Markerless Robot Detection and 6D Pose Estimation for Multi-Agent SLAM

Markus Rüggeberg[1], Maximilian Ulmer[1,2], Maximilian Durner[1],
Wout Boerdijk[1], Marcus Müller[1], Rudolph Triebel[1,2] and Riccardo Giubilato[1†]

***Abstract*— The capability of multi-robot SLAM approaches to merge localization history and maps from different observers is often challenged by the difficulty in establishing data association. Loop closure detection between perceptual inputs of different robotic agents is easily compromised in the context of perceptual aliasing, or when perspectives differ significantly. For this reason, direct mutual observation among robots is a powerful way to connect partial SLAM graphs, but often relies on the presence of calibrated arrays of fiducial markers (e.g., AprilTag arrays), which severely limits the range of observations and frequently fails under sharp lighting conditions, e.g., reflections or overexposure. In this work, we propose a novel solution to this problem leveraging recent advances in Deep-Learning-based 6D pose estimation. We feature markerless pose estimation as part of a decentralized multi-robot SLAM system and demonstrate the benefit to the relative localization accuracy among the robotic team. The solution is validated experimentally on data recorded in a test field campaign on a planetary analogous environment.**

I. INTRODUCTION

A necessary capability of robotic teams to perform collaborative actions, e.g., industrial inspection [1], rescue [2], or scientific exploration [3], is the one of mutual localization and collaborative mapping. Collaborative, or multi-robot Simultaneous Localization and Mapping approaches (SLAM), extend the ability of single robots to create maps of unknown environments, and localize themselves with respect to it, to multiple agents operating at the same time.

This conceptually simple addition brings a variety of complications to the table, from the perspective of many components of a traditional SLAM system. From an architectural point of view, multiple strategies to fuse data from the robots have been explored. Centralized SLAM approaches [4] vs. distributed SLAM approaches [5] bring considerable consequences both in term of algorithmic complexity as well as system behaviour and robustness to failure during operation. From the perceptual point of view, the task of associating measurements between robots, i.e., performing loop closures in visual SLAM systems, is severely affected by strong perspective and illumination changes, due to viewpoint differences between the systems [6]. Similarly, the problem of data association in LiDAR-based perception systems is affected by different noise patterns and pointcloud sparsity when robots operate at a distance. These issues are further exacerbated when operating on natural unstructured grounds, e.g., in the context of field robotic operations or planetary exploration, where no particular constraints are imposed on traversable paths.

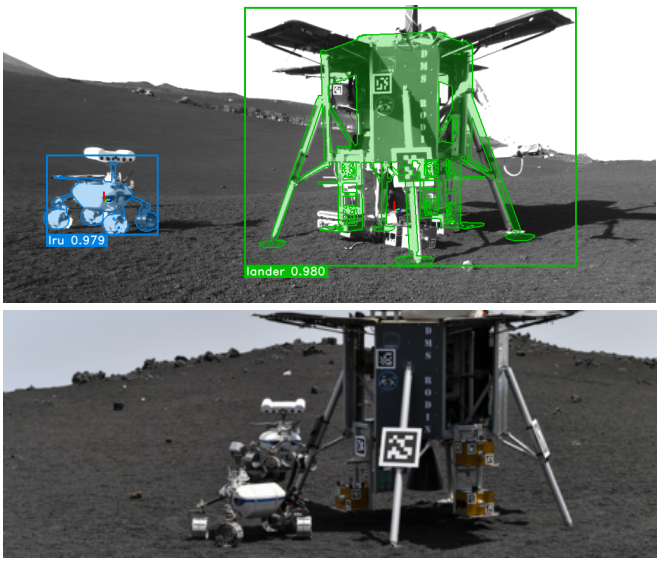

Fig. 1: (**top:**) Multi-robot detection of the LRU (Lightweight Rover Unit), and the Lander unit, from the perspective of the 2nd LRU unit during the [anonymous] field test campaign on Mount. Etna, Sicily. The figure shows the projection of the robot shapes, known a-priori, to demonstrate the quality of pose estimation. The distance of the LRU rover to the observer, the LRU2, as well as intense light reflections, would make it impossible through conventional fiducial markers to establish a robot detection. (**bottom:**) Members of the multi-robot team on Mt. Etna: LRU, LRU2 and the Lander unit.

Association between measurements from the robots in the team is of critical importance in collaborative SLAM, as it allows robots to join their environment representations into a single, and shared, formulation. Under aliased and ambiguous perceptual conditions, correspondence between views of the environments from multiple agents is challenging, therefore **direct observation** of other robots allow to establish immediate correspondences between, e.g., SLAM graphs, solving the association problem in a relatively straight-forward manner. While many existing collaborative

† Corresponding author

*This work was supported by the Helmholtz Association project iFOODis (contract number KA2-HSC-06) and by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG)

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Weßling, Germany fistname.lastname@dlr.de

[2]Rudolph Triebel is also with Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany rudolph.triebel@kit.de

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. DOI: 10.1109/ICRA57385.2026.11696223

TABLE I: Comparison of Selected Collaborative SLAM, Tracking or Localization Approaches by Inter-Robot Constraint

Method	Constraint Source	Detects Robots	Use Markers	Detection Type	Modality
Kimera-Multi [7]	VPR, BoW vectors	✗	-	-	Visual
DOOR-SLAM [5]	VPR, NetVLAD descriptors	✗	-	-	Visual
CVI-SLAM [8]	VPR, visual features	✗	-	-	Visual
CCM-SLAM [4]	VPR, visual features	✗	-	-	Visual
LAMP 2.0 [9]	LiDAR scan registration	✗	-	-	LiDAR
DiSCo-SLAM [10]	Scan Context descriptors	✗	-	-	LiDAR
Swarm-SLAM [11]	Scan Context descriptors	✗	-	-	LiDAR
MultiReg [12]	Generic measurements	✓	✓†	SE(2)	Generic
UVDAR [13]	Direct robot measurements	✓	✓(UV LEDs)	Range/Bearing	Visual
CREPES [14]	Camera, IMU and UWB fusion	✓(LEDs)	✓	SE(3)	Multiple
[15]	Camera tracking & target VIO fusion	✓	✗	SE(3)	Visual
[16]	DL-based keypoint extraction	✓	✗	SE(3)	Visual
MSL-Raptor [17]	DL-based bounding-box extraction	✓	✗	SE(3)	Visual

† Markers are specific geometries, e.g., square plates, but missing unique identifiers

SLAM approaches do not explicitly consider this case, the ones that do often utilize traditional visual markers, e.g. AprilTag [18] or ArUco [19]. This choice, however, brings significant limitations: the size of the markers limits the **maximum distance** at which a robot can be detected, the mounting positions limits **relative orientations** valid for detection, and **harsh lighting** conditions in outdoor contexts can prevent detection entirely.

We propose a solution to this problem taking inspiration from recent advances in 6D pose estimation techniques for arbitrary objects [20]. With the assumption that the robots that constitute a team have a known shape and appearance, we employ a deep learning-based approach to detection and relative 6D pose estimation. A simplified prior shape of each robot is aligned to detected instances of the robots in each other's camera image, estimating a full transformation that is directly used in a SLAM context. In summary, this manuscript presents the following contributions:

- To the best of our knowledge, we present for the first time a **markerless detection and pose estimation** approach to mutual robot detection, that relies on prior shape knowledge without handcrafted features or visual markers.
- We demonstrate the integration of our markerless robot detection scheme in an online **multi-robot SLAM system** for a team of planetary rovers.
- We highlight the performance of our system, and the benefit against traditional marker-based pose estimation on a large body of data recorded during a test campaign on a planetary analogous environment.

II. Related Works

In this section we provide an overview of related multi-agent SLAM works, regardless of the principal sensing modality, with the focus on methods that put the emphasis on the perceptual challenge in associating measurements from the members of the robotic team. Among these, there exist methods that consider direct mutual measurements between the agents, such as our proposed approach, or try to associate measurements from the environment shared across the team. Table I presents a synthetic overview of the works that are discussed in the next sections.

A. Indirect Multi-Robot Associations

In this category belong multi-agent SLAM systems that do not explicitly consider direct robot detection to join estimates from the team participants, but instead focus on the detection of *loop closures*, i.e., the association of measurable perceptual properties of the environment. Kimera-Multi [21] is a distributed multi-robot SLAM system that extend the visual-inertial and semantic mapping framework Kimera [7]. The multi-robot team shares Bag-of-Words vectors computed from traditional image features and attempts to find consistent associations using an incremental formulation of PCM (Pairwise Consistent Measurements), followed by a robust optimization scheme. Similar to Kimera-Multi is DOOR-SLAM [5], a decentralized multi-robot approach that too relies on careful selection of loop closure candidates through PCM. Differently from Kimera-Multi, however, visual keyframes are processed to extract global descriptors with NetVLAD, validating putative matches with geometrical verification. The collaborative visual SLAM CVI-SLAM [8], and its evolution CCM-SLAM [4] are centralized approaches where a number of agents run only a visual-inertial *front-end* and communicate keyframes and feature descriptors to a central server, which runs a loop closure detection and mapping engine, localizing agents on a shared feature map. All these methods rely on a visual modality to find correspondences between environment observations, which is prone to failure when the appearance is aliased and ambiguous [6]. Another family of multi-robot SLAM approaches focus instead on the LiDAR modality, emphasizing structure rather than texture. LAMP2.0 [9] is a centralized LiDAR-based multi-agent SLAM approach where a central server maintains and joins pose graphs from the participating agents based on scan registration. As the authors mention, however, the correspondences are prioritized based on an initial pose prior, which must be obtained by heuristics. LiDAR-based loop closure detections are instead used by the distributed SLAM approach DiSCo-SLAM [10]. By means of sharing

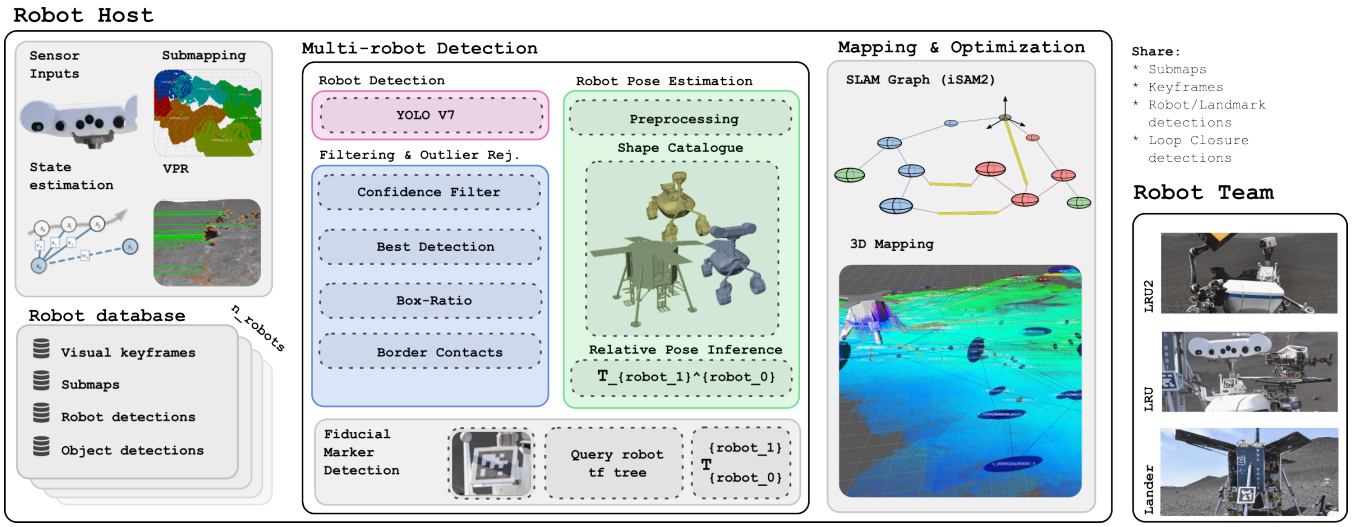

Fig. 2: Schematic overview of the employed decentralized SLAM system, with a focus on the multi-robot detection module capabilities. Each robot utilizes visual and inertial inputs to compute state estimation and partition robot states into submaps. Visual inputs are used to compute visual keyframes for place recognition, as well as detecting robots in the image, either through legacy AprilTag detection, or the proposed markerless approach. Results of submapping and robot detections from each robot in the team are embedded into a SLAM graph

and matching Scan Context descriptors [22], robots attempt to initialize and register their maps among a team. Although LiDAR sensing benefits from the robustness and invariance of structure to different perceptual conditions, the absence of salient structure, such as in the case of planetary-like scenes, rapidly leads to failures, as the authors of Swarm-SLAM [11] demonstrate in their field tests [23].

B. Direct Multi-Robot Associations

Directly measuring transformations between robots using onboard perception has been investigated for several purposes, either tracking systems from a base station, to provide absolute localization with respect to it, or to enable relative localization capabilities to a team of robots in the context of SLAM. The authors of [12] investigate the topic of relative localization between robots by means of mutual detection. Their system, named MultiReg, tackles the case where robot identities in a swarm are either ambiguous or unknown, and develop a theoretical framework to establish and validate detections. The approach is demonstrated with LiDAR-equipped systems, and, while theoretically applicable to different perception systems, visual fiducial markers could easily provide means of identification. Other approaches rely on visual perception systems and fiducial markers, which provide both an identity and means to establish transformations. The authors of [13] present UVDAR, a biologically inspired method for swarm localization among a team of UAVs. The author solve the problem of mutual localization between systems by means of observing LED patterns with specific blinking frequencies with UV-sensitive cameras. Similarly, the authors of CREPES [14] propose a relative localization system that rely on LEDs to define an identity, while using mutual state estimation inputs to compute relative poses. In order to avoid implementing specific hardware or preparing robots with fiducial markers, which especially for small agile system can be undesirable, the authors of MSL-Raptor [17] propose a Deep Learning architecture to infer the coordinates of a bounding box around a target drone observed from a camera. Similarly, the authors of [16] propose to learn instead drone-specific visual keypoints pointing to features such as rotor position. These approaches show multiple alternatives to detect and estimate the pose of known robots, but either require specific hardware, e.g. LEDs, or consider specific robot features that do not easily generalize. With this work we aim to provide a general solution to this problem that extends to arbitrary robots equipped with vision sensors, and for arbitrary contexts of operation.

III. Methods

In this section, we present the methods and components of our work, consisting of a multi-robot SLAM system (III-A), a learning-based markerless robot detection and 6D pose estimation pipeline (III-B), and a systematic training framework with planetary analogue synthetic training data (III-C).

A. Multi-Robot SLAM

This work is based upon a decentralized multi-robot SLAM system, developed at the [Anonymous Institution] for a team of heterogeneous robotic systems, such as the Lightweight Rover Unit (LRU) [24] and the ARDEA UAV [25], and described in detail in [26].

The SLAM system, in principle agnostic to the perception modality, is specialized for robots equipped with stereo vision. Local state estimation is computed as a Local Reference Filter [27] and fuses Visual Odometry (VO), IMU measurements, and odometric sources. State estimation is

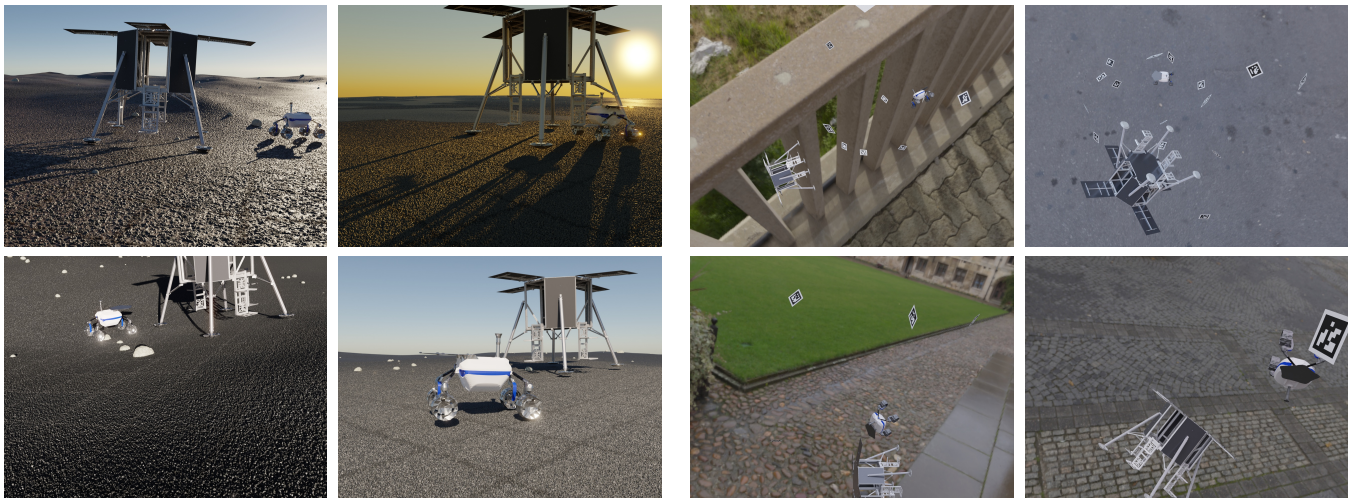

Fig. 3: Impressions of training samples. (**left:**) OAISYS images including LRU and Lander in an Etna-like setting. (**right:**) BlenderProc images, with LRU, Lander and random AprilTags for distractions, on a random background.

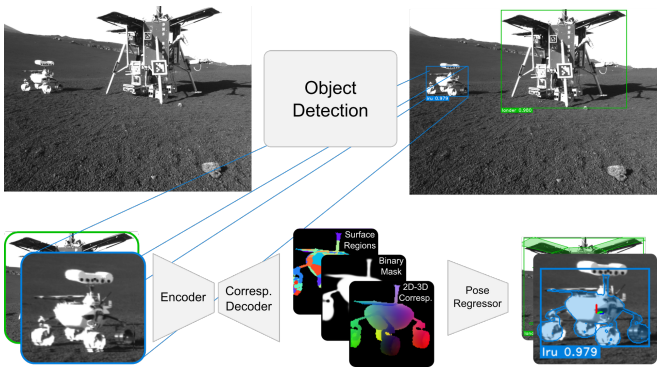

Fig. 4: Illustration of the markerless detection pipeline. A 2D object detection first generates an image crop, which is then processed by the pose estimator. The encoder-decoder architecture produces three outputs: 2D–3D correspondences, foreground-background masks, and surface regions as an auxiliary task. The 2D–3D correspondences are subsequently passed to a pose regression network to predict the object's 6D pose.

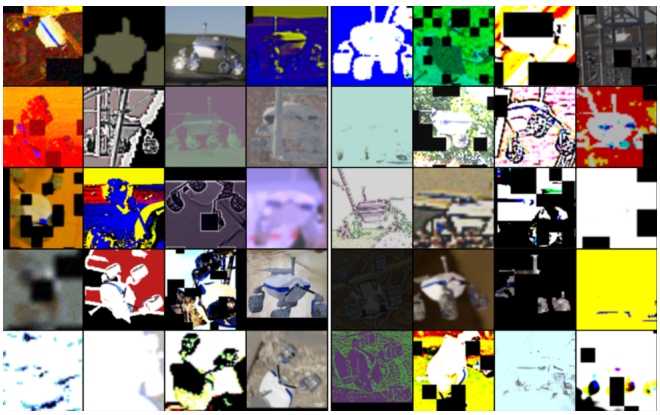

Fig. 5: Examples of heavy augmentations on LRU instances from the synthetic samples used for an effective training of the pose estimation network in Sec. III-B

therefore referenced with respect to local frames that switch to the current estimated position when the pose covariance grows over a certain limit, marginalizing past states. Frame switches signal the creation of new submaps, which partition the environment in rigid elements. A submap matching modules attempts to register overlapping submaps to create loop closures constraints [28]. Similarly, keyframes are extracted from camera views, and AKAZE features are used to insert visual words in a growing database and recalling matching keyframes, in order to establish visual loop closures.

In order to join estimates from multiple robots into a combined SLAM session, correspondences between robot states are estimated either from visual loop closures or direct inter-robot pose measurements. The detection of an array of fiducial markers (AprilTag), mounted on the robots, provides a 6D transformation between the robot's bodies, that initially aligns the individual map frames and subsequently provides additional robust loop closure constraints. Frame switch transformations, loop closure and robot detection constraints, from all the robots operating during a mapping session, are inserted in a pose graph in a decentralized manner. The iSAM2 algorithm, from the GTSAM library, computes then updated poses at quasi-constant computational effort. An architectural overview of the multi-robot SLAM system is shown in Fig. 2.

B. Markerless Robot Detection Pipeline

In this section we present our proposed markerless robot detection and pose estimation component, depicted in Fig. 4.

1) Detection and Instance Filtering: Known robot instances are searched for in all input images using the YOLO v7 object detector. Detection results are ordered by confidence, and only the one with the highest confidence value is retained for further processing. Excessive visual

TABLE II: Synthetic Test Results: Detection rate ρ, translation error t, rotational error θ for different training and test set combinations. The arrow after a variable indicates whether a higher (↑) or lower (↓) value is desirable.

Test Set	Performance Metric	Observed Robot	Training Set		
			BPROC	OAISYS	COMBINED
BPROC	ρ (↑)	Lander	0.160	0.260	**0.284**
		LRU	0.080	0.064	**0.324**
	t $[m]$ (↓)	Lander	0.310	1.497	**0.259**
		LRU	**0.082**	1.621	0.153
	θ $[rad]$ (↓)	Lander	**0.421**	2.207	0.514
		LRU	**0.099**	1.143	0.188
OAISYS	ρ (↑)	Lander	0.080	0.676	**0.792**
		LRU	0.020	0.184	**0.764**
	t $[m]$ (↓)	Lander	0.184	0.044	**0.040**
		LRU	**0.034**	0.115	0.098
	θ $[rad]$ (↓)	Lander	0.483	0.089	**0.048**
		LRU	0.128	0.092	**0.052**

occlusion or cases where the robot are cropped by image borders are identified by checking how the aspect ratio of the bounding boxes differ to a value of 1. Only detections whose bounding boxes have aspect ratios within a specified interval $\frac{1}{r} \leq \frac{height}{width} \leq r$ for a predefined aspect ratio $r > 1$ are accepted. Additionally, bounding boxes are discarded if they touch more than one image border, as in this case the aspect ratio check is insufficient to exclude partial detections.

2) Pose Estimation: We build upon the extended approach of Ulmer *et al.* [20]. It has demonstrated promising results in the domain of satellite pose estimation, a field that, like planetary exploration, has to deal with extreme lighting conditions. Our experiments showed that the method can be efficiently trained on modern hardware in an end-to-end fashion. Furthermore, it has sufficient capacity to handle the challenging conditions typical to space environments [29] and also shows promising performance on other benchmark datasets, confirming its suitability for our application. The approach employs a dense 2D-to-3D correspondence predictor, allowing the regression of 3D model coordinates for every pixel in the input image. To further improve robustness, we introduce several adaptations. We replace the decoder with a transformer-based architecture incorporating an attention mechanism. Inspired by GDR-Net [30], we substitute the classical PnP procedure with a neural network that directly regresses the target object's 6D pose. Additionally, in parallel with the standard modal instance mask prediction, we introduce a new prediction head to estimate amodal masks, enhancing the system's understanding of occluded object regions.

C. Training on Synthetic Data

Both models for instance detection (Sec.III-B.1) and pose estimation (Sec.III-B) were trained on synthetic data generated from simplified CAD models of the known robot types. Both variants of the Lightweight Rover Unit, LRU and LRU2, were treated as a single type of robot, with training performed on an intersection of their respective 3D geometries. A dataset was generated containing 8000 photo-realistic image samples, each annotated with binary object masks and reporting ground truth 6D poses with respect to the virtual camera. The first half of the training samples were generated using DLR's open-source rendering framework BlenderProc 2 [31], a Blender-based procedural pipeline to generate synthetic datasets, providing photo-realistic images and ground-truth annotations such as depth and object masks. To enhance robustness, BlenderProc supports generation of distractor objects and random 360°-backgrounds from both indoor and outdoor environments. The second half of the training samples were generated using the open-source rendering and simulation framework OAISYS [32], a Blender-based procedural simulator for unstructured outdoor environments with a specific focus on simulating terrains and natural scenes. Lastly, as previously experienced in [29], we implement a variety of strong data augmentations during training on synthetic input images, see Fig. 5, in order to enhance robustness to challenging visual conditions for real image samples.

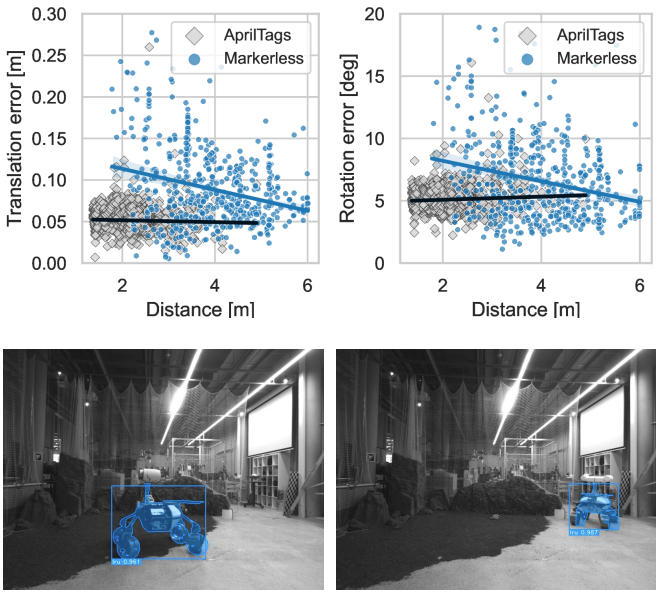

Fig. 6: Evaluation of translation (**top-left**) and rotation (**top-right**) errors of the proposed markerless pose estimation approach, and comparison with AprilTag detection, between the real LRU system, with LRU2 as observer, under VICON ground-truth measurements. (**bottom:**) Visual examples of markerless robot detection and re-projection of the CAD model on the image.

IV. EVALUATION AND DISCUSSIONS

A. Evaluation of Pose Estimation on Synthetic Data

Markerless detection and pose estimation performances were first validated on synthetic data using a desktop computer equipped with an Intel i9 and an NVIDIA RTX4000.

TABLE III: Multi-robot detection metrics during the SLAM tests comparing the baseline approach with AprilTag only, and the proposed addition of Markerless detection and pose estimation

Performance Metric	Tag	Tag + Markerl.	Improvement
Mission 1, LRU			
# Detections	49	87	**+87%**
Max. Det. Distance [m]	4.94	16.15	**+227%**
Max. Open-Loop Duration [s]	2300	1033	**-55%**
Traj. Error [m]	1.57	1.13	**-28%**
Mission 1, LRU2			
# Detections	154	180	**+17%**
Max Det. Distance [m]	5.63	9.12	**+62%**
Max Open-Loop Duration [s]	980	608	**-38%**
Traj. Error [m]	0.75	0.78	**+4%**
Mission 2, LRU			
# Detections	14	20	**+43%**
Max. Det. Distance [m]	3.42	17.14	**+402%**
Max. Open-Loop Duration [s]	440	203	**-54%**
Traj. Error [m]	0.76	0.52	**-32%**
Mission 2, LRU2			
# Detections	18	19	**+6%**
Max. Det. Distance [m]	7.63	7.63	–
Max. Open-Loop Duration [s]	490	490	–
Traj. Error [m]	0.96	1.05	**+9%**

Different models were trained separately on BlenderProc data, OAISYS data and combined data from both sources. The separate models were then evaluated on BlenderProc and OAISYS test datasets. The test set included 500 unseen BlenderProc and OAISYS samples, for which detection rates and pose accuracy were measured. Table II reports detection rate, translation and rotation accuracy of each model on two separate BlenderProc- and OAISYS-only test data, showing that the best performances are achieved on training with diverse data. For a collection of visual impressions of generated data from both simulation toolkits, see Fig. 3. In this evaluation, we measured the runtime of the combined detection and pose estimation steps achieving rates of up to 14 Hz, while for the detection alone up to 20 Hz.

B. Evaluation of Pose Estimation on Real Data

We now evaluate, on real image samples, the performances of the pose estimation approach trained in simulation (Sec. IV-A) using the generic LRU model without pan-tilt head. Both LRU systems, with LRU2 being the observer and LRU being the target, are used in an indoor laboratory settings equipped with a VICON system. With the observer placed in one corner of the VICON covered area, the target robot is manually driven in order to create a dataset of relative pose samples that cover as densely as possible ranges and relative orientations from the available space. The pose estimation method is then compared against a baseline multi-robot detection approach based on the detection of an array of AprilTags mounted around the neck of both rovers. Both the markerless- and AprilTag-based approaches estimate the transformation from the rover's body centers. Measurements are then compared with a VICON ground truth. Results in Fig. 6 highlight the distribution of rotation and translation errors separately against the true distance from the body centers. The first important insight is that the baseline approach fails to cover the entire range of distances, showing a distinct limit around the 4 meters range, which can vary with the marker size. Secondly, both translation and rotation accuracy of markerless pose estimations improves at larger distances. We assume that this behavior depends on a decreasing effect of perspective distortion at longer ranges, and may as well be compensate with carefully tuned synthetic data generation for training. Furthermore, the results highlight how both modalities complement each other with a clear benefit for inclusion in a SLAM system.

C. Evaluation in a Multi-Robot SLAM System

To validate the integrated system performance of the markerless SLAM approach in a real-world use case, the proposed markerless robot detection approach was interfaced as a ROS2 component in the multi-robot SLAM solution described in Sec. III-A. The SLAM system was evaluated offline on a variety of data recorded during the 2022 ARCHES field test campaign [3] on the analogous planetary surface of Mt. Etna, Sicily. In this evaluation, we consider the recordings of two multi-robot navigation experiments, denoted here Mission 1 and Mission 2, that lasted respectively 42 min and 14 min, during which the LRUs drove a wide trajectory in a 40m$\times$40m area around the Lander. The Lander, equipped with large AprilTags, functioned as the central reference for the multi-robot SLAM results. At the beginning of each mapping session, the LRUs glanced at the Lander to establish a connection to the elected global frame before proceeding to navigate to user-defined waypoints. Thanks to the distributed nature of the employed SLAM system, the LRUs communicated at all time any measurement that affected the structure of their shared SLAM graph, for instance AprilTag-based and markerless robot detections, visual loop closures, or additions of new submaps. For a detailed explanation of the SLAM system and multi-robot communication approach we refer to [26].

This evaluation focuses on *instantaneous* localization accuracy rather than traditional *a-posteriori* metrics, as the former is what directly impacts the performance of open-loop navigation toward POIs. We measure multi-robot detection frequency and maximum range while quantifying the reduction of the length of open-loop navigation sequences, that is, navigation sequences between detection events. Localization performance is summarized as the RMSE (Root Mean Square Error) between SLAM estimates and D-GNSS measurements.

As shown in Table III, our markerless approach significantly improves detection rates, ranges, and accuracy across both missions at the cost of potential minor fluctuations (see

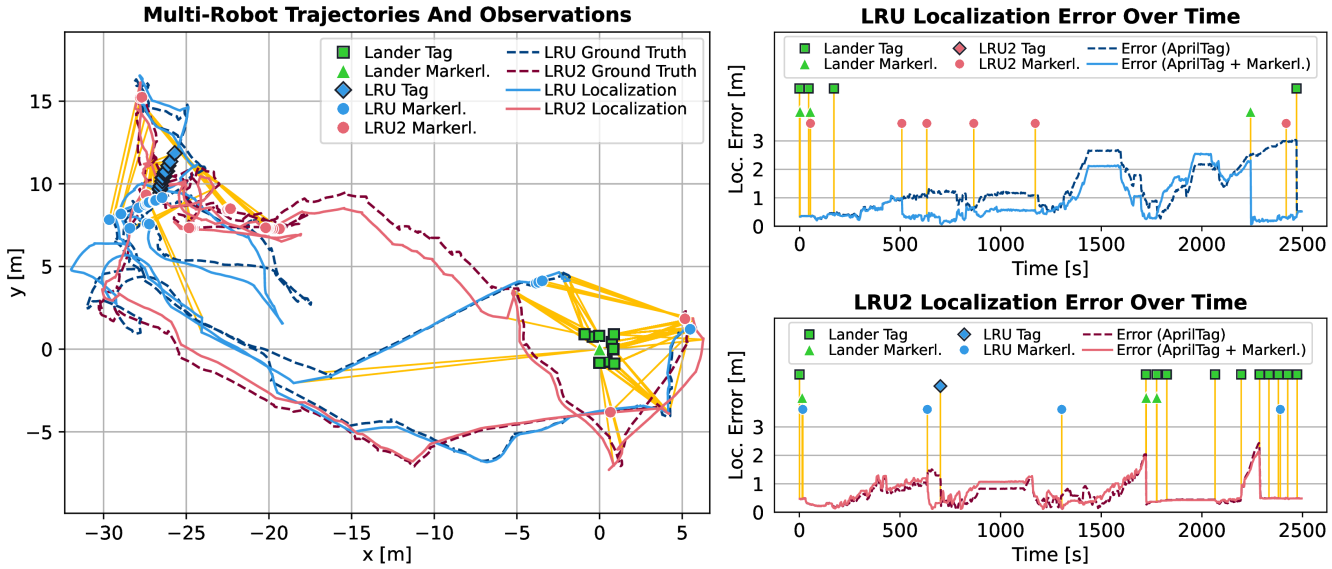

Fig. 7: Multi-robot SLAM results from **Mission 1**. (**left:**) Trajectory of the two rovers over the entire mission, localized via visual multi-robot SLAM using full input knowledge. (**top right:**) Localization error of LRU, when observing the Lander and LRU2. Error correction after markerless LRU2 observations at **516s**, **874s** and markerless Lander detection at **2250s**. (**bottom right:**) Localization error of LRU2, when observing the Lander and LRU. Error correction after markerless LRU detections at **636s**, **1304s**.

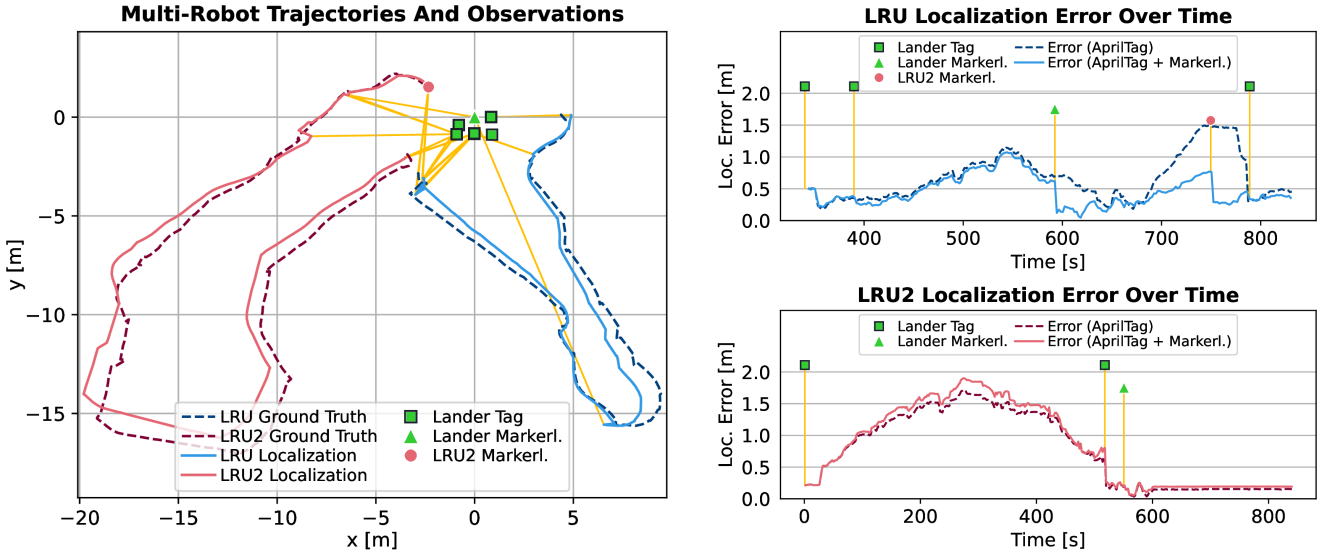

Fig. 8: Multi-robot SLAM results from **Mission 2**. (**left:**) Trajectory of the two rovers over the entire mission, localized via visual multi-robot SLAM using full input knowledge. (**top right:**) Localization error of LRU, when observing the Lander and LRU2. Error correction after markerless Lander observation at **593s** and markerless LRU2 observation at **751s**. (**bottom right:**) Localization error of LRU2, when observing the Lander and LRU. No significant differences with or without the markerless modality.

LRU2 trajectory error in Table III) due to sub-par rotational accuracy of markerless multi-robot detection at very close ranges compared to fiducial markers, effect that is already reported in Fig. 6. Finally, Fig. 7 and Fig. 8 visualize trajectories and their accuracy against a D-GNSS ground truth over mission time, highlighting sharp error reductions triggered by markerless detection events.

V. CONCLUSIONS

Markerless robot detection delivers a significant performance improvement in the context of multi-robot SLAM,

especially when operating in perceptually challenging environments, overcoming the limitations imposed by traditional approaches based on fiducial markers (e.g., AprilTag), which are limited in range and sensitive to lighting conditions. For robots whose geometry is known as full or partial CAD models the proposed approach increases the distance of robot detection up to 400% during our field tests, regardless of lighting conditions, leading to significant improvement on the instantaneous localization accuracy, which has paramount importance during live operations. Future work will explore solutions for robots as articulated ensembles of rigid components (head, torso, wheels, etc.), as well as the deployment of the approach into embedded GPUs for online operation.

REFERENCES

[1] M. Tavakoli, G. Cabrita, R. Faria, L. Marques, and A. T. de Almeida, "Cooperative multi-agent mapping of three-dimensional structures for pipeline inspection applications," *The International Journal of Robotics Research*, vol. 31, no. 12, pp. 1489–1503, 2012.

[2] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju, and T. Westerlund, "Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision," *Ieee Access*, vol. 8, pp. 191 617–191 643, 2020.

[3] M. J. Schuster, M. G. Müller, S. G. Brunner, H. Lehner, P. Lehner, R. Sakagami, A. Dömel, L. Meyer, B. Vodermayer, R. Giubilato *et al.*, "The arches space-analogue demonstration mission: Towards heterogeneous teams of autonomous robots for collaborative scientific sampling in planetary exploration," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5315–5322, 2020.

[4] P. Schmuck and M. Chli, "Ccm-slam: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams," *Journal of Field Robotics*, vol. 36, no. 4, pp. 763–781, 2019.

[5] P.-Y. Lajoie, B. Ramtoula, Y. Chang, L. Carlone, and G. Beltrame, "Door-slam: Distributed, online, and outlier resilient slam for robotic teams," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1656–1663, 2020.

[6] R. Giubilato, W. Stürzl, A. Wedler, and R. Triebel, "Challenges of slam in extremely unstructured environments: The dlr planetary stereo, solid-state lidar, inertial dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8721–8728, 2022.

[7] 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.

[8] M. Karrer, P. Schmuck, and M. Chli, "Cvi-slam—collaborative visual-inertial slam," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2762–2769, 2018.

[9] Y. Chang, K. Ebadi, C. E. Denniston, M. F. Ginting, A. Rosinol, A. Reinke, M. Palieri, J. Shi, A. Chatterjee, B. Morrell *et al.*, "Lamp 2.0: A robust multi-robot slam system for operation in challenging large-scale underground environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9175–9182, 2022.

[10] Y. Huang, T. Shan, F. Chen, and B. Englot, "Disco-slam: Distributed scan context-enabled multi-robot lidar slam with two-stage global-local graph optimization," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1150–1157, 2021.

[11] P.-Y. Lajoie and G. Beltrame, "Swarm-slam: Sparse decentralized collaborative simultaneous localization and mapping framework for multi-robot systems," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 475–482, 2023.

[12] A. Franchi, G. Oriolo, and P. Stegagno, "Mutual localization in a multi-robot system with anonymous relative position measures," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 3974–3980.

[13] P. Petráček, V. Walter, T. Báča, and M. Saska, "Bio-inspired compact swarms of unmanned aerial vehicles without communication and external localization," *Bioinspiration & Biomimetics*, vol. 16, no. 2, p. 026009, 2020.

[14] Z. Xun, J. Huang, Z. Li, Z. Ying, Y. Wang, C. Xu, F. Gao, and Y. Cao, "Crepes: Cooperative relative pose estimation system," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5274–5281.

[15] F. Steidle, S. Boche, W. Stürzl, and R. Triebel, "A temporal perspective n-point problem with model uncertainties for cooperative pose estimation in a heterogeneous robot team," in *2023 European Conference on Mobile Robots (ECMR)*. IEEE, 2023, pp. 1–7.

[16] R. Jin, J. Jiang, Y. Qi, D. Lin, and T. Song, "Drone detection and pose estimation using relational graph networks," *Sensors*, vol. 19, no. 6, p. 1479, 2019.

[17] B. Ramtoula, A. Caccavale, G. Beltrame, and M. Schwager, "Msl-raptor: A 6dof relative pose tracker for onboard robotic perception," in *International Symposium on Experimental Robotics*. Springer, 2020, pp. 520–532.

[18] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4193–4198.

[19] 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.

[20] M. Ulmer, M. Durner, M. Sundermeyer, M. Stoiber, and R. Triebel, "6d object pose estimation from approximate 3d models for orbital robotics," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 10 749–10 756.

[21] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, 2022.

[22] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4802–4809.

[23] P.-Y. Lajoie, K. Soma, H. M. Bong, A. Lemieux-Bourque, R. Zhang, V. S. Varadharajan, and G. Beltrame, "Multi-robot decentralized collaborative slam in planetary analogue environments: Dataset, challenges, and lessons learned," *IEEE Transactions on Field Robotics*, 2025.

[24] M. J. Schuster, S. G. Brunner, K. Bussmann, S. Büttner, A. Dömel, M. Hellerer, H. Lehner, P. Lehner, O. Porges, J. Reill *et al.*, "Towards autonomous planetary exploration: The lightweight rover unit (lru), its success in the spacebotcamp challenge, and beyond," *Journal of Intelligent & Robotic Systems*, vol. 93, no. 3, pp. 461–494, 2019.

[25] P. Lutz, M. G. Mueller, M. Maier, S. Stoneman, T. Tomić, I. von Bargen, M. J. Schuster, F. Steidle, A. Wedler, W. Stürzl *et al.*, "Ardea—an mav with skills for future planetary missions," *Journal of Field Robotics*, vol. 37, no. 4, pp. 515–551, 2020.

[26] M. J. Schuster, K. Schmid, C. Brand, and M. Beetz, "Distributed stereo vision-based 6d localization and mapping for multi-robot teams," *Journal of Field Robotics*, vol. 36, no. 2, pp. 305–332, 2019.

[27] K. Schmid, F. Ruess, and D. Burschka, "Local reference filter for life-long vision aided inertial navigation," in *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014, pp. 1–8.

[28] C. Brand, M. J. Schuster, H. Hirschmüller, and M. Suppa, "Submap matching for stereo-vision based indoor/outdoor slam," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5670–5677.

[29] M. Ulmer, L. Klüpfel, M. Durner, and R. Triebel, "How important are data augmentations to close the domain gap for object detection in orbit?" in *2025 IEEE Aerospace Conference*, 2025, pp. 1–12.

[30] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 611–16 621.

[31] M. Denninger, D. Winkelbauer, M. Sundermeyer, W. Boerdijk, M. Knauer, K. H. Strobl, M. Humt, and R. Triebel, "Blenderproc2: A procedural pipeline for photorealistic rendering," *Journal of Open Source Software*, vol. 8, no. 82, p. 4901, 2023. [Online]. Available: https://doi.org/10.21105/joss.04901

[32] M. G. Müller, M. Durner, A. Gawel, W. Stürzl, R. Triebel, and R. Siegwart, "A Photorealistic Terrain Simulation Pipeline for Unstructured Outdoor Environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.