Artifacts Mapping:
Multi-Modal Semantic Mapping Extension of Geometric Maps

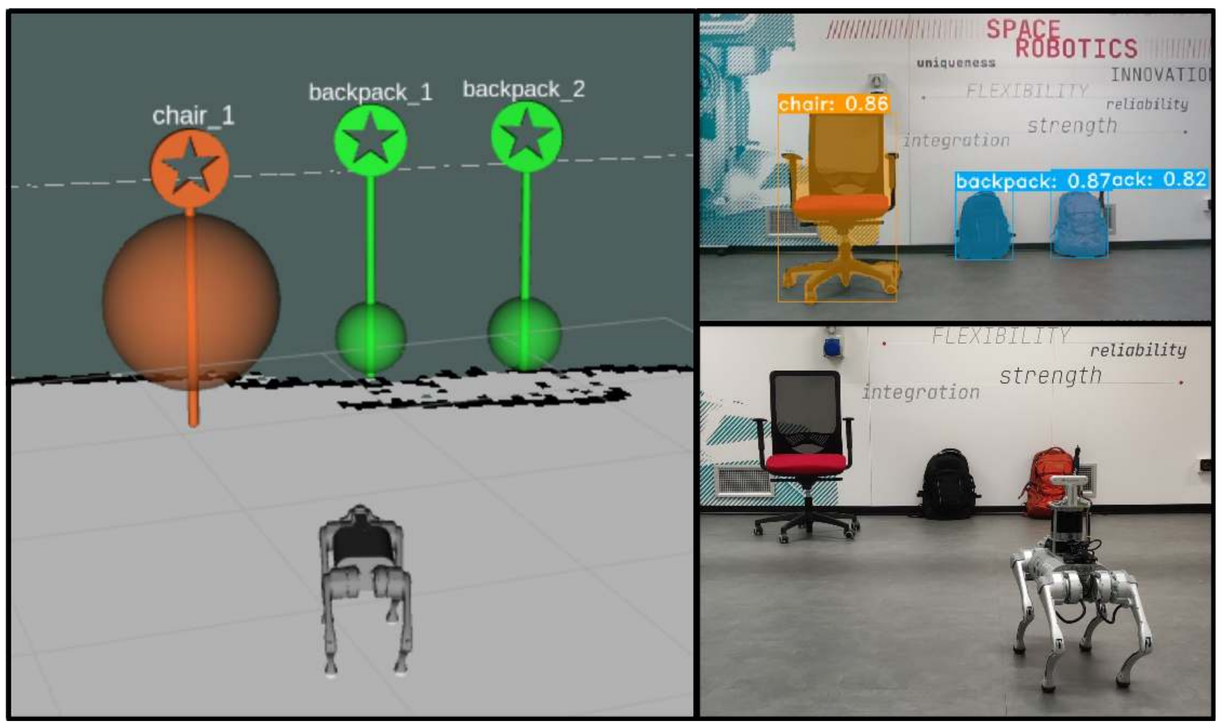
Federico Rollo[1,2,3], Gennaro Raiola[1,2], Andrea Zunino[1,2], Nikolaos Tsagarakis[2], Arash Ajoudani[2]

[1]*Intelligent and Autonomous Systems, Leonardo Labs, Genoa, Italy*

[2]*HHCM & HRII, Istituto Italiano di Technology, Genoa, Italy*

[3]*Industrial Innovation, DISI, Università di Trento, Trento, Italy*

Geometric navigation is now a well-established field in robotics, and the focus of research is shifting towards higher-level scene understanding, such as Semantic Mapping. When a robot needs to interact with its environment, it must be able to comprehend the contextual information of its surroundings. This work focuses on the classification and localization of objects within a map, whether it is in the process of being built (SLAM) or already constructed. To further explore this direction, we propose a framework that can autonomously map predefined objects in a known environment using a multi-modal sensor fusion approach (combining RGB and depth data from an RGB-D camera and a lidar).



Application

In the figure on the right, it is presented an application scenario where an autonomous legged robot accessorized with an RGB-D camera and a lidar is able to correctly identify and localize three objects. The robot is able to autonomously move in the whole environment and perceive objects providing an exact absolute position on the map. On the left, is the visual application where objects are shown with a landmark and a spherical region for the location. Such visual objects provide a simple menu user interaction (e.g. with a mouse right click you can select the *GoTo* option and the robot autonomously navigates towards it). On the top right, is the instance segmentation inference of the image taken from the robot camera while on the bottom right is the external representation of the experimental scene.

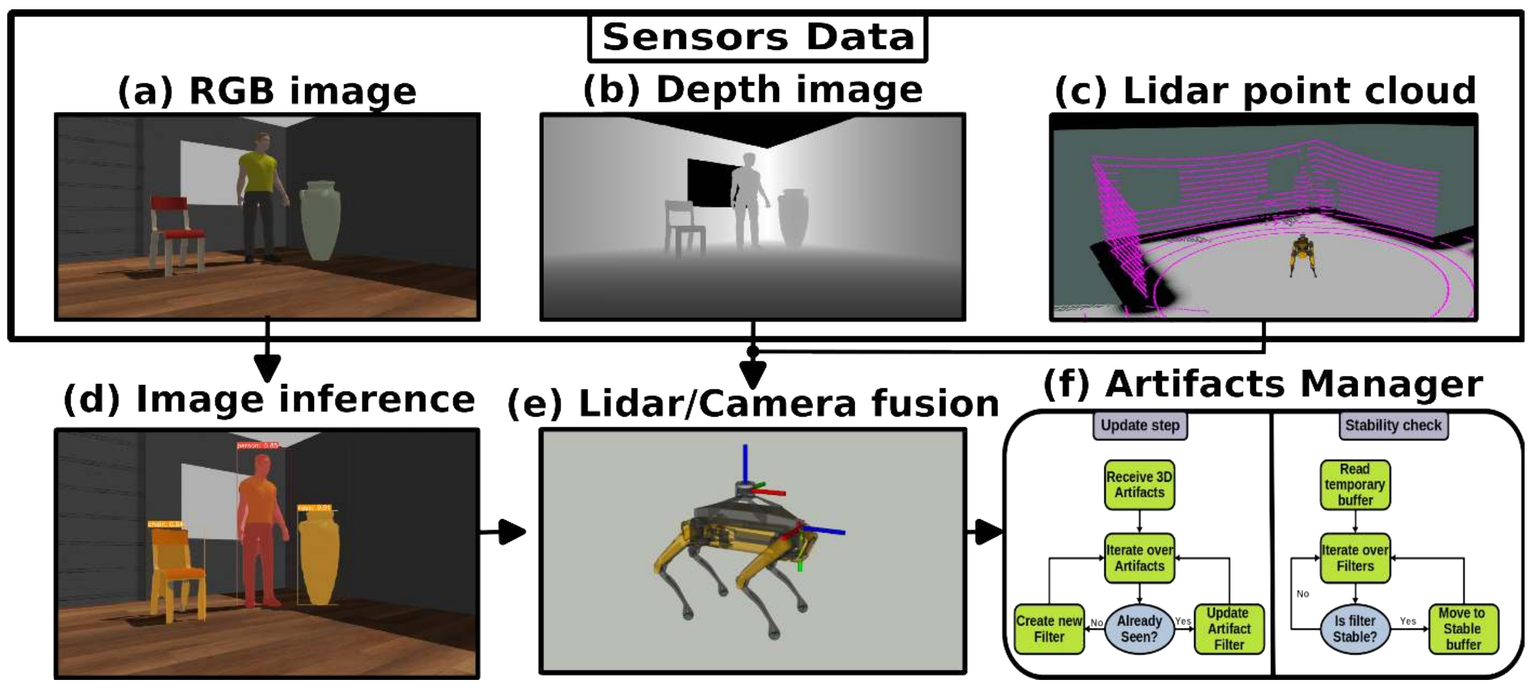
Framework Pipeline

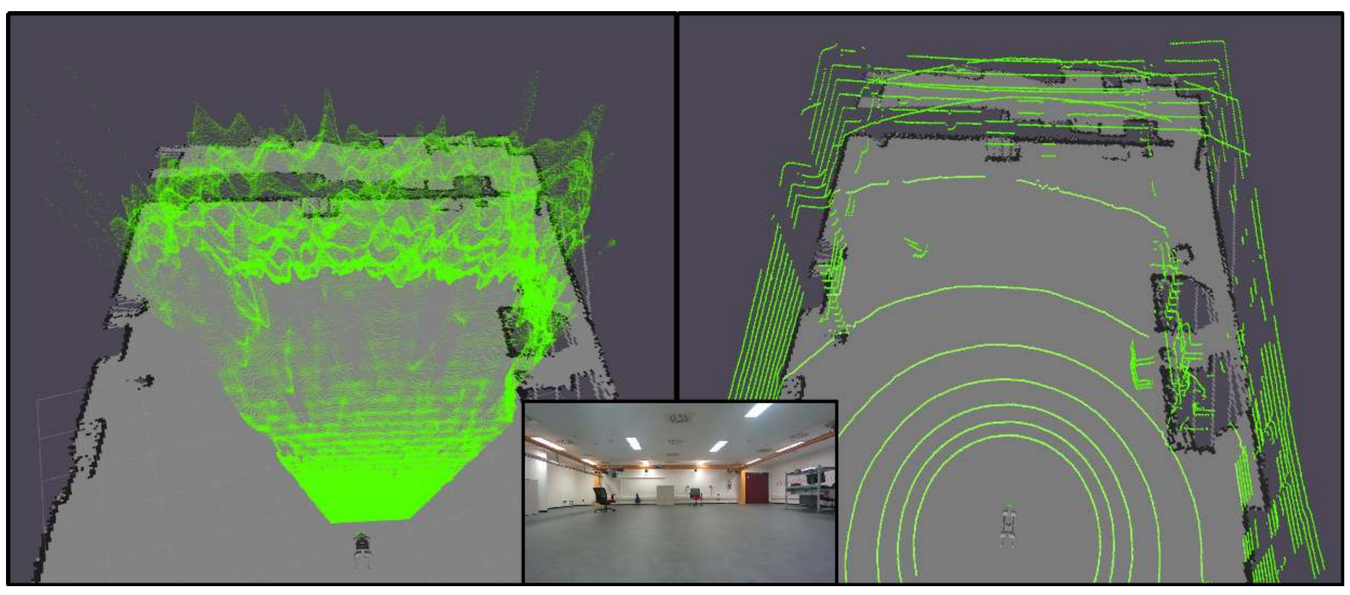
With an Instance segmentation algorithm, Fig. (d), we can extract semantic information from a 2D image which will be then used to perceive and filter the 3D position of the objects with respect to the robot, Fig. (e). Camera and lidar pose estimation are filtered using the formula:

$$X = \begin{cases} 0 & If\ dist_C < min_C \\ X_C & If\ min_C \leq dist_C \leq acc_C \\ \xi X_C + (1-\xi)X_L & If\ acc_C \leq dist_C \leq max_C \\ X_L & If\ dist_C > max_C \end{cases}$$

$$\text{Where } \xi = \frac{1}{max_C - acc_C}(dist_C\text{-}acc_C) + 1$$

The Multi-Modal sensor fusion gives as output the 3D object centroid which is then passed to the artifacts manager, Fig. (f), to perform data association, stabilize detection and filter out noises.





Motivation

Recent works focus only on RGB-D cameras Semantic Mapping techniques which are limiting for wide spaces or outdoor environments. The Lidar-Camera sensor allows the robot to accurately detect both near and far obstacles, which would have been noisy or imprecise in a purely visual or laser-based approach. As can be noted from the image on the left, Cameras and lidars are complementary sensors. RGB-D cameras extract high texture information and perceive precise depth only at low distances while lidars have sparser data but can reach 150m maintaining good accuracy.

UNIVERSITÀ DI TRENTO — **Prof. Marco Roveri**

Leonardo Labs — **Dr. Enrico Mingo Hoffman**

ISTITUTO ITALIANO DI TECNOLOGIA — **Dr. Arash Ajoudani**, **Dr. Nikolaos Tsagarakis**

Contacts
Federico Rollo
federico.rollo.ext@leonardo.com
federico.rollo@unitn.it