If the budget permits addition of the best SLC candidate, we further compare the scaled (f_{sc}) IG of the latter with the utility (i.e. cost-benefit index) of the upcoming viewpoint. The comparison result determines the candidate SLC candidate should be inserted into the refined exploration path. We only allow one SLC candidate to be inserted between consecutive exploration viewpoints, which also bounds the total running time complexity.

2) *Complexity Analysis:* Suppose that we have E exploration nodes, C loop closure candidates, S semantic landmarks, and P robot poses. The outer loop in line 2 requires

Algorithm 1 Active Uncertainty Reduction Planning

LC: loop closure candidate; BLC: best loop closure candidate; VP: frontier viewpoint;
Input: ExpPath ▷ exploration path from COP
Output: ReExpPath ▷ refined exploration path

```

1: ReExpPath ← [];
2: for i ← 1 to ExpPath.size() do
3:   VP ← ExpPath(i);
4:   VP.IG ← ComputeRemainingIG(ReExpPath, VP);
5:   for j ← 1 to LC.size() do
6:     Estimate Cost & IG for LC(j);
7:     UpdateCurrentBLC;
8:   if Scale(BLC.IG) > (VP.IG/VP.Cost) and BudgetEnough(BLC) then
9:     Insert BLC to ReExpPath;
10:    Insert VP to ReExpPath if BudgetEnough(VP);
11:   else if BudgetEnough(VP) then
12:     Insert VP to ReExpPath;
13:   else
14:     return;

```

$\mathcal{O}(E)$ iterations. In each iteration, the dominant cost comes from lines 4 to 7. The computational complexity of line 4 is $\mathcal{O}(E)$. The inner loop in line 5 is executed $\mathcal{O}(C)$ times. In every iteration of the inner loop, we run two sub-procedures. The first procedure, which estimates the cost of the SLC candidate, runs in time required to complete an A^* search - say T_{A^*} . The second procedure, which computes the IG, runs in time $\mathcal{O}(\max(P, S)^{1.5})$ [29]. Collecting the latter, the running time of our algorithm is $\mathcal{O}(E^2 + EC(T_{A^*} + \max(P, S)^{1.5}))$.

In the case of traditional SLAM, where dense geometric features are used, hundreds of features are tracked for each key pose. In this case, $\max(P, S) = S$, which is at the order of $100P$ or even larger. However, in our case, S will usually be no larger than P since semantic landmarks are sparse. Thus, $\max(P, S) = P$. We can further reduce it by, for example, only adding pose nodes whenever we observe a semantic landmark. Practically, given $E \ll P$ and $C \ll P$, the complexity of the algorithm is reduced to $\mathcal{O}(T_{A^*} + P^{1.5})$. This is essentially doing A^* searches and solving semantic SLAM problem with loop closures multiple times. By searching over a low-resolution map, the A^* search is manageable. Again, since the number of landmarks is much smaller in the semantic SLAM problem than in a traditional SLAM problem, this can be done efficiently online onboard the robot.

E. Drift Compensation and Trajectory Planning

Compared to VIO, our semantic SLAM algorithm outputs pose estimates with higher accuracy. However, the smoothness may be sacrificed due to intermittent drift correction induced by SLC. To solve this, we employ a drift compensation module similar to our previous work [20]. It transforms the next local planning goal from the SLAM reference frame to the odometry reference frame, using the difference between the VIO and semantic SLAM pose estimates.

By this design, the robot's controller, local mapper, VIO, and trajectory planner operate in the odometry reference frame. The exploration planner, semantic SLAM, and global mapper operate in the SLAM reference frame. We use [32] for local trajectory planning and the yaw optimization approach introduced in our previous work [26].

VII. RESULTS AND ANALYSIS

To evaluate the efficiency and performance of the entire system and critical modules, we conducted four sets of real-

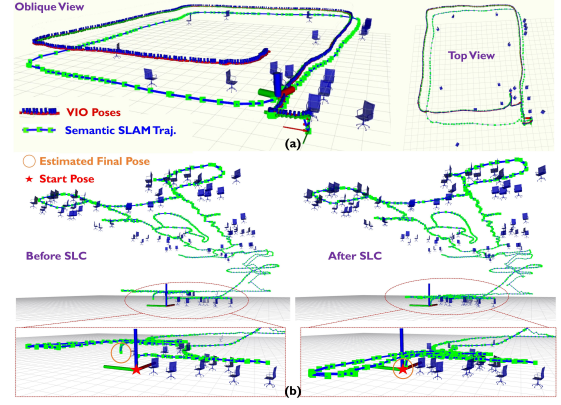


Figure 5: **Robot trajectories and semantic maps with and without SLC for one-floor (a) and three-floor (b) experiments.** The robot starts and ends at the exact same location. SLC significantly improves both the semantic map and the robot trajectory. A detailed analysis is provided in Sec. VII-B.

world experiments: (1) we evaluated the CPU utilization to empirically estimate the computational requirements of the software stack; (2) we studied the effects of the semantic loop closure module on datasets collected by surveying the building and establishing SLCs by revisiting the places with clusters of semantic landmarks at the end; (3) we carried out autonomous exploration and metric-semantic mapping experiments where the robot autonomously explored a multi-floor environment; (4) we benchmarked our system with state-of-the-art SLAM methods. In all experiments, we set $f_{r1} = 15\%$, $f_{r2} = 10\%$, $f_{sz} = 1.2m$, $f_{sr} = 5m$, $f_{sc} = 6$, $f_{cs} = 4$, $f_{gr} = 0.25m$, $f_{lr} = 0.1m$, $f_{lx} = 15m$, $f_{ly} = 15m$, $f_{lz} = 4m$.

A. Computational Requirements

We empirically evaluated the CPU utilization of our system using our UAV's onboard computer as mentioned in Sec. IV. The total CPU utilization is 42.2-53.3% for the full stack. The majority of the computation is taken by the semantic SLAM front end, which includes an instance segmentation neural network and a point cloud processing module, taking in total $\sim 34\%$ of the CPU. Note that we used the medium version of the YOLO V8, i.e. yolov8m, and we limited the inference rate to 2 Hz. The backend of the semantic SLAM, i.e. the optimization of the factor graph, took 0.88%. This is an average load, which may include surges when loop closures are triggered. The COP-based exploration module took 0.58%. The SLC module utilized 0.23%. The rest of the CPU utilization was taken by the remaining modules in our navigation stack, including the voxel mapper, viewpoint sampler, trajectory planner and tracker, state machine, controller, etc. VIO was done on the VOXL board. An important aspect to note is that different modules of our stack execute asynchronously. The delay in one module does not propagate to the other modules.

B. Semantic Loop Closure

We carried out multiple loop closure experiments inside a cluttered three-story building. Two examples are shown in Fig. 5. In the one-floor experiment (a), the robot traveled a squared loop on a single floor. On the left panel of (a), it is clear that the VIO drifted along the Z direction, which

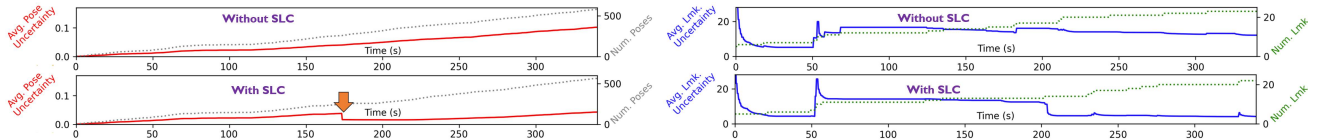


Figure 6: **Uncertainty without (top) and with (bottom) active SLC.** The red and blue lines represent average uncertainties in robot pose and semantic landmarks, respectively. The uncertainty is defined as the trace of covariance matrix in semantic factor graph. The orange arrow shows when SLC takes place.

Table I: **Quantitative results on error reduction in position and yaw estimation.**

Mission	Position Error (m)			Yaw Error (deg)			Traj. Len. (m)
	VIO(X/Y/Z)	Ours(X/Y/Z)	Reduction	VIO	Ours	Reduction	
Loop 1	2.50 (-1.95, -1.52, 0.34)	0.18 (-0.12, -0.11, 0.08)	92.68%	2.96°	-1.77°	39.99%	179.33
Loop 2	2.63 (2.36, 0.39, 0.80)	0.41 (0.02, -0.18, 0.37)	83.84%	7.05°	-2.69°	61.82%	454.85
Loop 3	4.15 (2.56, 3.16, 0.82)	0.67 (0.52, 0.42, -0.03)	83.78%	-12.98°	-7.81°	39.79%	497.83

Table II: **Quantitative results on uncertainty reduction (U. Red.) in robot poses and semantic landmarks.**

Mission	U. Red. of Avg. Pose upon SLC	U. Red. (w/ v.s. w/o SLC)		Traj. Len. (m)
		Avg. Pose	Avg. Lmk.	
Auto 1	56.67%	52.06%	68.53%	227.47
Auto 2	45.72%	54.87%	26.37%	72.62
Auto 3	52.98%; 14.62%	70.99%	23.93%	185.17

was significantly corrected by the SLC. The right panel of (a) indicates the drift of VIO along X-Y axes, while the semantic SLAM was able to close the loop with SLC. In the three-floor experiment (b), the robot took off and landed at the same position, and traveled across the entire three-story building. Before SLC, the final pose estimates were far away from the start pose, and the chairs were reconstructed at different altitudes. The SLC was able to correct the pose estimation drift and close the loop. Such drastic drift correction was backward propagated in the semantic factor graph to correct robot poses and the semantic map, which is illustrated in the zoomed-in views (red boxes). After SLC, the robot poses and chairs were at the same altitude with the ground plane as expected.

Next, we quantitatively compared our system against the commercial VIO solution on position and yaw estimation errors. As shown in Table. I, in the three measurements from our experiments, the position estimates from the VIO system produced errors up to 1.4% of total trajectory length. With the SLC happening at the end of each experiment, the position errors were reduced by 83.78-92.68% and the yaw errors were reduced by 39.79-61.82%. These measurements demonstrate the performance of our SLC algorithms. Such drastic drift reduction is critical for the robot to construct high-fidelity maps as well as navigate safely and accurately. We refer the reader to our demo video for more animation on SLC.

C. Autonomous Exploration and Metric-Semantic Mapping

To evaluate the effectiveness and robustness of our proposed system, the robot performed autonomous exploration missions in the multi-floor indoor environment. Fig. 1 (d-e) shows the final metric-semantic map constructed from these experiments. Our system is able to explore the environment autonomously and generate 3D maps that contain not only geometric but also semantic information about the environment. Although we are only concerned about one specific class of semantic objects (in this case, chairs) in these experiments, our algorithm can directly work with any other classes of objects that can be detected by the instance segmentation model.

Next, we will quantitatively analyze the uncertainty reduction achieved by our proposed system. We employ the average uncertainty, in terms of the trace of covariance matrices, of

robot poses and semantic maps (i.e. semantic landmarks) as evaluation metrics. The results of one autonomous exploration experiment are shown in Fig. 6. As the robot explored the environment, the uncertainty of robot poses and the semantic map gradually increased. When new semantic landmarks were observed, the uncertainty surged. The uncertainty of landmarks decreased as more observations accumulated. During the exploration, the robot actively navigated to establish SLC. Upon SLC, the pose uncertainty dropped sharply by 56.67%. The subsequent observations of landmarks further reduced the uncertainty of landmarks as shown in the bottom right panel. We compared the results with and without the SLC module, by turning off the SLC in the latter. At the end of the exploration, the average robot pose uncertainty was reduced by 52.06%, and the average landmark uncertainty was reduced by 68.53%. The total trajectory length of this mission was 227.47m, where SLC happened in the middle of the mission when the robot traveled 85.6m. Results show that the SLC module reduced position errors by 17.07% and yaw errors by 74.46%.

To demonstrate the robustness of our proposed system, the results of multiple autonomous exploration experiments are presented in Table. II. The second column shows the uncertainty reduction upon SLC, which is calculated based on the difference in uncertainties before and after the SLC event. The SLC effectively reduced the average uncertainty of robot pose by 46-57%. It is worth noting that, in instances where consecutive SLCs took place over a short travel distance, there was a diminishing return in terms of uncertainty reduction, as expected. The third and fourth columns show the uncertainty reduction of poses and semantic maps, which is derived based on the difference in uncertainties with and without SLC module after the entire mission. This is achieved by simply turning on and off the SLC module. Results demonstrate that our system achieves up to 71% and 69% reduction in uncertainties of robot poses and semantic maps. The consistent reduction of errors and uncertainties in robot poses and semantic maps indicates that our system is robust and effective.

D. Benchmark

Finally, we conduct benchmark experiments with state-of-the-art SLAM methods. Due to drastic viewpoint changes in our datasets, Kimera [6] and ORB-SLAM3 [33] cannot detect loop closure as they require matching image features between keyframes. In addition, within the six evaluated benchmark datasets, both Kimera and ORB-SLAM3 encountered failures on certain ones. On the datasets where they both succeeded,

Kimera and ORB-SLAM3 result in an odometry drift of 1.93-3.71% and 0.45-1.51%, respectively. However, our SLC algorithm is robust to viewpoint changes, contributing to the superior performance of our system. As a result, our system has an odometry drift that is consistently under 0.5%.

VIII. CONCLUSION

In this paper, we developed a system for 3D exploration and metric-semantic mapping of GPS-denied indoor environments with autonomous UAVs. Our system features core algorithms, including metric-semantic SLAM, COP-based exploration planning, active SLC, and active uncertainty reduction planning. It leverages the abstractions of the environment, including exploration viewpoints extracted from the metric map, and the sparse semantic map, to significantly reduce computational load for real-time exploration and active localization. Through extensive real-world experiments, we show the effectiveness of our proposed system in enabling the UAV to plan long-horizon paths, trading off exploration and exploitation. Qualitative results demonstrate that our system empowered the UAV to not only explore the multi-floor environment and construct metric-semantic maps, but also intermittently establish SLC to improve the quality of the map. The quantitative evaluation shows that our SLC module can help the robot significantly reduce position and orientation estimation errors and uncertainties. We envision that such a system can be deployed to solve various real-world problems.

REFERENCES

- [1] S. W. Chen, G. V. Nardari, E. S. Lee, C. Qu, X. Liu, R. A. F. Romero, and V. Kumar, "SLOAM: Semantic lidar odometry and mapping for forest inventory," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 612–619, 2020.
- [2] S. Yang and S. Scherer, "CubeSLAM: Monocular 3-D object SLAM," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.
- [3] 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.
- [4] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic SLAM," in *IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.
- [5] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Dual quadratics from object detections as landmarks in object-oriented SLAM," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.
- [6] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From SLAM to spatial perception with 3d dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021.
- [7] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. Towards New Computational Principles for Robotics and Automation*. IEEE, 1997, pp. 146–151.
- [8] C. Wang, D. Zhu, T. Li, M. Q.-H. Meng, and C. W. de Silva, "Efficient autonomous robotic exploration with semantic road map in indoor environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2989–2996, 2019.
- [9] K. Saulnier, N. Atanasov, G. J. Pappas, and V. Kumar, "Information theoretic active exploration in signed distance fields," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4080–4085.
- [10] C. Gomez, M. Fehr, A. Millane, A. C. Hernandez, J. Nieto, R. Barber, and R. Siegwart, "Hybrid topological and 3d dense mapping through autonomous exploration for large indoor environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9673–9679.
- [11] S. Shen, N. Michael, and V. Kumar, "Autonomous indoor 3d exploration with a micro-aerial vehicle," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 9–15.
- [12] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon "next-best-view" planner for 3d exploration," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1462–1468.
- [13] Z. Meng, H. Qin, Z. Chen, X. Chen, H. Sun, F. Lin, and M. H. Ang, "A two-stage optimized next-view planning framework for 3-D unknown environment exploration, and structural reconstruction," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1680–1687, 2017.
- [14] B. Zhou, Y. Zhang, X. Chen, and S. Shen, "Fuel: Fast uav exploration using incremental frontier structure and hierarchical planning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 779–786, 2021.
- [15] X. Liu, A. Prabhu, F. Cladera, I. D. Miller, L. Zhou, C. J. Taylor, and V. Kumar, "Active metric-semantic mapping by multiple aerial robots," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3282–3288.
- [16] A. Asgharivaskasi and N. Atanasov, "Active bayesian multi-class mapping from range and semantic segmentation observations," *CoRR*, vol. abs/2112.04063, 2021.
- [17] J. A. Placed, J. Strader, H. Carrillo, N. Atanasov, V. Indelman, L. Carlone, and J. A. Castellanos, "A survey on active simultaneous localization and mapping: State of the art and new frontiers," *planning*, vol. 2, 2022.
- [18] J. Yu, M. Schwager, and D. Rus, "Correlated orienteering problem and its application to persistent monitoring tasks," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1106–1118, 2016.
- [19] S. Agarwal and S. Akella, "The correlated arc orienteering problem," in *Algorithmic Foundations of Robotics XV*, S. M. LaValle, J. M. O’Kane, M. Otte, D. Sadigh, and P. Tokekar, Eds. Cham, Switzerland: Springer, 2023, pp. 402–418.
- [20] X. Liu, G. V. Nardari, F. C. Ojeda, Y. Tao, A. Zhou, T. Donnelly, C. Qu, S. W. Chen, R. A. F. Romero, C. J. Taylor, *et al.*, "Large-scale autonomous flight with real-time semantic SLAM under dense forest canopy," *IEEE Robotics and Automation Letters*, 2022.
- [21] I. D. Miller, F. Cladera, T. Smith, C. J. Taylor, and V. Kumar, "Stronger together: Air-ground robotic collaboration using semantics," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9643–9650, 2022.
- [22] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," *Robotics: Science and Systems XVIII*, 2022.
- [23] J. Yu and S. Shen, "Semanticloop: Loop closure with 3d semantic graph matching," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 568–575, 2023.
- [24] Y. Zhang, B. Zhou, L. Wang, and S. Shen, "Exploration with global consistency using real-time re-integration and active loop closure," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9682–9688.
- [25] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [26] Y. Tao, Y. Wu, B. Li, F. Cladera, A. Zhou, D. Thakur, and V. Kumar, "SEER: Safe efficient exploration for aerial robots using learning to predict information gain," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1235–1241.
- [27] "ModalAI VOXL Technical Docs." [Online]. Available: <https://docs.modalai.com/docs/datasheets/voxl-datasheet>
- [28] F. Dellaert and GTSAM Contributors, "borglab/gtsam," May 2022. [Online]. Available: <https://github.com/borglab/gtsam>
- [29] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "isam2: Incremental smoothing and mapping using the bayes tree," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.
- [30] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [31] I. Spasojevic, X. Liu, A. Prabhu, A. Ribeiro, G. J. Pappas, and V. Kumar, "Robust localization of aerial vehicles via active control of identical ground vehicles," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3048–3055.
- [32] Z. Wang, X. Zhou, C. Xu, and F. Gao, "Geometrically constrained trajectory optimization for multicopters," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 3259–3278, 2022.
- [33] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, p. 1874–1890, Dec. 2021.