Real-time 6 DOF Pose Estimation with Limited Priors

Student: Dhruv Sheth

Collaborators: Aurelio Noca, Jonathan Becktor

Co Mentor: Dr. Ersin Das

Mentor: Prof. Joel Burdick

Caltech

Agenda

- Background and Motivation
- Objectives
- Methods
  - Simulation
  - Perception
- Results and Discussion
- Future Work
- Acknowledgements
- Q&A

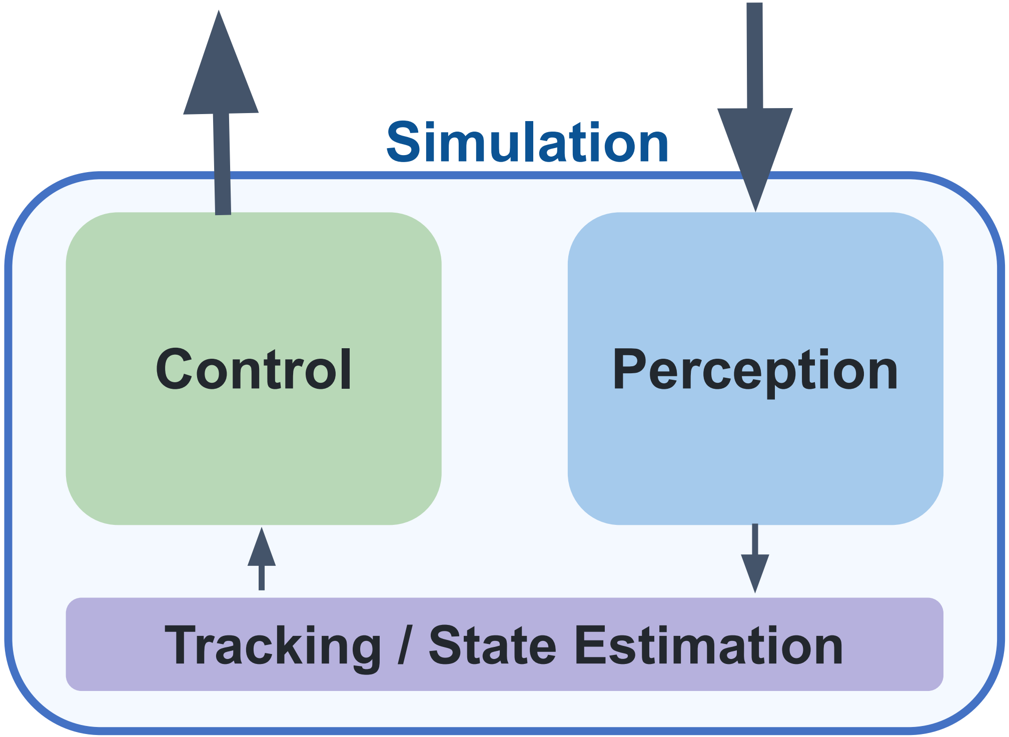
Background

tldr; this is a complex problem

DARPA LINC Phase 1

- Precise payload placement on ships
- Respond to events not predicted at design time
- Robustness to:
  - uncertain state estimates
  - perception error in dynamic environments
- Design ML models with **safety guarantees**

from Burdick Lab slack



Simulation

Control

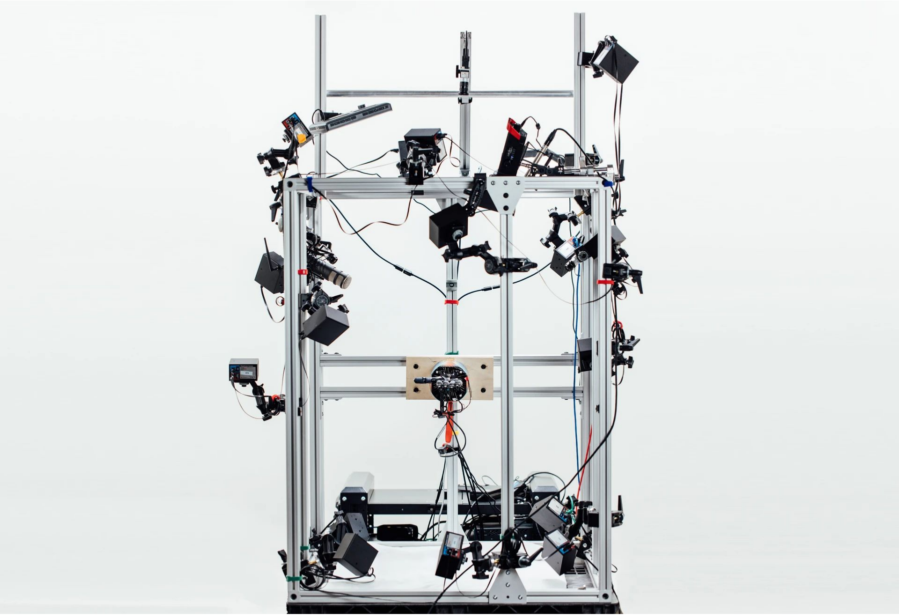
Perception

Tracking / State Estimation

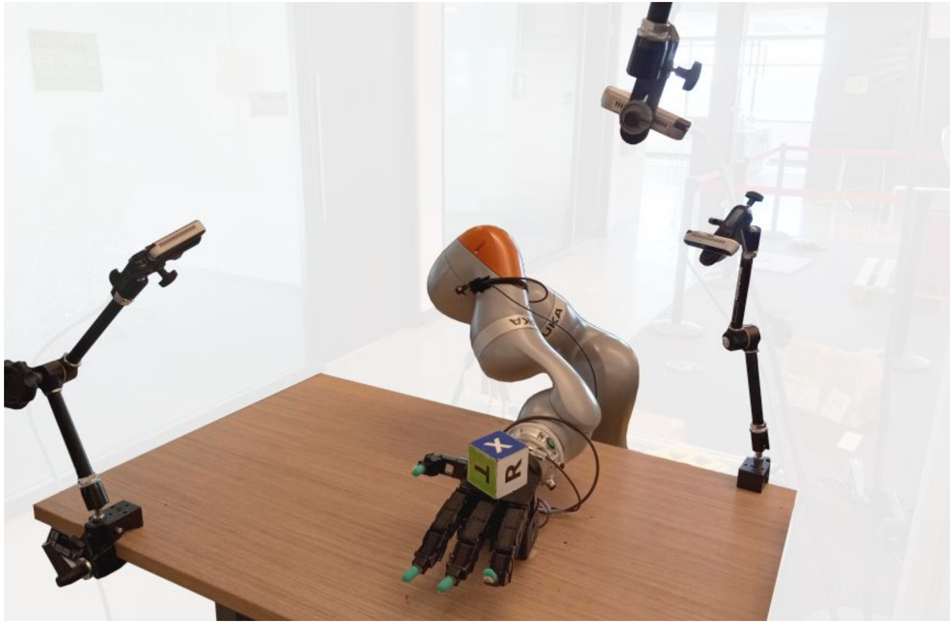
Project Objectives

- Implement a robust 6 DoF pose estimation method for objects with unknown geometry
  - Robust: Make it recover from errors in complete occlusion
  - Optimize: Make it run real time on AGX Orin with low compute
- Develop a high-fidelity simulation stack for testing control+perception
- Extend the algorithm to multiple cameras
- Create a ROS2 wrapper for plug and play support

Other Applications



Dactyl lab setup with Shadow Dexterous Hand, PhaseSpace motion tracking cameras, and Basler RGB cameras
OpenAI Dactyl Experiment, 2018



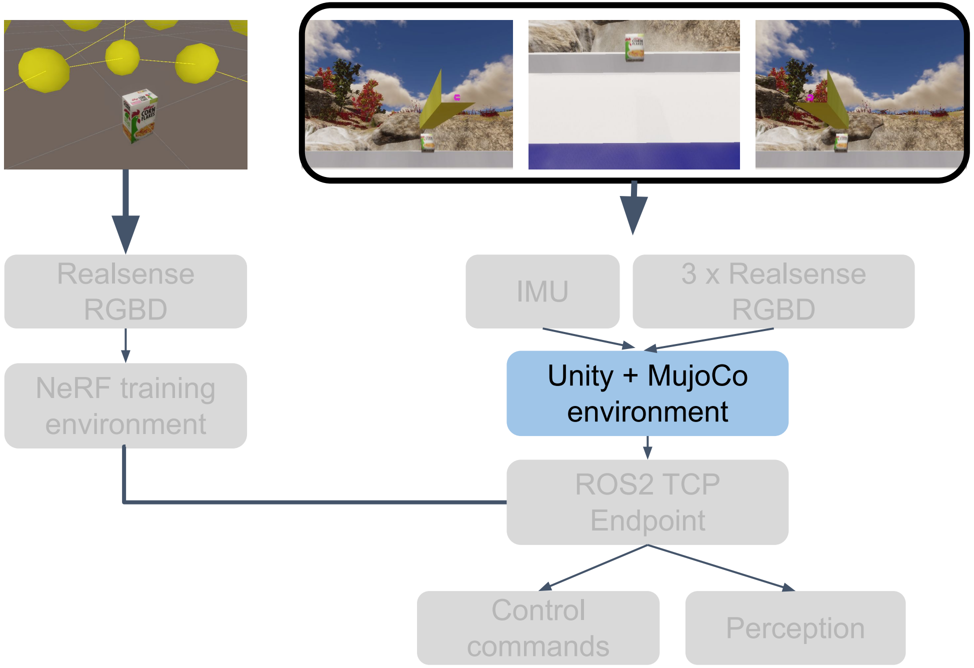
DeXtreme: Transfer of Agile In-Hand Manipulation from Simulation to Reality, 2024
OpenAI hand replication experiment

Caltech

Methods

Simulation Stack

- High Fidelity environment replicating testbed
- Unity + MuJoCo integration
    - MuJoCo physics: control
    - Unity photorealism: perception



Realsense RGBD

NeRF training environment

IMU

3 x Realsense RGBD

Unity + MujoCo environment

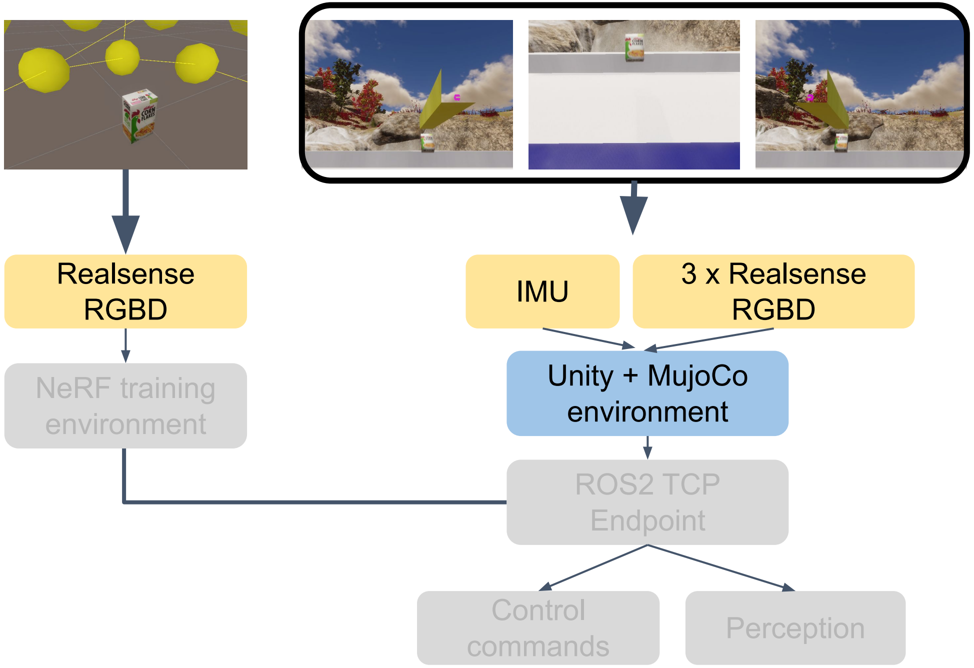
ROS2 TCP Endpoint

Control commands

Perception

Simulation Stack

- High Fidelity environment replicating testbed
- Unity + MuJoCo integration
  - MuJoCo physics: control
  - Unity photorealism: perception
- RGBD + IMU sensor data integration for testing



Realsense RGBD

NeRF training environment

IMU

3 x Realsense RGBD

Unity + MujoCo environment

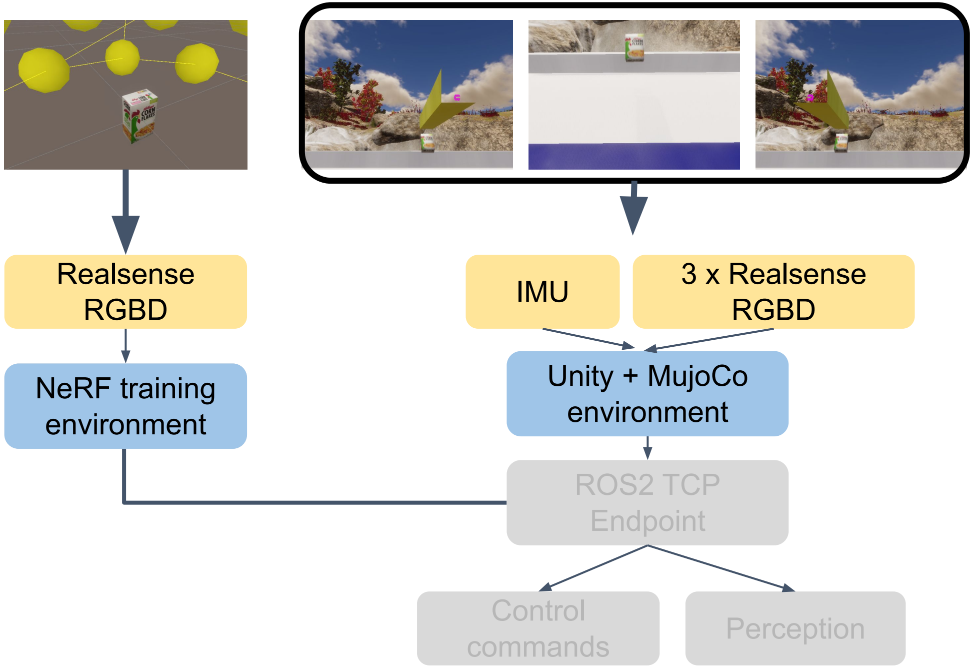
ROS2 TCP Endpoint

Control commands

Perception

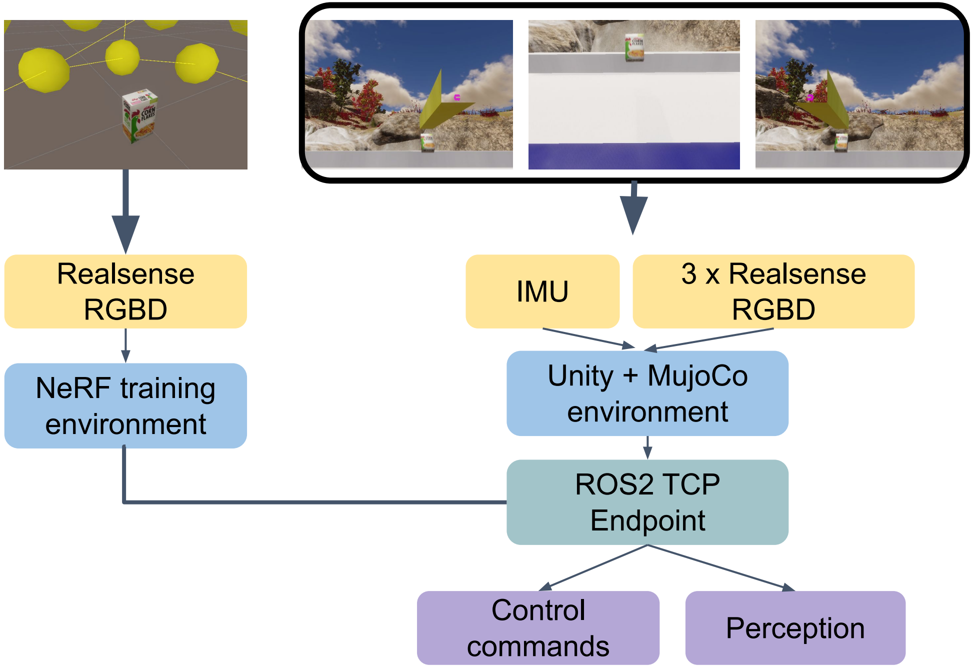
Simulation Stack

- High Fidelity environment replicating testbed
- Unity + MuJoCo integration
  - MuJoCo physics: control
  - Unity photorealism: perception
- RGBD + IMU sensor data integration for testing
- NeRF training environment



Simulation Stack

- High Fidelity environment replicating testbed
- Unity + MuJoCo integration
  - MuJoCo physics: control
  - Unity photorealism: perception
- RGBD + IMU sensor data integration for testing
- NeRF training environment
- ROS2 wrapper
  - communication with sim
  - swapping with hardware



Realsense RGBD

NeRF training environment

IMU

3 x Realsense RGBD

Unity + MujoCo environment

ROS2 TCP Endpoint

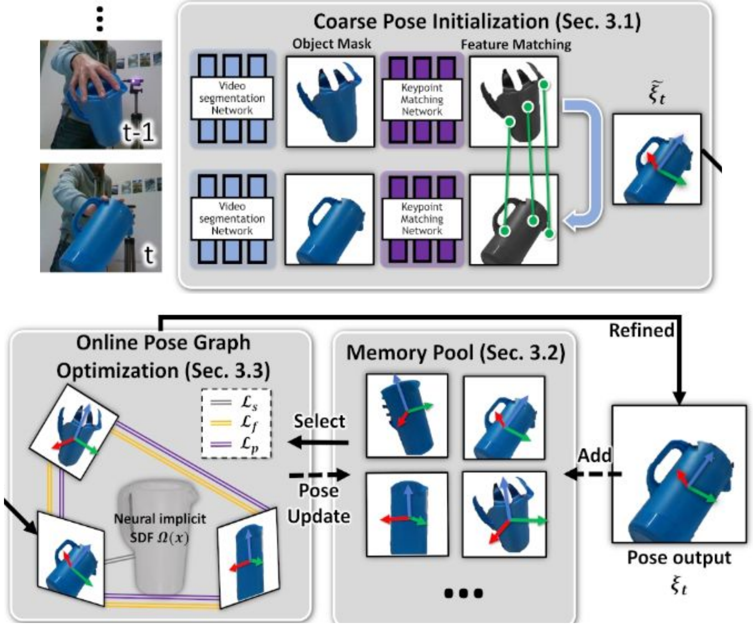
Control commands

Perception

Perception Stack

Previous Methods

- Feature Tracking approaches
  - Classical Approaches:
    - Inverse SLAM
    - LoFTR feature tracking
  - Deep Learning:
    - BundleSDF
- Transformer approaches:
  - FoundationPose
    - Not suited for longer videos

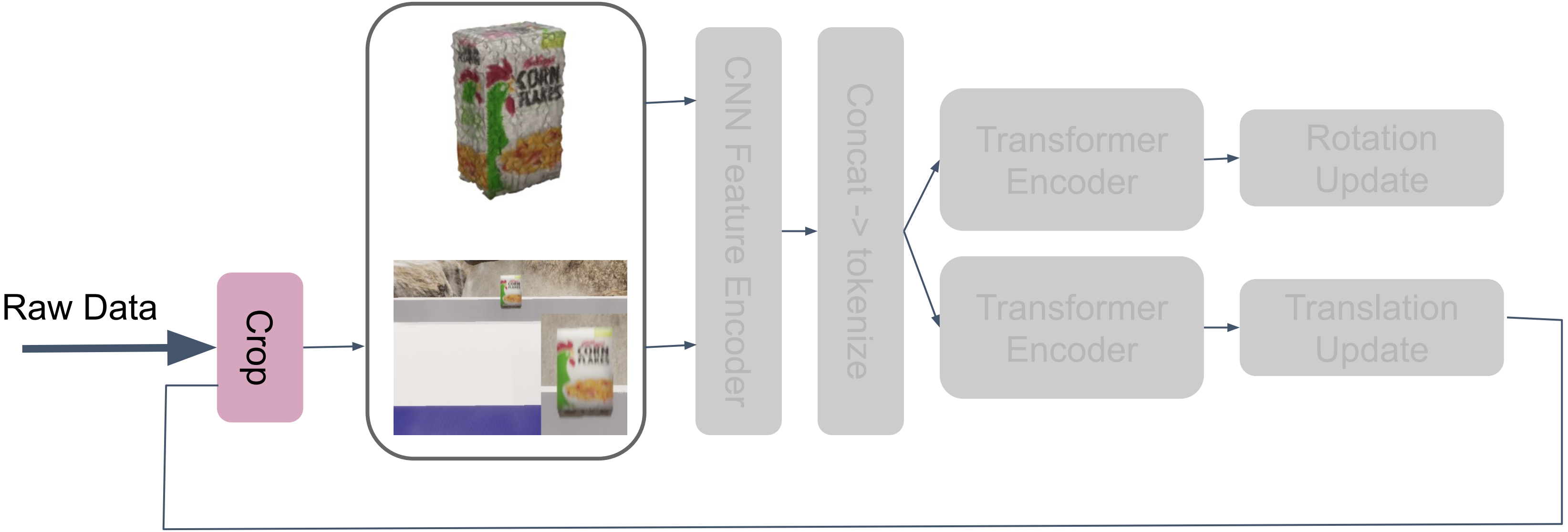

t-1
t
Coarse Pose Initialization (Sec. 3.1)
Object Mask
Feature Matching
Video segmentation Network
Keypoint Matching Network
Video segmentation Network
Keypoint Matching Network
$\tilde{\xi}_t$
Refined
Online Pose Graph Optimization (Sec. 3.3)
$\mathcal{L}_s$
$\mathcal{L}_f$
$\mathcal{L}_p$
Neural implicit SDF $\Omega(x)$
Select
Pose Update
Memory Pool (Sec. 3.2)
Add
Pose output
ξ_t

BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects
NVIDIA, 2023

Caltech

Perception Stack

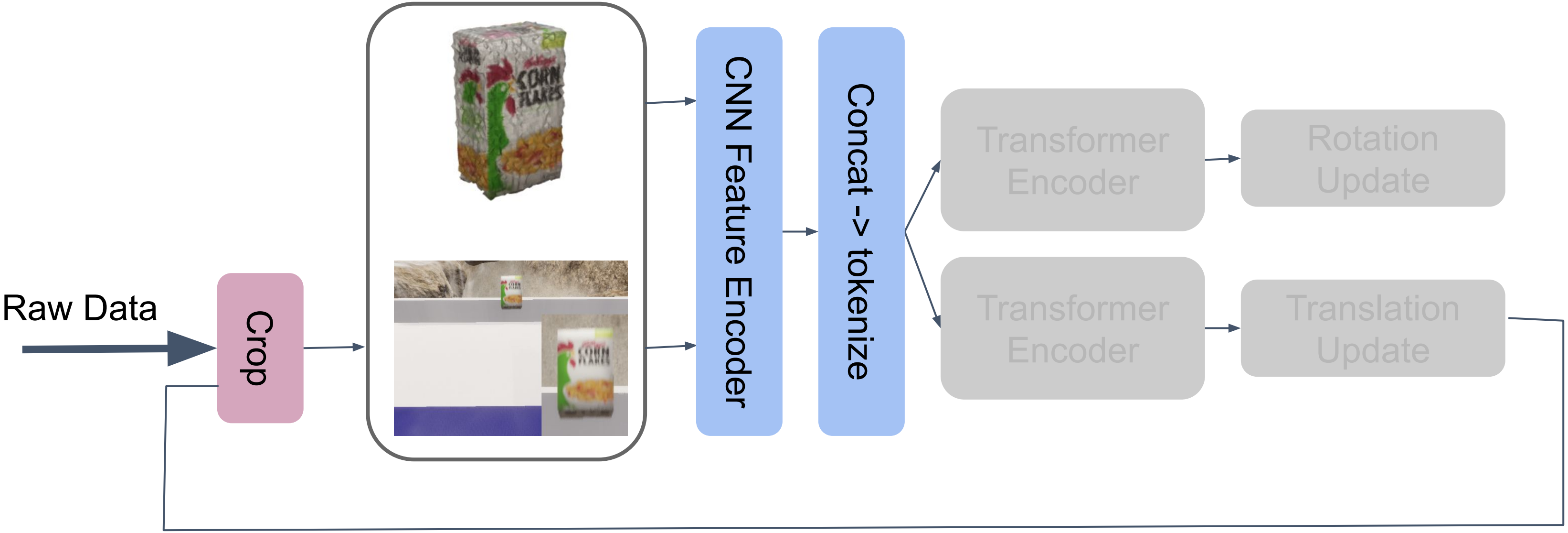
FoundationPose



FoundationPose Architecture

Perception Stack

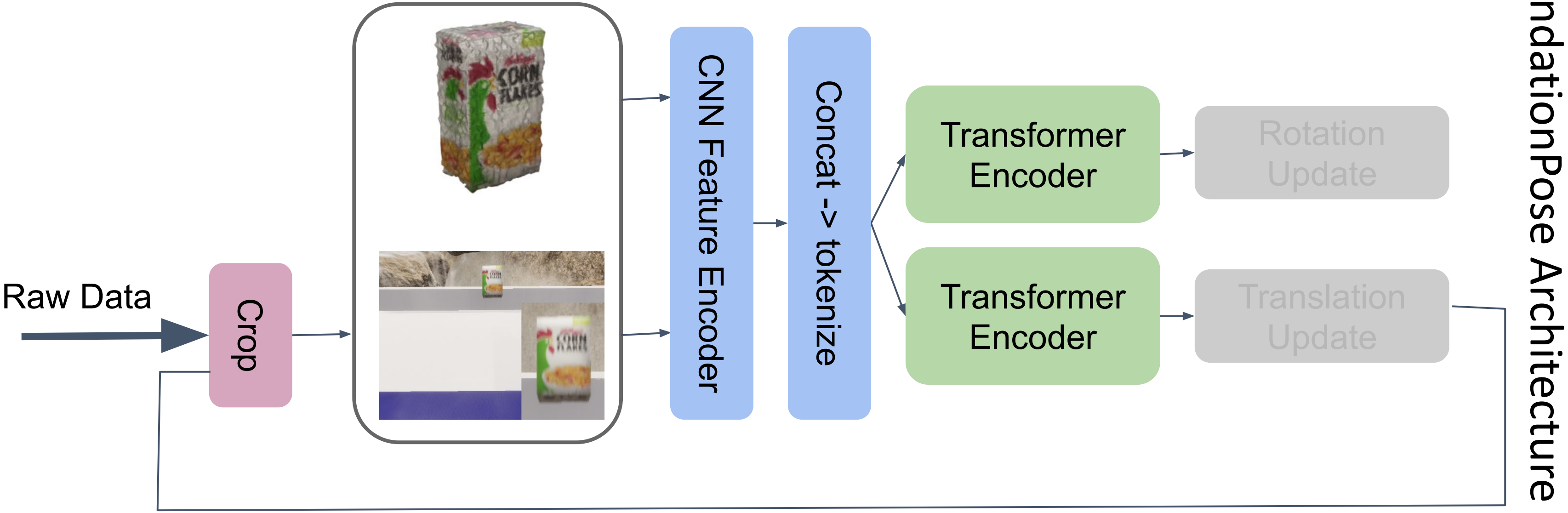
FoundationPose



FoundationPose Architecture

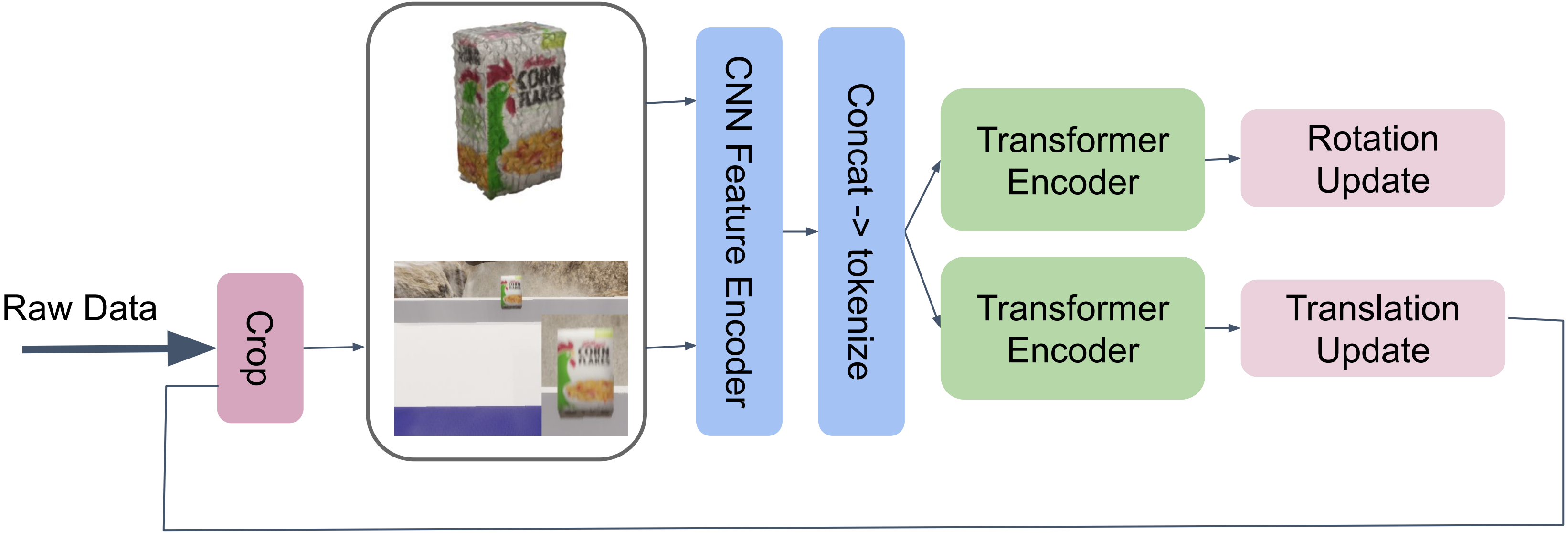
Perception Stack

FoundationPose



Perception Stack

FoundationPose



FoundationPose Architecture

Caltech

Perception Stack

Video Object Segmentation

- Cutie: Putting the Object Back Into Video Object Segmentation
- Why segmentation?
  - 2D data is easier than 3D data
  - Current 2D models outperform 3D
  - Provides a good coarse input

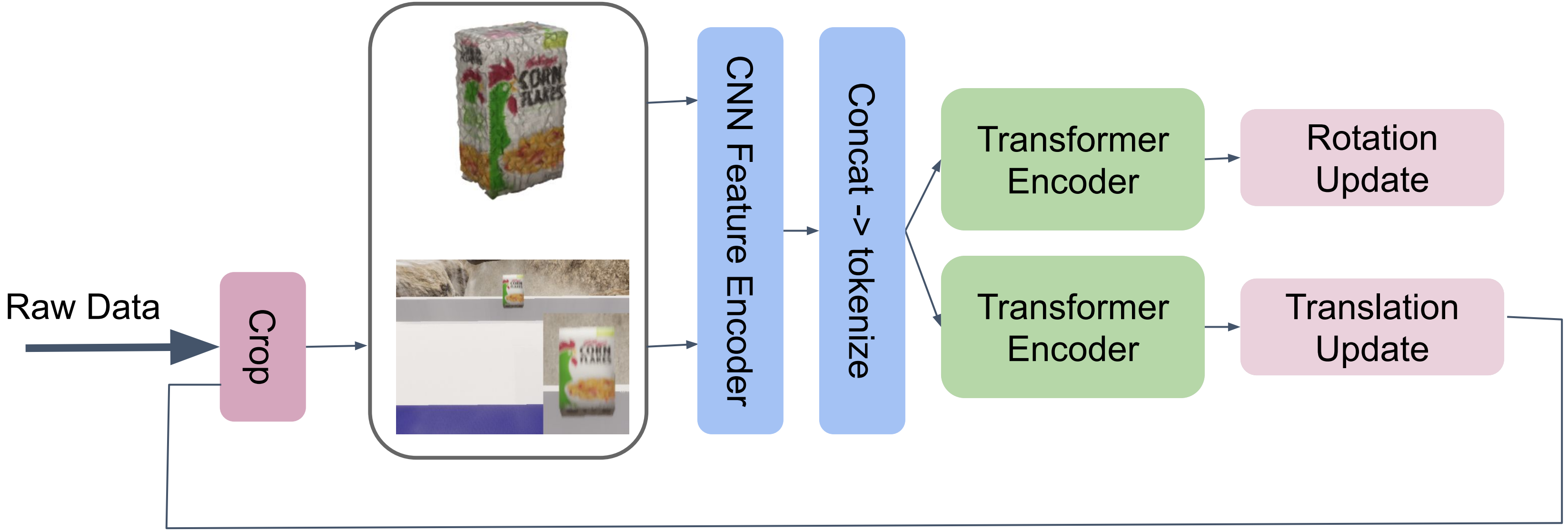
Perception Stack

Video Object Segmentation

- Cutie: Putting the Object Back Into Video Object Segmentation
- Why segmentation?
  - 2D data is easier than 3D data
  - Current 2D models outperform 3D
  - Provides a good coarse input
- Why Cutie?
  - Memory model
  - Long horizon segmentation
  - Recovers from occlusions

Perception Stack

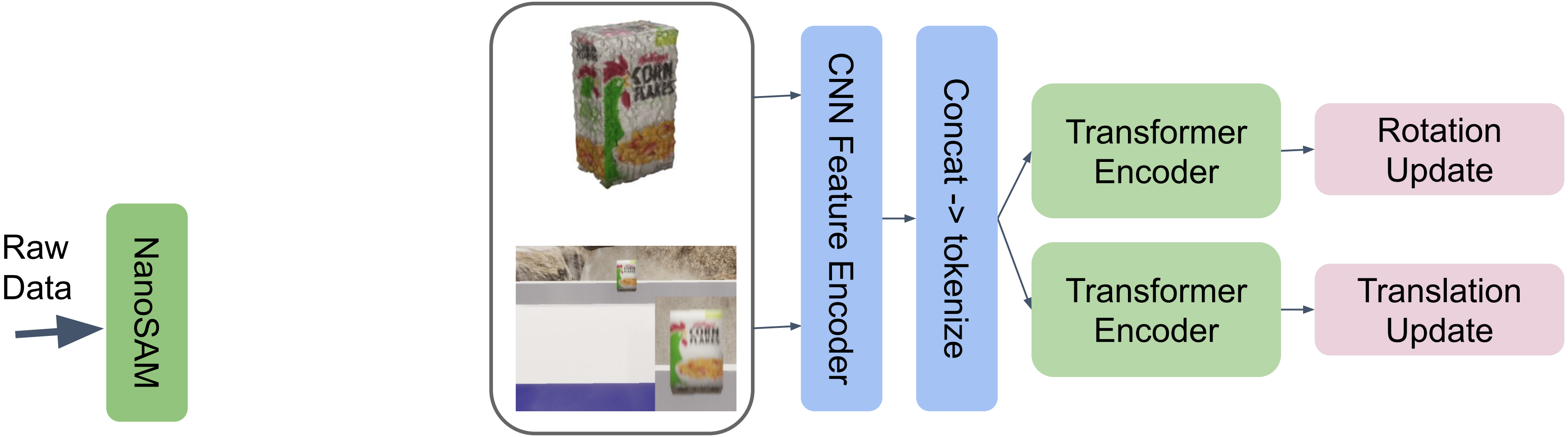
FoundationPose



FoundationPose Architecture

Perception Stack

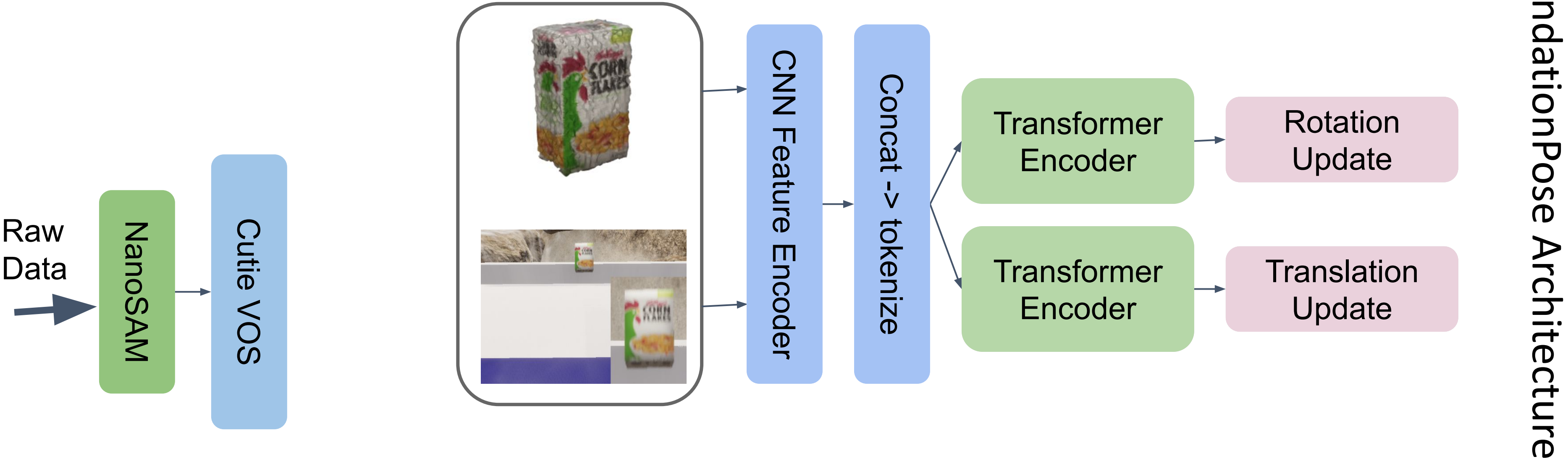
FoundationPose Improvement



FoundationPose Architecture

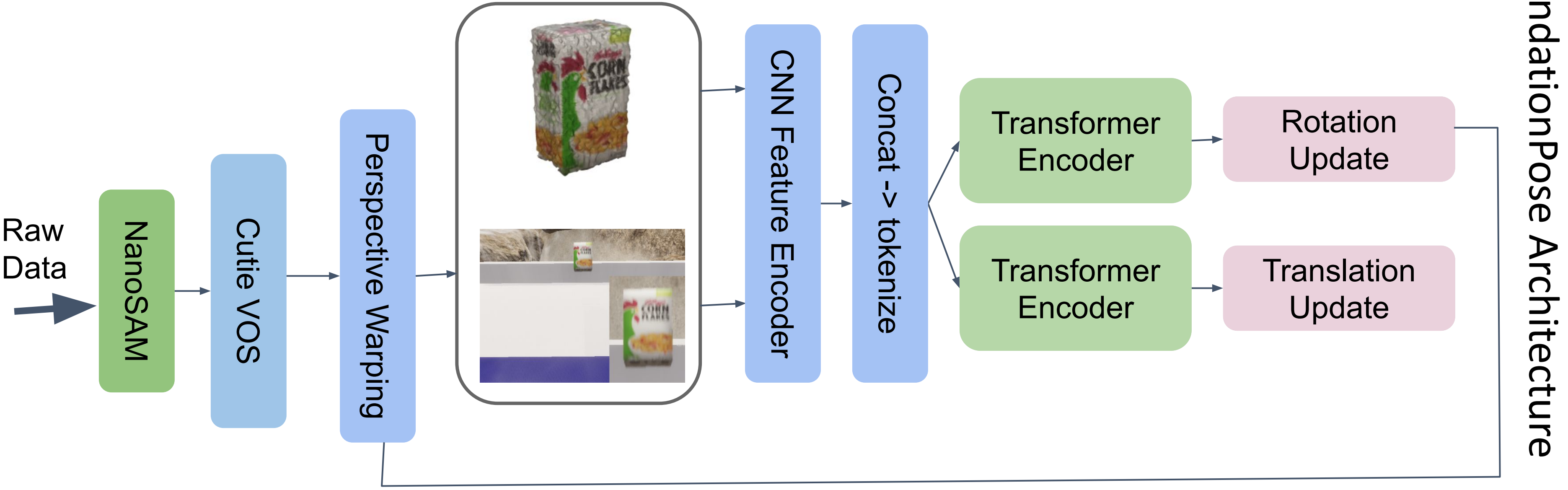
Perception Stack

FoundationPose Improvement



Perception Stack

FoundationPose Improvement

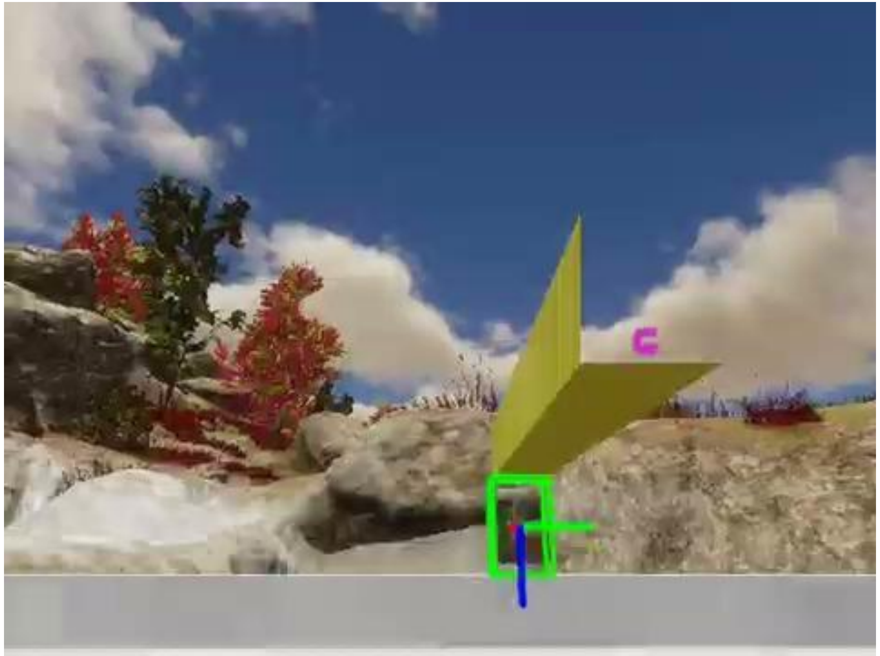


Caltech

Perception Stack

FoundationPose + VOS results

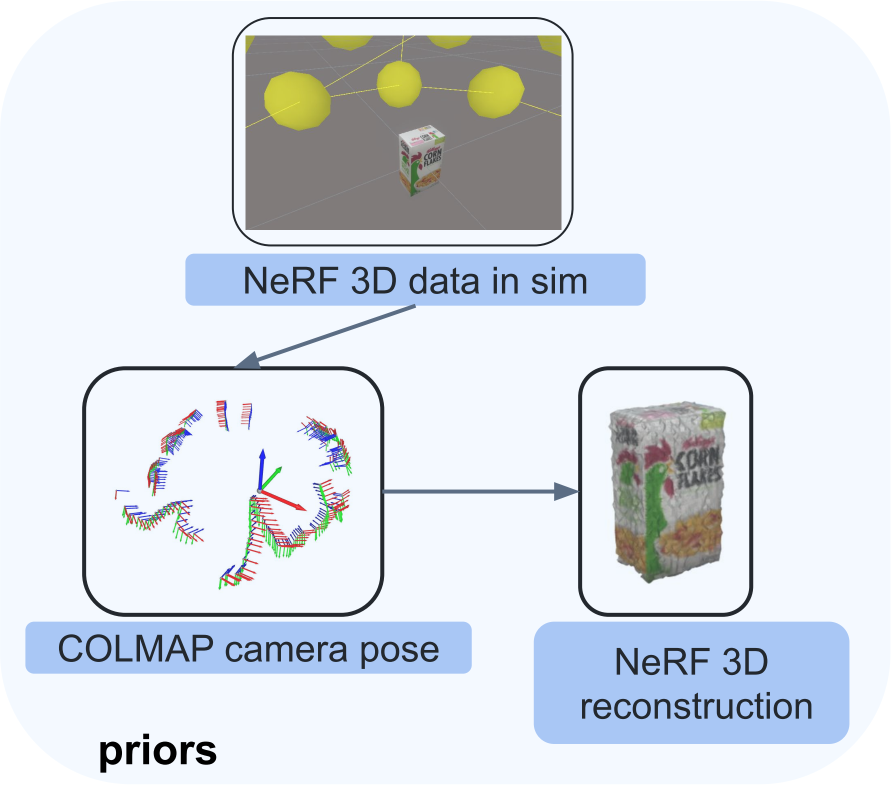
- High Disturbance
- Low visibility of object



Caltech

Perception Stack

FoundationPose pipeline



NeRF 3D data in sim

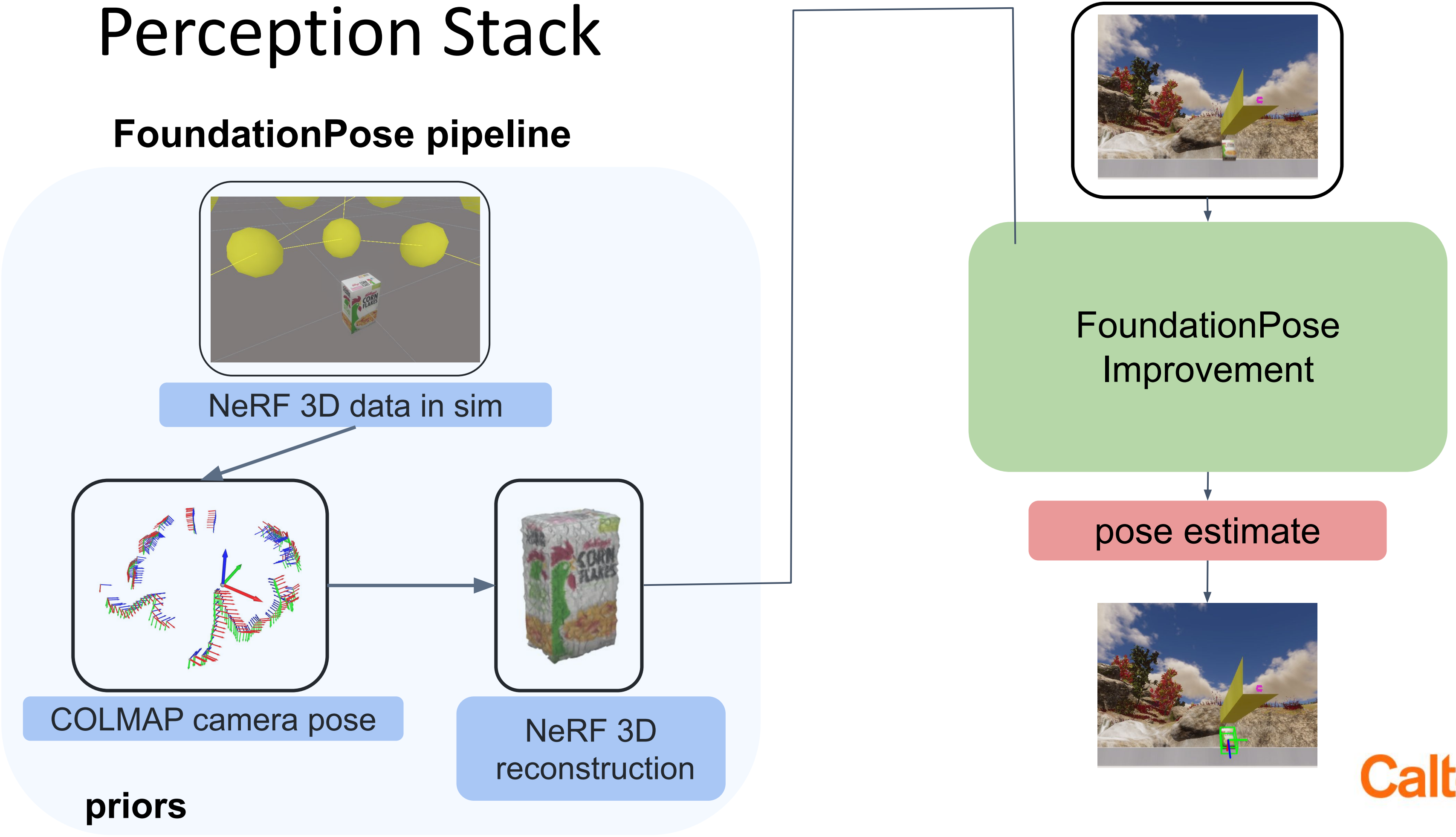
COLMAP camera pose

NeRF 3D reconstruction

priors

Perception Stack

FoundationPose pipeline



NeRF 3D data in sim

COLMAP camera pose

NeRF 3D reconstruction

priors

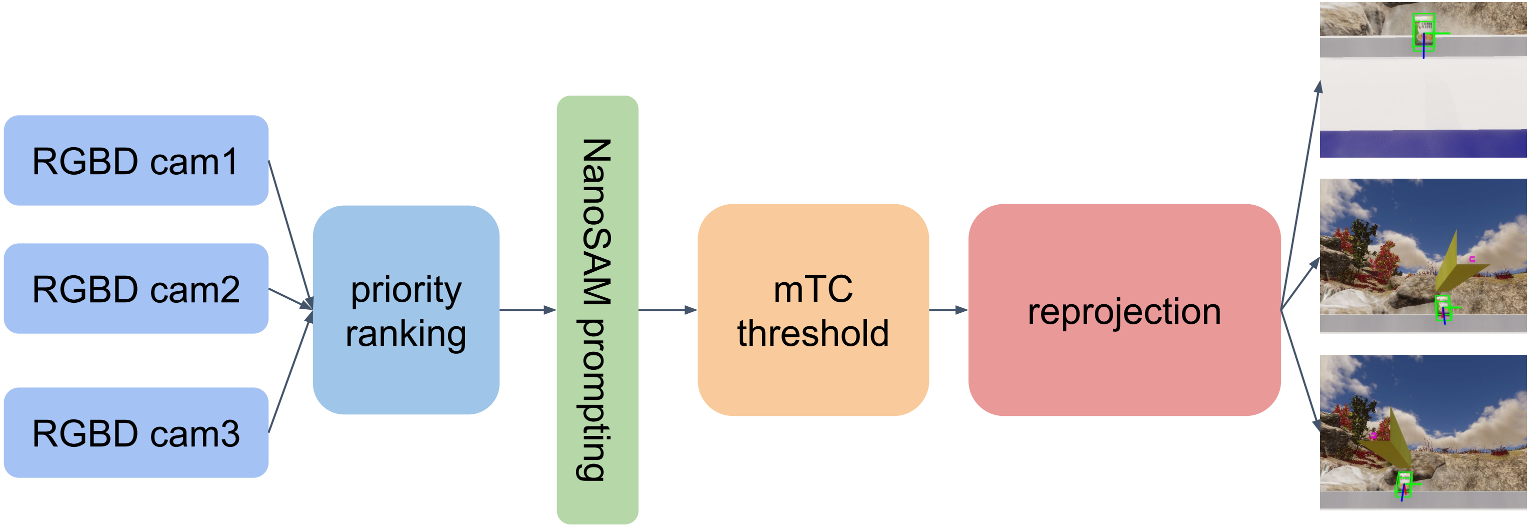
FoundationPose Improvement

pose estimate

Perception Stack

Multi-Cam integration & deployment

- Under complete occlusion in one camera:
    - Switch between cameras with priority ranking
    - mTC to know when object recovered from occlusion



RGBD cam1

RGBD cam2

RGBD cam3

priority ranking

NanoSAM prompting

mTC threshold

reprojection

Perception Stack

Multi-Cam integration & deployment

- Under complete occlusion in one camera:
  - Switch between cameras with priority ranking
  - mTC to know when object recovered from occlusion
- Implement ROS2 interface for:
  - Multicam data
  - Pose tracking
  - Simulation to perception

Perception Stack

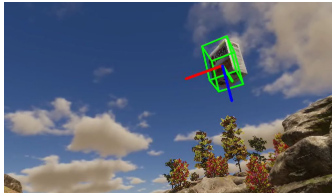
Multi-Cam integration & deployment

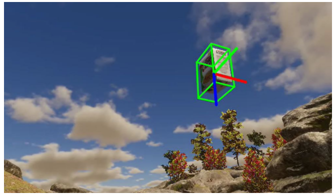
- Under complete occlusion in one camera:
  - Switch between cameras with priority ranking
  - mTC to know when object recovered from occlusion
- Implement ROS2 interface for:
  - Multicam data
  - Pose tracking
  - Simulation to perception
- TensorRT Optimization on VOS
  - 20% improvement
    - 3 modules

Perception Stack

Safety-Critical use

Which pose estimate do we trust more?
How we do know?



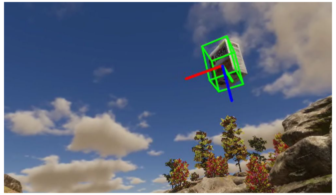


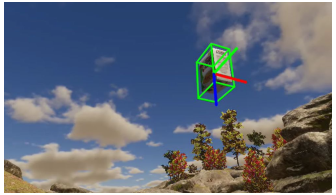
Caltech

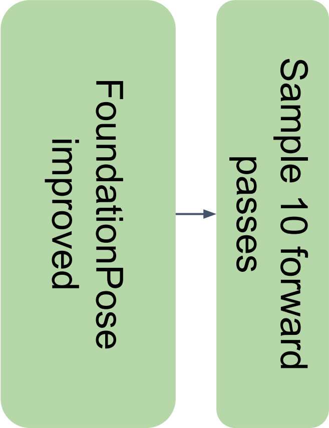
Perception Stack

Safety-Critical use

Which pose estimate do we trust more?
How we do know?







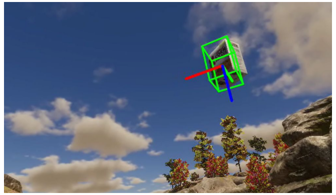
FoundationPose improved

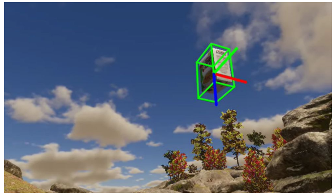
Sample 10 forward passes

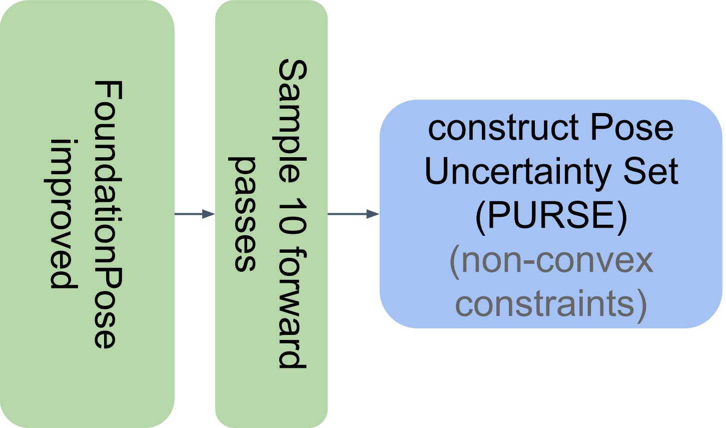
Perception Stack

Safety-Critical use

Which pose estimate do we trust more?
How we do know?







FoundationPose improved

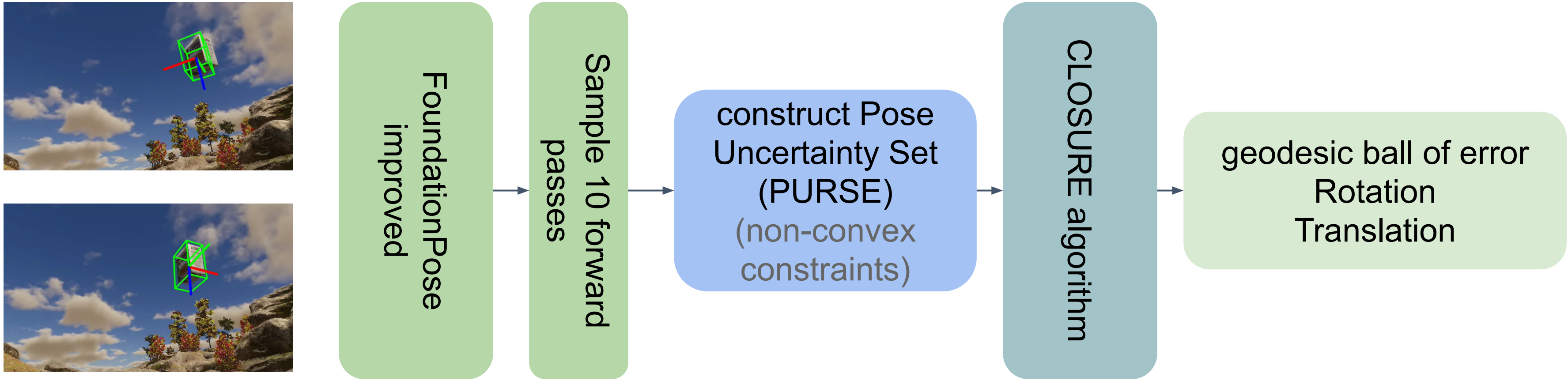
Sample 10 forward passes

construct Pose Uncertainty Set (PURSE)
(non-convex constraints)

Perception Stack

Safety-Critical use

Which pose estimate do we trust more?
How we do know?



FoundationPose improved

Sample 10 forward passes

construct Pose Uncertainty Set (PURSE) (non-convex constraints)

CLOSURE algorithm

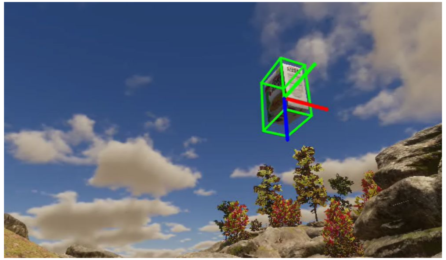
geodesic ball of error
Rotation
Translation

Perception Stack

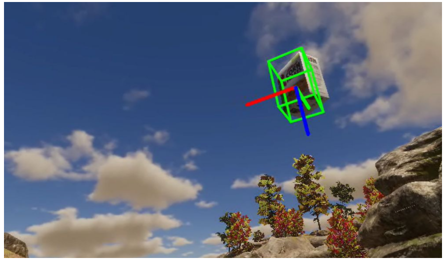
Safety-Critical use

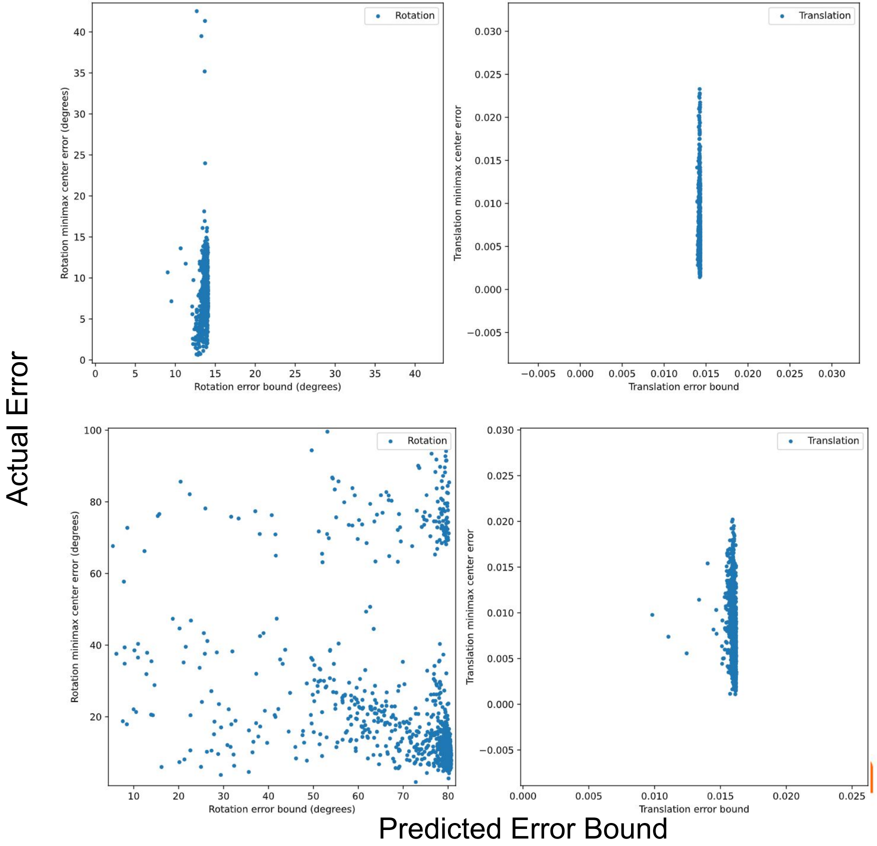
Results from **CLOSURE**

complete NeRF reconstruction



partial NeRF reconstruction





Results and Conclusion

- FoundationPose + VOS pose estimation
- Multi-Cam integration to FoundationPose
- ROS2 interface for perception
- High-fidelity simulation environment for testing
- Optimization for low compute
- Uncertainty Quantification for safety

Future Work

- More Benchmarks
- Use less priors (No 3D reconstruction of object)
- Integrate a state estimator for filtering

Acknowledgements

- Prof. Joel Burdick for giving me the opportunity to SURF with the lab, resources and guidance along the way
- Dr. Ersin Das for his guidance and support and help whenever needed
- Aurelio, Jonathan and Thomas for help with code, hardware and debugging
- Dr. Jane Chen for SURF fellowship to support my research this summer

Q&A

caltech.edu