AEROSPACE ENGINEERING

Leonardo S.p.A. - Unified multi-modal feature space for autonomous aircraft: localization and scene understanding

Funded By	LEONARDO S.p.A. (Roma) [P.iva/CF:00881841001]
Supervisor	CAPELLO ELISA - elisa.capello@polito.it
Contact	PRIMATESTA STEFANO - stefano.primatesta@polito.it
Context of the research activity	This PhD project aims to advance 3D reconstruction and scene understanding for autonomous aircraft, focusing on generalizable multi-modal approaches for challenging environments like wildfires and GNSS-denied scenarios.
Objectives	This PhD project aims to advance 3D reconstruction and scene understanding for autonomous aircraft, focusing on generalizable multi-modal approaches for challenging environments like wildfires and GNSS-denied scenarios. The research is motivated by several key trends and challenges: 1. Limitations of current approaches: Conventional methods often rely on heavy sensors like LiDAR. While lightweight imaging sensors are gaining interest, vision-based techniques struggle with fast motions, textureless environments, and insufficient view coverage. These limitations are particularly emphasized in aerial scenarios, where the field of view and sensor performance are constrained by the aircraft altitude and speed, as well as by aircraft dynamics. Performance degrades in challenging conditions like smoke or wildfires. 2. Advancements in data-driven approaches: Recent works have shown impressive performance in 3D reconstruction by learning generalizable priors across diverse scenarios. Leveraging these approaches for localization and navigation in GNSS-denied environments is still and open challenge. 3. Insights from depth estimation research: Small-quantity, high-quality synthetic datasets can suffice for training generalizable 3D reconstruction models, with fine-tuning on real-world data for robustness. 4. Multi-modal approaches: Robust aerial autonomy requires extending vision-based methods to multiple modalities like thermal and RADAR imaging. However, the efficacy of multi-modal features for geometric understanding remains limited and unexplored. Moreover, the use of multi-modal data needs to be investigated to enable a precise 6 DOF pose

	estimation. The project will employ a unified transformer encoder-decoder backbone to ensure scalability across modalities. Each modality will be represented by discretized tokens, with modality-specific information encoded in the token embedding. This approach alleviates the need for synchronized multi-modal data across all modalities, which is often challenging to collect, particularly for fine-tuning on real-world data. The research will focus on training a discriminative model where pose and consistency-driven supervision will be the key binding factor across different modalities. Multi-modal latent feature maps will be projected into a canonical global 3D volume and used as descriptors for pose estimation. By using differentiable pose estimation techniques to align latent features across modalities, the model is expected to learn robust multi-modal priors suitable for downstream tasks like 3D reconstruction and localization. Following recent trends in vision-based geometric perception research, the project will first pre-train a large teacher model on high-quality multi-modal synthetic data. Then, a student model will be trained on real-world data using pseudo-labels from the teacher model and noisy labels from real-world 3D data collection pipelines. This approach leverages the benefits of synthetic data for precise supervision while ensuring robustness through real-world fine-tuning.

Skills and competencies for the development of the activity	Basic knowledge of Visual Inertial Odometry (VIO) algorithm , Gazebo and robotic systems