SS 2025-2026, SMT-ROB, DISAL-MP62 Project Summary

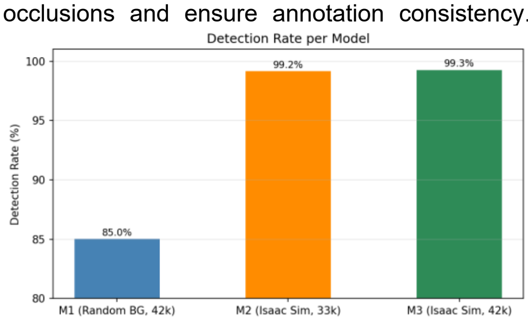
Bridging the Sim-to-Real Gap Using a Photorealistic Environment

Diana Celestina Bejan

Professor : Alcherio Martinoli
Assistant(s) : Yacine Derder

Training vision-based robotics models requires large amounts of labelled data, but collecting and annotating real-world datasets is expensive and time-consuming. Simulation provides a scalable alternative, yet models trained in simulation often perform poorly in reality due to differences in lighting, textures, and sensor conditions (the Sim-to-Real gap). This project investigates whether reconstructing a real robotics laboratory in 3D and using it as a photorealistic simulation background can reduce this gap for 6D drone pose estimation on the Crazyflie 2.1 Brushless nano-quadrotor.

The lab was photographed with a smartphone and reconstructed using the Gaussian Splatting methods 3DGRUT and SuGaR. After evaluating reconstruction quality, 3DGRUT was selected for dataset generation. Synthetic training images were generated in NVIDIA Isaac Sim using the reconstructed environment, with automatic pose annotations exported in BOP format. A custom visibility-based filtering pipeline was developed to handle occlusions and ensure annotation consistency.

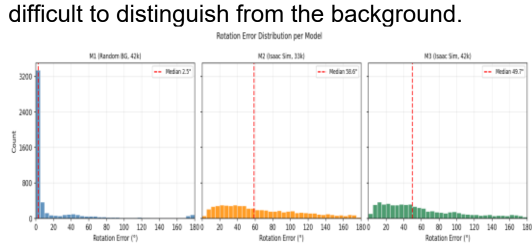


Detection Rate on Synthetic images

Three HCCEPose models were trained. M1 used 42'000 images with random textured backgrounds following the original pipeline, while M2 and M3 used 33'000 and 42'000 images respectively with photorealistic Isaac Sim backgrounds. All models employ a YOLO11 detector followed by a correspondence-based pose regression network.

Evaluation was performed on 6'000 synthetic test images and 21 real images captured in the lab. On synthetic data, the fully trained M1 achieved a median rotation error of 2.46° and translation error of 24 mm. M2 and M3, still training at the time of evaluation, reached detection rates of 99.2% and 99.3%, compared to 85.0% for M1, but had not yet converged in pose accuracy.

On real images, M2 and M3 detected the drone in 71.4% of frames versus 61.9% for M1, demonstrating improved detection generalisation from photorealistic backgrounds. However, M1 achieved the lowest mean translation error (68 mm) due to complete training. Detection failures were concentrated at distances greater than 1.2 m and near-horizontal viewpoints, where the drone's 27 mm profile is difficult to distinguish from the background.



Rotation error distribution by model

The results show that photorealistic simulation backgrounds improve detection generalisation to real scenes, while pose estimation accuracy depends primarily on training convergence. A definitive comparison awaits completion of M2 and M3 training. The full pipeline requires approximately 6–11 days depending on reconstruction complexity. Future work includes completing training, acquiring real-image ground truth with motion capture, and extending the system to real-time closed-loop drone control.