Ricoh Project 2: Robust Object Detection from Physically Generated Video Using Foundation Models

Background & Context

Current perception relies on clustering point clouds using DBSCAN, which lacks semantic understanding and robustness. By leveraging simulation and NVIDIA's Cosmos world foundation models, we aim to train a vision model that reliably identifies robot arms, paper stacks, conveyors, and people.

Goals and Approach

Goal: Develop and deploy an object detector that can identify relevant elements in the robot's environment in real-time using RGB-D input from a fixed camera.

Approach:

- Use Isaac Sim and OpenUSD to create labeled RGB-D video datasets.
- Use Cosmos to diversify the dataset.
- Train modern detectors like YOLO and DETR.
- Evaluate on real-world data and optimize for real-time inference.

Tools & Resources

- Isaac Sim Replicator for synthetic dataset generation
- OpenUSD for dynamic scene scripting
- Cosmos world foundation models
- YOLOv8, DETR, Mask R-CNN, and segmentation pipelines
- Real-time inference on AGX Orin or RTX 4000

Expected Outcomes

- Accurate object detector trained on simulated data
- Dataset and training scripts
- Real-time detector deployment on physical robot
- Documentation for future use and retraining

References

- NVIDIA Cosmos
- Isaac Sim + Replicator
- YOLOv5/YOLOv8
- DETR
- Segment Anything Model (SAM)
- OpenUSD Overview