Introduction to Computer Vision

Lecture 15 - Embodied AI

Prof. He Wang

Logistics

- Final exam
  - Time: 6/19
  - Scope: all the lectures after midterm including today's lecture.
  - Question types: similar to midterm exam.
  - 1-page A4-size cheat sheet is allowed .

Embodied AI

Embodied AI

Embodied = Em + body

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.
>
> ——Alan Turing

Computing Machinery and Intelligence. 1950.

Classic Disembodied AI

Classical:
"intelligence as computation"

Illustration: Shun Iwasawa, from Pfeifer, R: How the body shapes the way we think, 2007

Problems?

What remains?

- AI still heavily biased toward representation and computation.

- vs. natural (also human) intelligence:
  - Embodied
  - emergent from sensory-motor and interaction processes

Embodiment

- “intelligence requires a body”
- Interplay / task distribution
  - Brain
  - Body (morphology – shape, materials, …)
  - Environment
- Principal of ecological balance
  - match in complexity of sensory, motor, and neural system

Human Visual System

- The visual system comprises eyes as sensor organ, which are connected to visual cortex in the brain.

- Visual tasks:
  - sensation
  - processing
  - perception
  - cognition
  - visuomotor coordination
  - and more.

Visual Motor Coordination/Integration

- Visual motor control is the ability to coordinate visual information with motor output, where the eyes provide sensory feedback to adjust body motion.

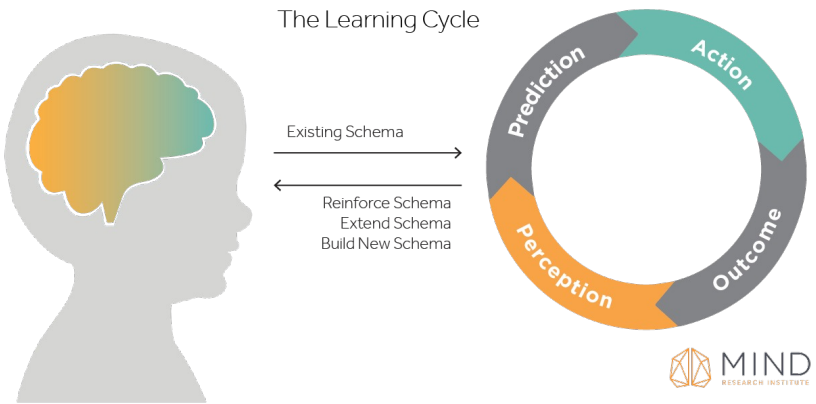
- It is crucial for coordinating the hands, legs, and the rest of the body's movements with what the eyes perceive.

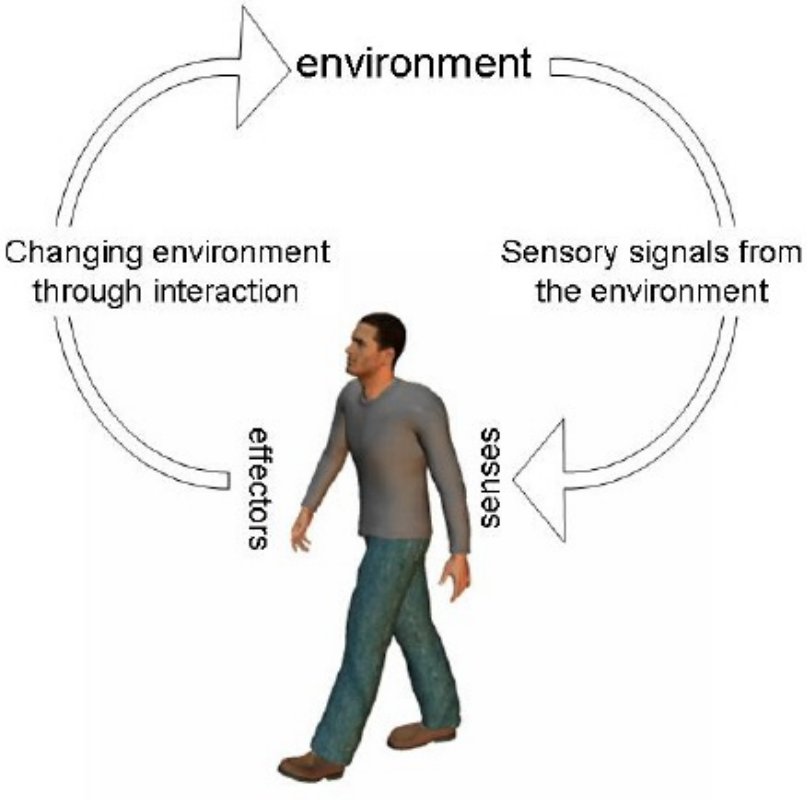
Examples of Visuomotor coordination: Eye-Hand Coordination

More Examples: Running, Balancing, etc.

How Humans Learn: Perception-Action Loop

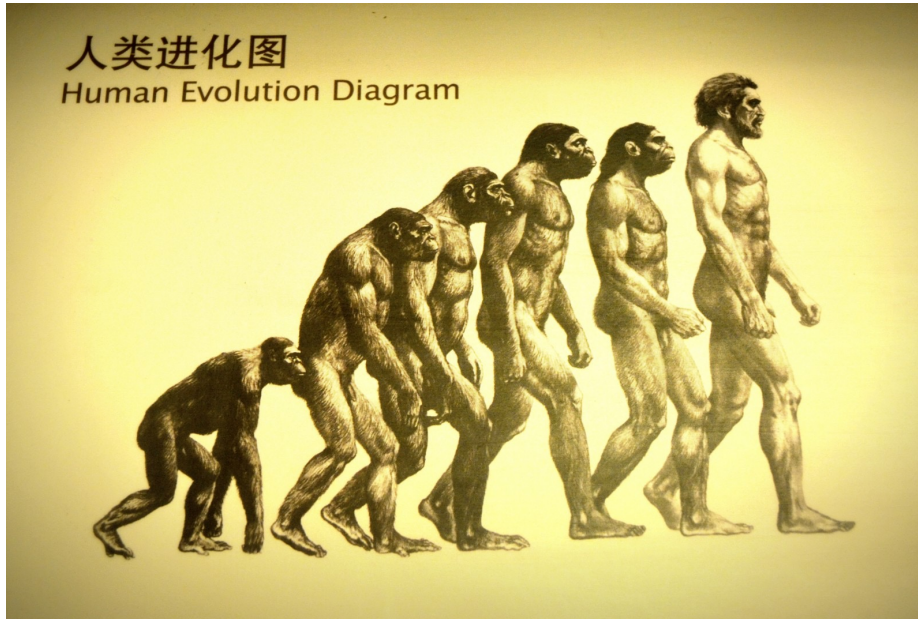
- Perceive, forms hypotheses, and then take action to examine.
- Our brain makes sense of the world around us by creating and testing hypotheses about the way the world works.



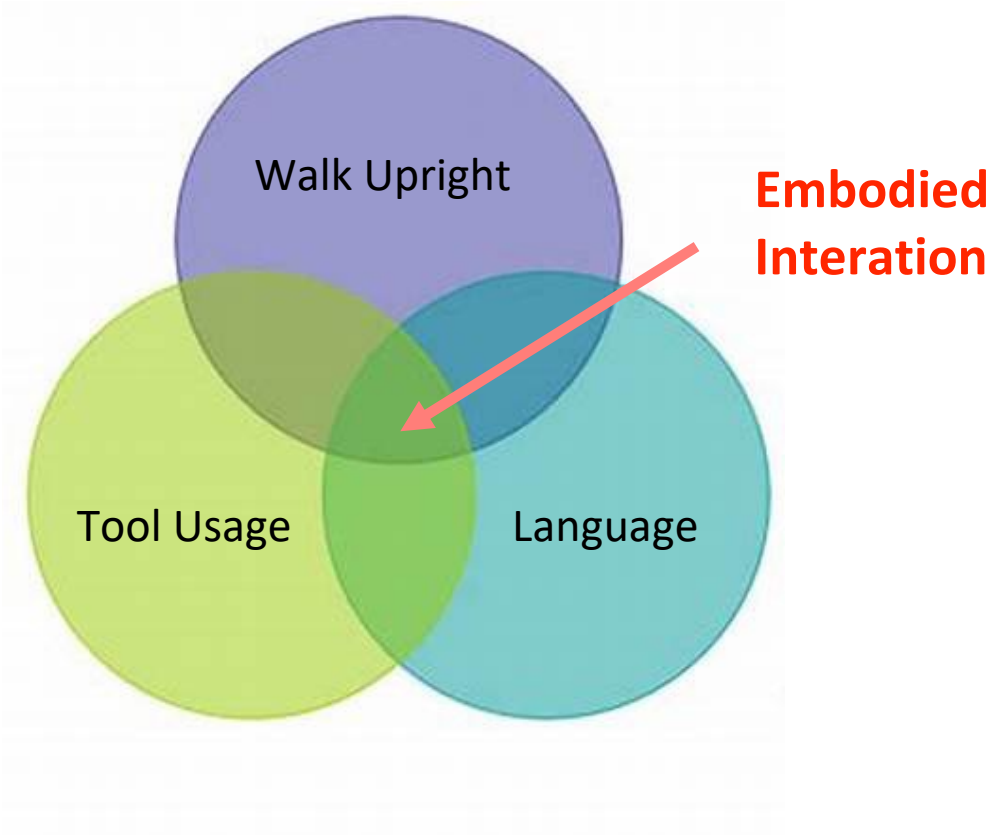


Perception-Action Loop

Evolution of Human Intelligence



The keys to evolution of human intelligence:



Research Questions

- Classical AI
  - Thinking, reasoning, abstract problem solving
- Embodied AI
  - Movement, physical interaction with the real world

"Why do plants not have brains? The answer is actually quite simple: they don't have to move."

Lewis Wolpert, UCL

Morphology Facilitating Perception

Active perception

"We begin not with a sensory stimulus, but with a sensory-motor coordination [...] In a certain sense it is the movement which is primary, and the sensation which is secondary, the movement of the body, head, and eye muscles determining the quality of what is experienced. In other words, the real beginning is with the act of seeing; it is looking, and not a sensation of light." ("The reflex arc in psychology," John Dewey, 1896)

"Since all the stimulations which the organism receives have in turn been possible only by its preceding movements which have culminated in exposing the receptor organ to external influences, one could also say that behavior is the first cause of all the stimulations." ("The structure of Behavior," Maurice Merleau-Ponty, 1963)

"Problems that are ill-posed, nonlinear, or unstable for a passive observer become well-posed, linear, or stable for an active observer." (Ruzena Bajcsy, 1988) (similar points: Aloimonos, 1990; Ballard, 1991)

quotations courtesy Olaf Sporns

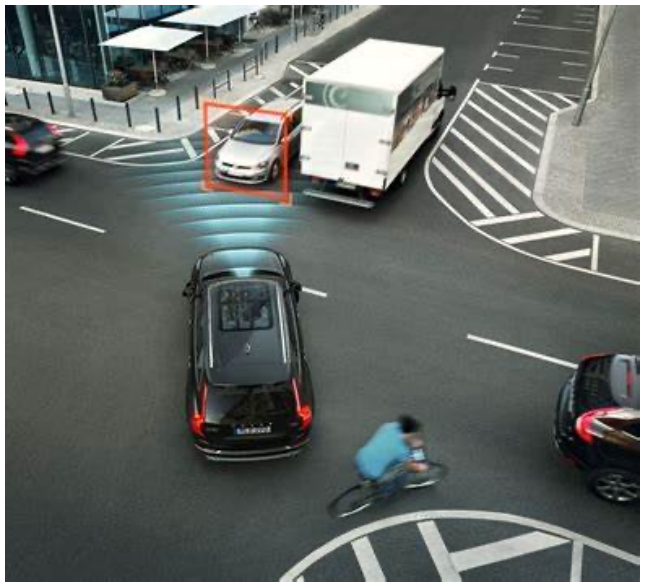
Trends in Computer Vision Field

- Moving to Interaction/Robotic tasks

- Encourage active/interactive perception

Now and Future for Embodied AI

Industrial Robots



Autonomous Driving

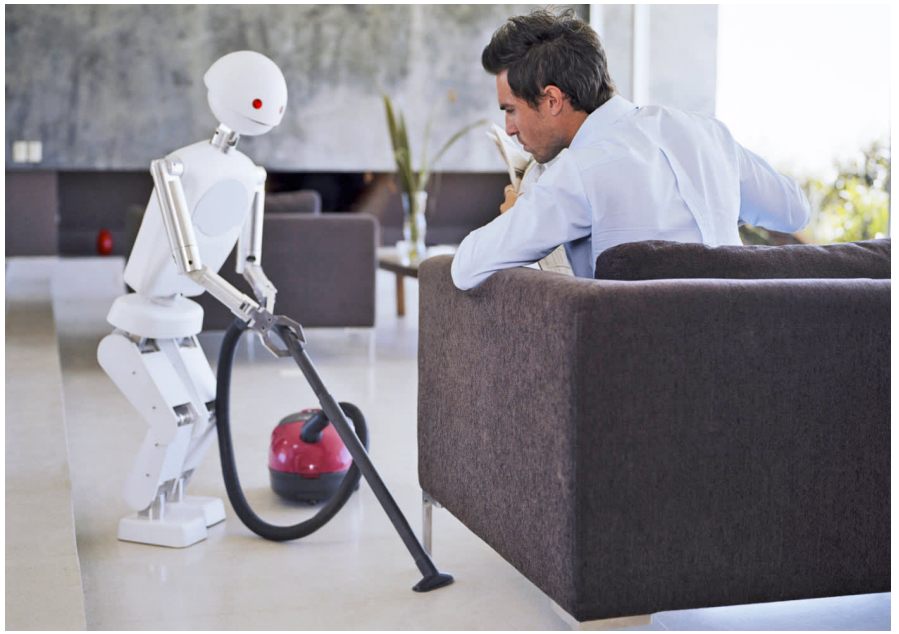
Now

Home Robots

Future

Generalist Robot

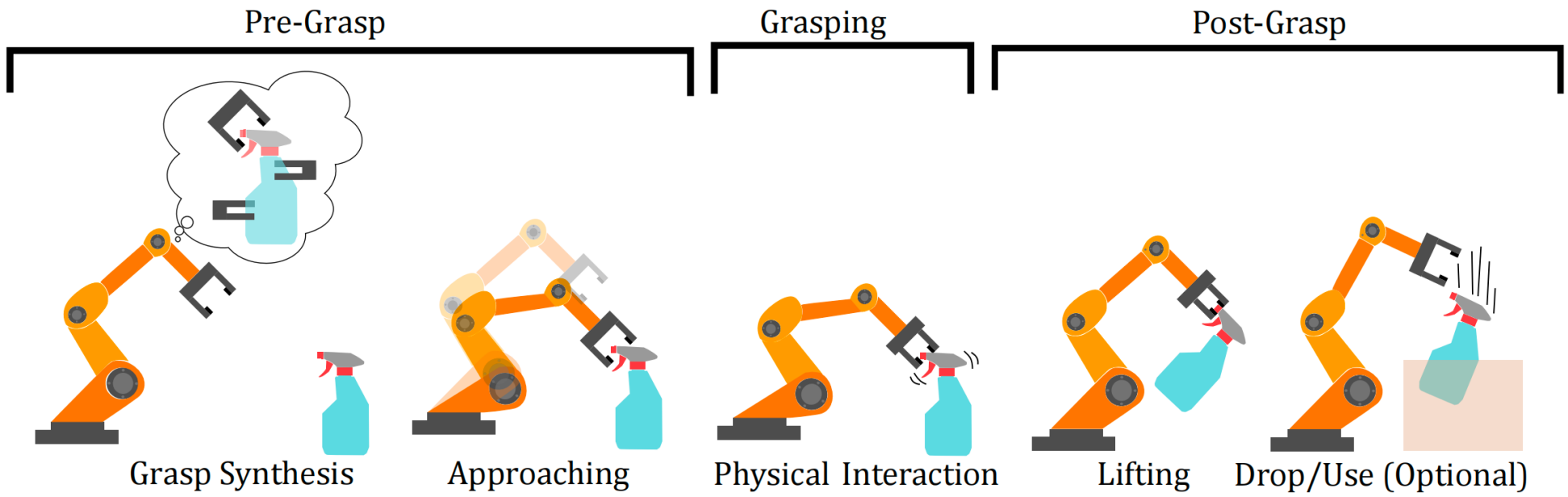
Goal: a generalist robot



Object Grasping

What is Grasp Synthesis?

- **Grasping** is the process of restraining an object's motion in a desired way by applying forces and torques at a set of contacts
- **Grasp Synthesis** is a high-dimensional search or optimization problem to **find gripper poses** or joint configurations

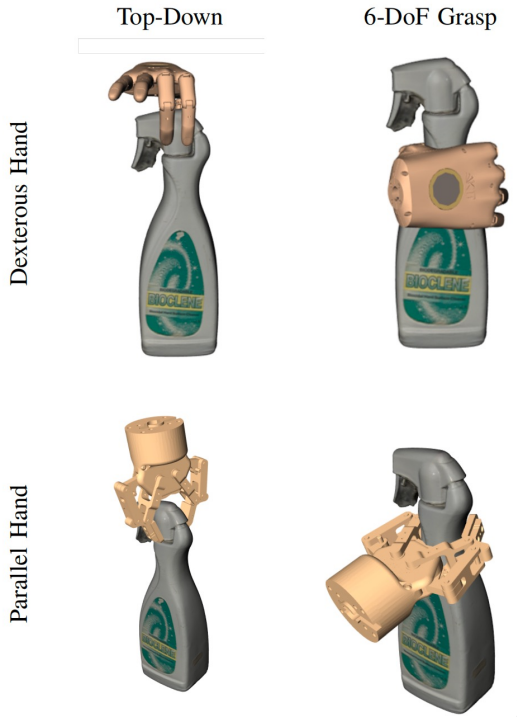


R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun. Deep Learning Approaches to Grasp Synthesis: A Review. arXiv 2207.02556

Terminology

- **Grasp Pose** defines the position and orientation of a hand
  - **4-DoF grasp**: a 3D position and 1D hand orientation aligned with the direction of gravity, a.k.a. “top-down grasping”
  - **6-DoF grasp**: defined by a 3D position and 3D orientation

- **Force Closure** means if the forces that a grasp can be applied at the set of frictional contacts are sufficient to compensate for any external wrench applied to the object



Force Closure

- In physics,
  - given a set of frictional contacts acting on a body, it is in **force closure** if the positive span of the wrench cones is the entire wrench space
  - if a rigid body is fully immobilized by a set of rigid stationary fixtures, we say it is in **form closure**
- The conditions for **force closure** are identical to the conditions for **first-order form closure** if all contacts are frictionless.
- When planning a grasp by a robot hand, **force closure** is a good minimum requirement. **Form closure** is usually too strict, requiring too many contacts.
- Simply, form closure => force Closure => successful grasp

https://modernrobotics.northwestern.edu/nu-gm-book-resource/12-2-3-force-closure

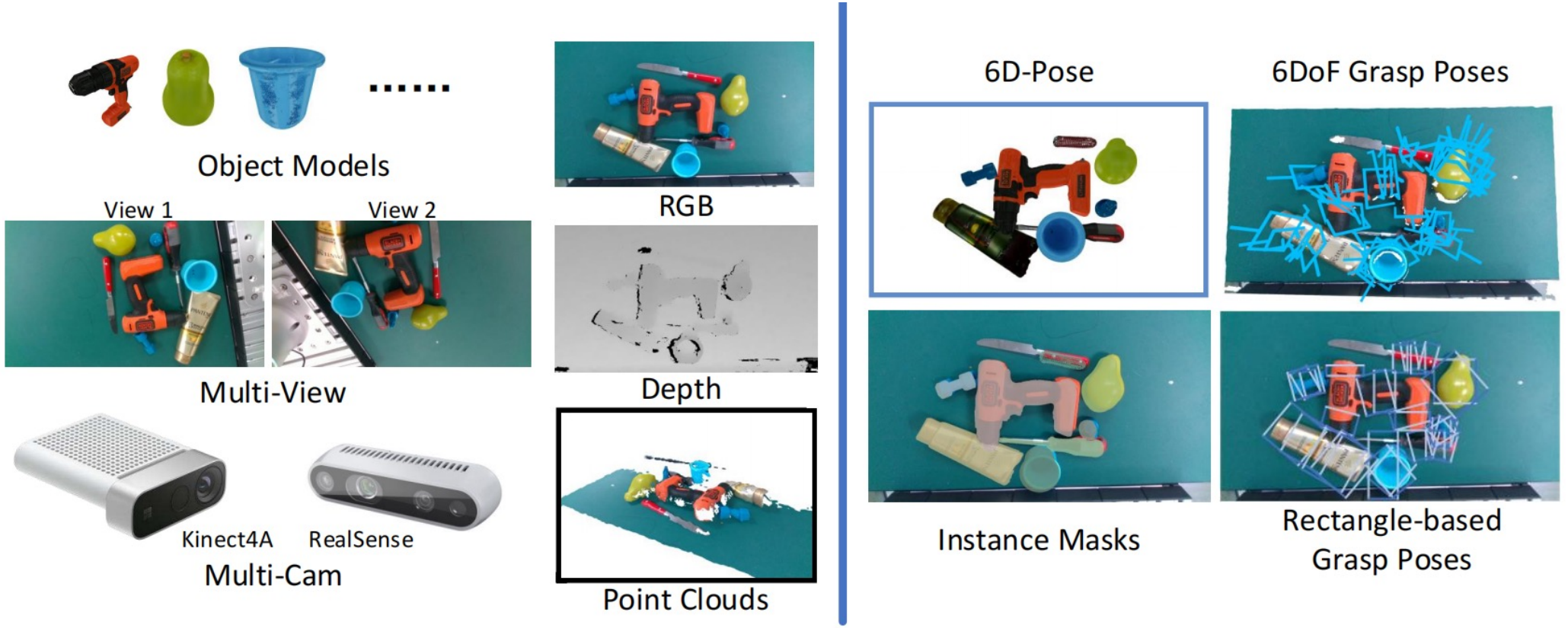
Force Closure

- Watch: https://www.youtube.com/watch?v=6RWFFMtD5k8

Dataset

Real dataset

- **GraspNet-1Billion**: with grasping annotation

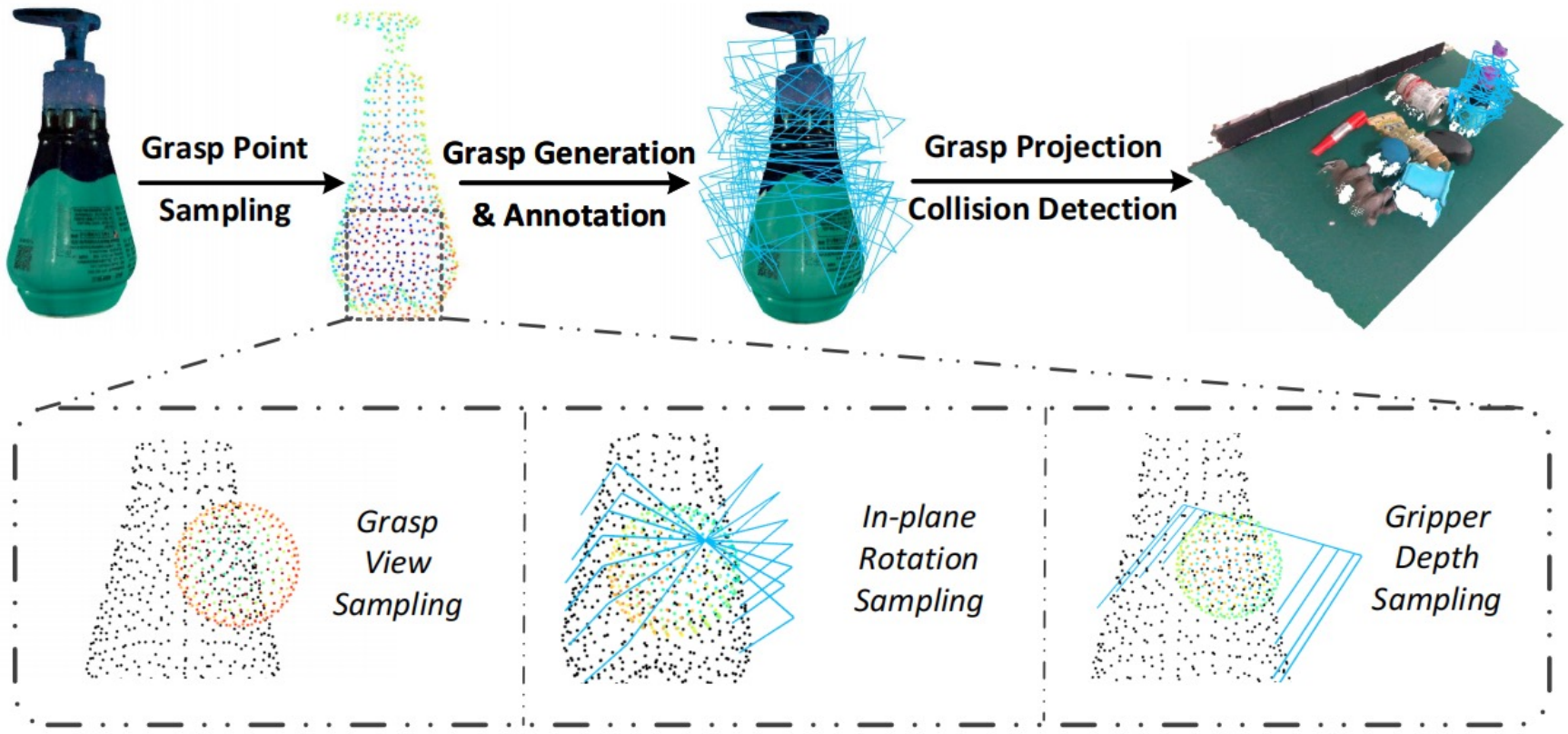


H.-S. Fang, C. Wang, M. Gou, and C. Lu. Graspnet-1billion: a largescale benchmark for general object grasping. CVPR 2020.

GraspNet-1Billion: How Grasp Annotations Generated?

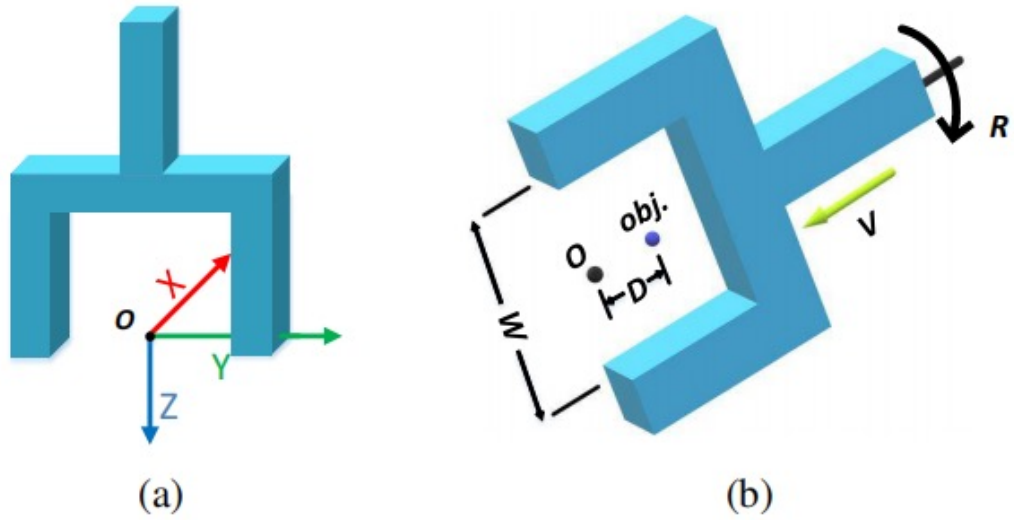
Grasp pose annotation pipeline

- The grasp point is firstly sampled from point cloud.
- Then the grasp view, the in-plane rotation and the gripper depth are sampled and evaluated. Finally, the grasps are projected on the scene using the 6D pose of each object.
- Collision detection is also conducted to avoid the collision between grasps and background or other object.



GraspNet-1Billion: How Grasp Annotations Generated?

- (a) The coordinate frame of the gripper.
- (b) Our new representation of grasp pose. “obj.” denotes object point. Our network needs to predict i) the approaching vector V , ii) the approaching distance from grasp point to the origin of gripper frame D, iii) the in-plane rotation around approaching axis R and iv) the gripper width W.



GraspNet-1Billion: How Grasp Annotations Generated?

- **Step 1:** Grasp poses are sampled and annotated for **each single object**
  - high quality mesh models are downsampled such that the sampled points (called grasp points) are uniformly distributed in voxel space.
  - For each grasp point, we sample V views uniformly distributed in a spherical space.
  - Grasp candidates are searched in a two dimensional grid D × A, where D is the set of gripper depths and A is the set of in-plane rotation angles. Gripper width is determined accordingly such that no empty grasp or collision occurs.
  - Each grasp candidate will be assigned a confidence score (by force-closure) based on the mesh model.

GraspNet-1Billion: How Grasp Annotations Generated?

- **Step 2**: **For each scene**, we project these grasps to the corresponding objects based on the annotated 6D object poses.

$$\mathbf{P}^i = \mathbf{cam_0}\mathbf{P}_0^i,$$

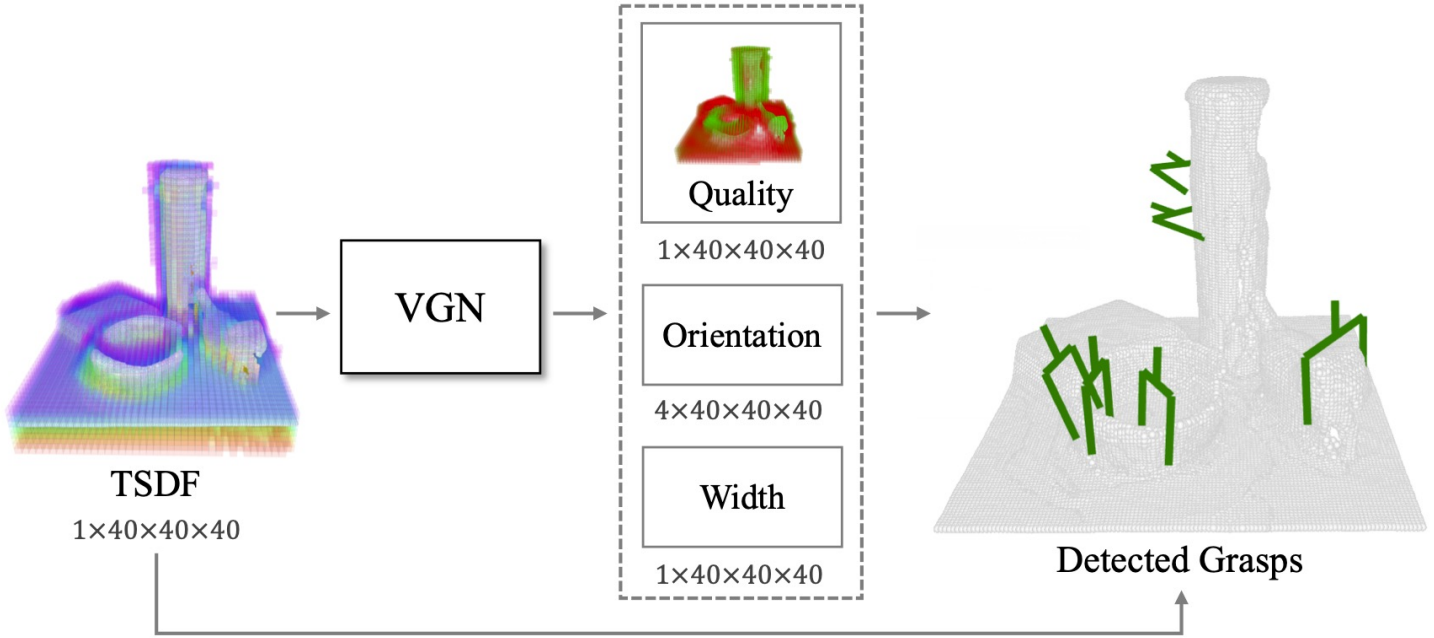
$$\mathbb{G}_{(w)}^i = \mathbf{P}^i \cdot \mathbb{G}_{(o)}^i$$

Collision check is performed to avoid invalid grasps.

Scene Representation

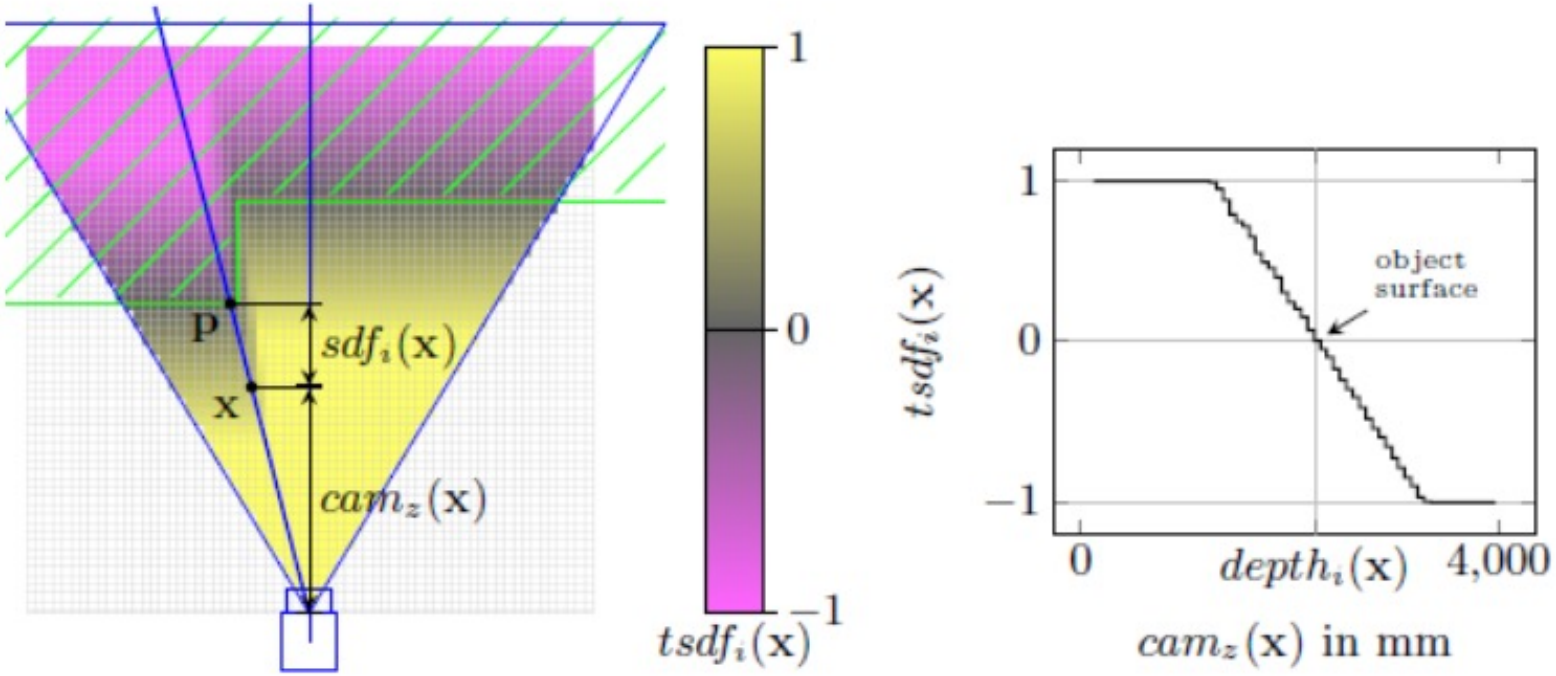
Voxel Grids

- A voxel represents a value in a regular 3D grid. Voxel Grids are analogous to images in 3D space, where a voxel is similar to a pixel over a 3D grid instead of a 2D image
- Network: VGN
- Explicit geometry; limited by the volume resolution



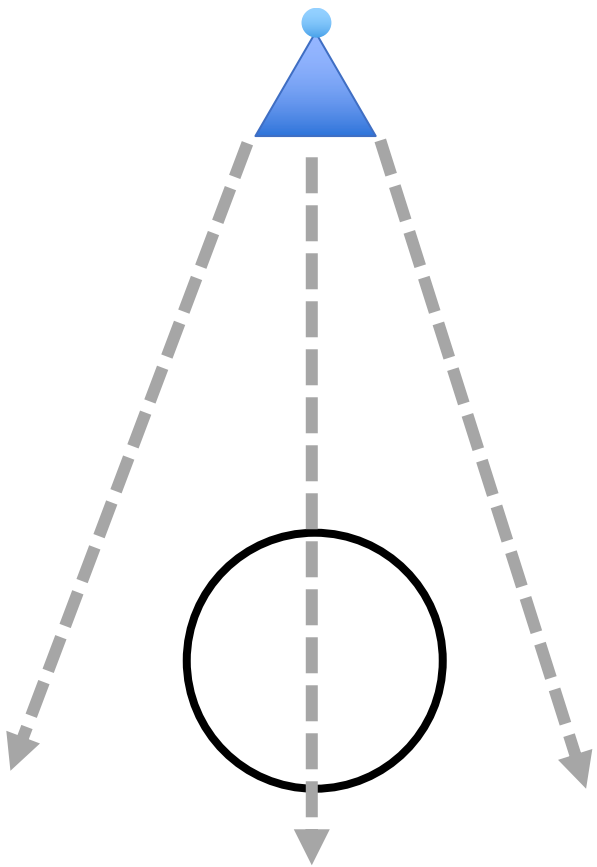
M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto. Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter. CoRL 2020.

TSDF

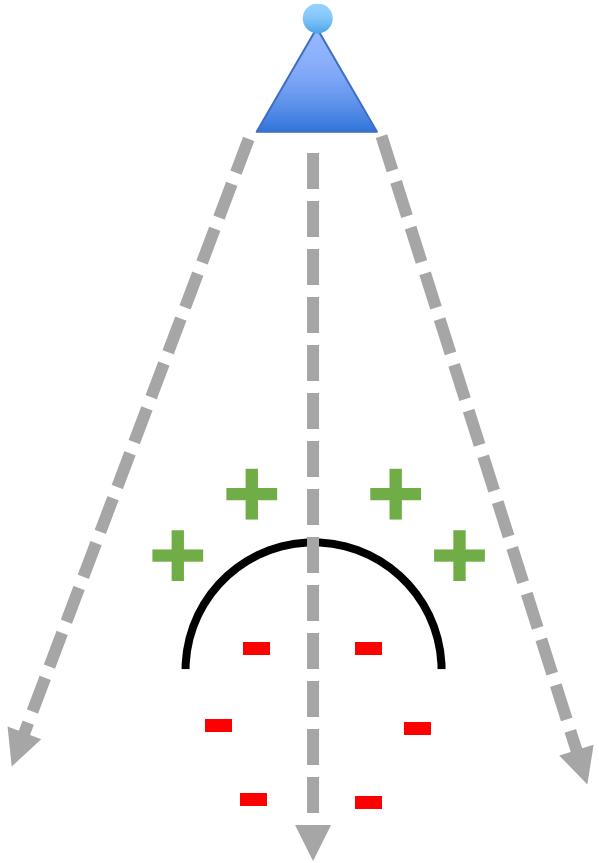


TSDF: Truncated **S**igned **D**istance **F**unction

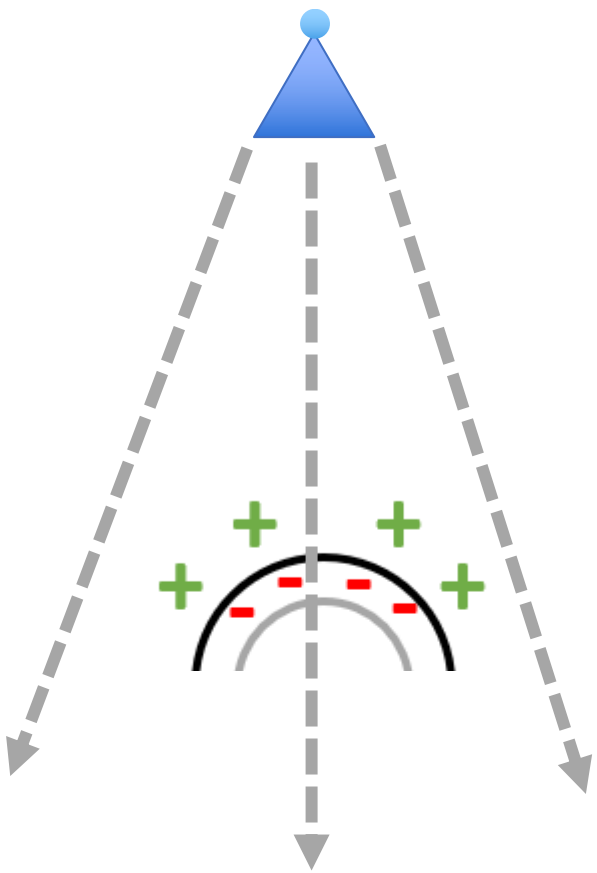
TSDF vs SDF



Sphere Reconstruction

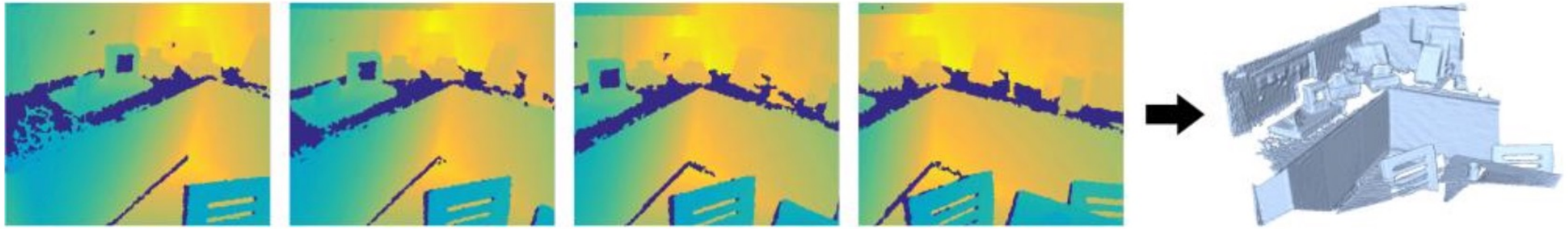


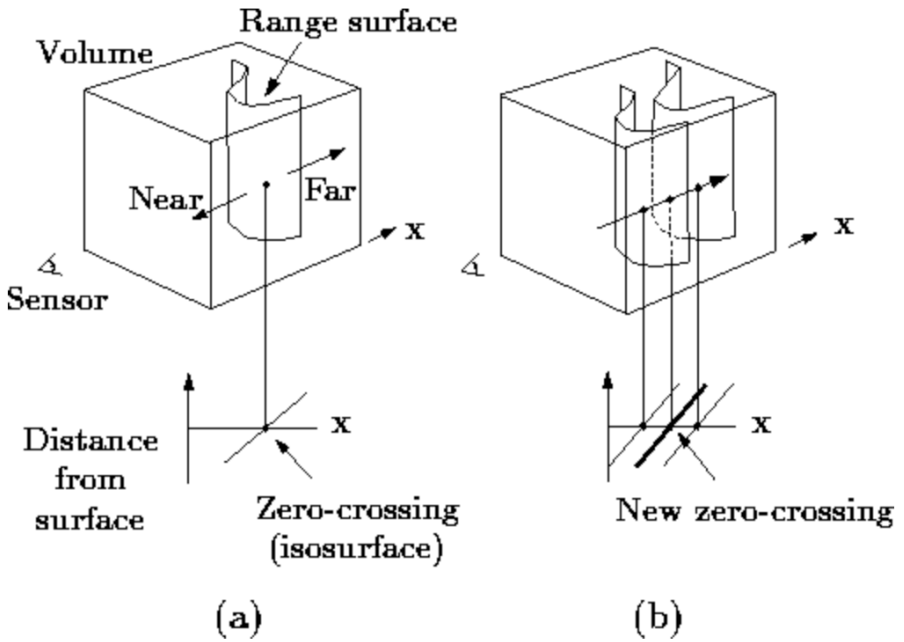
Signed Distance Function



Truncated Signed Distance Function

Volumetric TSDF Fusion of Multiple Depth Maps





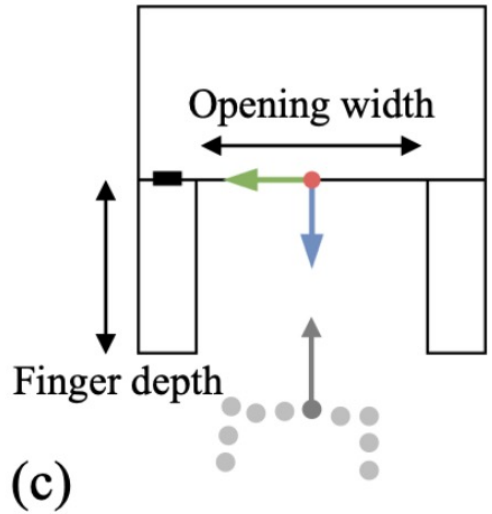
Reference:

1. https://zhuanlan.zhihu.com/p/469539351
2. A Volumetric Method for Building Complex Models from Range Images.
3. https://github.com/andyzeng/tsdf-fusion

VGN

How to define grasps?

- Define a grasp g by:
  - The position t (index in the volume)
  - Orientation r (quaternion)
  - The opening width w of the gripper

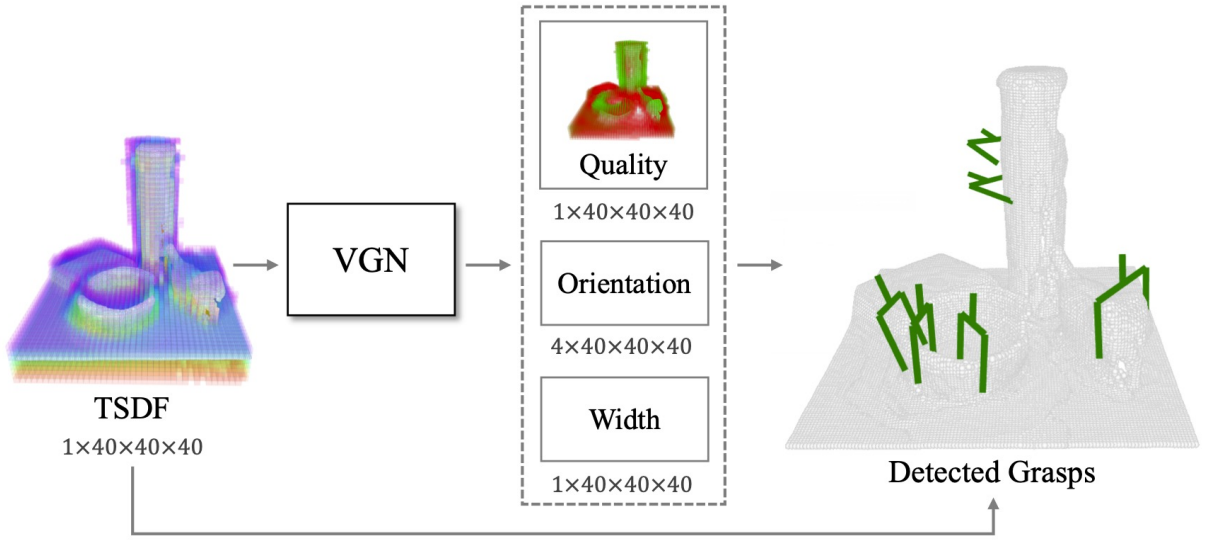


- Each pose is associated with a scalar quantity $q \in [0, 1]$ capturing the **probability of grasp success**
- The edges of the volume correspond to the boundaries of the grasping workspace

VGN

Detection Network

- Input is represented as a TSDF volume fused from multi-view depth maps
- The output is 6-channel volume grid, with each voxel containing quality(1D), orientation (quaternion, 4D) and width (1D), respectively.
- Follows a 3D FCN encoder-decoder architecture.



VGN

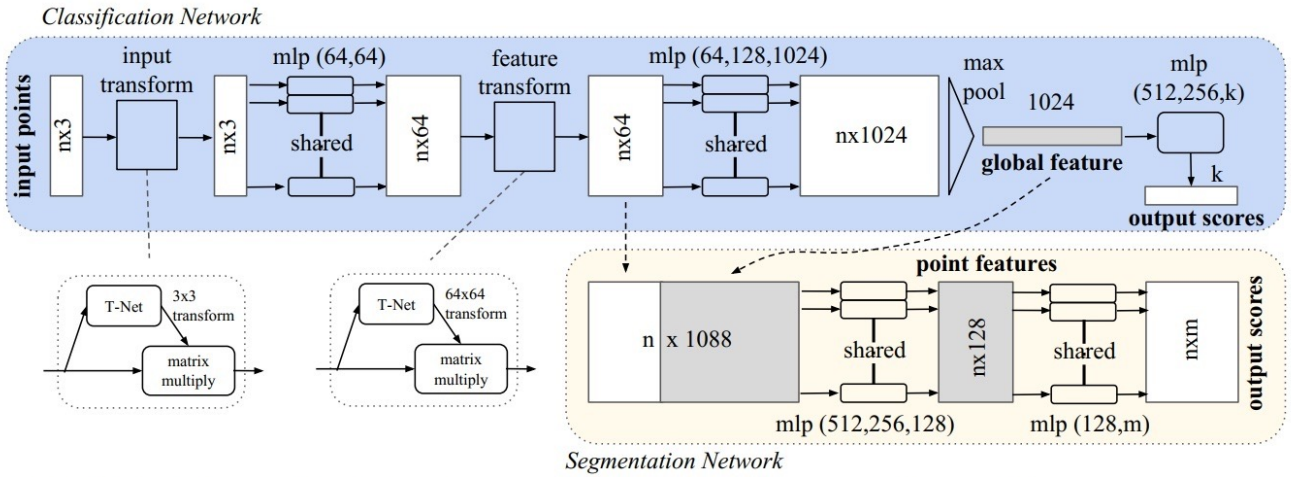
Post processing

- The grasp quality tensor is smoothed with a 3D Gaussian kernel which favors grasps in regions of high grasp quality
- Mask out voxels whose distance to the nearest surface is smaller than the finger depth
- Apply non-maximal suppression

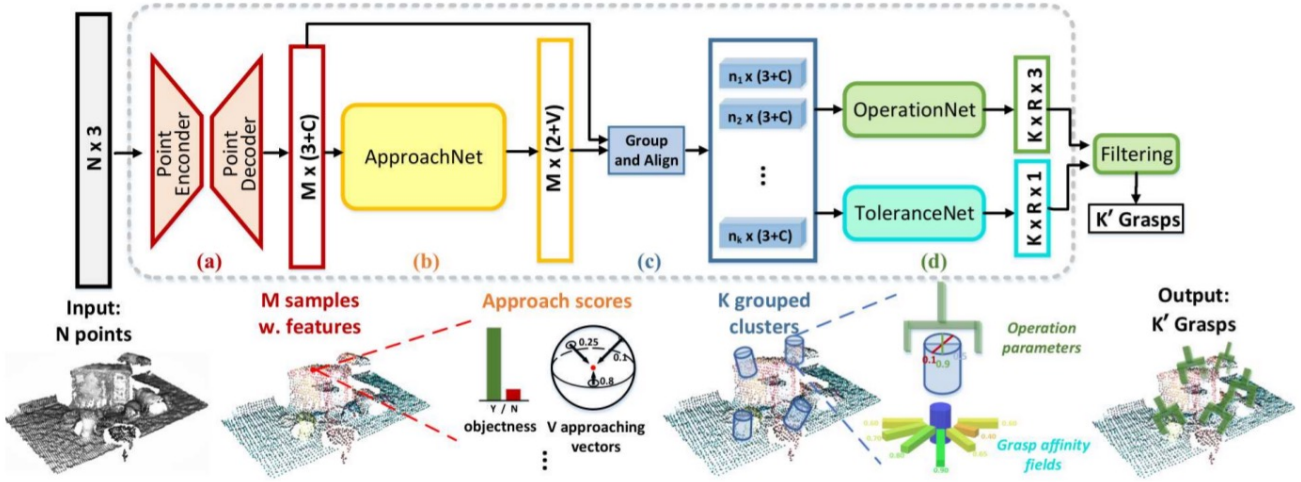
Data Representation

Point Cloud

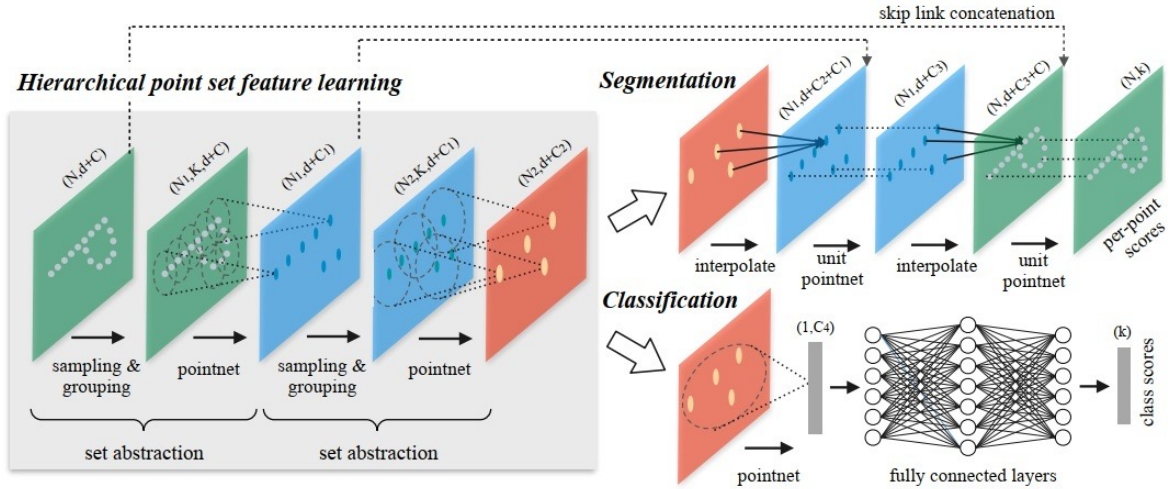
- Backbone: PointNet, PointNet++
- Network: GraspNet-baseline
- Explicit geometry; sensor noise, depth error (specular/transparent)



PointNet



GraspNet-baseline

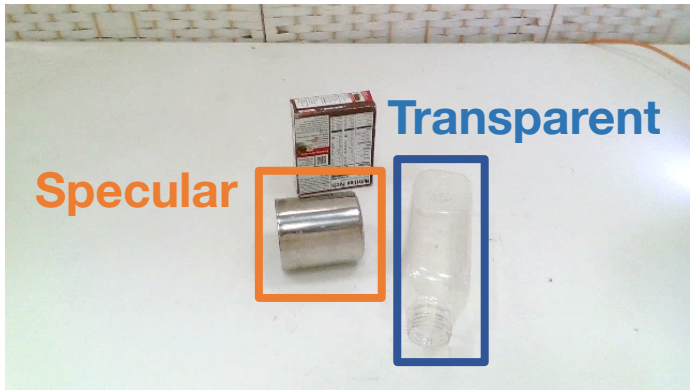


PointNet++

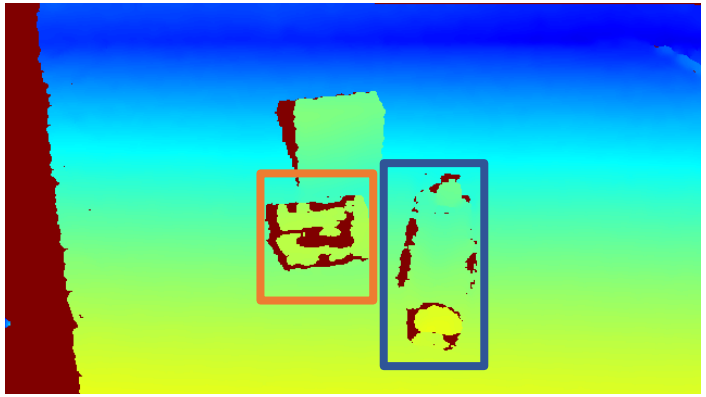
Depth Sensing Problem

Grasping transparent and specular objects

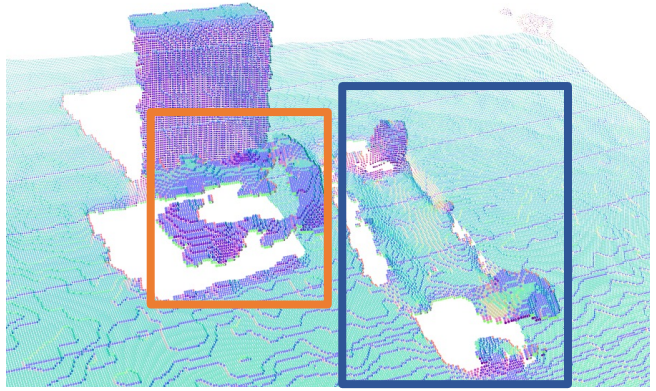
Depth-based grasping: not robust to material



RGB



Depth



Point cloud

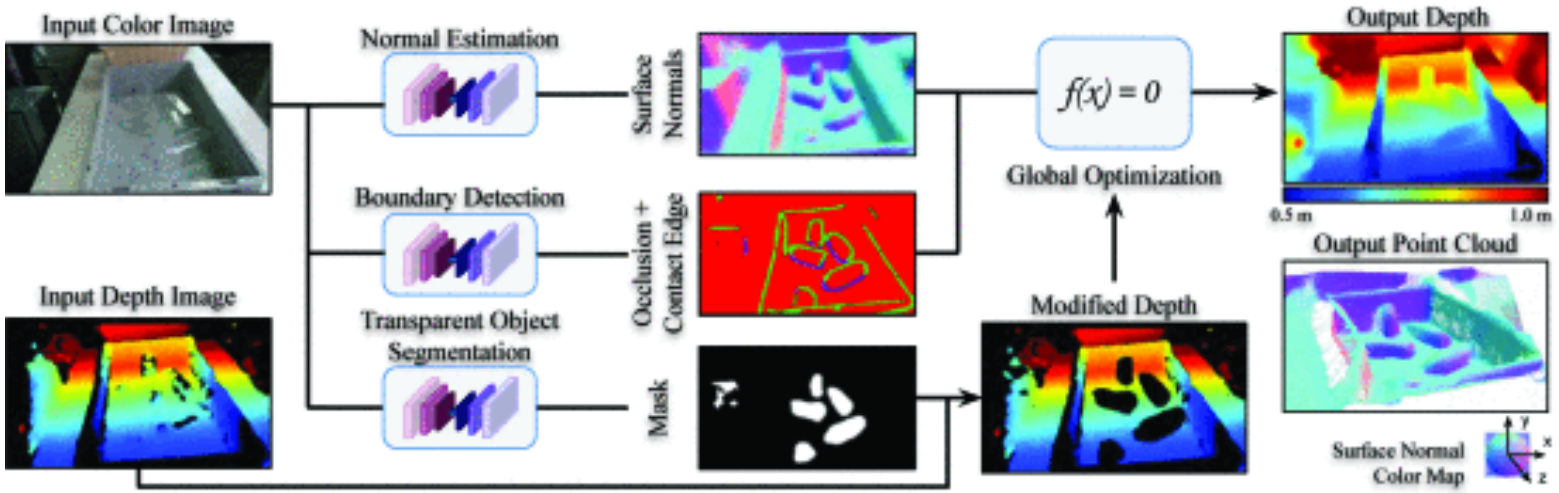
Goal: Generalizable Material-Agnostic Object Grasping

- **Material-Agnostic**: grasping algorithm works for any material, including diffuse, transparent, specular, metallic, etc.

- **Generalizable**: works for unseen objects and unseen categories with arbitrary object poses.

Different Approaches

- Approach 1:First restoring depth, then grasping

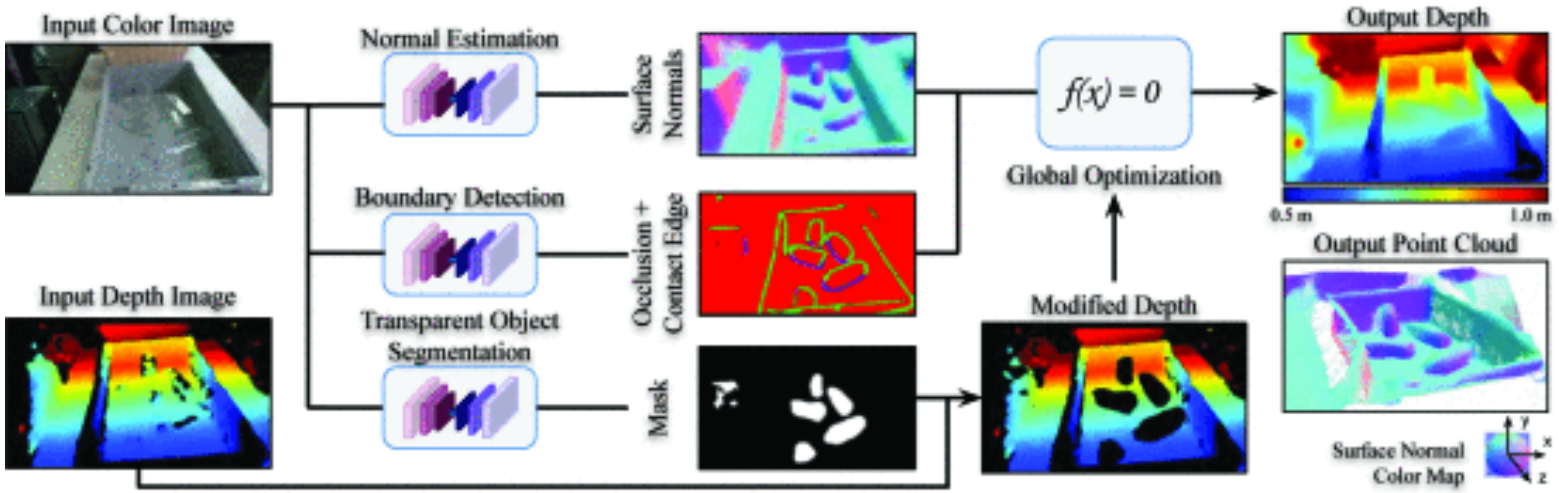


Sajjan, Shreeyak, et al. "Clear grasp: 3d shape estimation of transparent objects for manipulation." *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

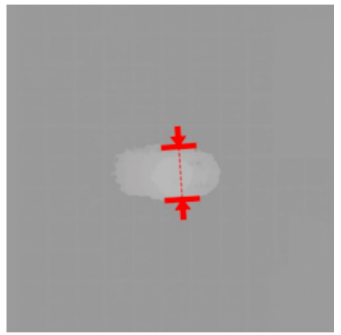
*J. Ichnowski, et. al, Dex-NeRF: Using a neural radiance field to grasp transparent objects, CoRL 2021

Different Approaches

- Approach 1:First restoring depth, then grasping



- Approach 2: Multiview RGB-based method (depth-free)



Dex-NeRF + Grasp

Sajjan, Shreeyak, et al. "Clear grasp: 3d shape estimation of transparent objects for manipulation." *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

*J. Ichnowski, et. al, Dex-NeRF: Using a neural radiance field to grasp transparent objects, CoRL 2021

GraspNeRF:

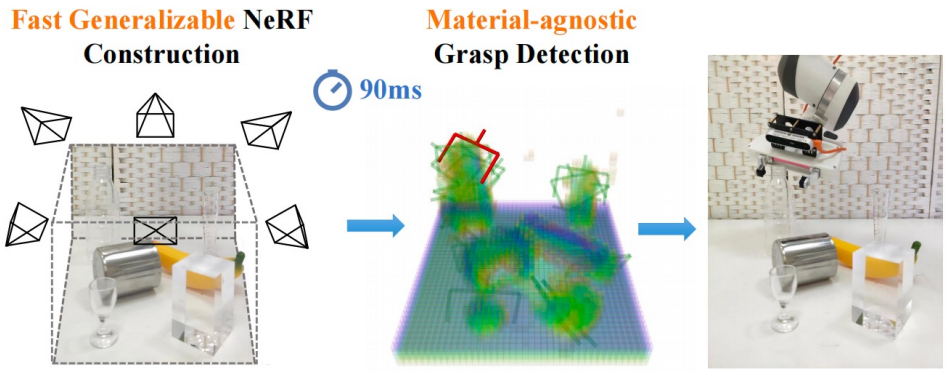
Multiview-based 6-DoF Grasp Detection for Transparent and Specular Objects Using Generalizable NeRF

Qiyu Dai*, Yan Zhu*, Yiran Geng, Ciyu Ruan,
Jiazhao Zhang, He Wang†

*: equal contributions, †: corresponding author

NeRF for Object Manipulation

Summary of GraspNeRF



- Sparse inputs
- Generalize to novel scene
- Real-time speed
- 6-DoF grasping
- End-to-end differentiable

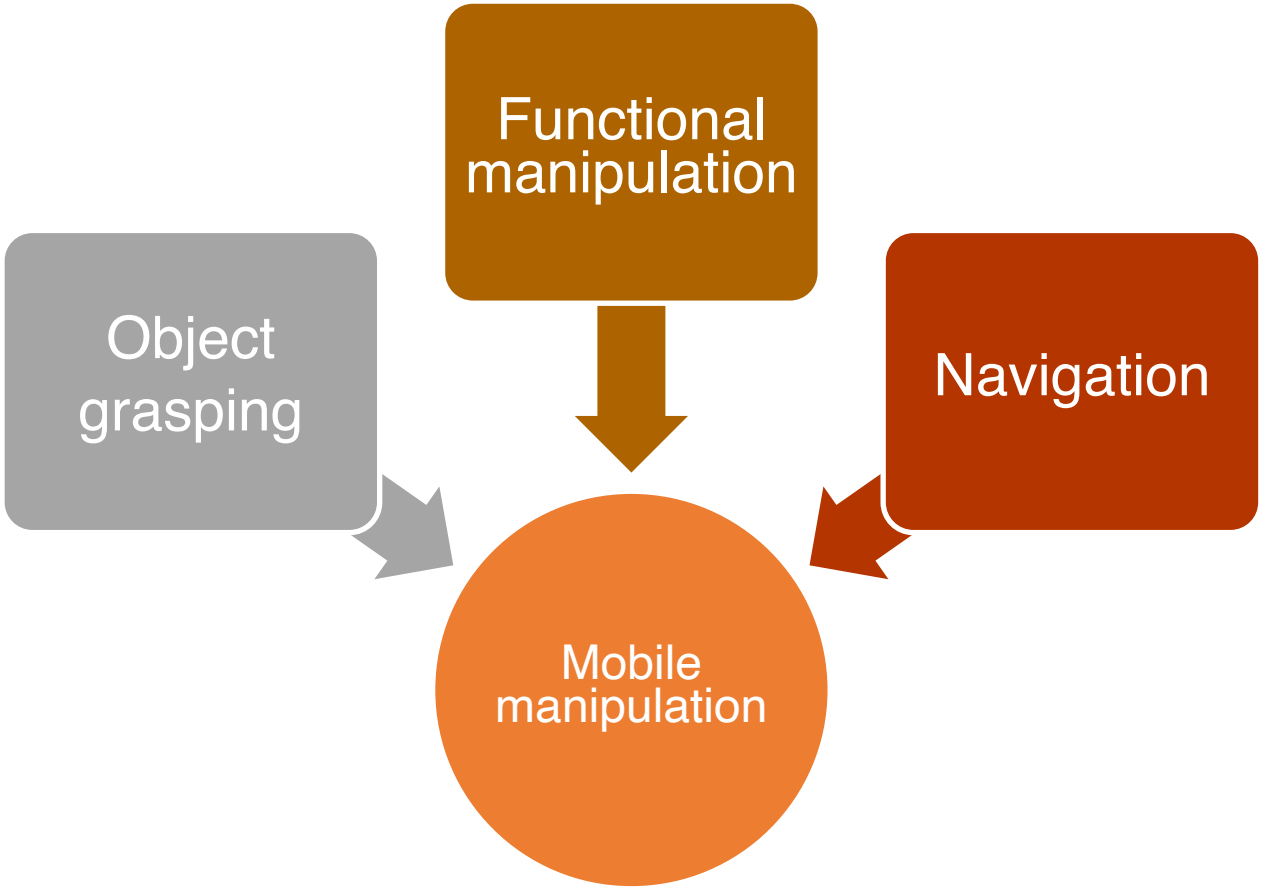
Previous NeRF-based grasping methods	GraspNeRF
✗ 49 views*	✓ 6 views
✗ Training needed	✓
✗ 7s**	✓ 90 ms
✗ 3-DoF	✓
✗ Stage-wise	✓

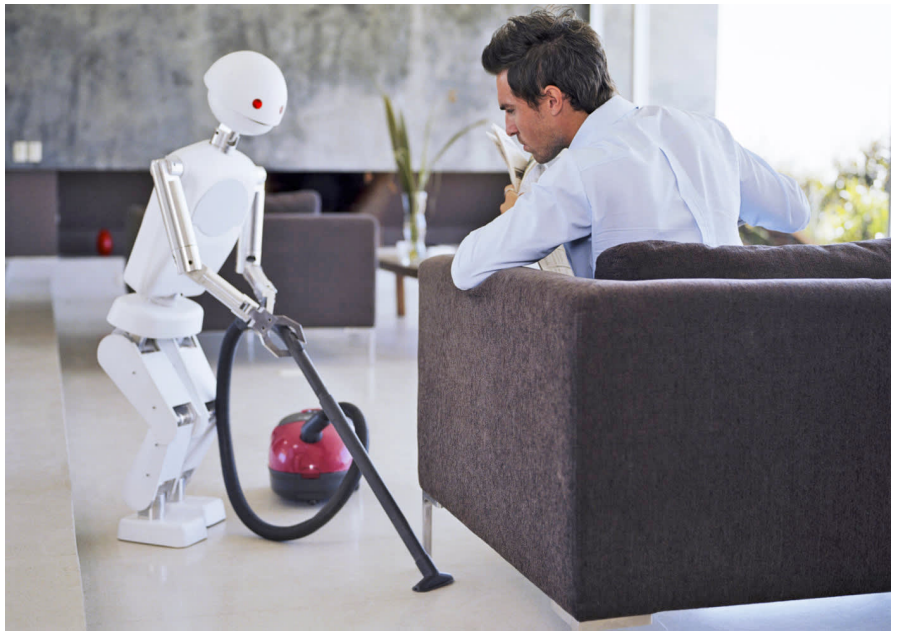
*J. Ichnowski, et. al, Dex-NeRF: Using a neural radiance field to grasp transparent objects, CoRL 2021

**J. Kerr, et. al, Evo-NeRF: Evolving NeRF for sequential robot grasping, CoRL 2022

Progress

Goal: a scalable **3D-aware** home robot





Object Manipulation

Introduction: Manipulation Tasks

- Grasping
- Ungrasping
- Object rearrangement
- In-hand manipulation
- Tool use
  - Harmers, scissors …

Introduction: Different Type of Grippers

Suction-type gripper

Parallel gripper

Dexterous hand

Category-Level Generalizable Object Manipulation

Learning Category-Level Generalizable Object Manipulation Policy via Generative Adversarial Self-Imitation Learning from Demonstrations

Hao Shen*, Weikang Wan*, He Wang

IEEE Robotic Automation Letter (RAL)

IROS KYOTO 2022

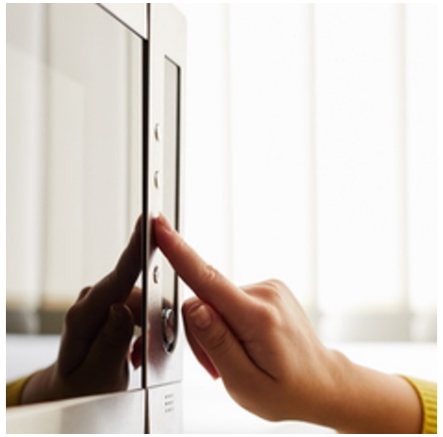
Category-Level Generalizable Imitation Learning

IEEE RAL + IROS 22

1st place, ManiSkill Challenge 2021 “no external annotation” track

Part classes are more elementary and fundamental than object categories.

Buttons on Remote



Buttons on Microwave

Handles on Furniture

Handles on Refrigerator

GAPartNet:

Cross-Category Domain-Generalizable Object Perception and Manipulation via Generalizable and Actionable Parts

CVPR 2023 Highlight

Haoran Geng*, Helin Xu*, Chengyang Zhao*,
Chao Xu, Li Yi, Siyuan Huang, He Wang+



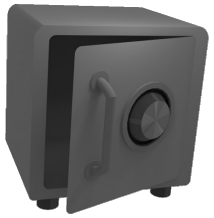

PartNet-Mobility
Object-Centric

define cross-category parts and their poses

GAPartNet
Part-Centric

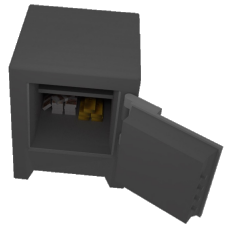
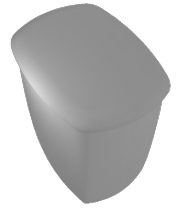












PartNet-Mobility
Object-Centric

GAPartNet
Part-Centric

9 part classes
8,489 parts

from

1166 objects
27 object categories

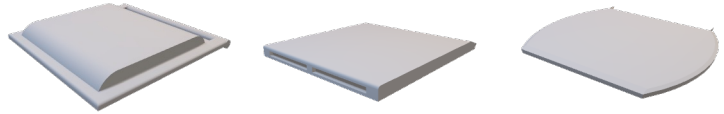
Part Definition

Similar geometry
Generalizable in visual recognition

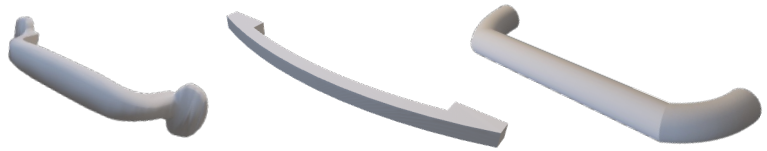
Similar actionability
Generalizable in manipulation

Slider Lid

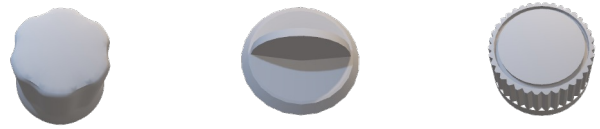
Round Handle



Hinge Lid

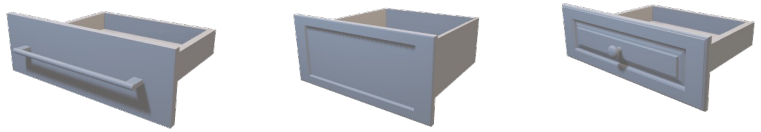


Line Handle



Hinge Knob

Slider Button



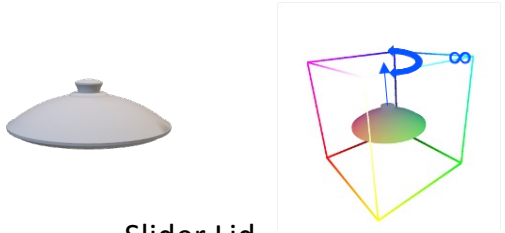
Slider Drawer

Hinge Door

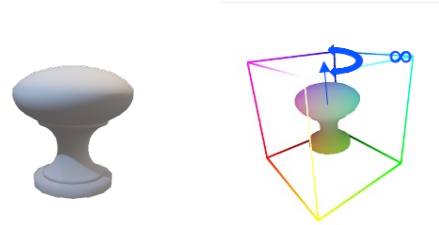
Part Definition

Similar geometry
Generalizable in visual recognition

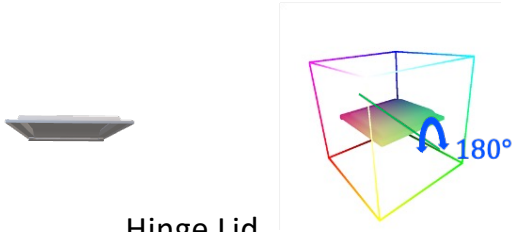
Similar actionability
Generalizable in manipulation



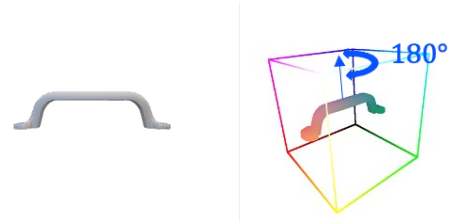
Slider Lid



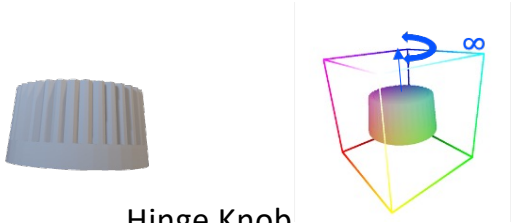
Round Handle



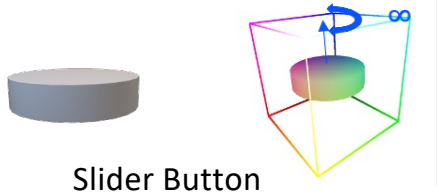
Hinge Lid



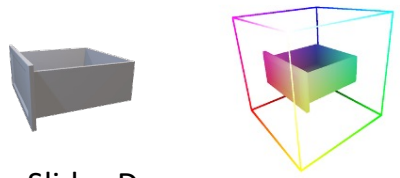
Line Handle



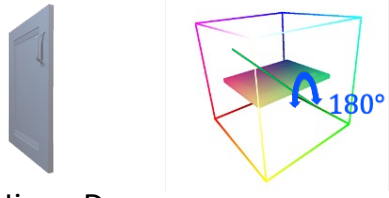
Hinge Knob



Slider Button



Slider Drawer



Hinge Door

GAPartNet

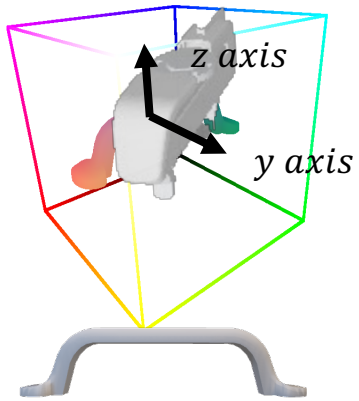
Generalizable
Actionable

Class-Level Part Pose

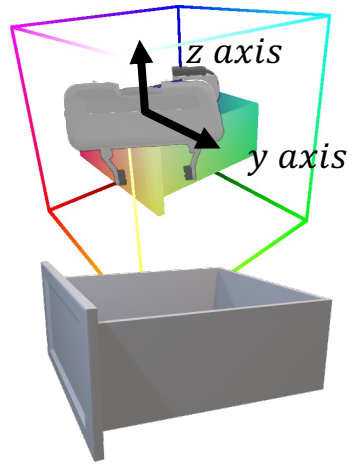
Symmetric Part Pose
Approaching Direction
from +z to -z

Asymmetric Part Pose
Specify y direction

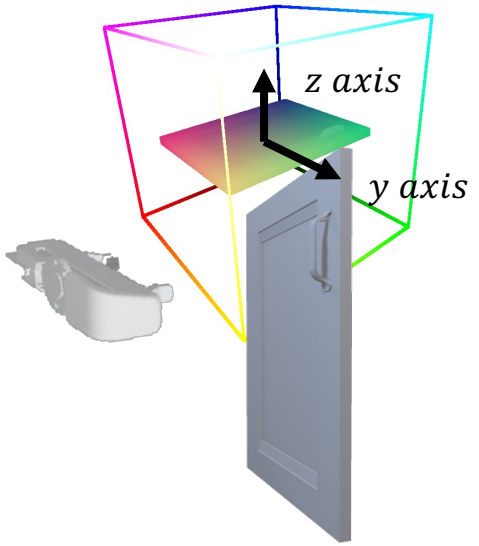
Line Handle



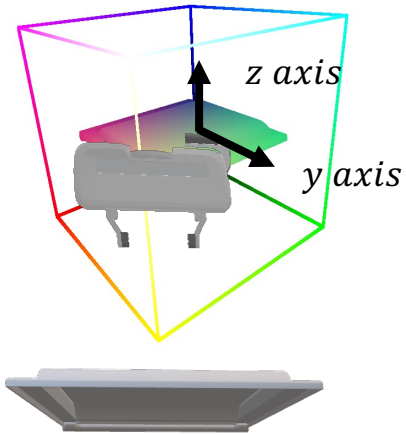
Slider Drawer



Hinge Door



Hinge Lid



Cross-Category Tasks

Inputs

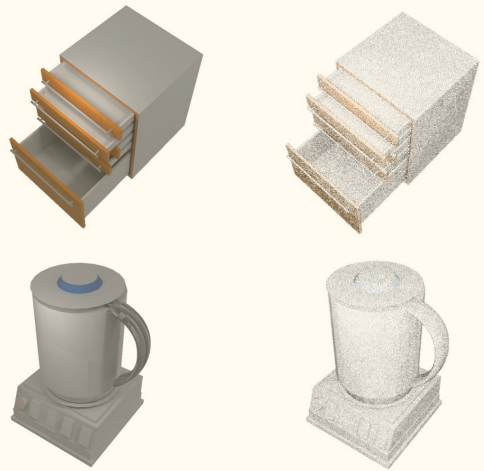
Part Segmentation

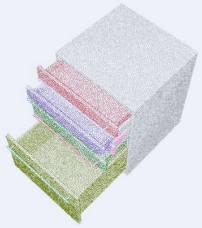
Part Pose Estimation

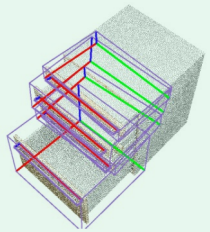
Part-based Object Manipulation

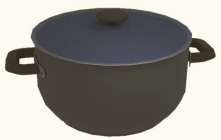
Seen Categories

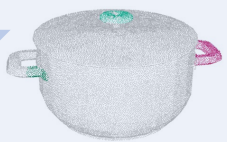
Unseen Categories

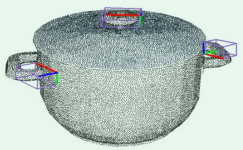


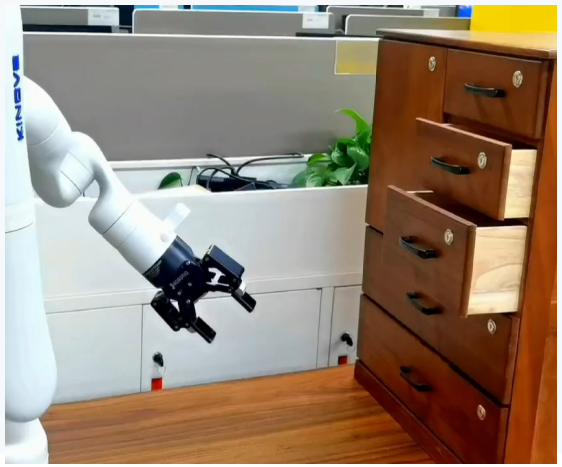






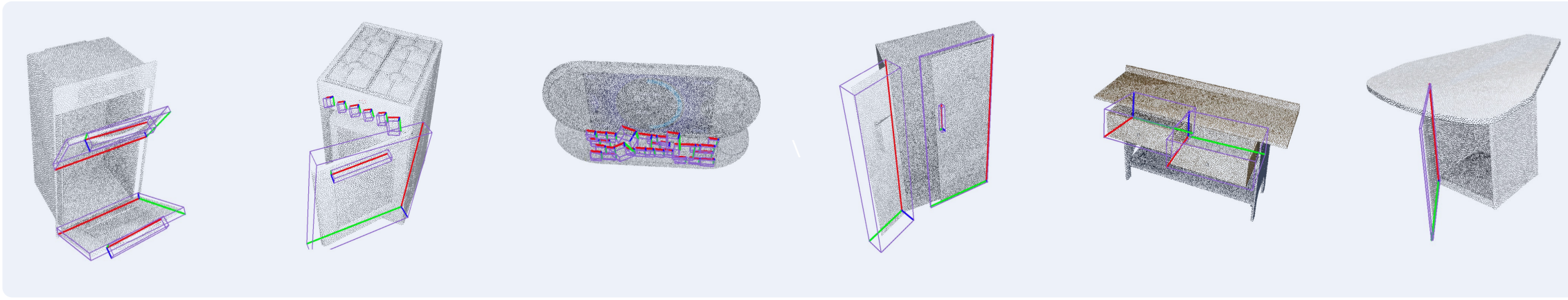




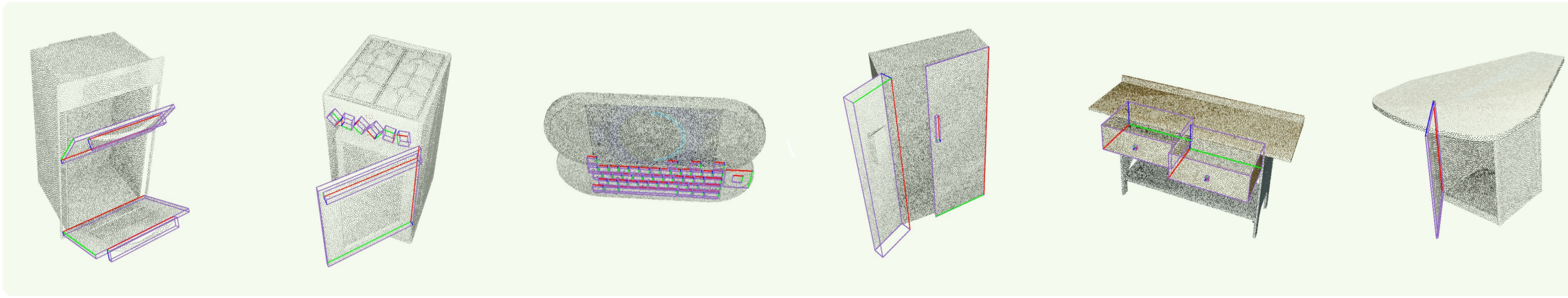


Cross-Category Part Pose Estimation
Unseen Categories

Prediction



GT



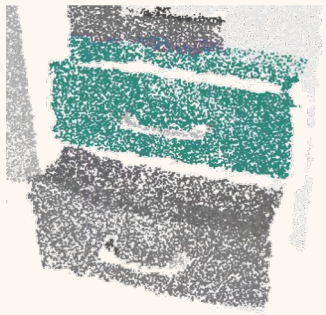
Real-World Experiment

Inputs

point cloud

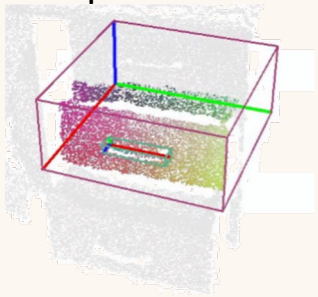
RBG image

Predictions



segmentation

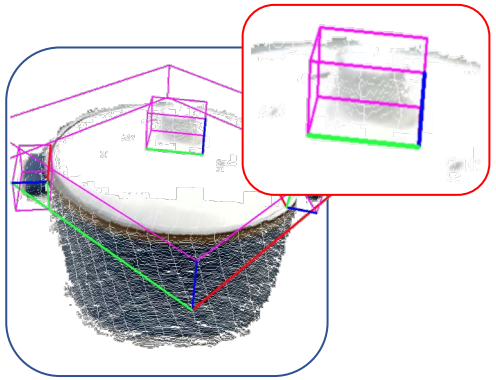
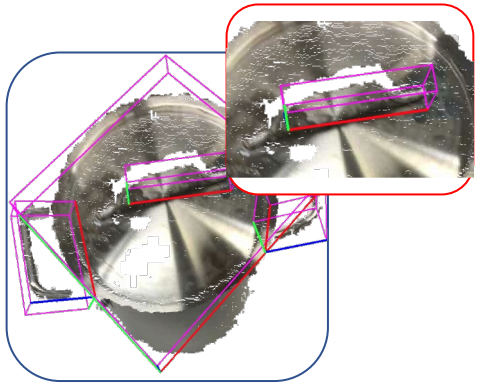
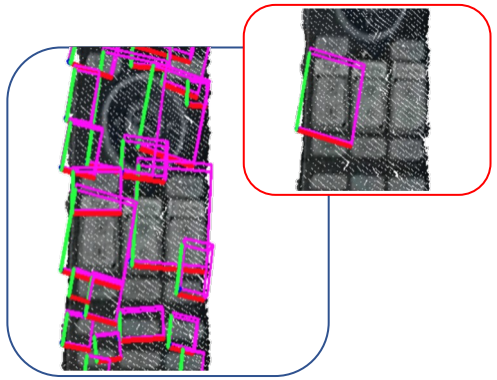
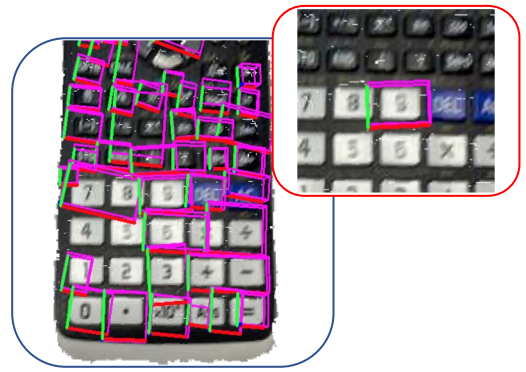
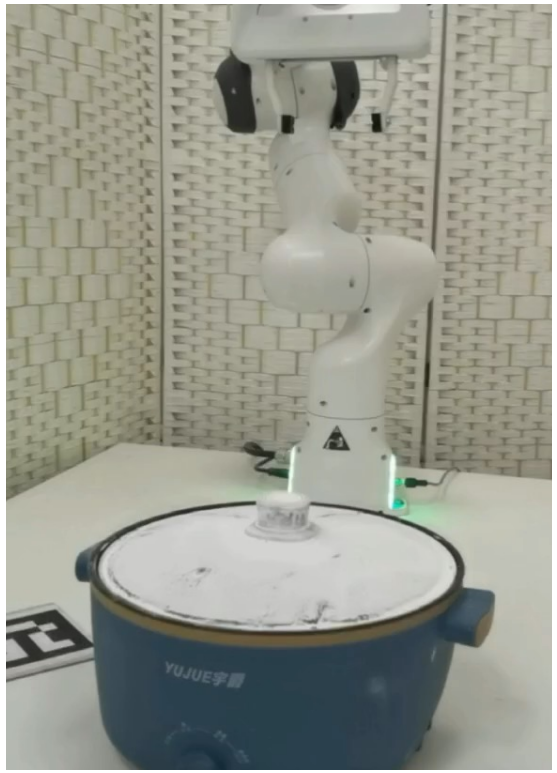
NPCS prediction

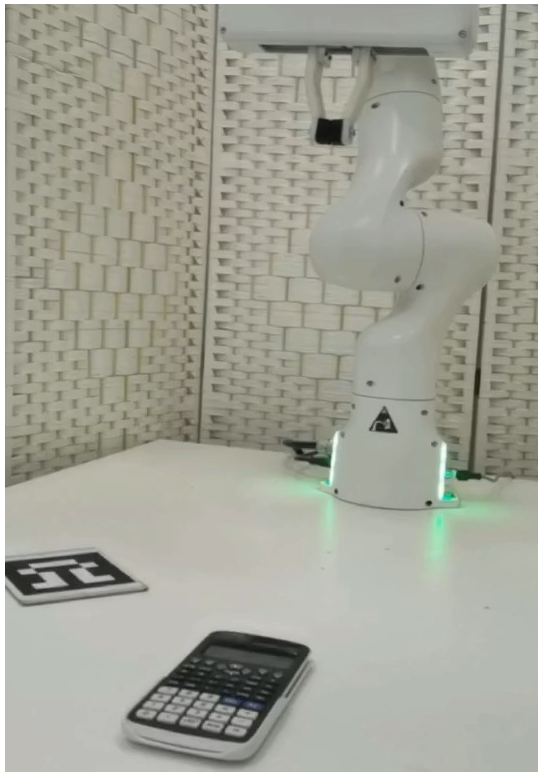
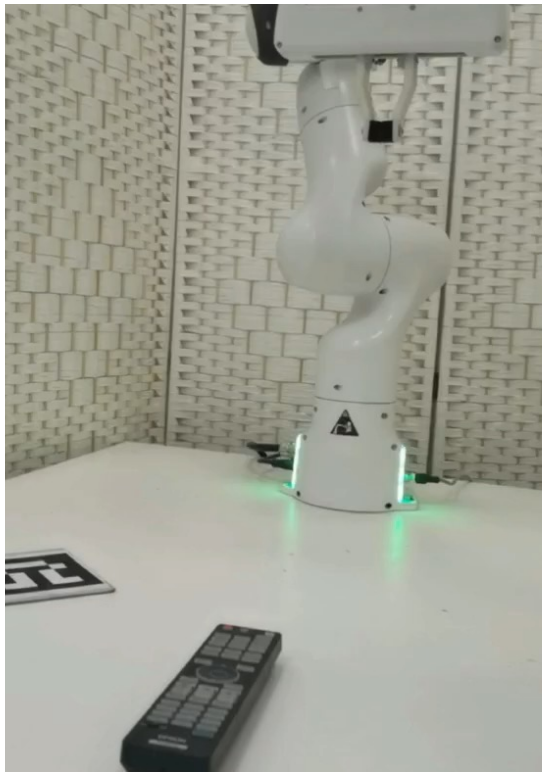


bounding box

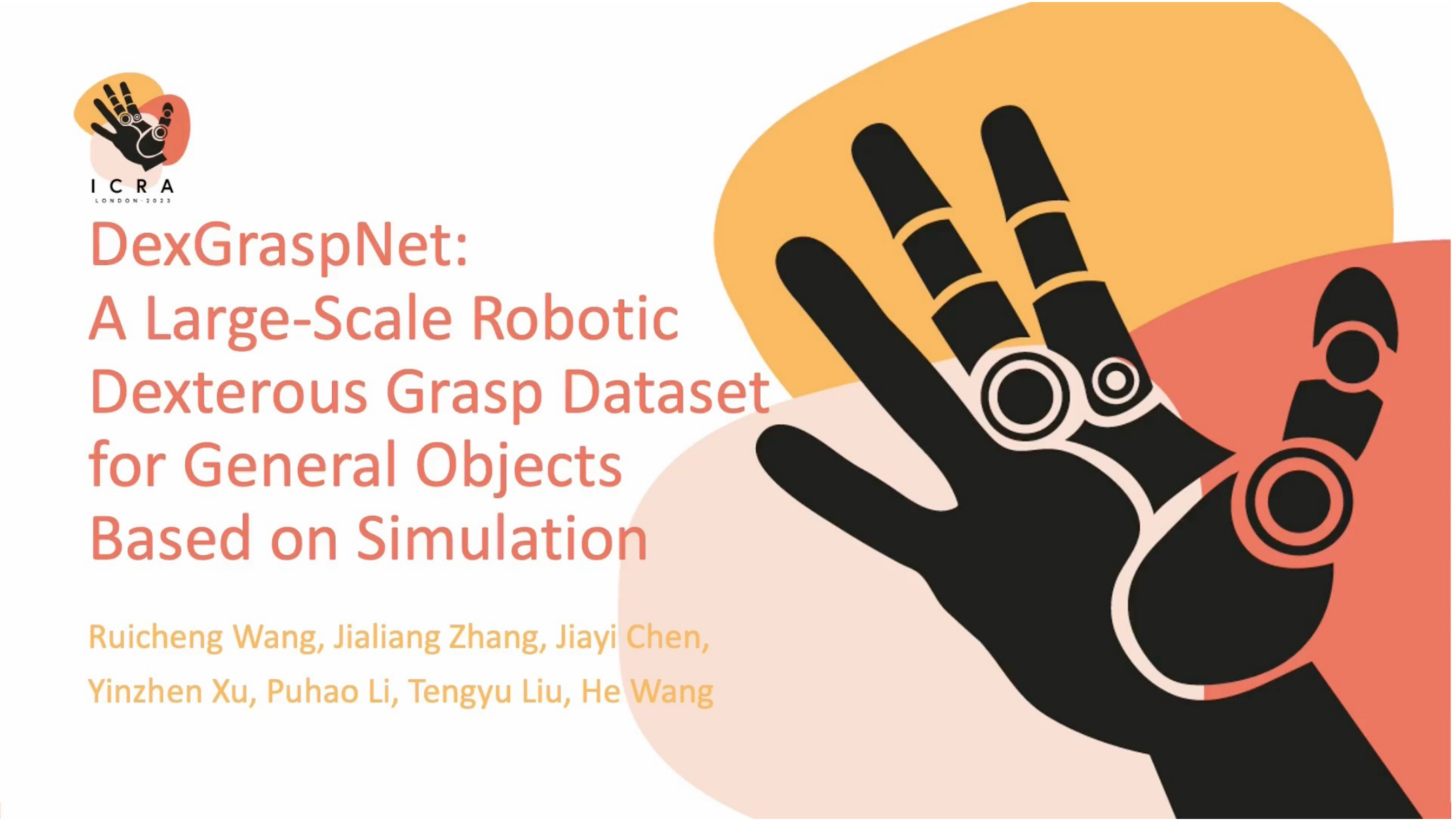
Interaction Video



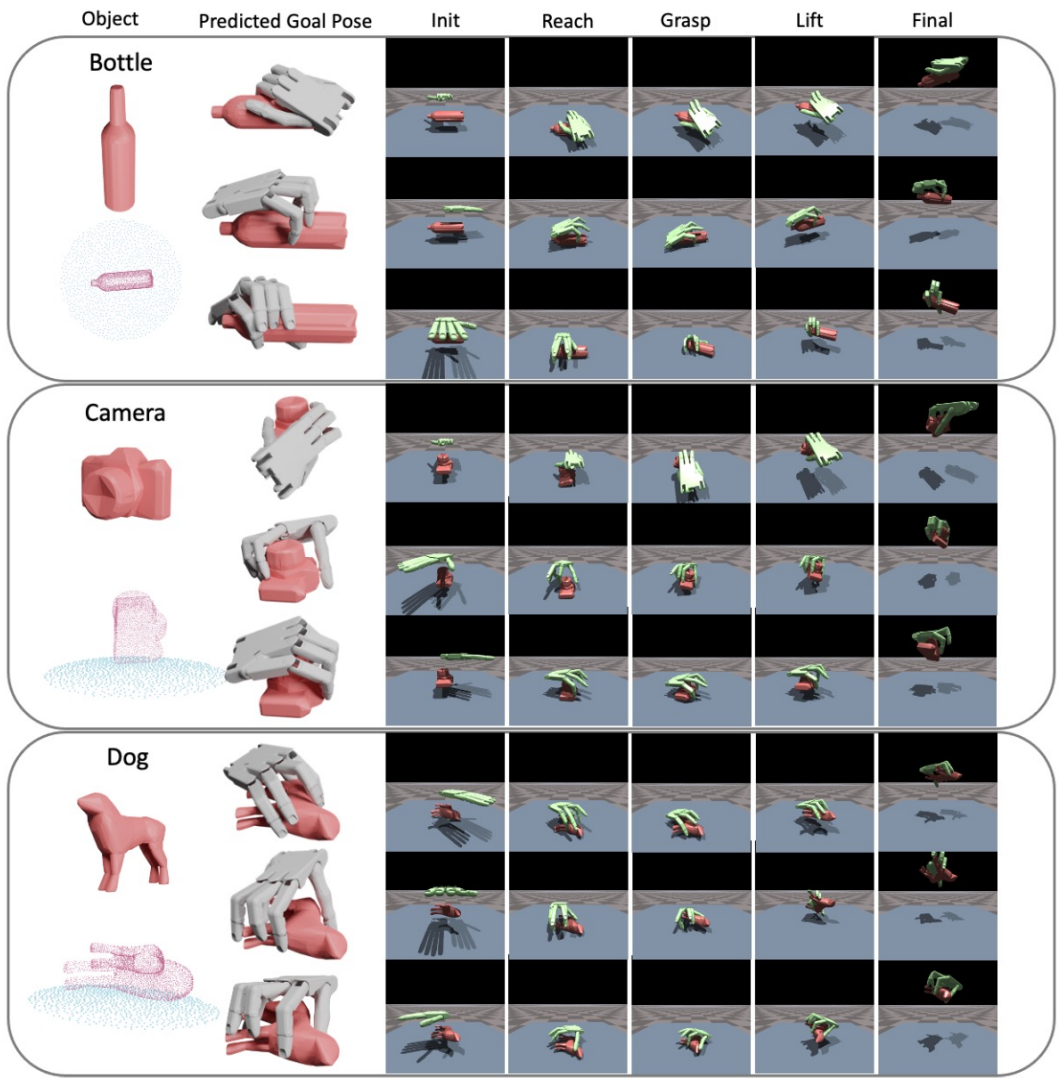







DexGraspNet **(ICRA 2023 Best Manipulation Finalist)**



UniDexGrasp (CVPR 2023)



Language-guided Dexterous Grasping

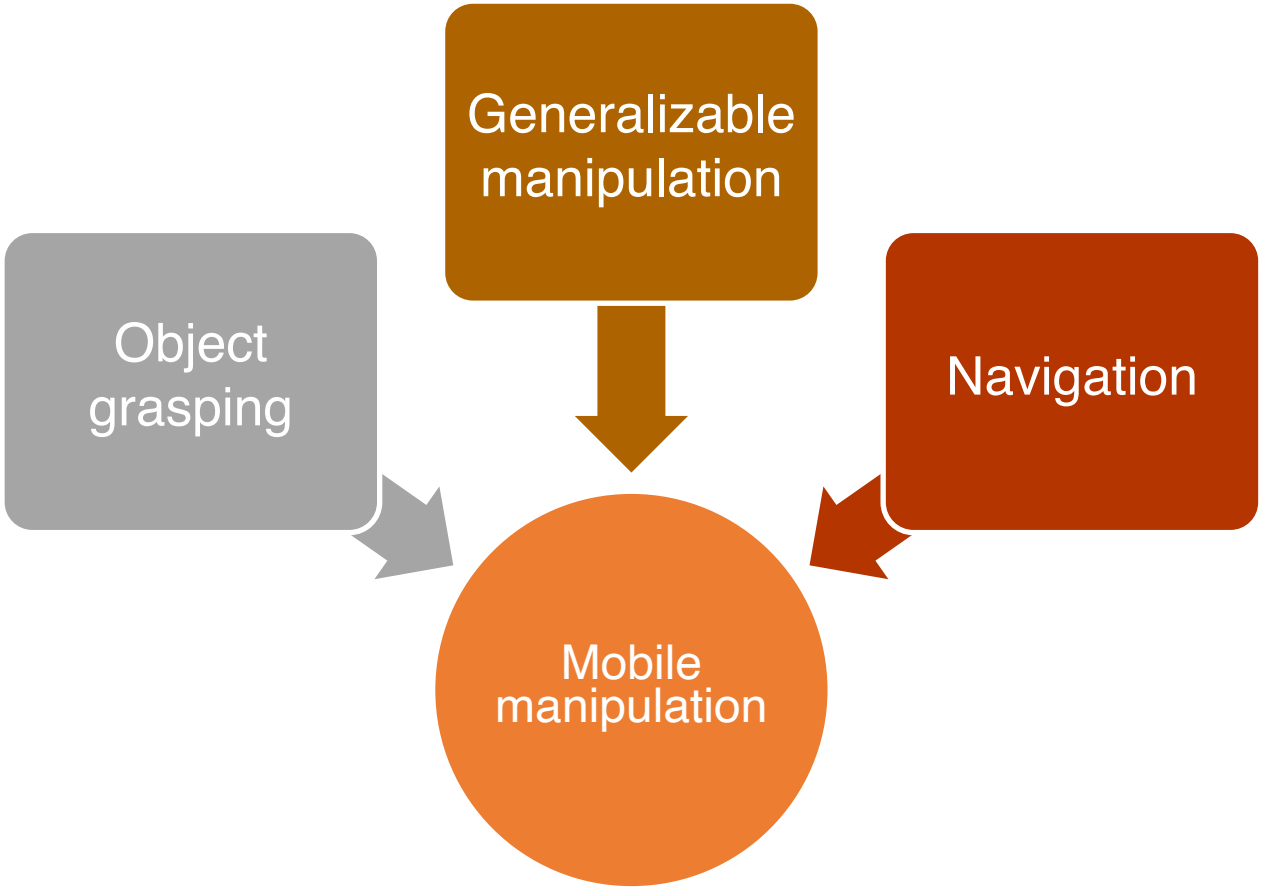
(a) Grasp a bottle by the body/bottleneck, or from below

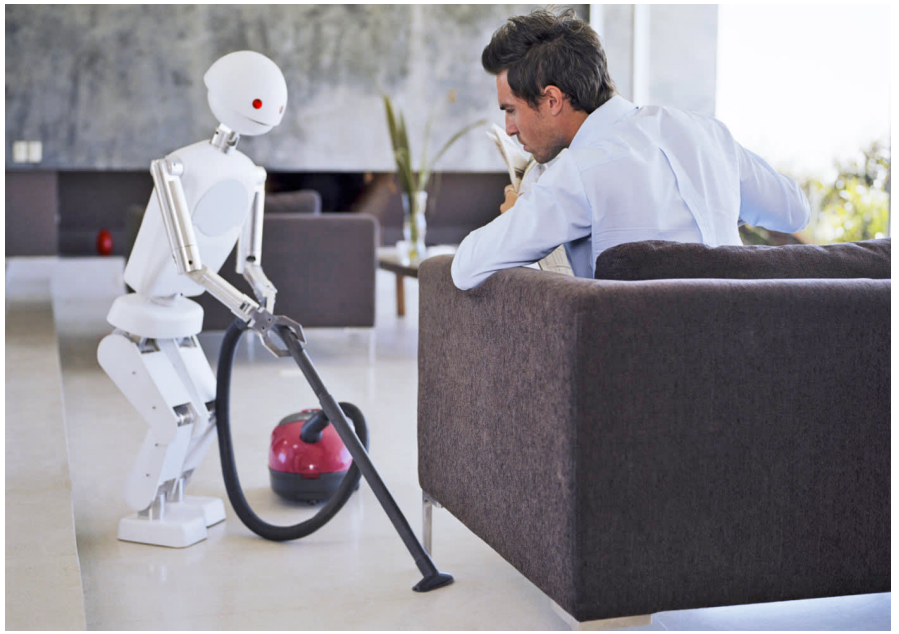
(b) Grasp a hammer by the head/handle

Qualitative results of language-guided grasp proposal selection. CLIP can select proposals complying with the language instruction, allowing goal-conditioned policy to execute potentially functional grasps.

Progress

Goal: a scalable **3D–aware** home robot





Locomotion and Navigation

Legged robot

Boston human-like robot

Four legs are more stable and robust!

Unitree quadruped robot

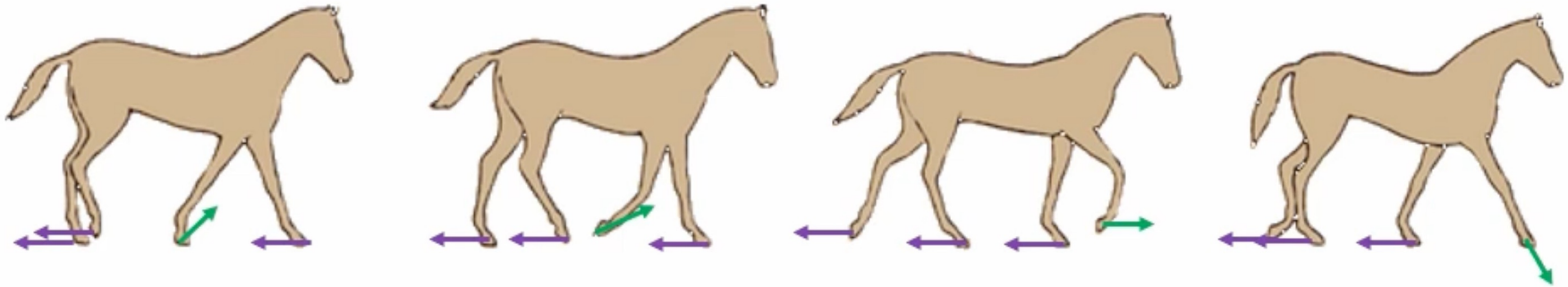
The most robust one with certain limitations.

Wheeled Robots

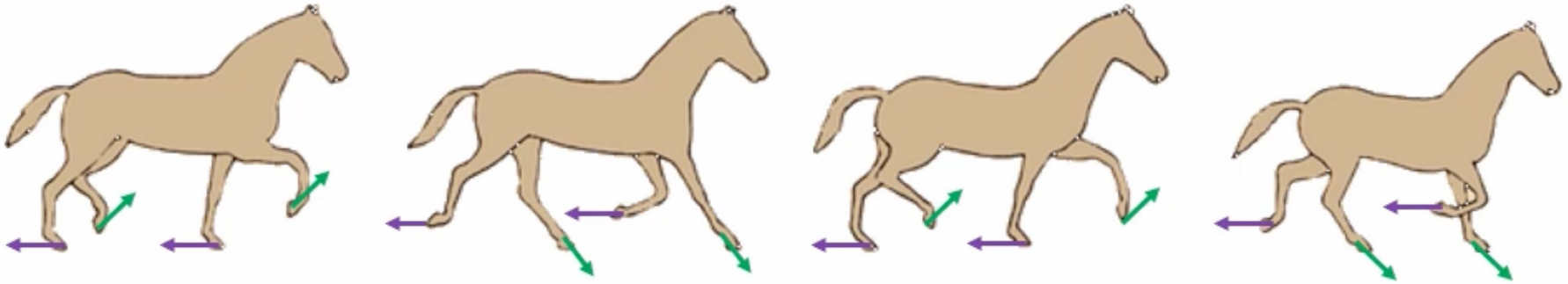
Gait

Gait : the pattern of how does a person/animal walk

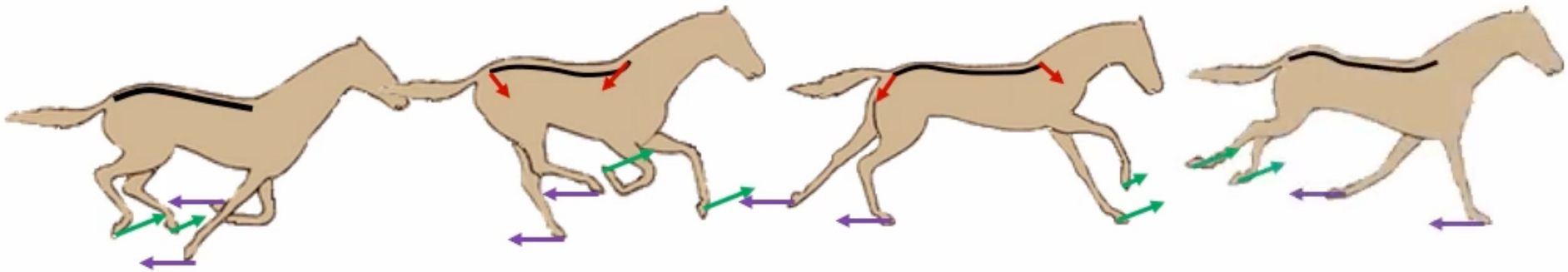
crawl (walk)



trot



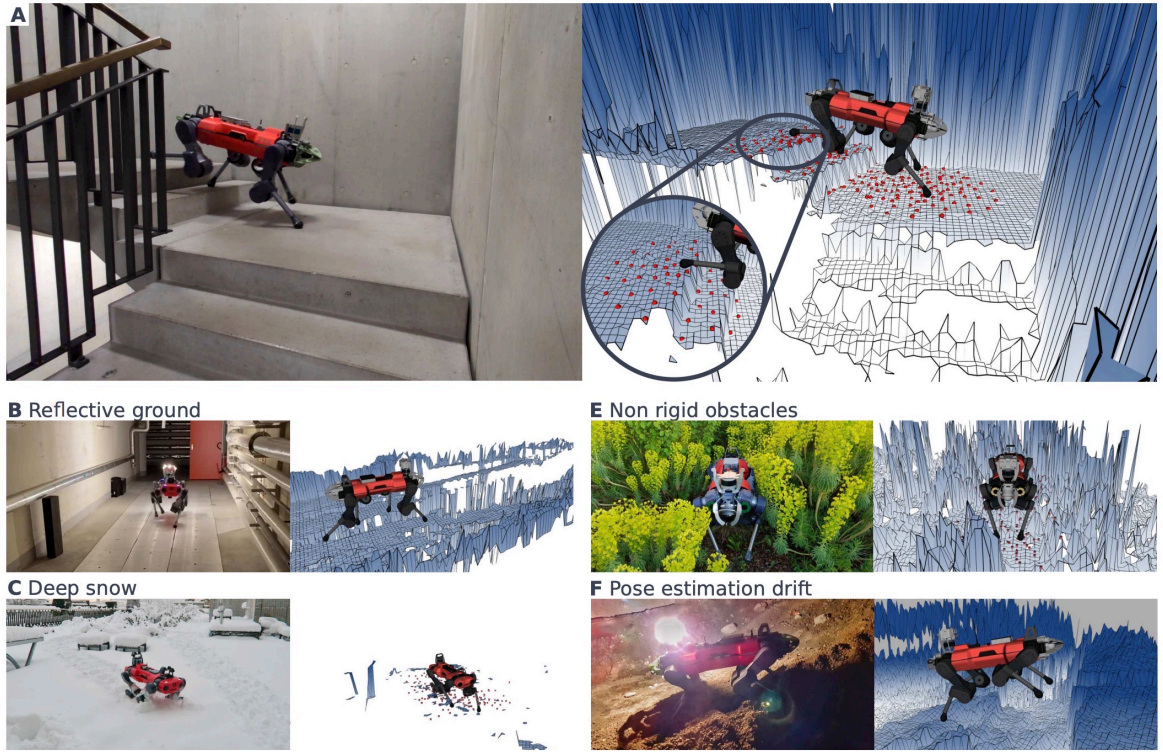
gallop



Challenge and Recent Works

Robust to terrain
(only proprioception)

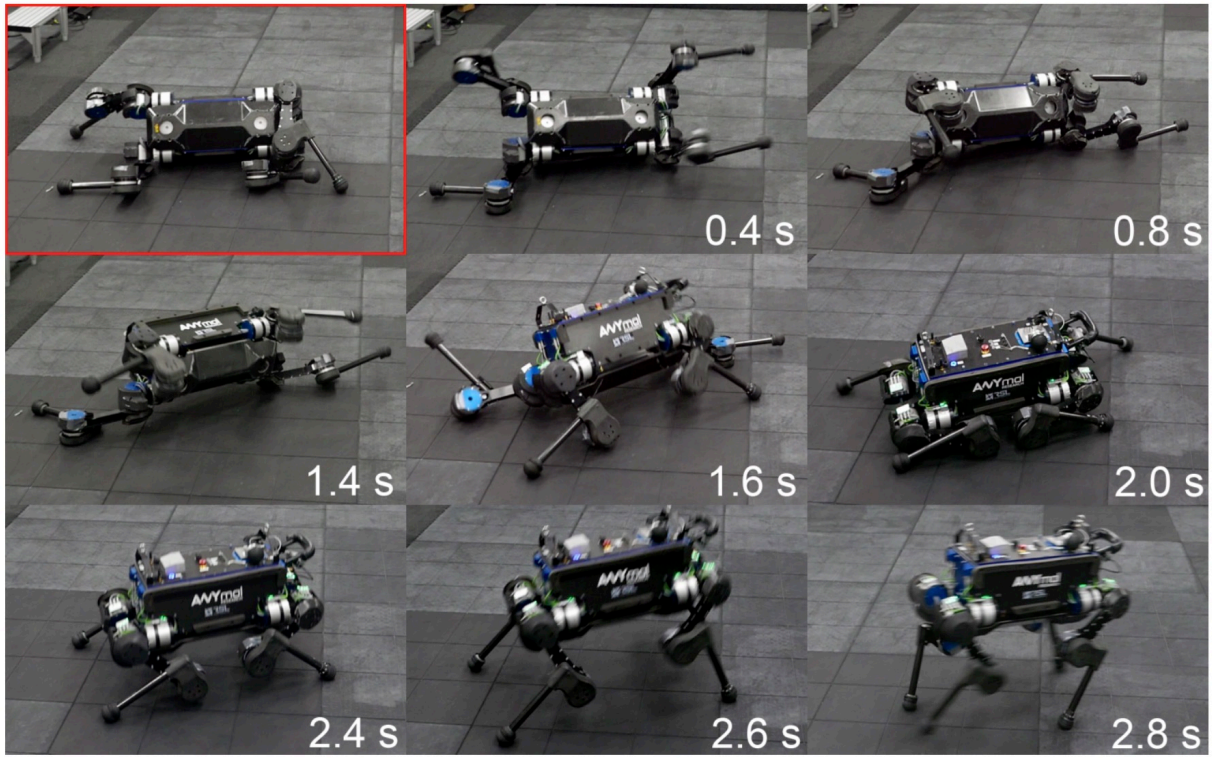
[Lee, Joonho, et al. 2020]



Robust to terrain
(perceive env by vision,radar)

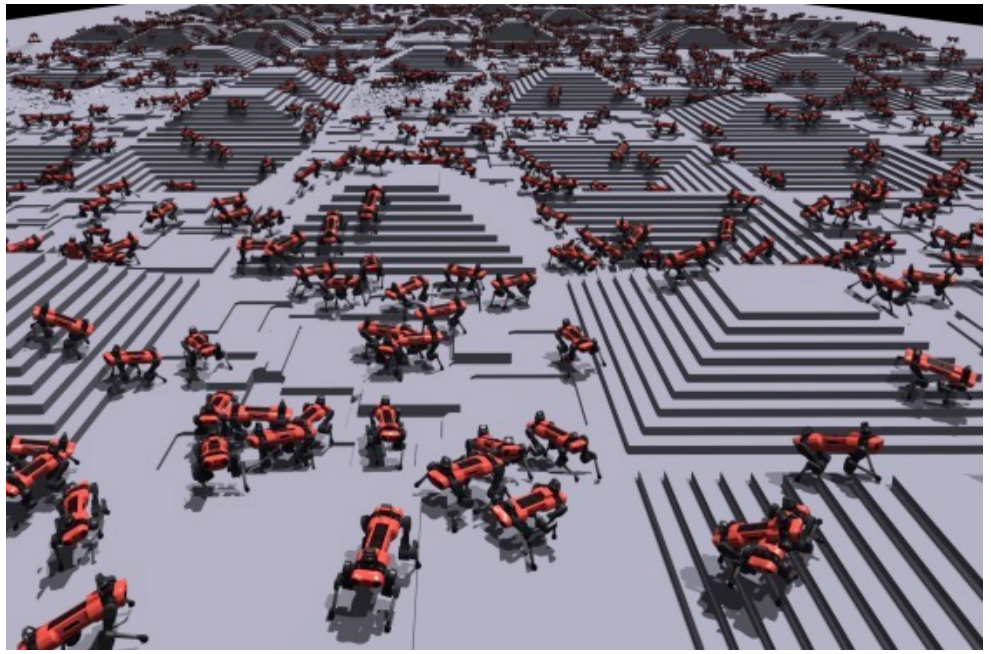
[TAKAHIRO MIKI et al. 2022]

Challenge and Recent Works



Robust to a fall

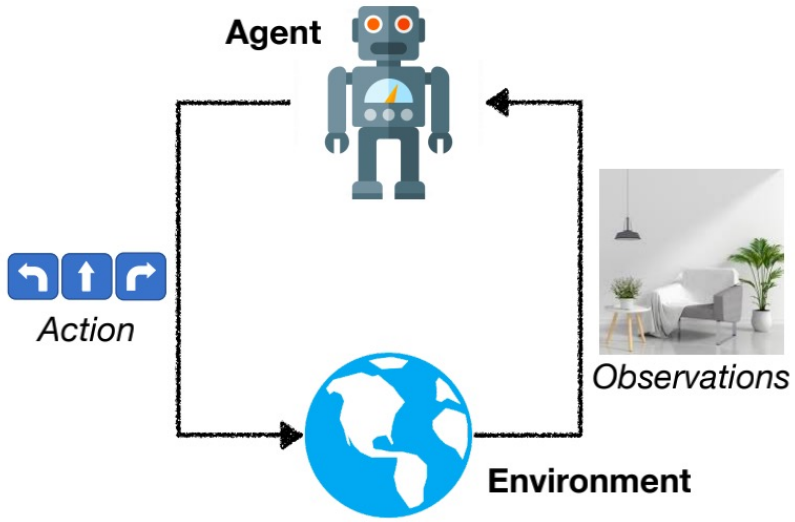
[Lee, Joonho, et al. 2019]



Learn fast

[Nikita Rudin et al. 2020]

Embodied AI & Navigation

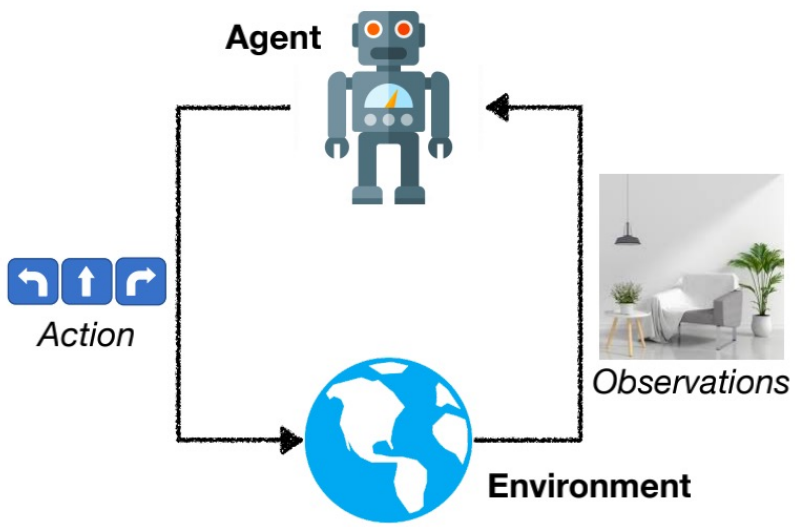


Embodied AI

· Agent physically exists.
· Actions can affect observations

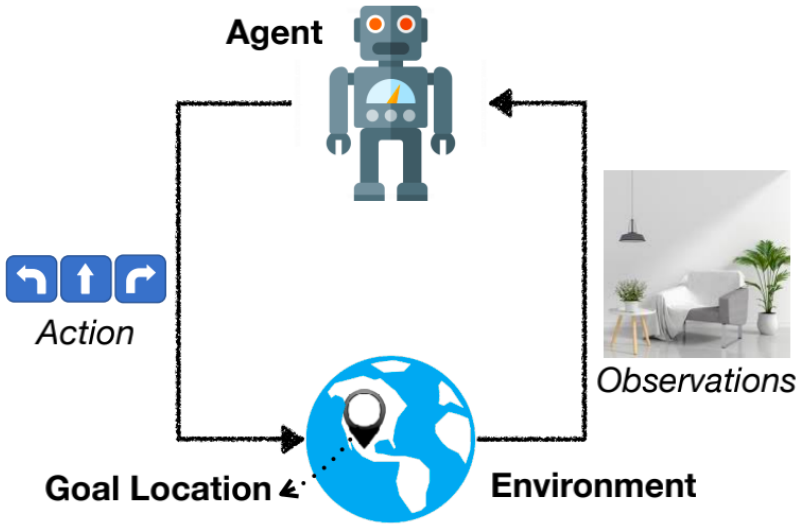
Chaplot, Devendra Singh. "Building Intelligent Autonomous Navigation Agents." Defense of Ph.D. Thesis

Embodied AI & Navigation



Embodied AI

· Agent physically exists.
· Actions can affect observations

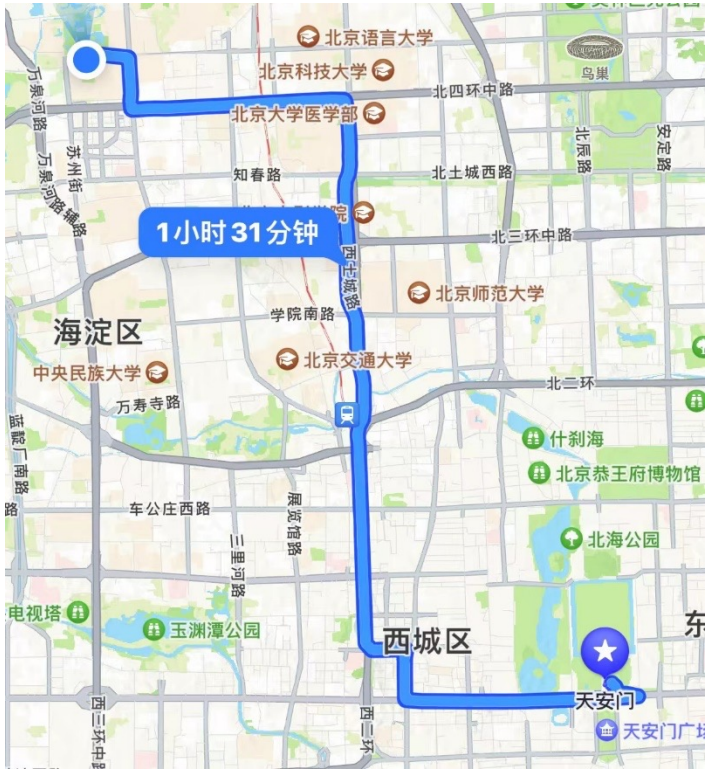


Navigation
A typical task of **Embodied AI**

Chaplot, Devendra Singh. "Building Intelligent Autonomous Navigation Agents." Defense of Ph.D. Thesis

Introduction to Navigation

- **Definition** of **Navigation**: drive the agent to find the target.



Ourdoor GPS

VS

Indoor Navigation

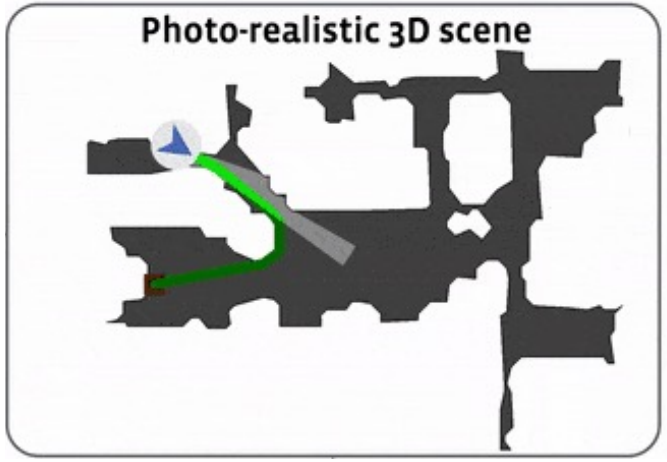
Challenges:

a) Novel scene
b) Noisy indoor positioning
c) Poor scene understanding

Navigation Tasks

- **Definition** of **Navigation**: drive the agent to find the target.

Point Goal

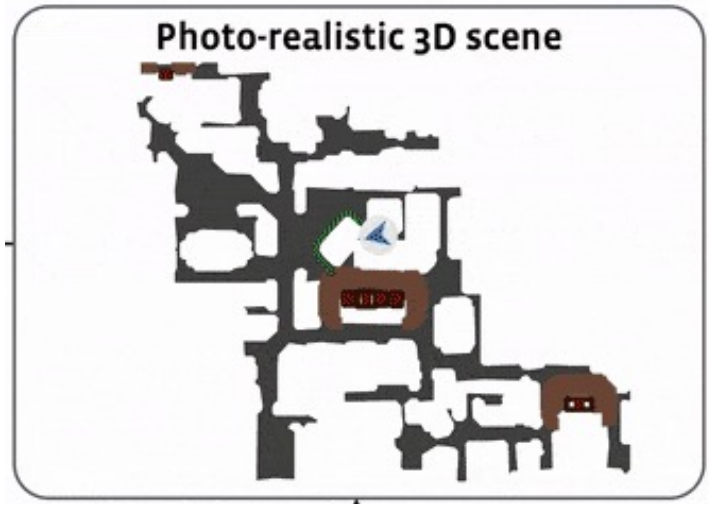


Goal:
Go 5m south, 3m west relative to start.

Image Goal

Goal:
Go where the photo was taken.

Object Goal



Goal:
Go find a sofa.

Two Types of Navigation Methods

Classical Modular Navigation

- ☺ Good generalizability to novel scenes
- ☺ Satisfying performance
- ☹ Hard to implement

vs

End-to-end Reinforcement Learning

- ☺ Easy to implement
- ☺ Satisfying performance
- ☹ Require extensive training time (1k-10k training hours)
- ☹ Poor generalizability to novel scenes

Two Types of Navigation Methods

Classical Modular Navigation

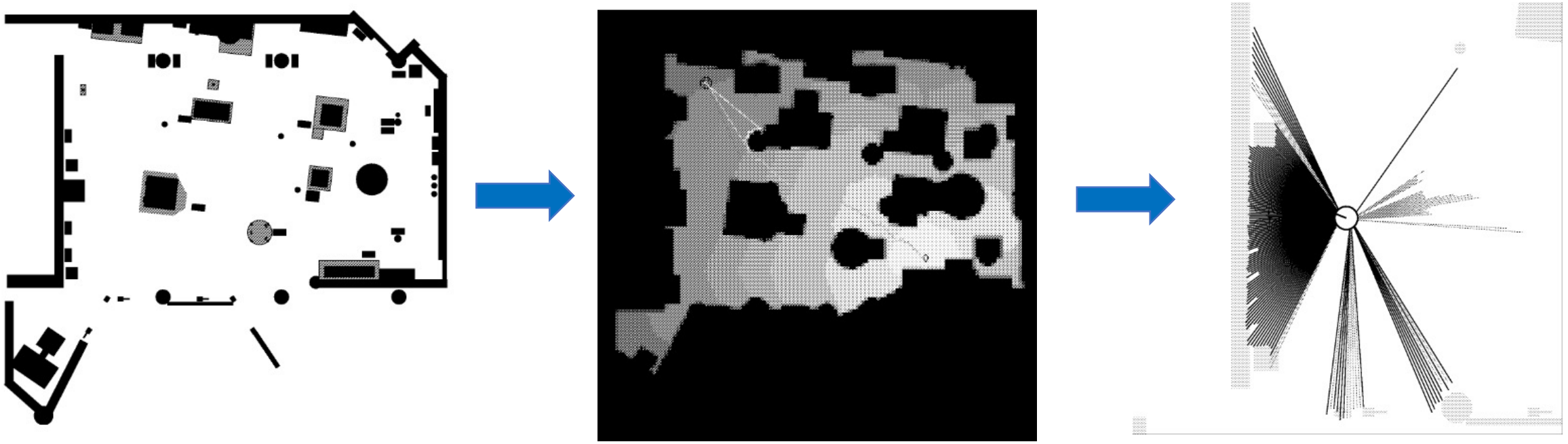
- ☺ Good generalizability to novel scenes
- ☺ Satisfying performance
- ☹ Hard to implement

vs

End-to-end Reinforcement Learning

- ☺ Easy to implement
- ☺ Satisfying performance
- ☹ Require extensive training time (1k-10k training hours)
- ☹ Poor generalizability to novel scenes

Navigation: Classical Modular Navigation



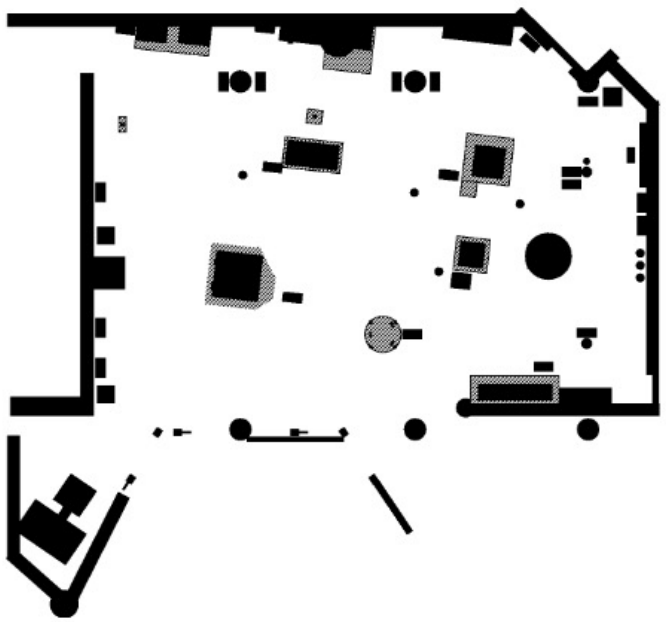
Mapping

Planning

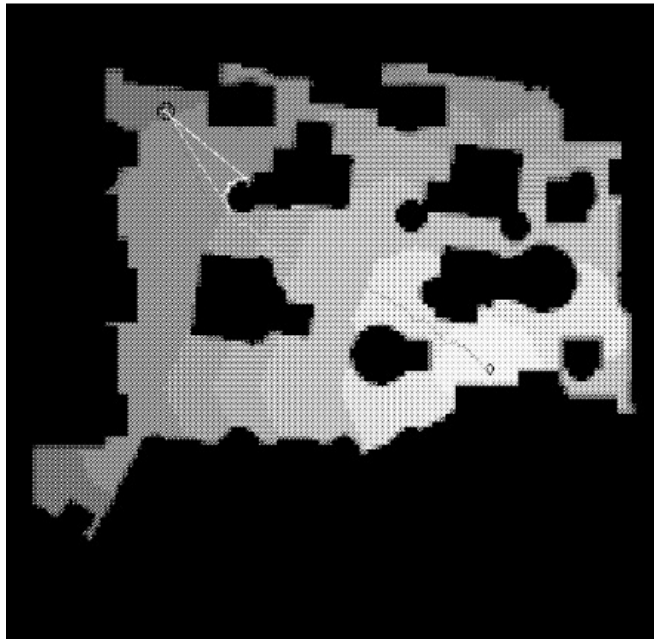
Execution

Burgard, Wolfram, et al. "Experiences with an interactive museum tour-guide robot." Artificial intelligence 114.1-2 (1999): 3-55.

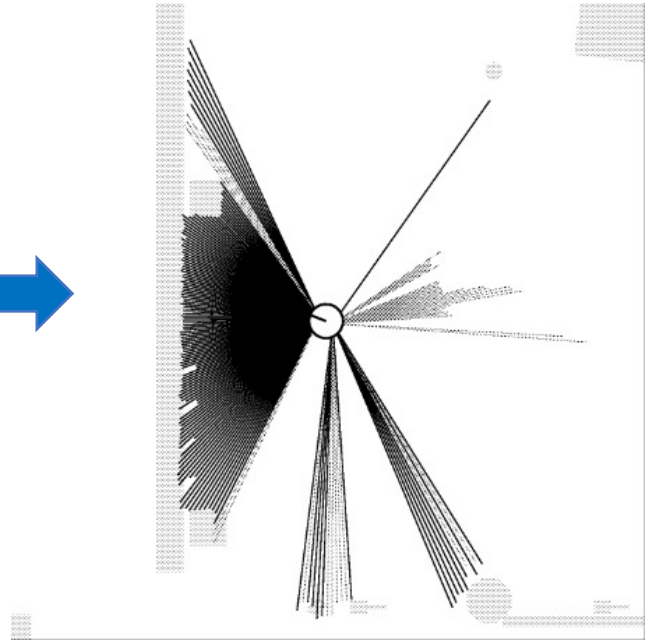
Navigation: Classical Modular Navigation



Mapping



Planning



Execution

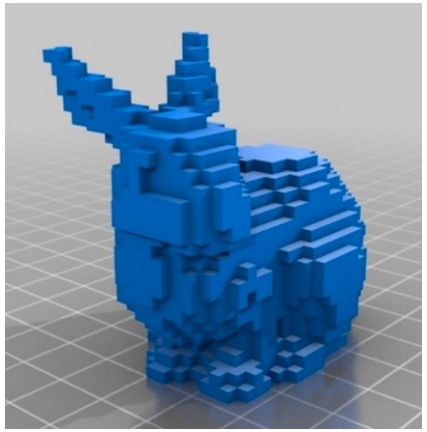
Burgard, Wolfram, et al. "Experiences with an interactive museum tour-guide robot." Artificial intelligence 114.1-2 (1999): 3-55.

Mapping & Representation

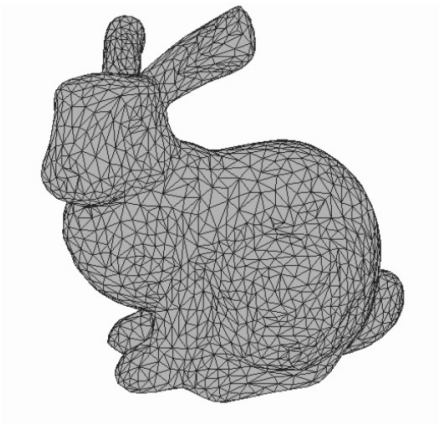
- 3D Representations
  - Explicit: voxel, mesh, point cloud, octomap
  - Implicit: TSDF, voxel hashing

- Mapping Methods
  - Panoptic fusion
  - Kimera (Real-Time Metric-Semantic Localization and Mapping)
  - 3D Dynamic Scene Graphs

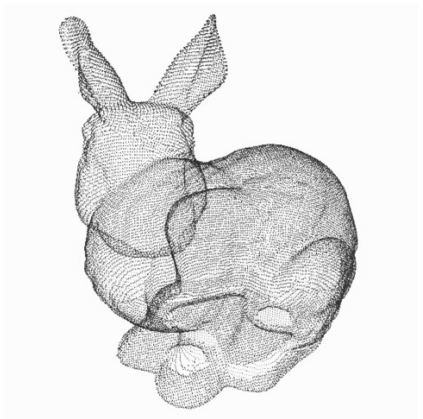
3D Representations: Explicit



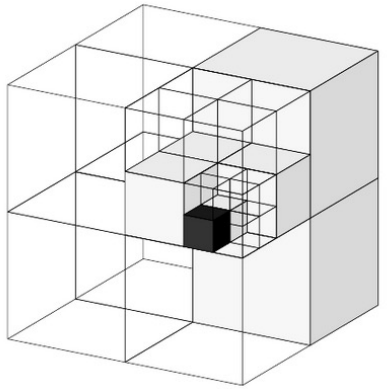
Voxel



Mesh



Point Cloud



OctTree

Mapping Methods

PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things

Gaku Narita, Takashi Seno, Tomoya Ishikawa, Yohsuke Kaji[1]

Panoptic Fusion

IROS 2020

Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping

Antoni Rosinol, Marcus Abate, Yun Chang, Luca Carlone

Kimera

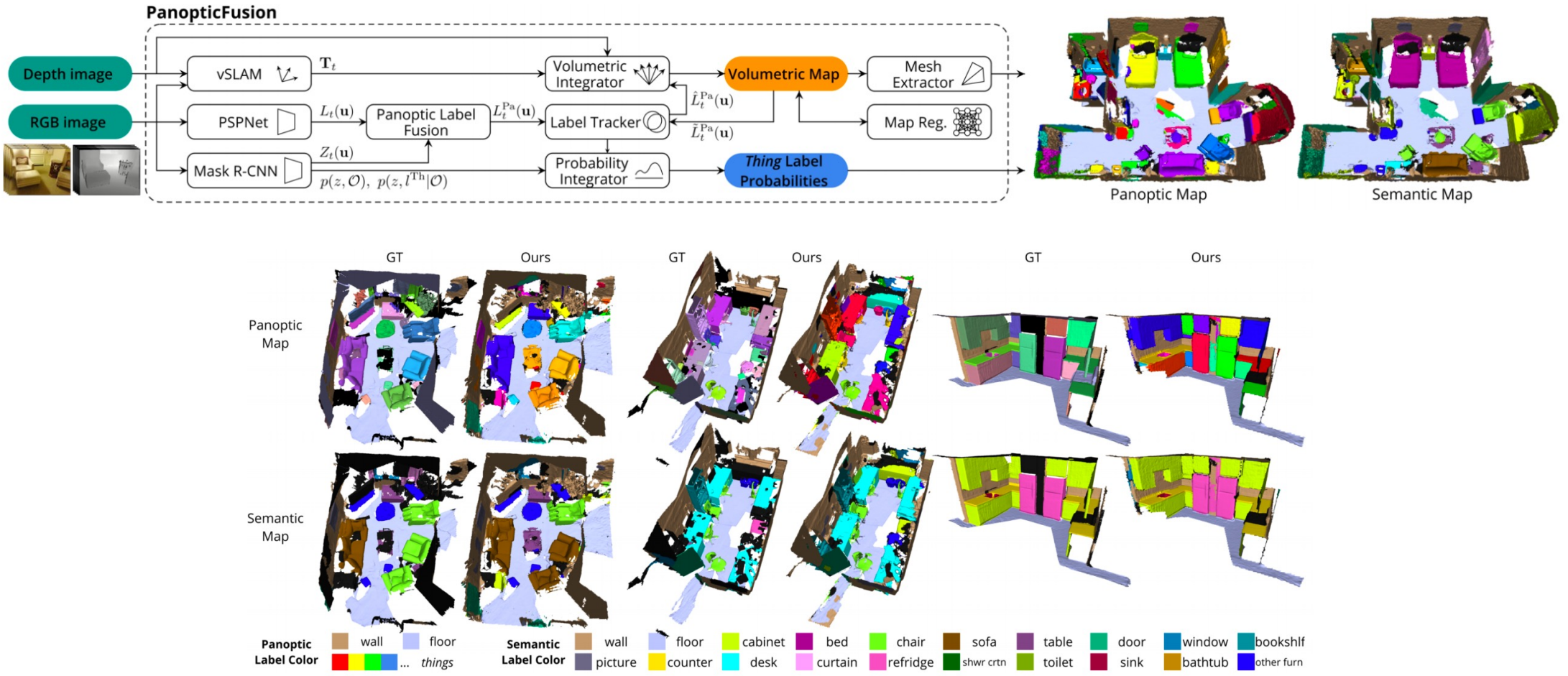
ICRA 2020

3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans

Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, Luca Carlone
Laboratory for Information & Decision Systems (LIDS)
Massachusetts Institute of Technology
{arosinol,agupta,mabate,jnshi,lcarlone}@mit.edu

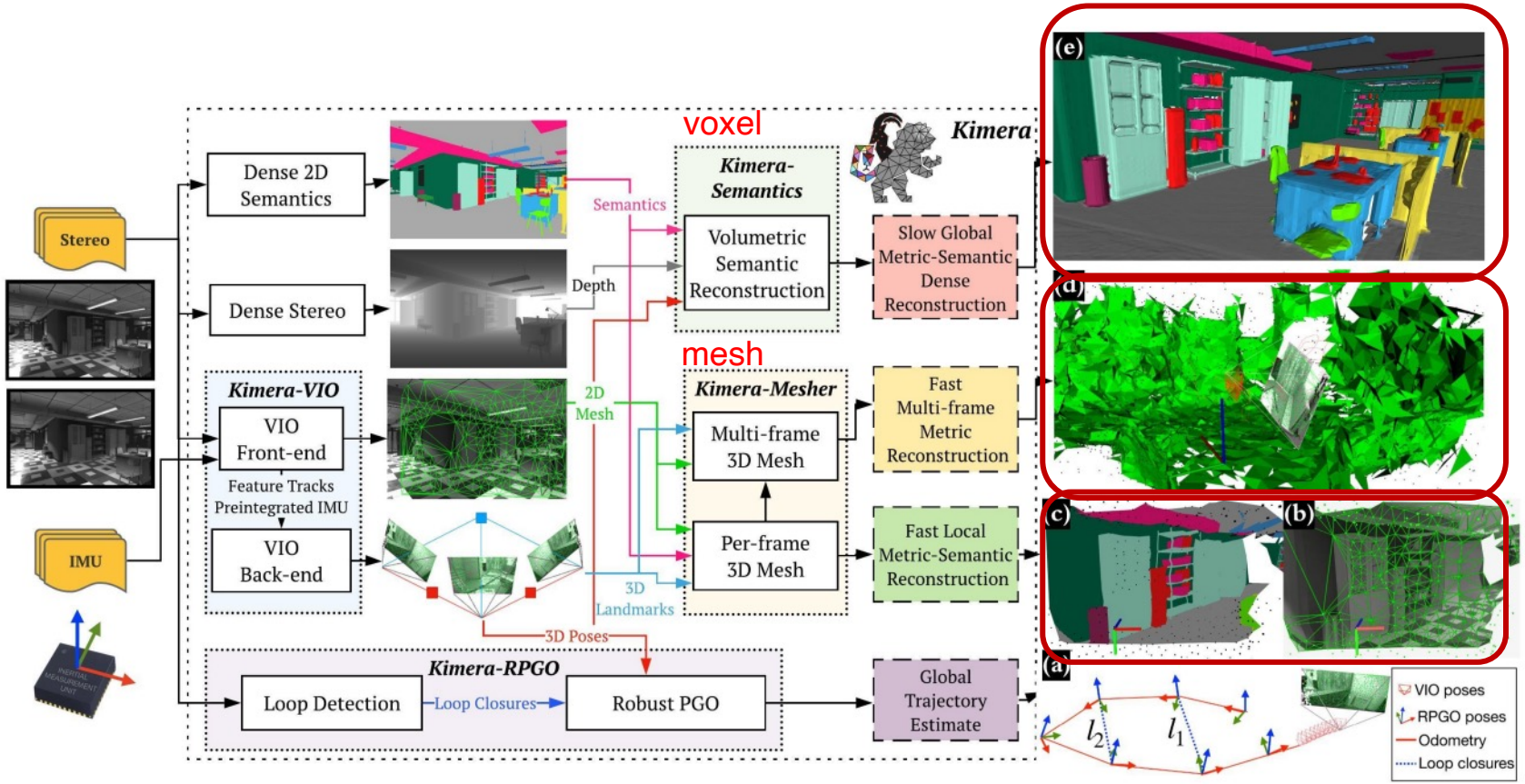
3D Scene Graph

Mapping Methods: Panoptic Fusion



Narita, Gaku, et al. "Panopticfusion: Online volumetric semantic mapping at the level of stuff and things." 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019.

Mapping Methods: Kimera



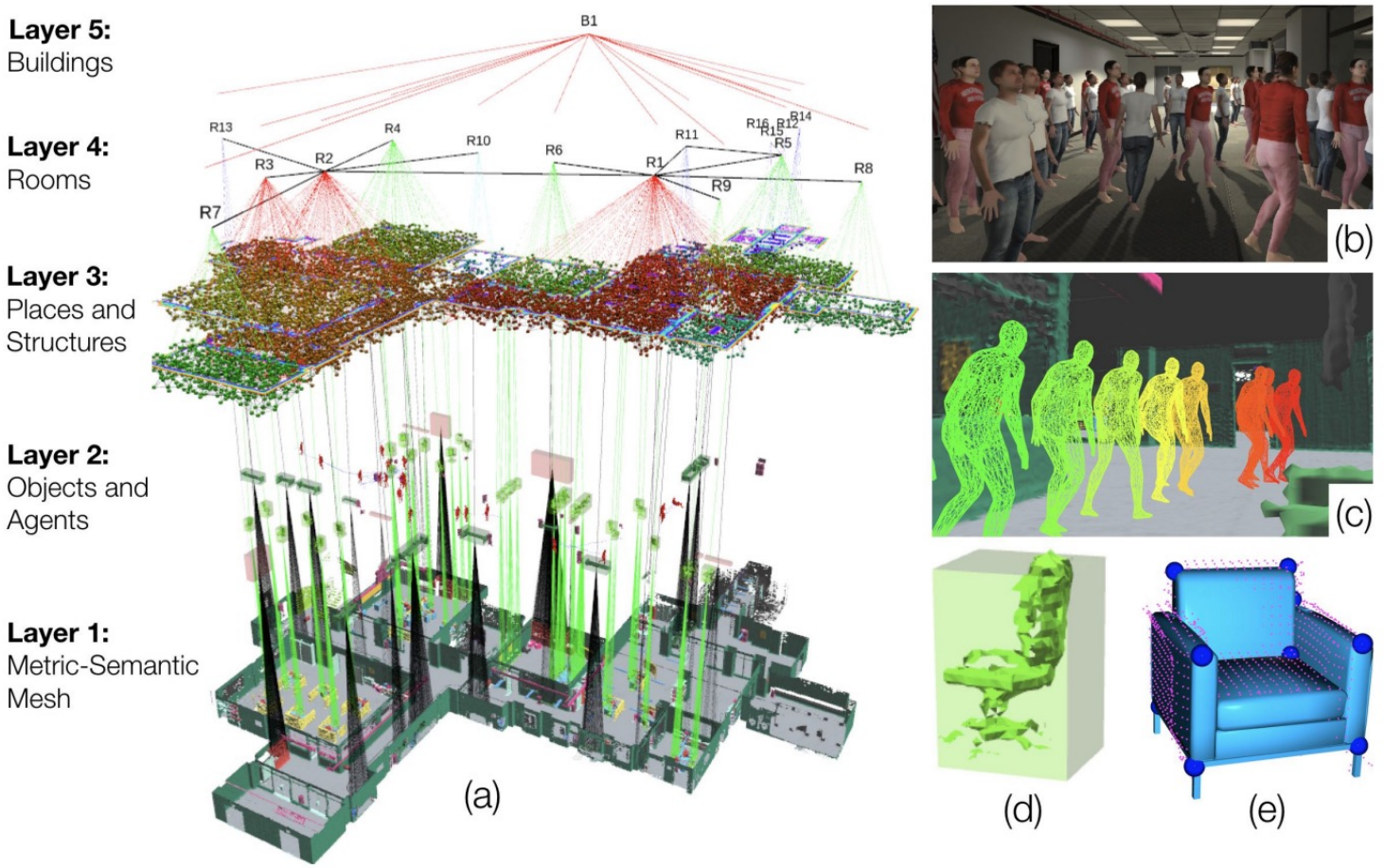
Panoptic, Slow

Panoptic, Fast without Semantic

Local, Fast

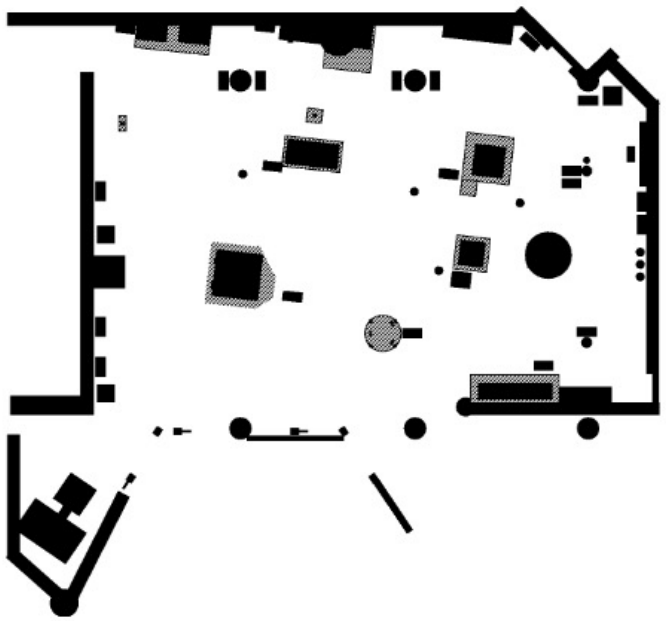
Rosinol, Antoni, et al. "Kimera: an open-source library for real-time metric-semantic localization and mapping." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

Mapping Methods: 3D Dynamic Scene Graphs

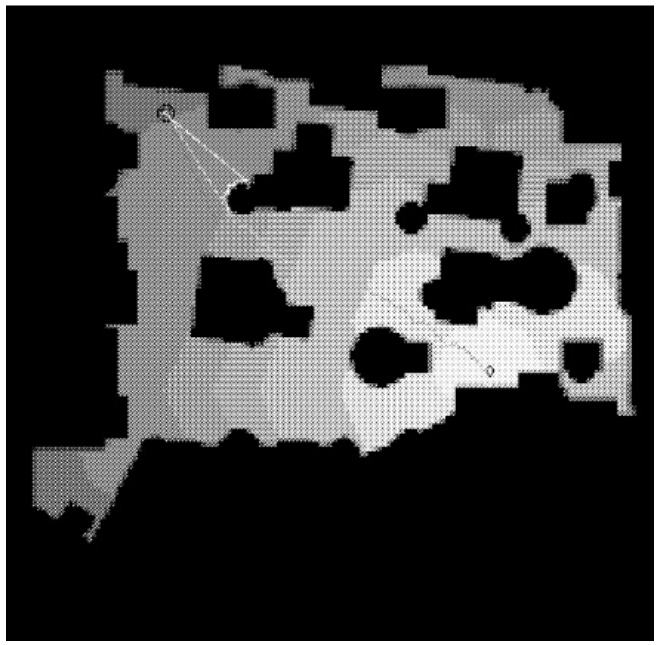


Rosinol, Antoni, et al. "3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans." arXiv preprint arXiv:2002.06289 (2020).

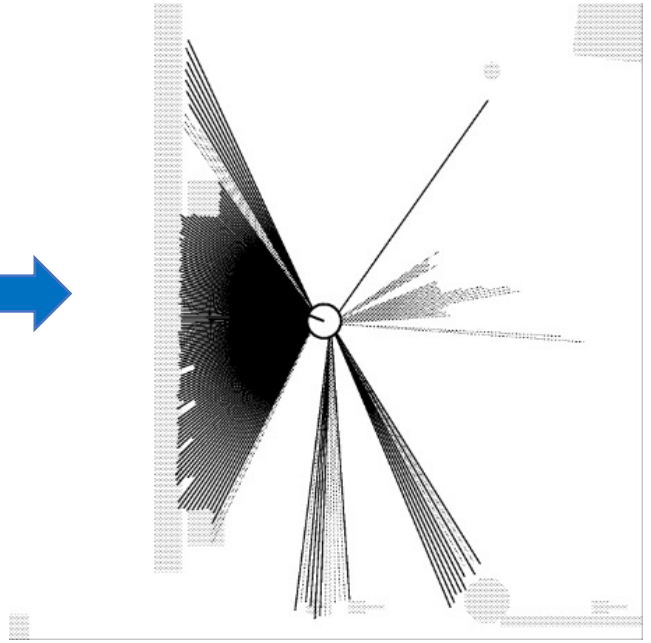
Navigation: Classical Modular Navigation



Mapping



Planning

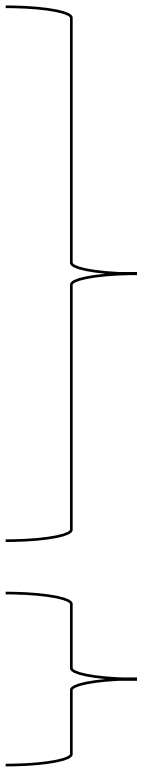


Execution

Burgard, Wolfram, et al. "Experiences with an interactive museum tour-guide robot." Artificial intelligence 114.1-2 (1999): 3-55.

Robot Control: Path Planning

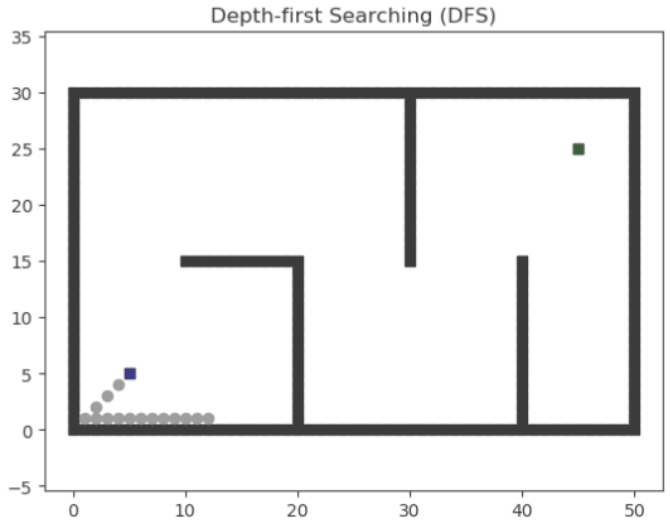
- Search-based Methods
  - DFS & BFS
  - A* & variations
- Sample-based Methods
  - RRT & variations
  - Fast Marching
- Learning-based Methods
  - Imitation Learning



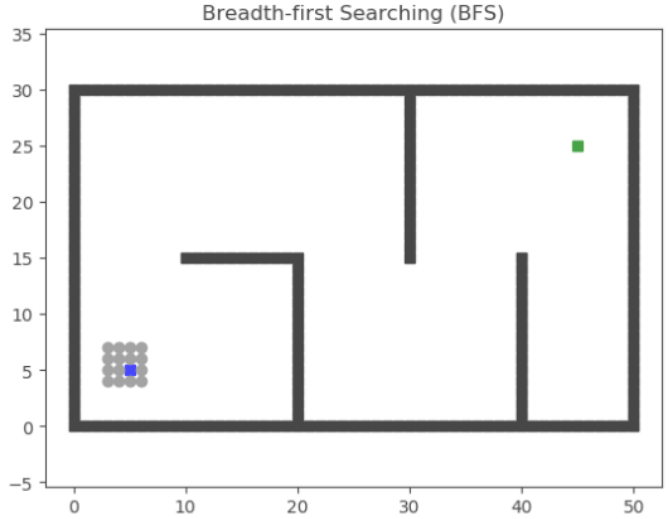
Rule-based
local planner

Learning-based
non-local
planner

Path Planning: Search-based Methods

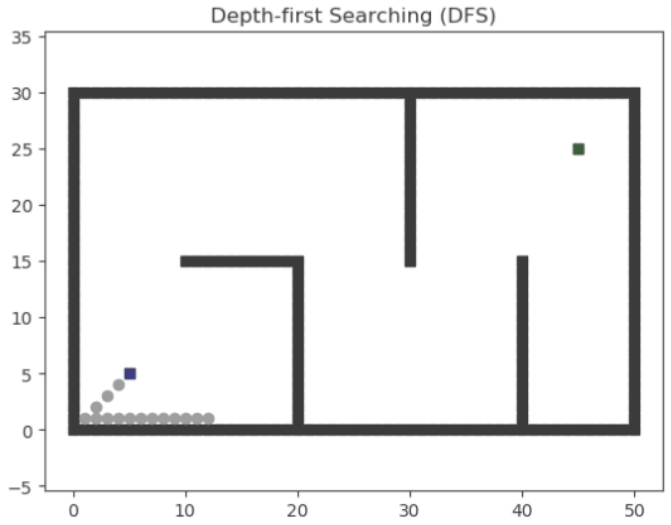


DFS

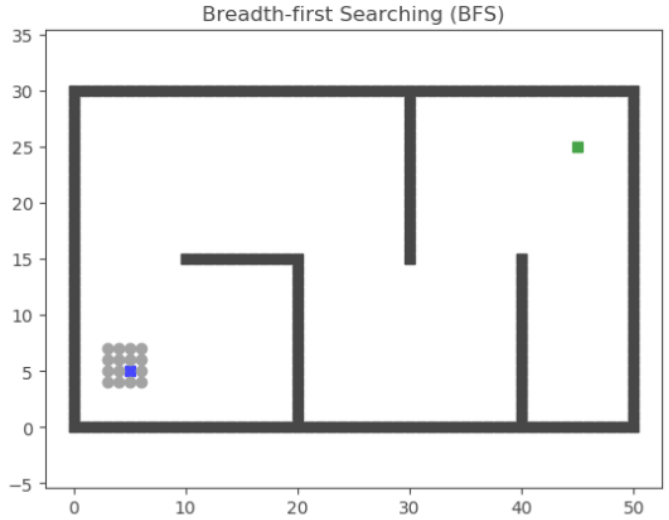


BFS

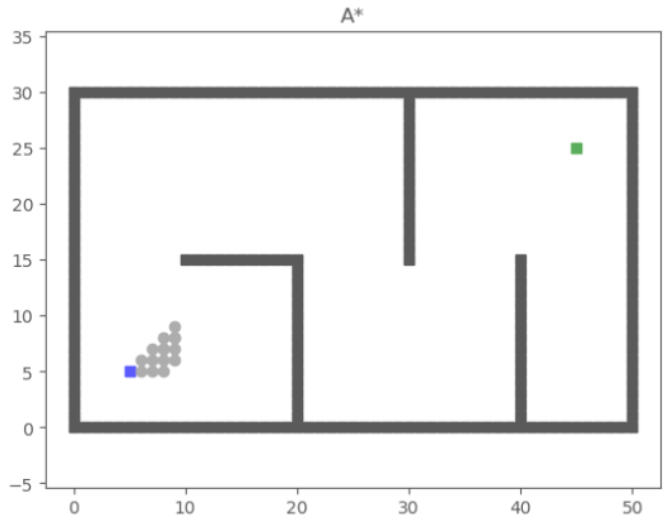
Path Planning: Search-based Methods



DFS



BFS

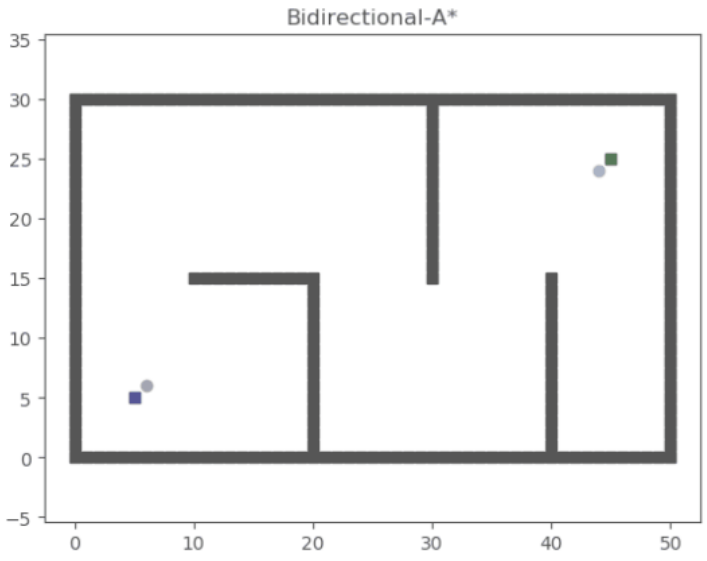


A*

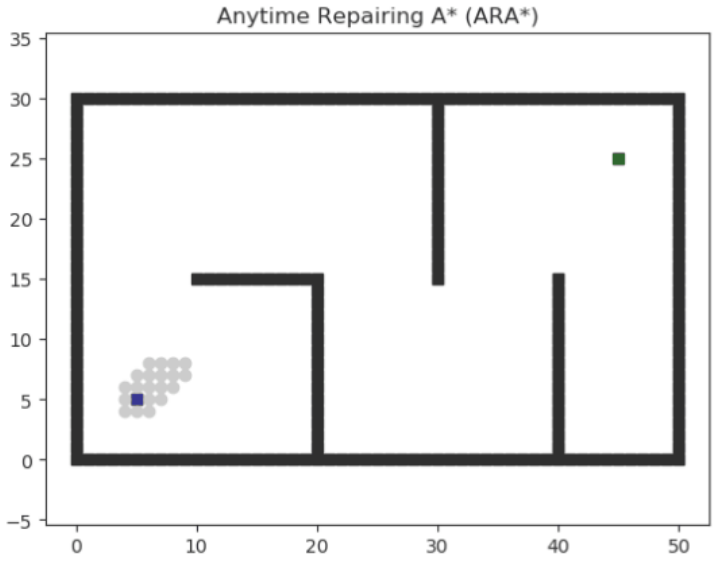
$$f(i) = g(i) + h(i)$$

- Reduce the Complexity by Cutting Branches
- Still Waste Time Searching

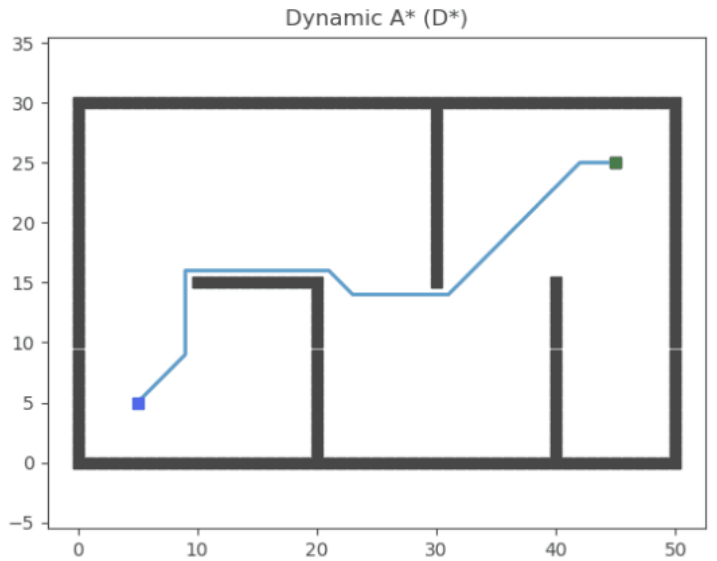
Path Planning: Search-based Methods



Bidirectional A*



Anytime Repairing A*



Dynamic A* (D*)

Path Planning: Search-based Methods

Occupancy Anticipation for Efficient Exploration and Navigation

PointNav Challenge Leaderboard (sorted by SPL)

Rank	Team	SPL	SOFT_SPL	DISTANCE_TO_GOAL	SUCCESS
1	OccupancyAnticipation	0.21	0.50	2.29	0.28
2	ego-localization	0.15	0.60	1.82	0.19
3	DAN	0.13	0.24	4.00	0.25
4	Information Bottleneck	0.06	0.43	2.72	0.09
5	cogmodel_team	0.01	0.33	4.27	0.01
6	UCULab	0.001	0.11	5.97	0.002

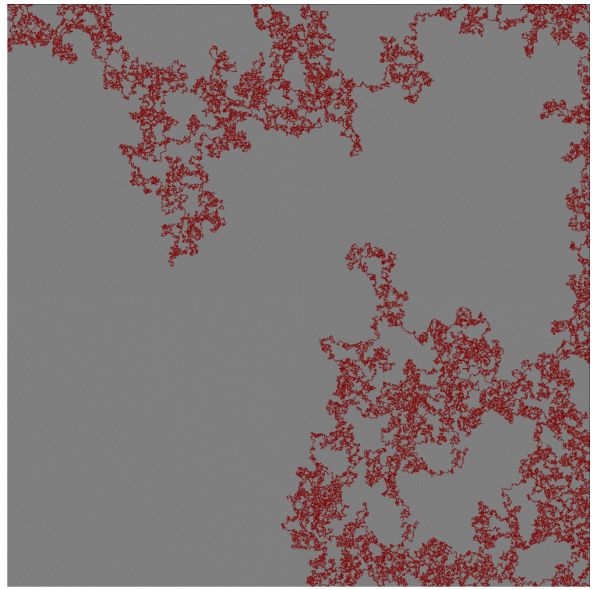
Winner of Habitat Challenge 2020 PointNav

README.md

build passing | coverage 100% | pypi package 1.0.6

PyAstar2D

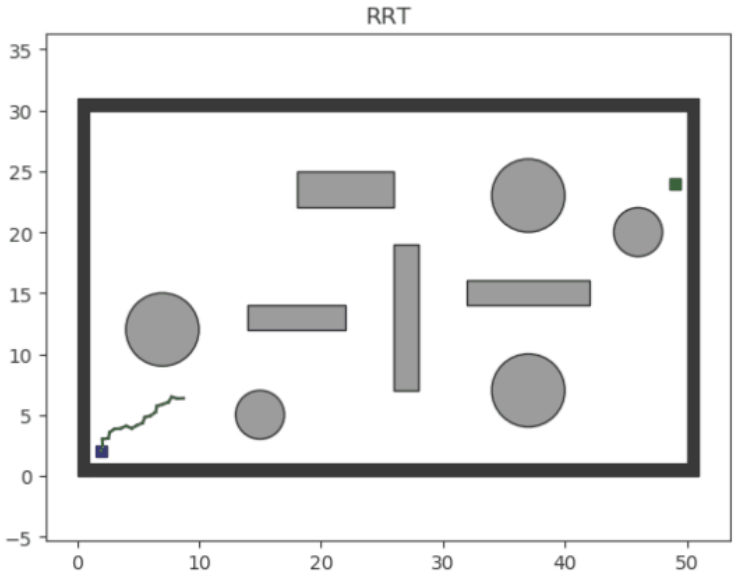
This is a very simple C++ implementation of the A* algorithm for pathfinding on a two-dimensional grid. The solver itself is implemented in C++, but is callable from Python. This combines the speed of C++ with the convenience of Python.



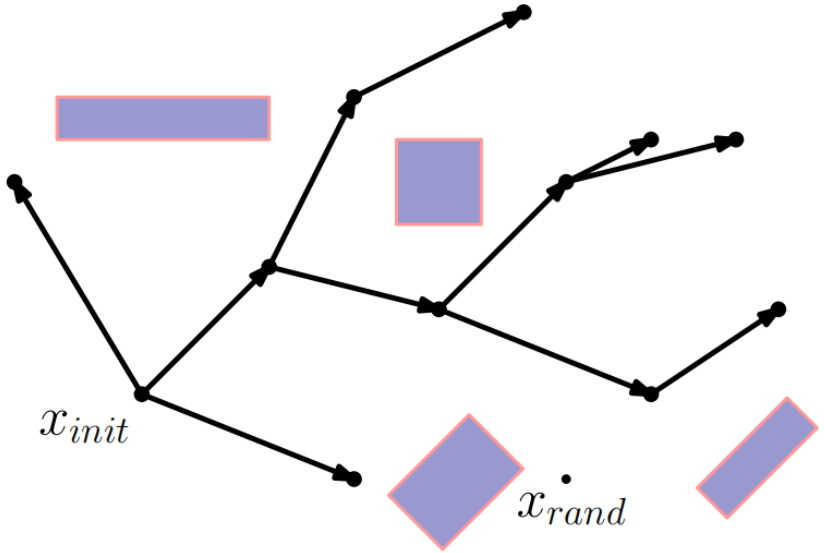
https://github.com/hjweide/pyastar2d

Ramakrishnan, Santhosh K., Ziad Al-Halah, and Kristen Grauman. "Occupancy anticipation for efficient exploration and navigation." European Conference on Computer Vision. Springer, Cham, 2020.

Path Planning: Sample-based Methods

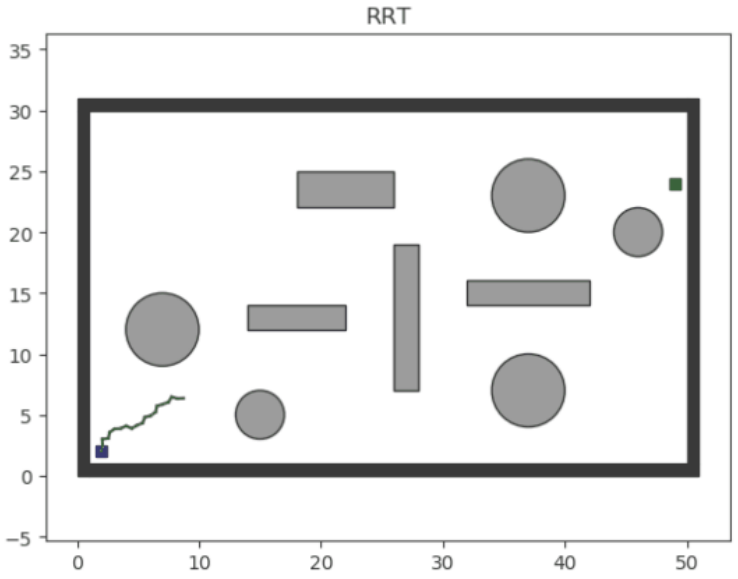


RRT

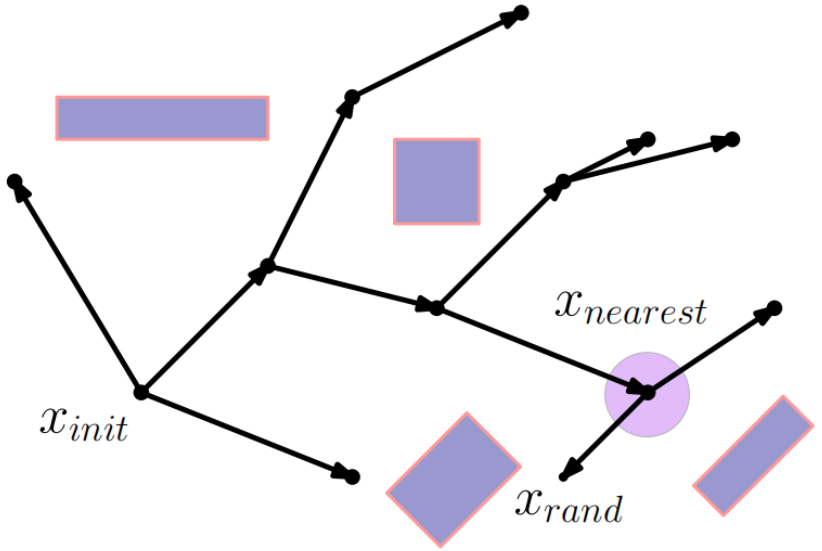


random sampling

Path Planning: Sample-based Methods

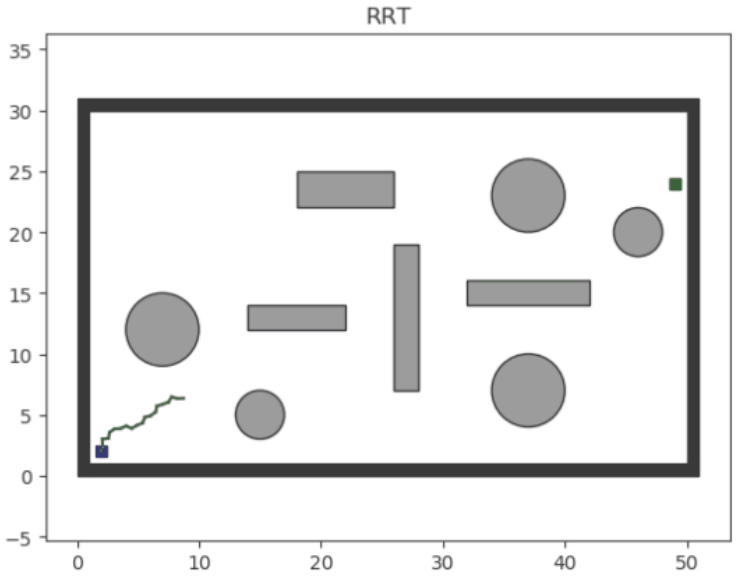


RRT

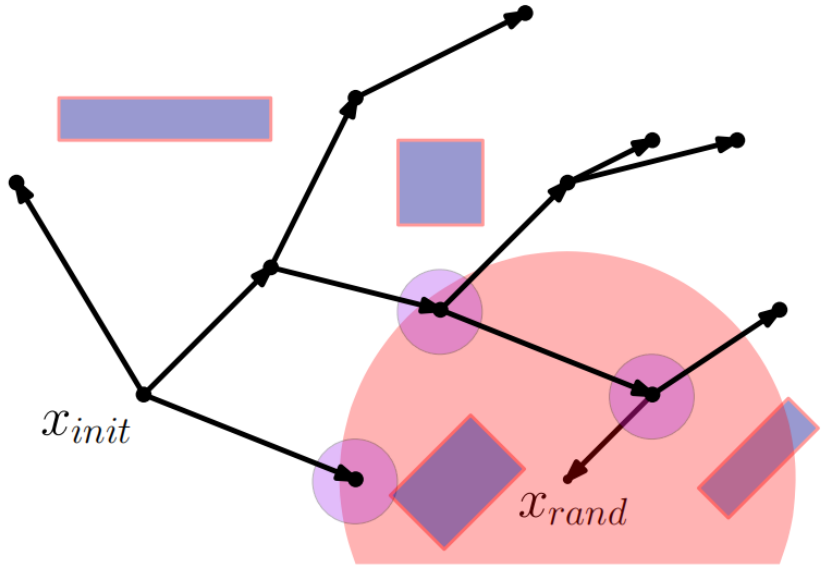


find the nearest point

Path Planning: Sample-based Methods

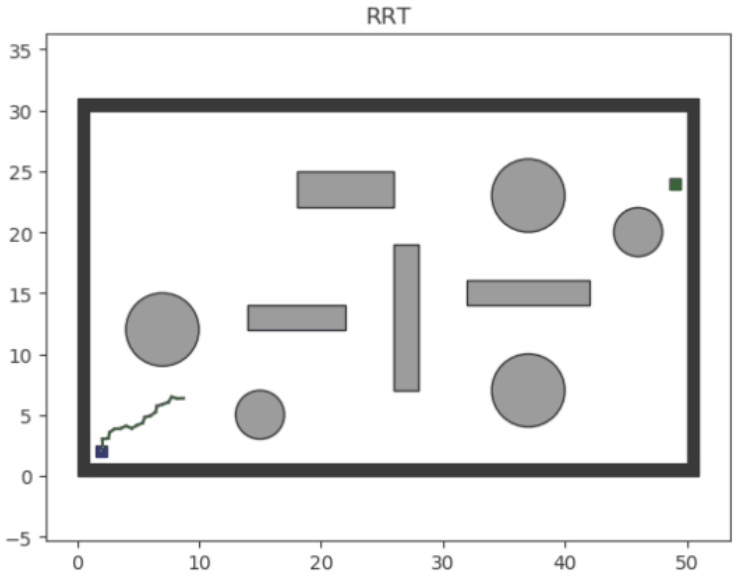


RRT

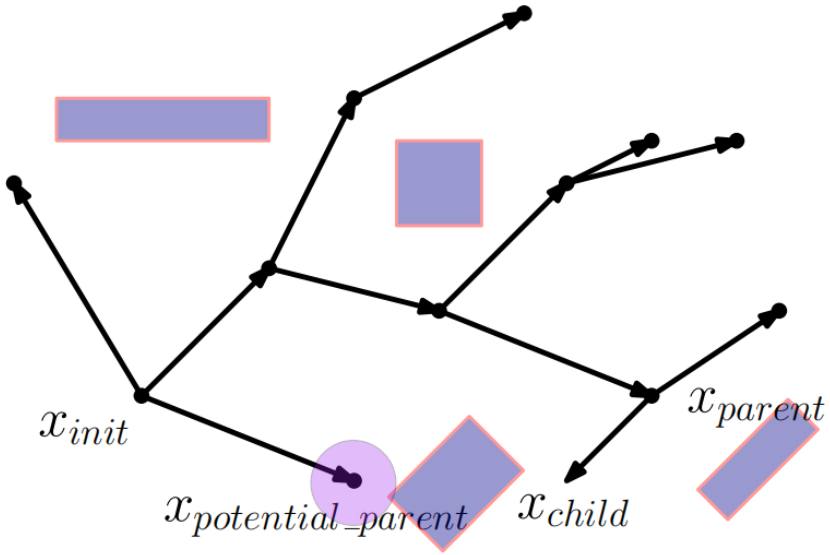


find potential parent-nodes

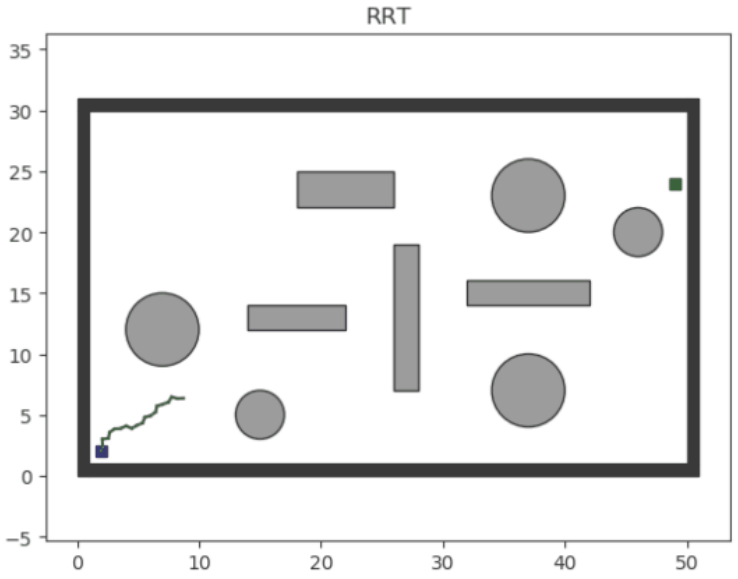
Path Planning: Sample-based Methods



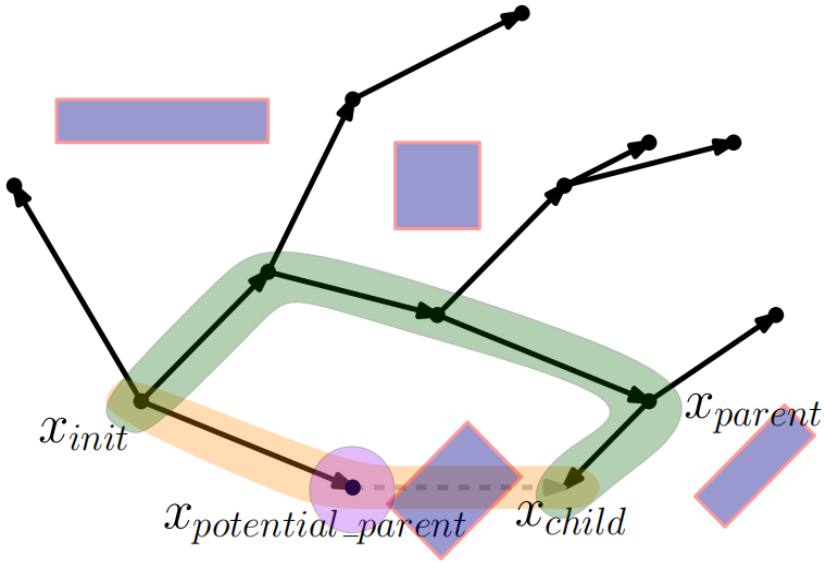
RRT



Path Planning: Sample-based Methods

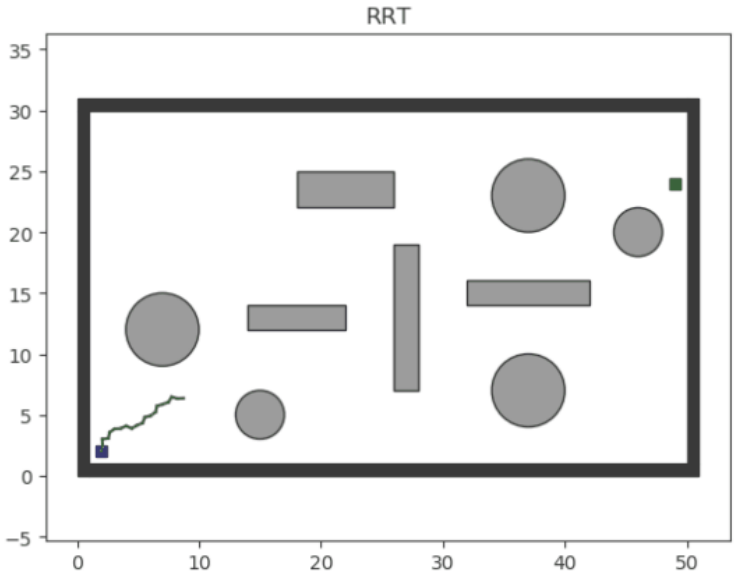


RRT

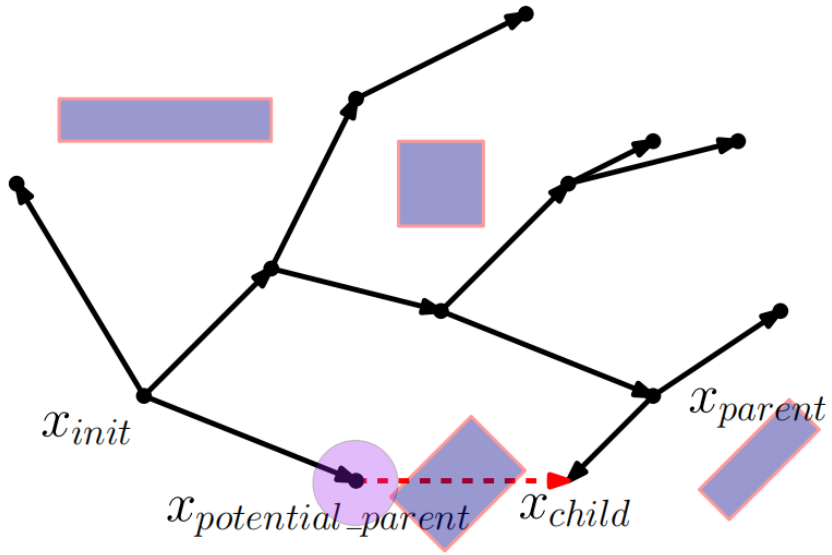


compare path &
collision detection

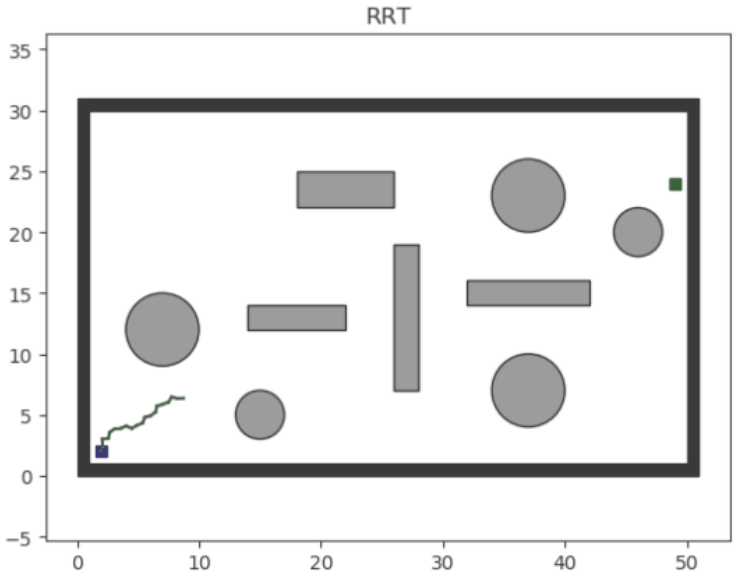
Path Planning: Sample-based Methods



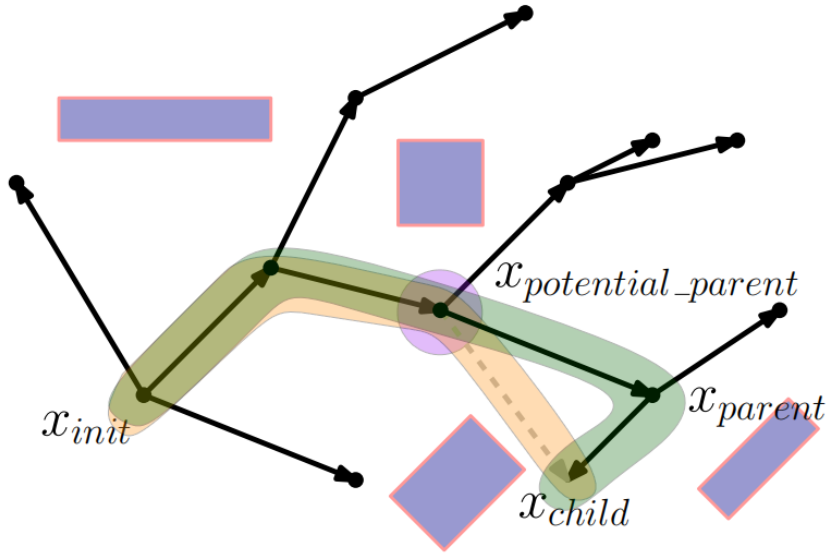
RRT



Path Planning: Sample-based Methods

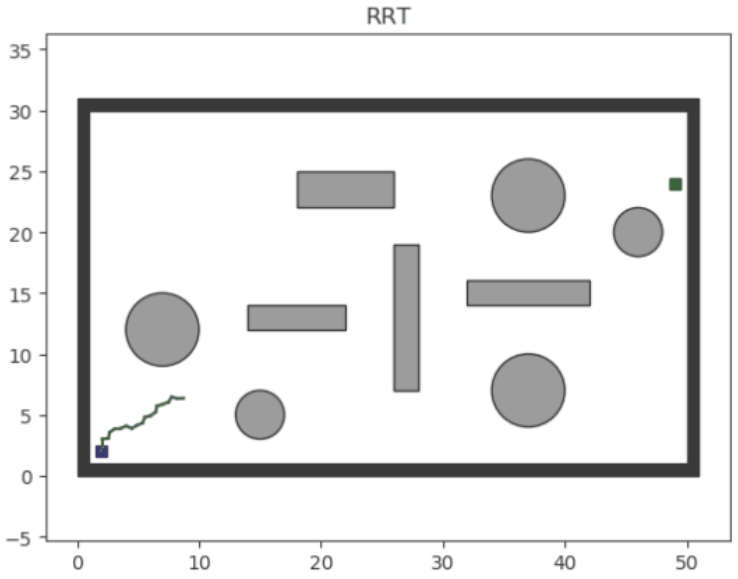


RRT

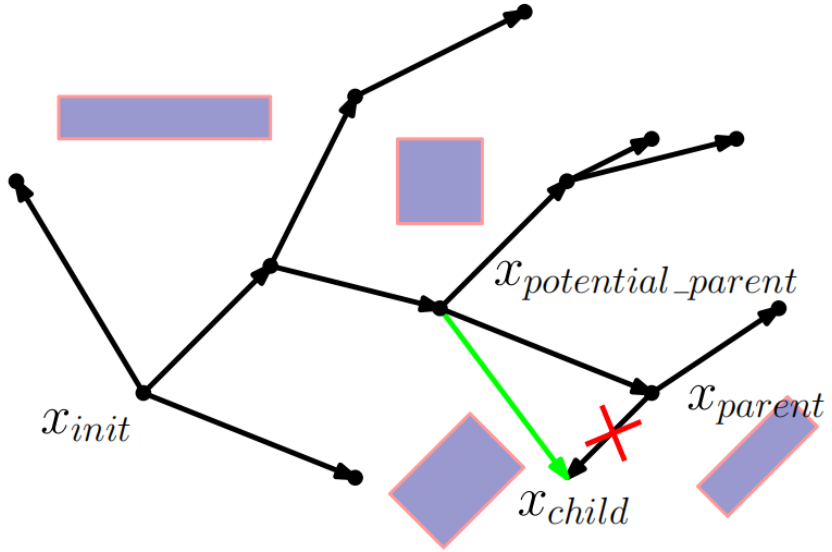


compare path &
collision detection

Path Planning: Sample-based Methods

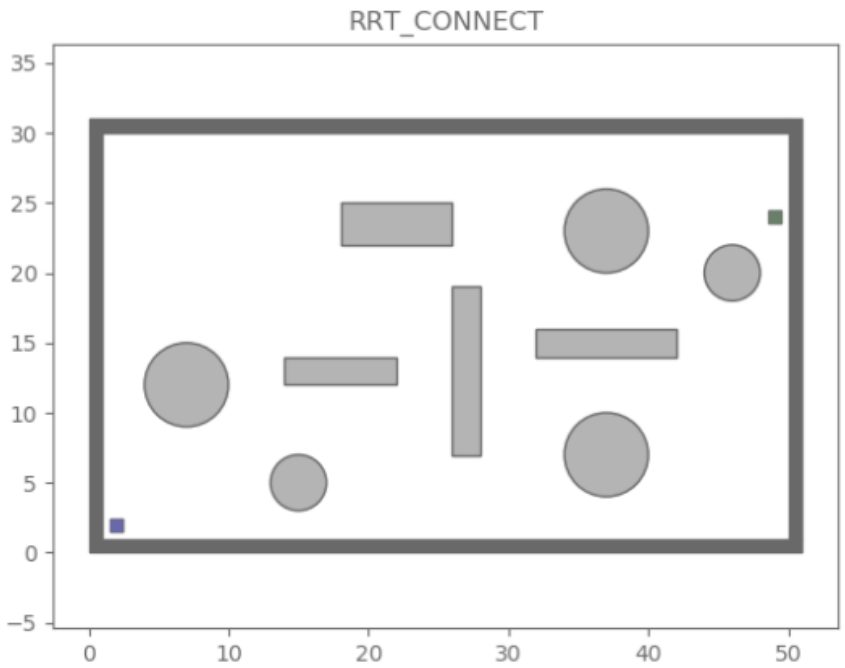


RRT

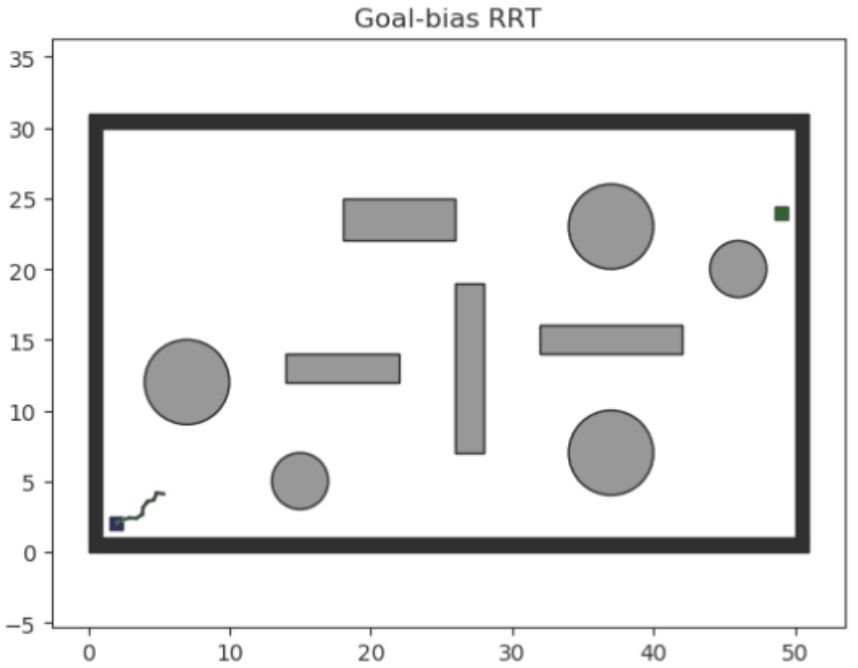


adjust the path

Path Planning: Sample-based Methods



RRT Connect



Goal-bias RRT

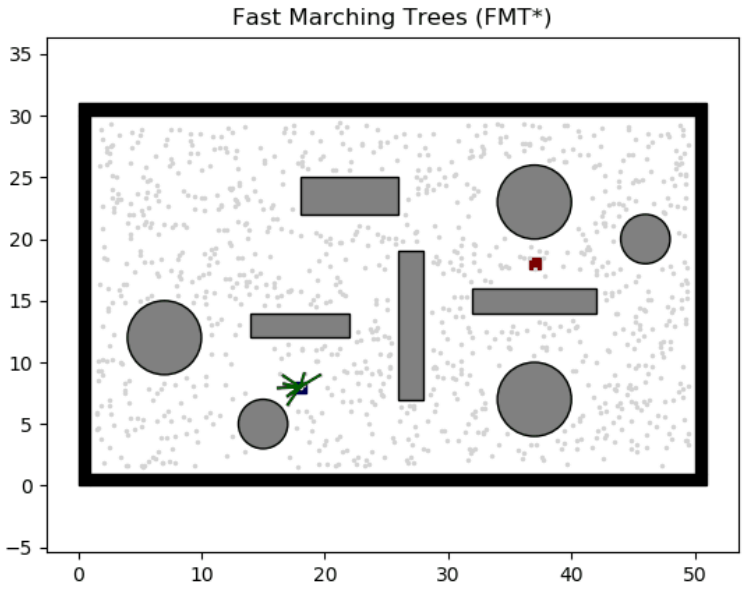
Path Planning: Sample-based Methods

A fast marching level set method for monotonically advancing fronts

J. A. SETHIAN

Department of Mathematics, University of California, Berkeley, CA 94720

Communicated by Alexandre J. Chorin, University of California, Berkeley, CA, November 16, 1995 (received for review October 20, 199



Fast Marching

Mostly used method in navigation

LEARNING TO EXPLORE USING ACTIVE NEURAL SLAM

Devendra Singh Chaplot[1†], Dhiraj Gandhi[2], Saurabh Gupta[3*], Abhinav Gupta [1,2*], Ruslan Salakhutdinov[1*]
[1]Carnegie Mellon University, [2]Facebook AI Research, [3]UIUC

SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency

Devendra Singh Chaplot[1*], Murtaza Dalal[2], Saurabh Gupta[3], Jitendra Malik[1,4], Ruslan Salakhutdinov[2],
[1]Facebook AI Research, [2]Carnegie Mellon University, [3]UIUC, [4]UC Berkeley

PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning

Santhosh Kumar Ramakrishnan[1,2], Devendra Singh Chaplot[1], Ziad Al-Halah[2], Jitendra Malik[1,3], Kristen Grauman[1,2]
[1]Meta AI [2]UT Austin [3]UC Berkeley

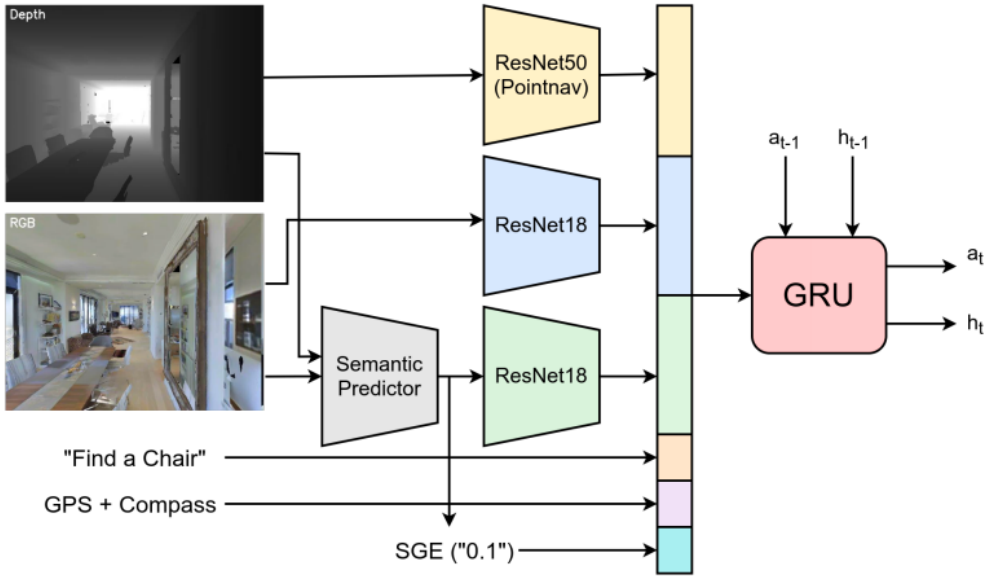
Active Nerual SLAM
The Winner of
Habitat Challenge 2019

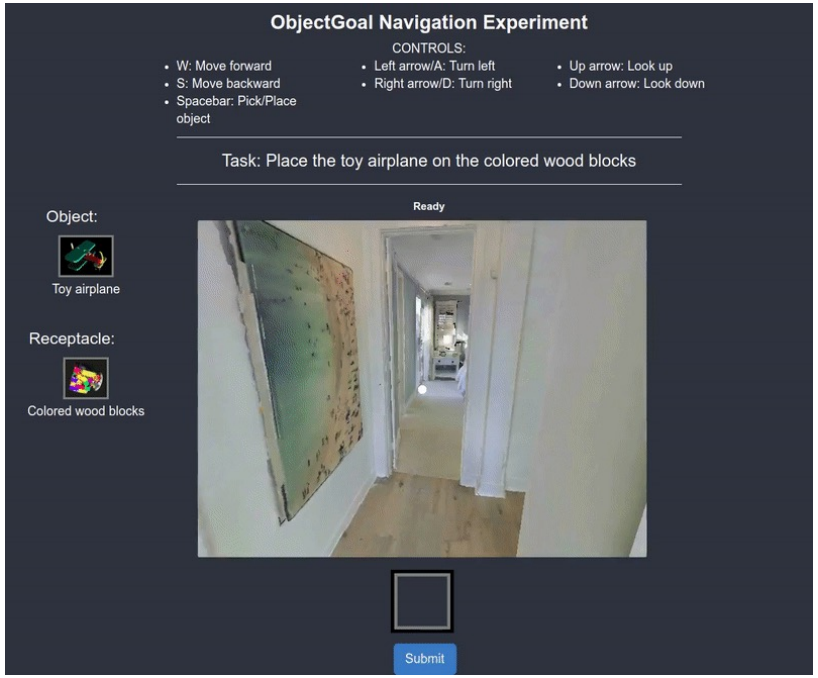
SEAL
Baseline of the Winner of
Habitat Challenge 2020

PONI
SOTA of ObjectNav 2022

Sethian, James A. "A fast marching level set method for monotonically advancing fronts." Proceedings of the National Academy of Sciences 93.4 (1996): 1591-1595.

Path Planning: Trajectory Imitation Learning

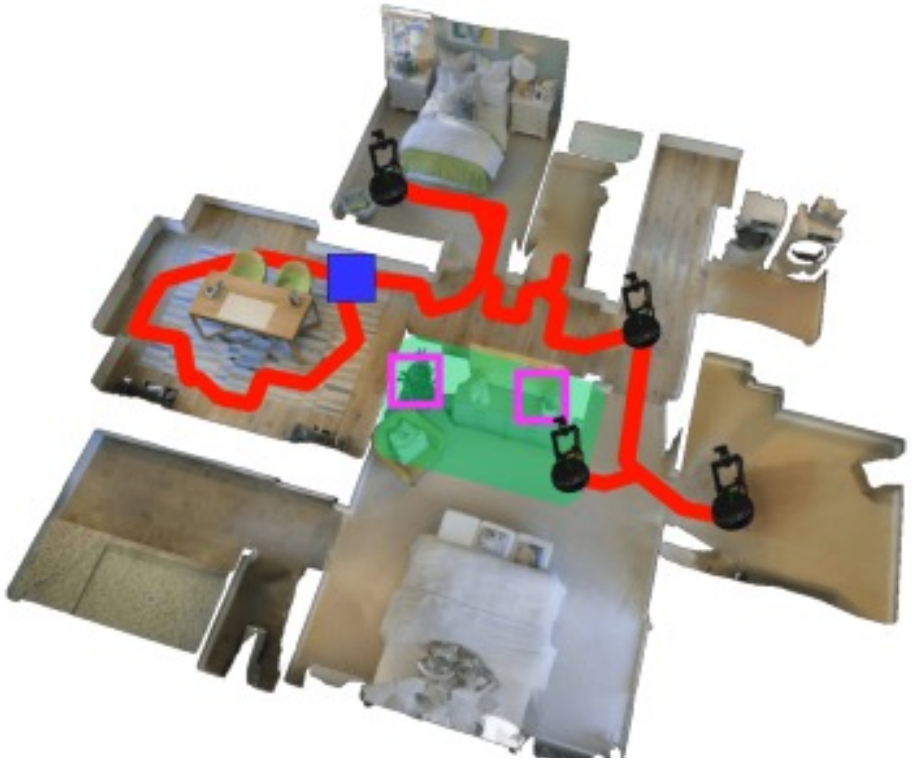


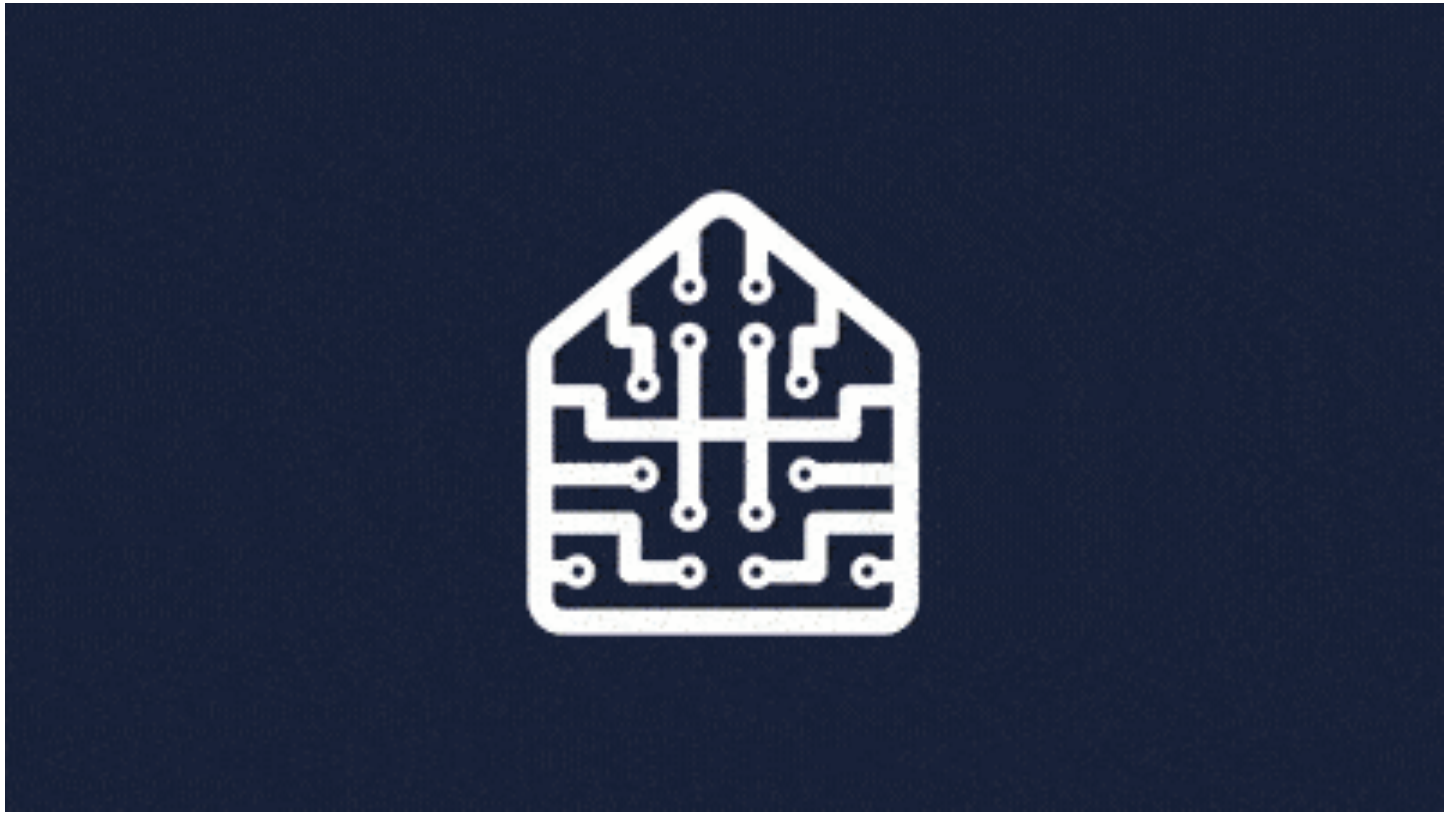


Path Collecting System

Ramrakhya, Ram, et al. "Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Execution in Simulator





Noise-free localization and mapping

HABITAT

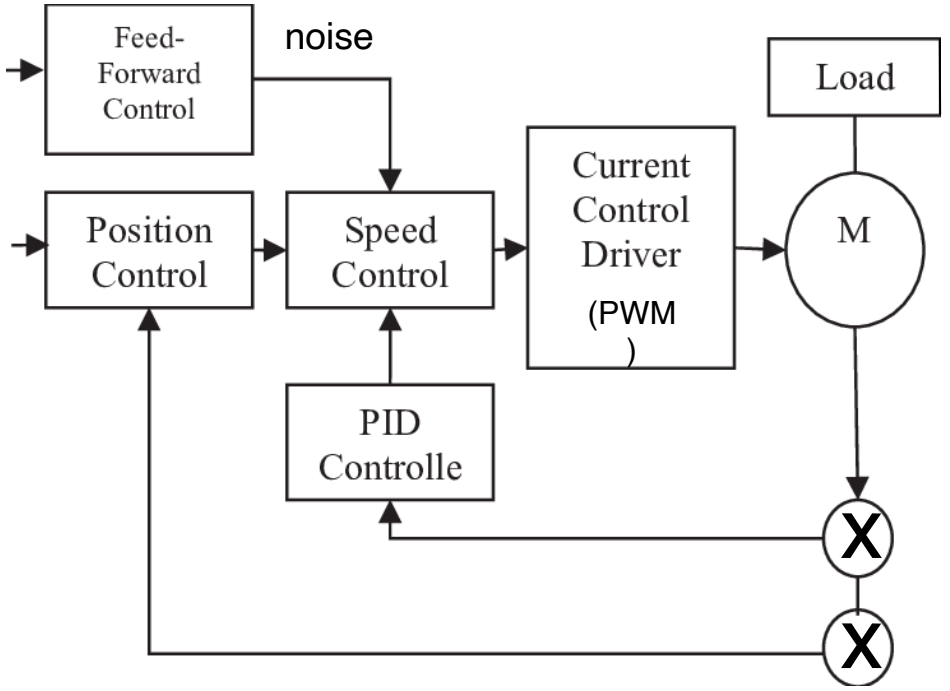
Meta

Execution in the Real World

How to control the motor ?
Take **servo motor** as example
· By a double loop control system.

Inducing noises to localization and mapping

Basic Control



$$u(t) = Kp[e(t) + \frac{1}{Ti}\int_0^t e(t)dt + Td\frac{de(t)}{dt}]$$

$$u_k = Kp[e_k + \frac{T}{Ti}\sum_{j=0}^{k} e_j + Td\frac{e_k - e_{k-1}}{T}]$$

Two Types of Navigation

Classical Modular Navigation

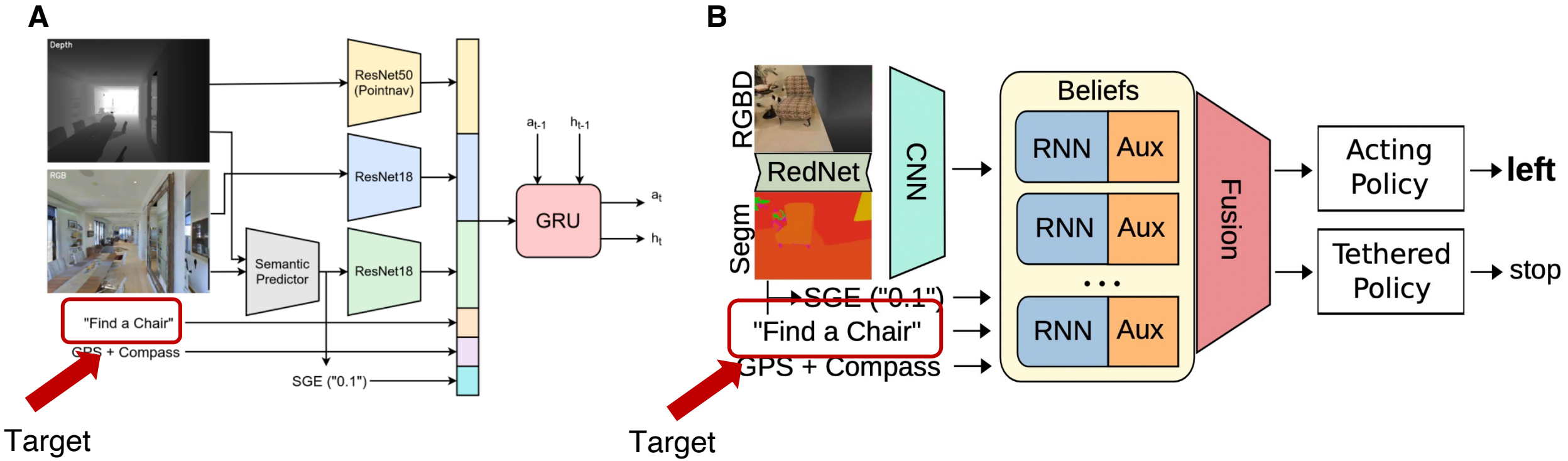
- ☺ Good generalizability to novel scenes
- ☺ Satisfying performance
- ☹ Hard to implement

vs

End-to-end Reinforcement Learning

- ☺ Easy to implement
- ☺ Satisfying performance
- ☹ Require extensive training time (1k-10k training hours)
- ☹ Poor generalizability to novel scenes

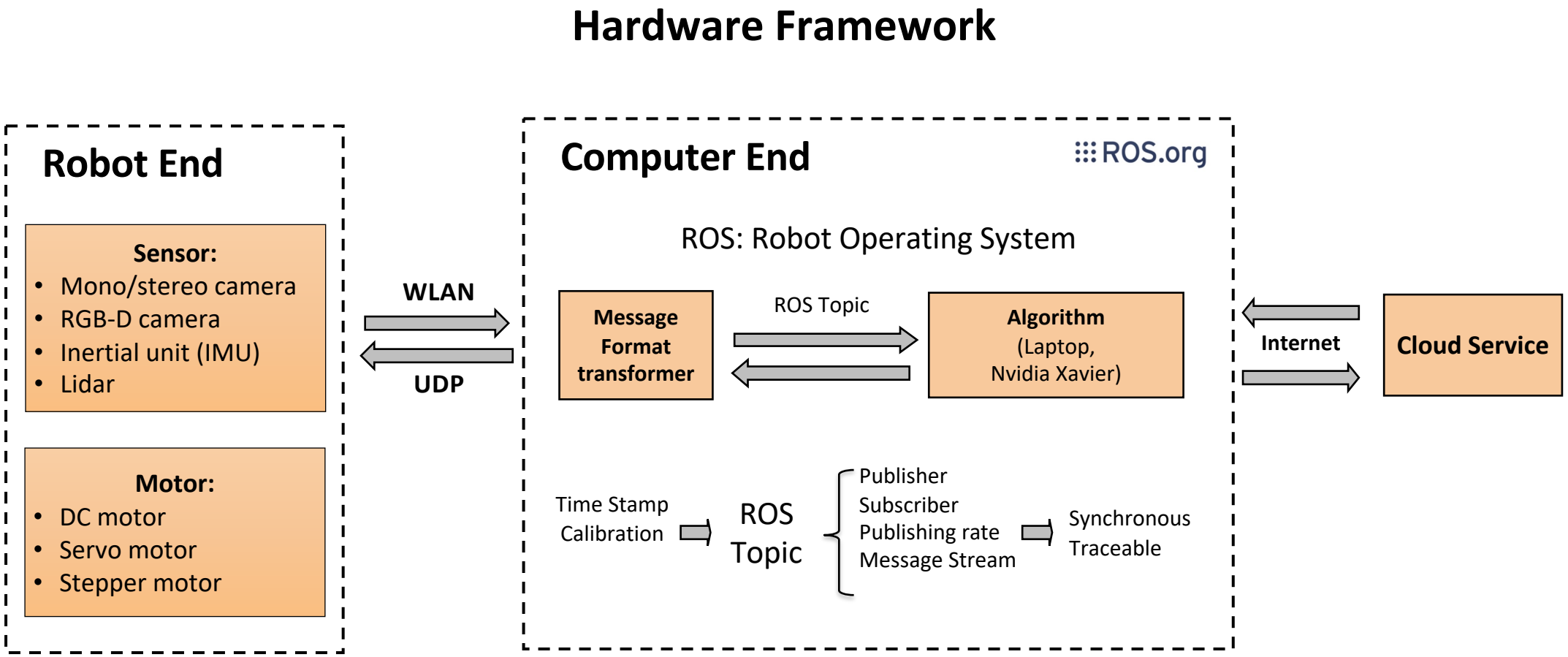
End-to-end RL



(A) Ramrakhya, Ram, et al. "Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

(B) Auxiliary Tasks and Exploration Enable ObjectGoal Navigation, Joel Ye et al, ICCV2021

Real-world Navigation is more complex



Generalizable Object Goal Navigation

Object Goal: dining table

Semantic Scene Understanding

Object detection

Learning Semantic Priors

Where is 'dining table' more likely to be found?

Episodic Memory

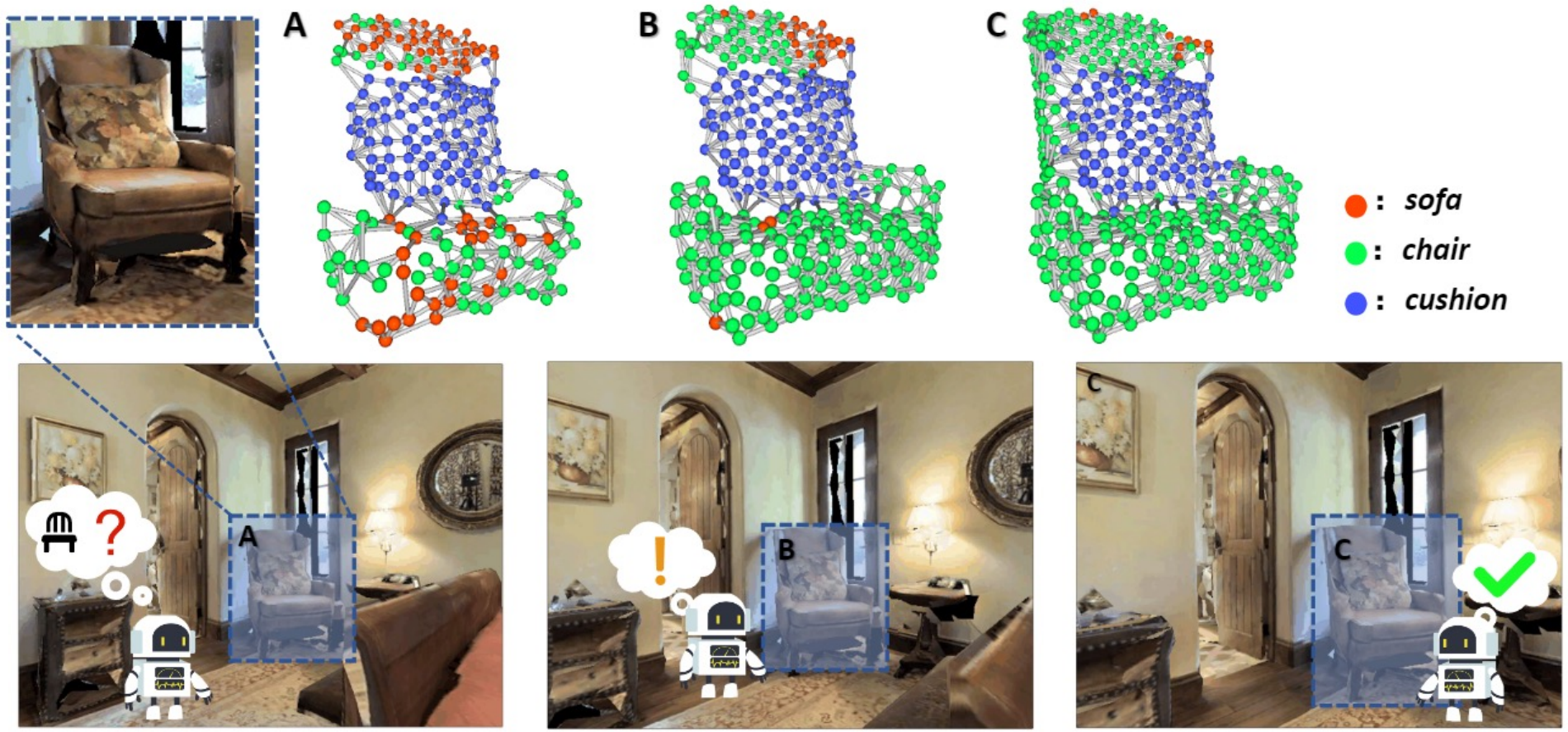
Keeping track of explored and unexplored areas

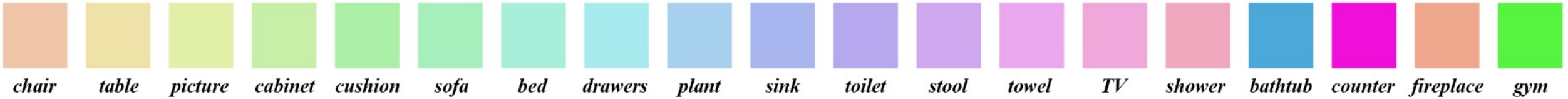
Leveraging Online Semantic Point Fusion for 3D-aware Object Goal Navigation

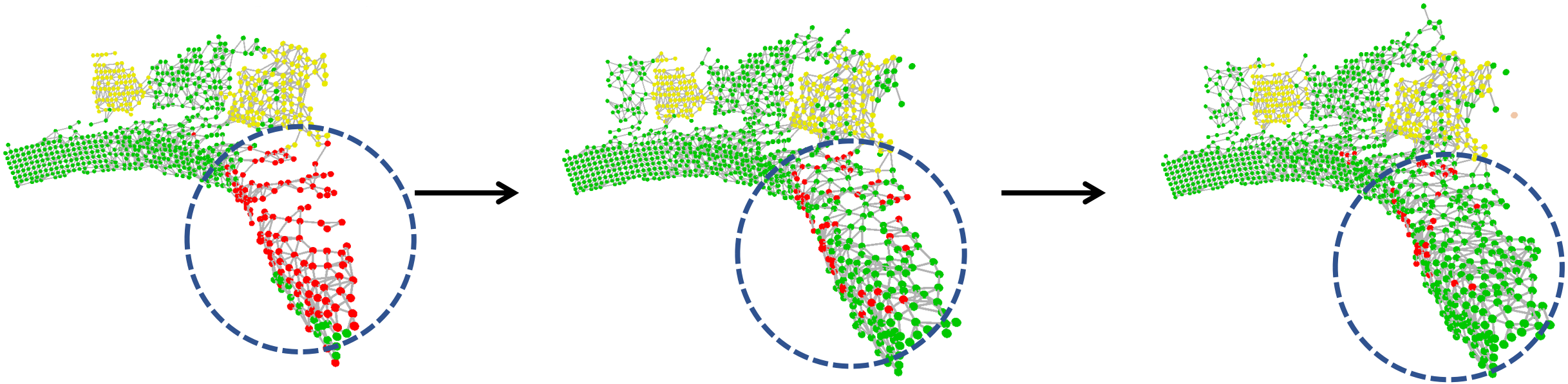
CVPR 2023

Jiazhao Zhang*, Liu Dai*, Fanpeng Meng,
Qingnan Fan, Xuelin Chen, Kai Xu, He Wang†

Online 3D Reconstruction + Understanding





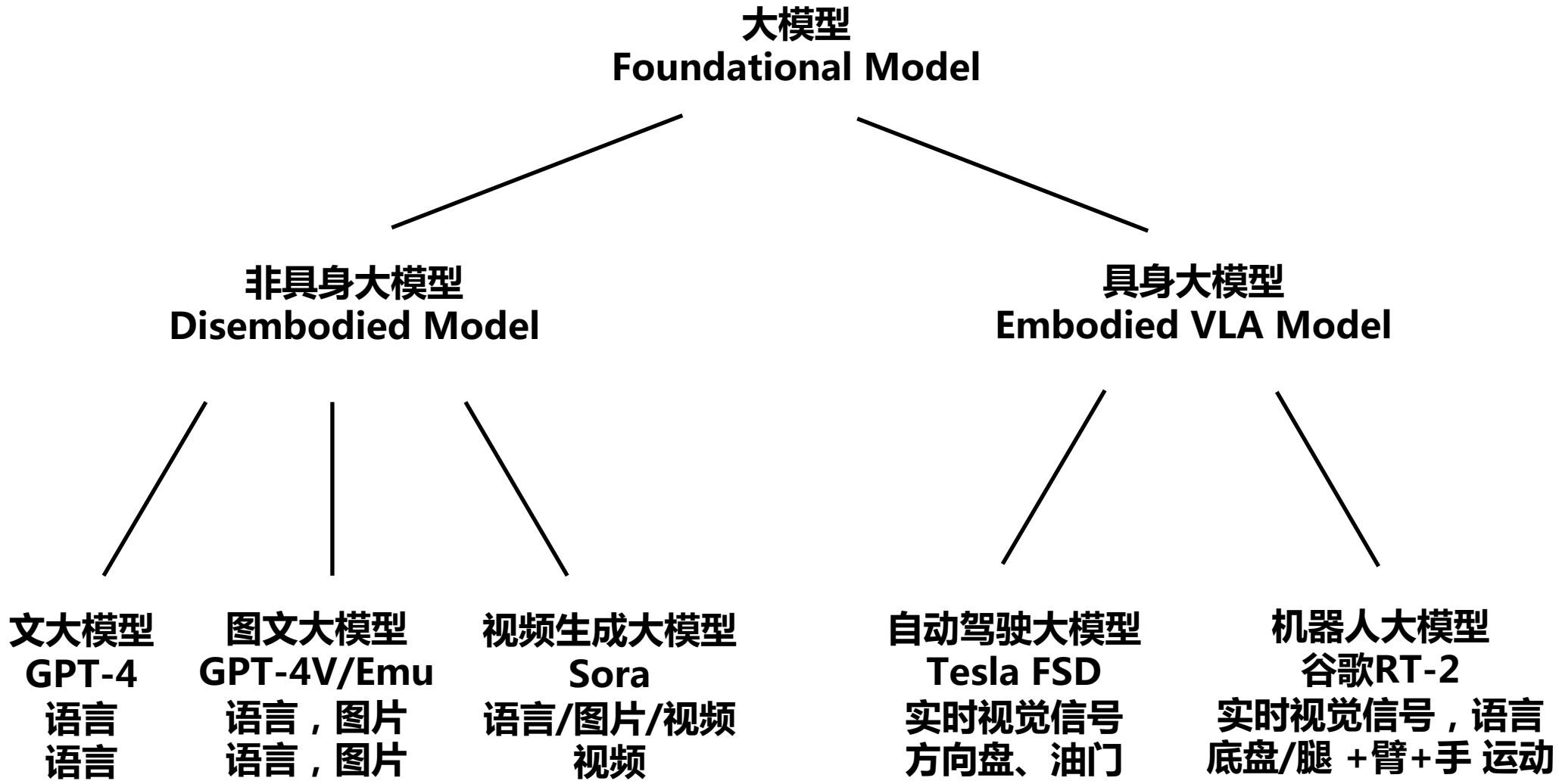


bed sofa cushion

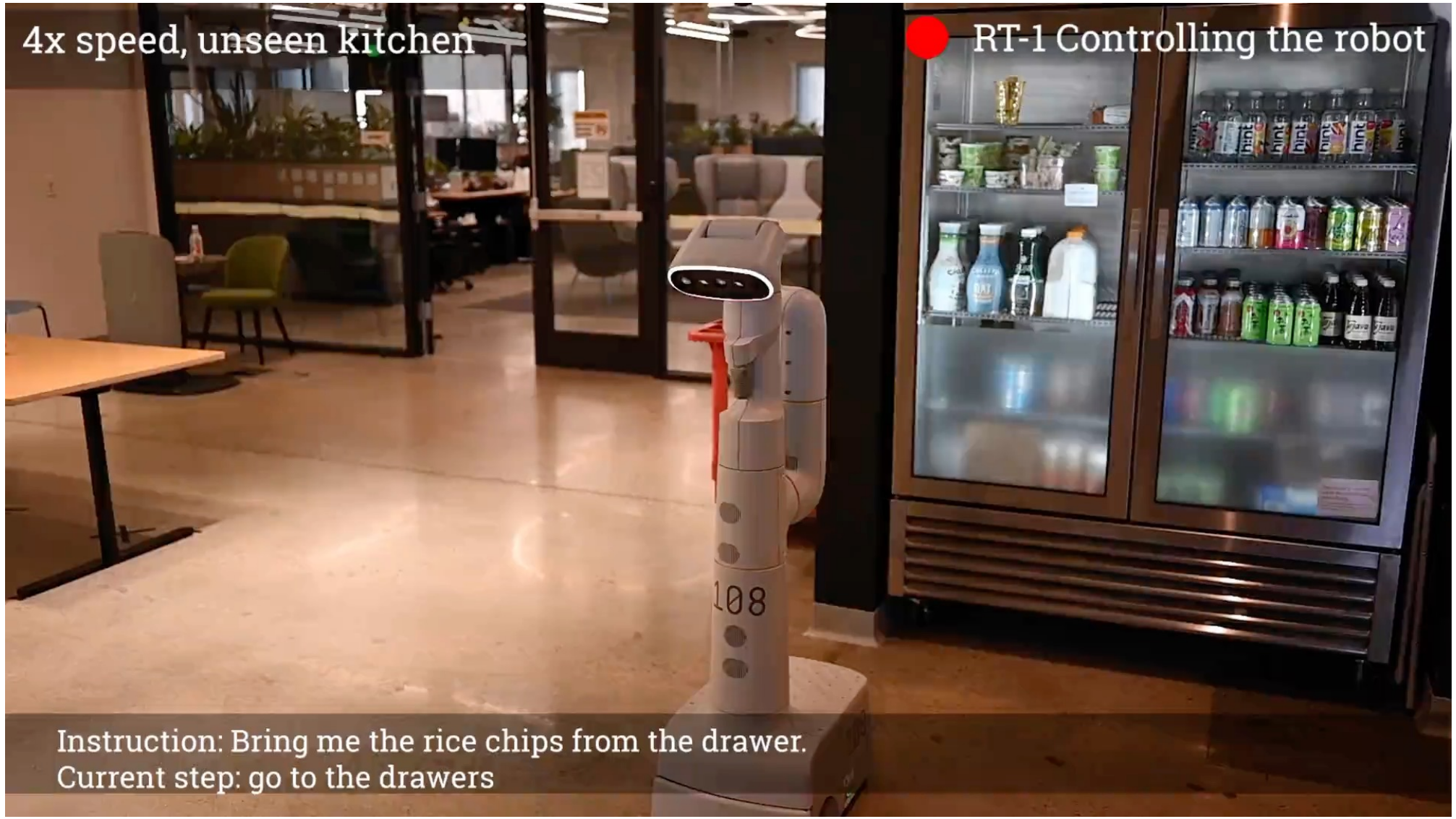
More Correct and Consistent!

Embodied Multimodal Large Model

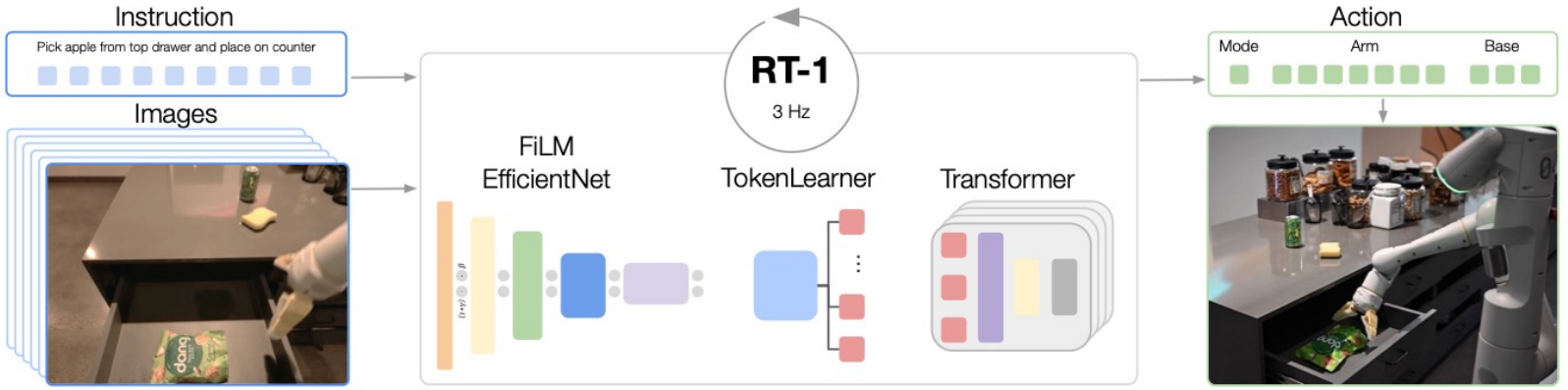
Large Models



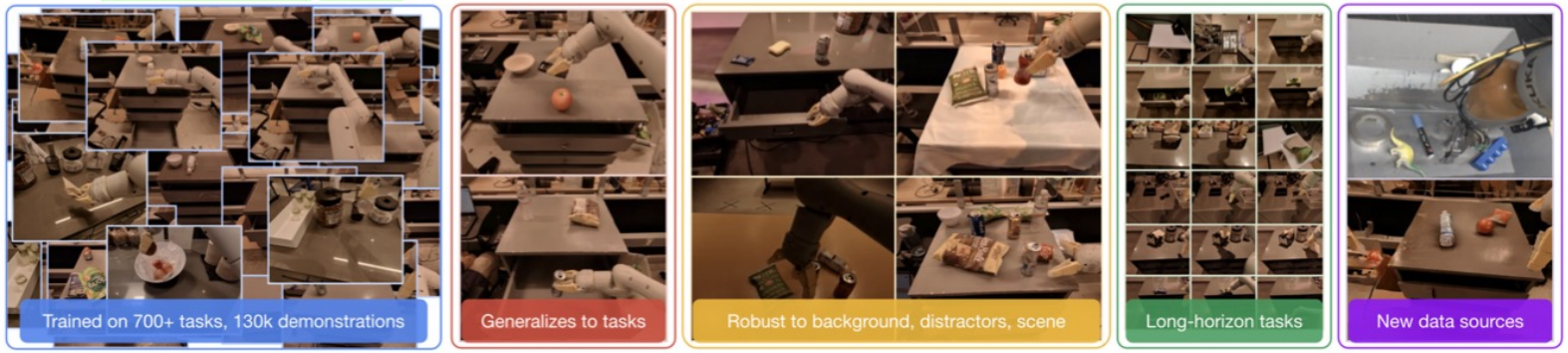
Related Work: RT-1



Related Work: RT-1



(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).



(b) RT-1's large-scale, real-world training (130k demonstrations) and evaluation (3000 real-world trials) show impressive generalization, robustness, and ability to learn from diverse data.

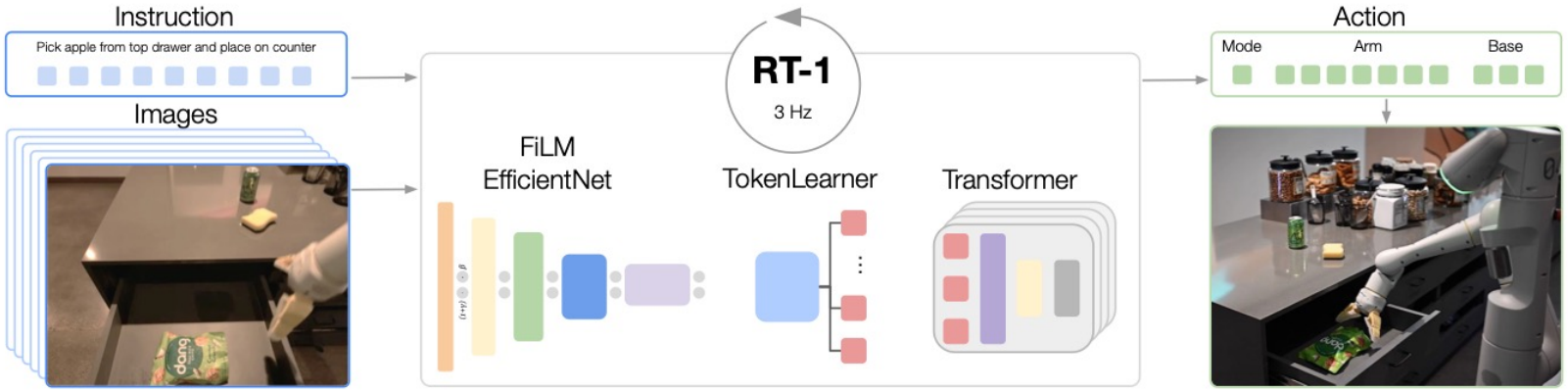
Figure 1: A high-level overview of RT-1's architecture, dataset, and evaluation.

Limitations

Skill	Count	Description	Example Instruction
Pick `Object`	130	Lift the object off the surface	pick iced tea can
Move `Object` Near `Object`	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place `Object` Upright	8	Place an elongated object upright	place water bottle upright
Knock `Object` Over	8	Knock an elongated object over	knock redbull can over
Open `Drawer`	3	Open any of the cabinet drawers	open the top drawer
Close `Drawer`	3	Close any of the cabinet drawers	close the middle drawer
Place `Object` into `Receptacle`	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick `Object` from `Receptacle` and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Section 6.3 and 6.4 tasks	9	Skills trained for realistic, long instructions	open the large glass jar of pistachios pull napkin out of dispenser grab scooper
Total	744		

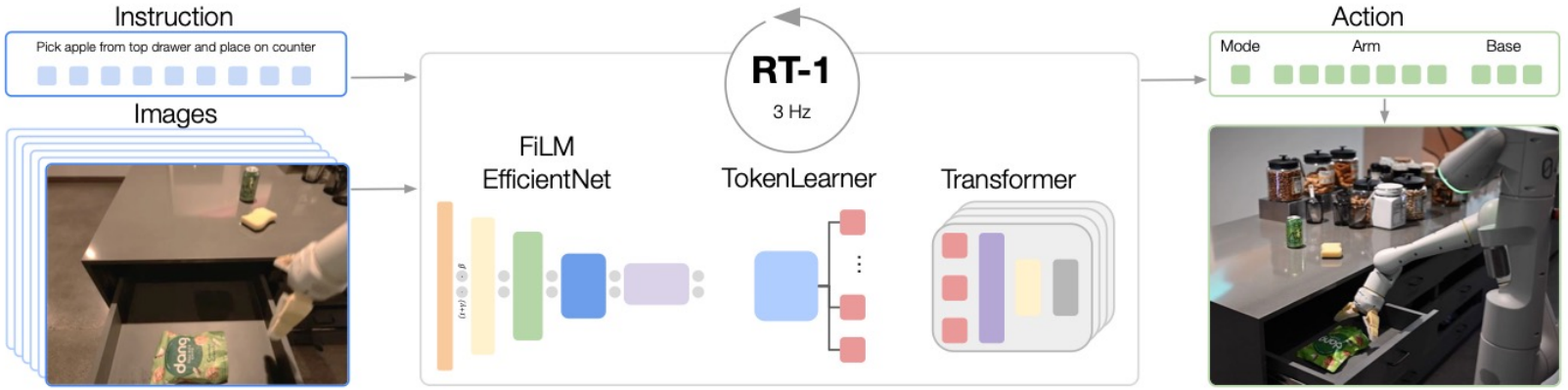
- Limited task diversity, still mainly pick and place

Limitations



- Limited task diversity, still mainly pick and place
- No 3D vision

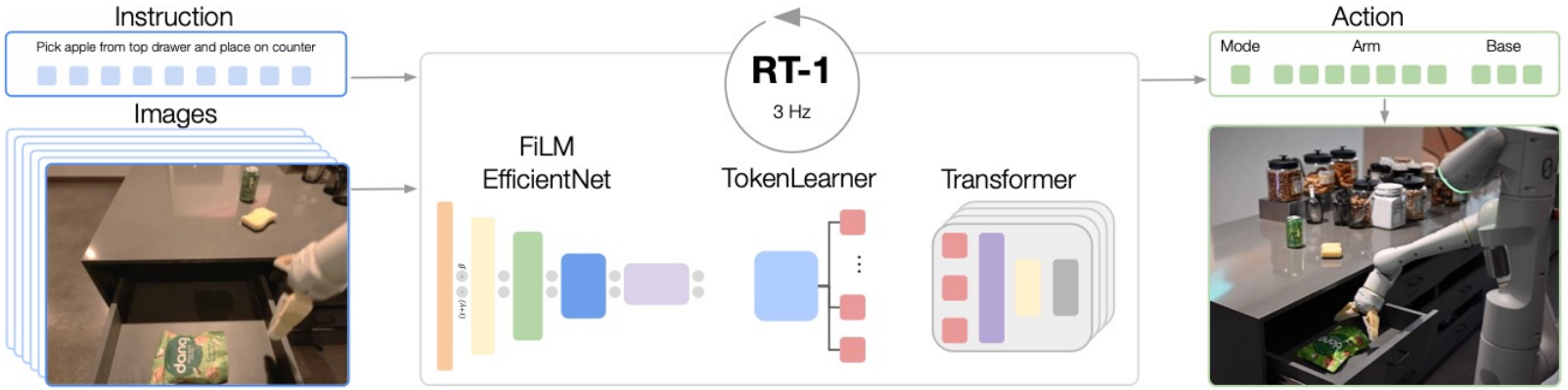
Limitations



- No 3D vision

1. Lack of geometry –> performance? generalization?

Limitations



- No 3D vision

1. Lack of geometry –> performance? generalization?

2. Lack of 3D scene modeling –> scene memory? generalization?

Limitations

- **Data collection and evaluations**: Noah Brown, Justice Carbajal, Joseph Dabis, Tomas Jackson, Utsav Malla, Deeksha Manjunath, Jodily Peralta, Emily Perez, Jornell Quiambao, Grecia Salazar, Kevin Sayed, Jaspiar Singh, Clayton Tan, Huong Tran, Steve Vega, and Brianna Zitkovich.

- Limited task diversity, still mainly pick and place
- No 3D vision
- Demonstrations are costly (130K demo, 17 months, 16 people, 13 robots) and the scalability is questionable

Exemplar Work: RT-2 from Google

Exemplar Work: RT-2 from Google

Body :
A Low-cost General Mobile Manipulation Platform

Mobile base: Wheeled Robot

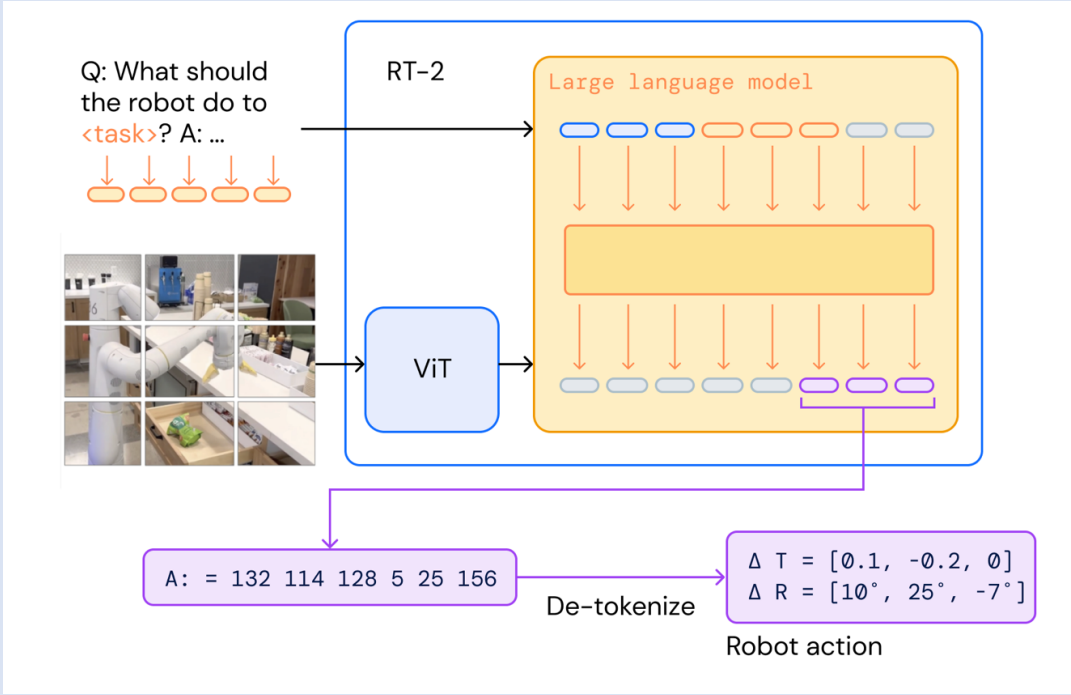
Single arm

End effector: parallel gripper

Equipped with teleoperation

Embodiment

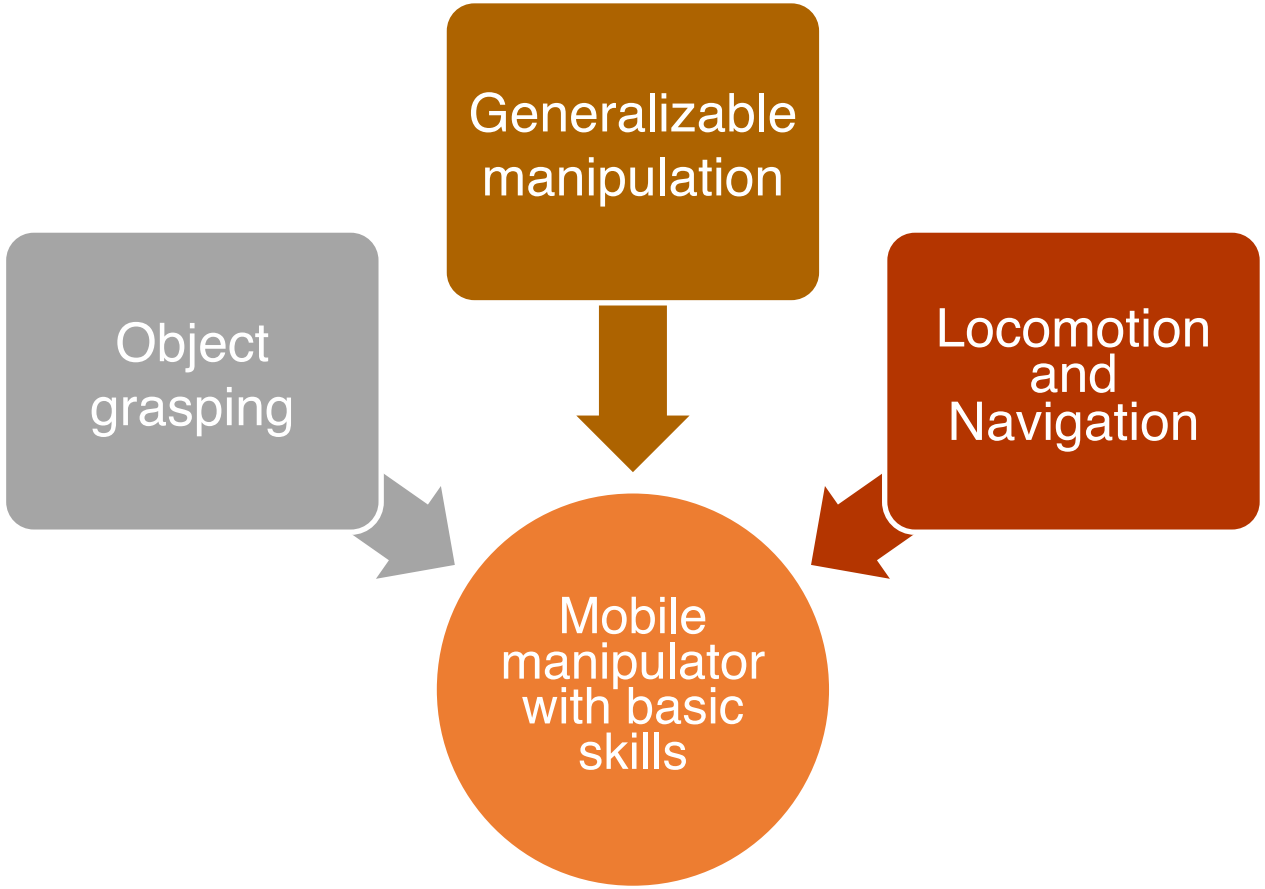
Brain :
Embodied Vision-Language-Action (VLA) Model



Visual Perception + Task Planning + Action Generation

Summary of Home Robots

Goal: a scalable **3D-aware** home robot



Summary of Computer Vision

- Compared to human vision, computer vision deals with the following tasks:
  - visual data acquisition (similar to human eyes but comes with many more choices)
  - image processing and feature extraction (mostly low-level)
  - analyze local structures and then 3D reconstruct the original scene (from mid-level to high-level)
  - understanding (mostly high-level)
  - generation (beyond the scope of human vision system)
  - and further serving embodied agents to make decisions and take actions.

Email: hewang@pku.edu.cn
Homepage: https://hughw19.github.io

PKU EPIC
北京大学具身感知与交互实验室
Embodied Perception and Interaction Lab, PKU

Thanks for Taking This Course!
All the best with your final!