UNIVERSITY OF MICHIGAN

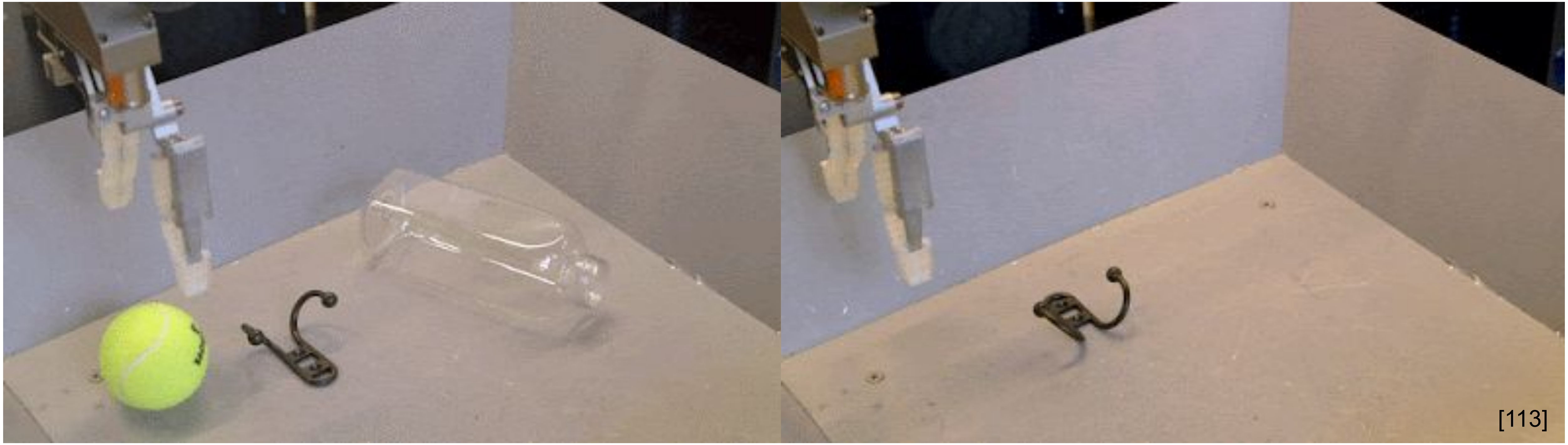
Deep Learning Approaches to Grasp Synthesis: A Review

Rajiv Govindjee | Zixuan Huang

2022-03-27

Introduction

- **grasping**: controlling an object by applying forces and torques
- high-dimensional search: pose, joint angles, contact points
- quality of **grasp hypothesis** evaluated on task-specific metrics (e.g. stability)



[113]

Introduction

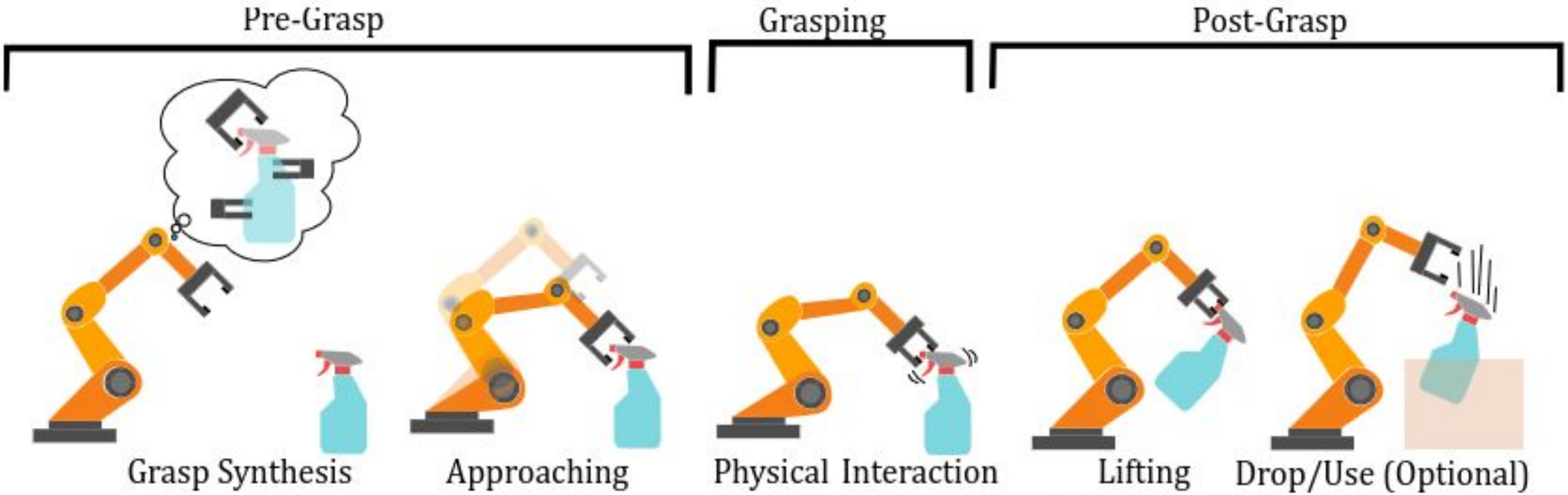
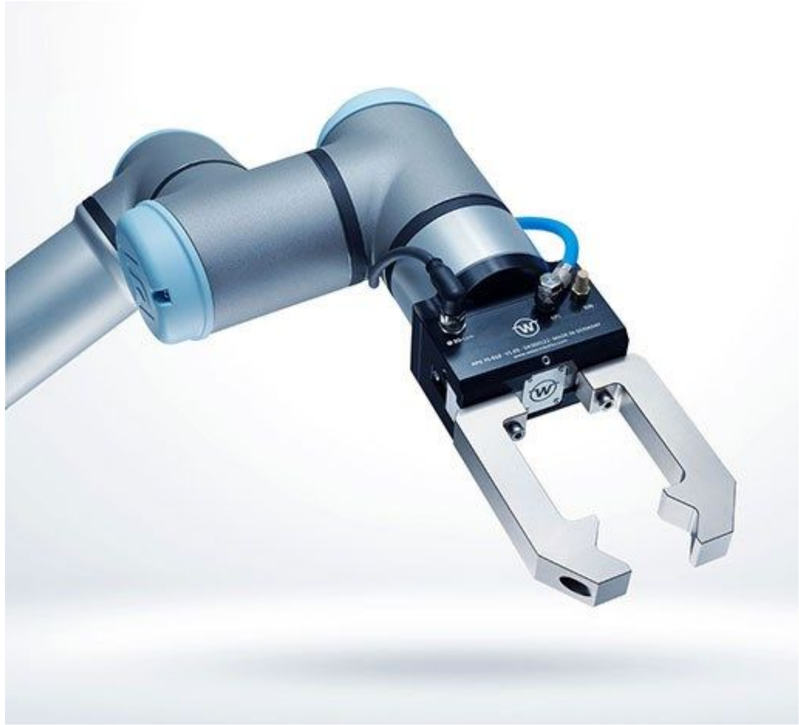
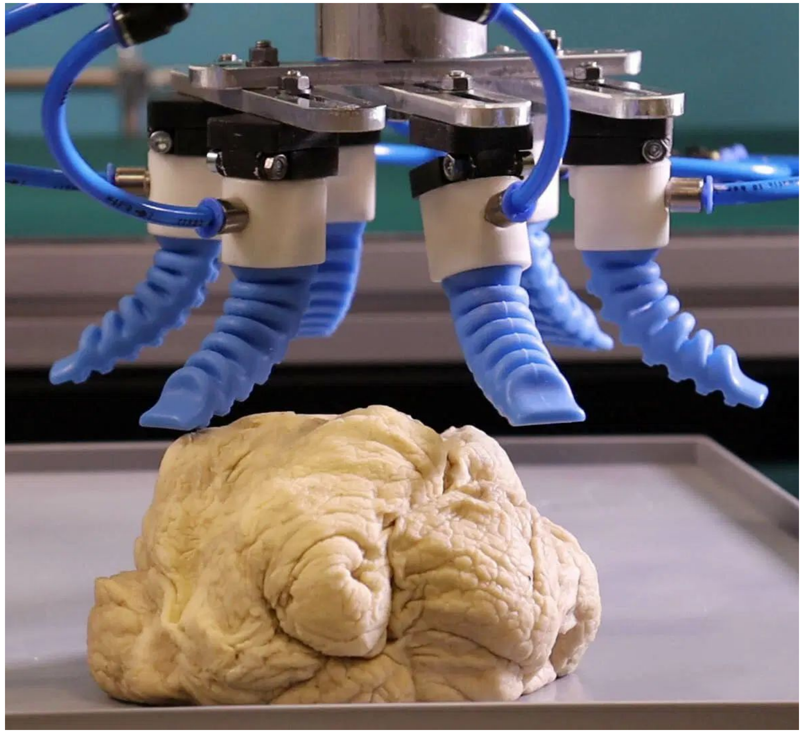


Fig. 3: Typical stages for grasping an object. Our review focuses on grasp synthesis, the first stage in the grasping process.

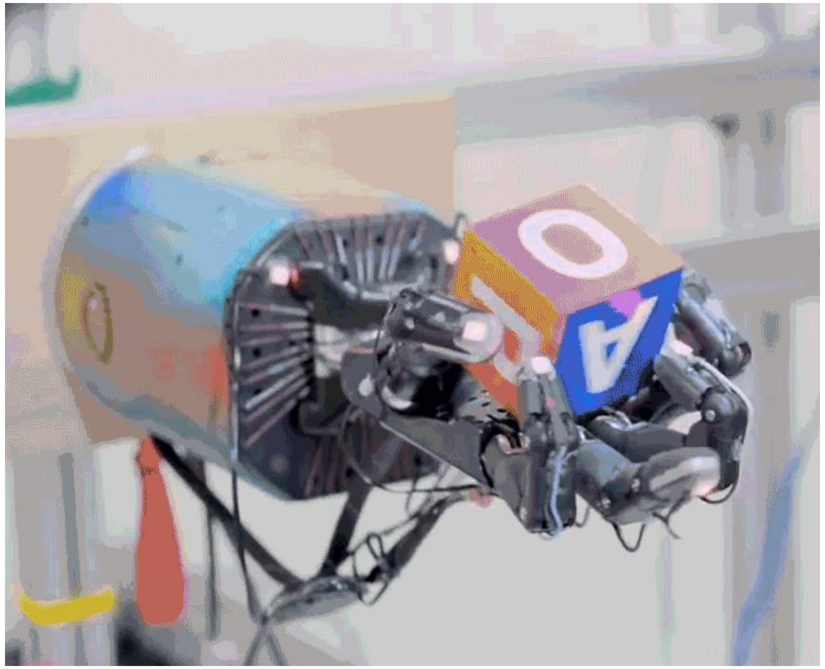
Robot hardware: end effectors



Parallel jaws



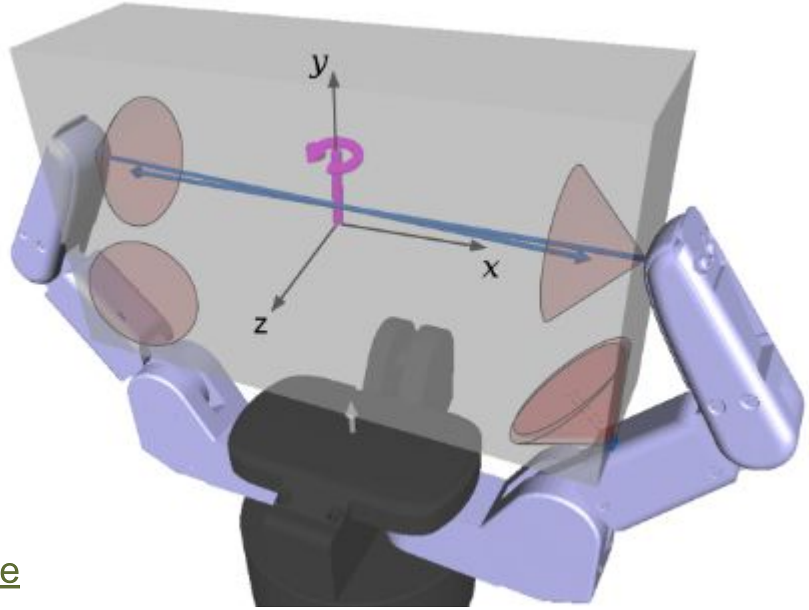
Soft grippers



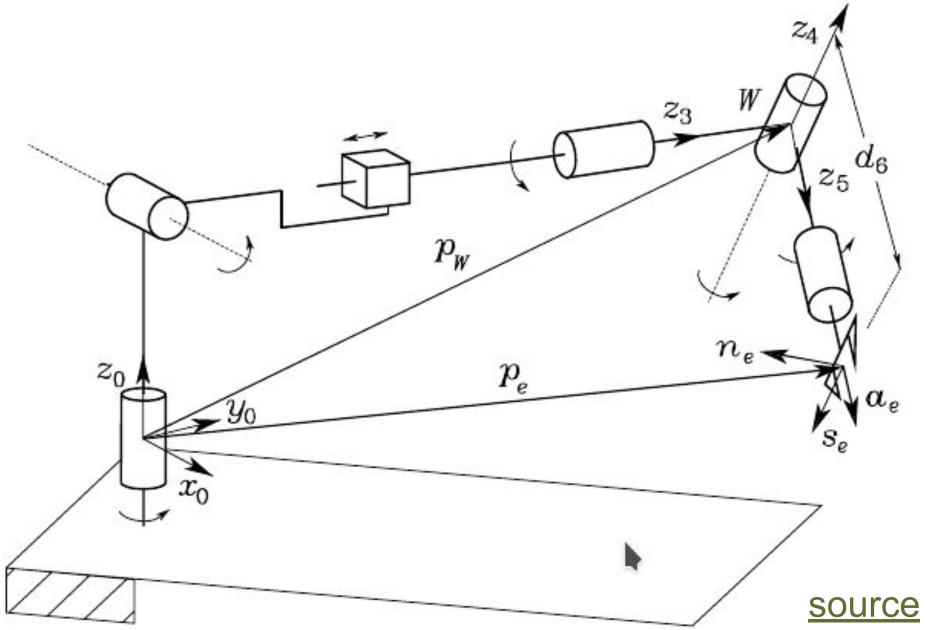
Dextrous articulated hand

Analytical approaches

- **grasp**: set of forces and torques on an object
- **fixturing**: find a grasp that keeps the object in equilibrium
- **manipulation**: find a grasp that moves the object in a specific way
- analytical approaches often require full knowledge of object properties



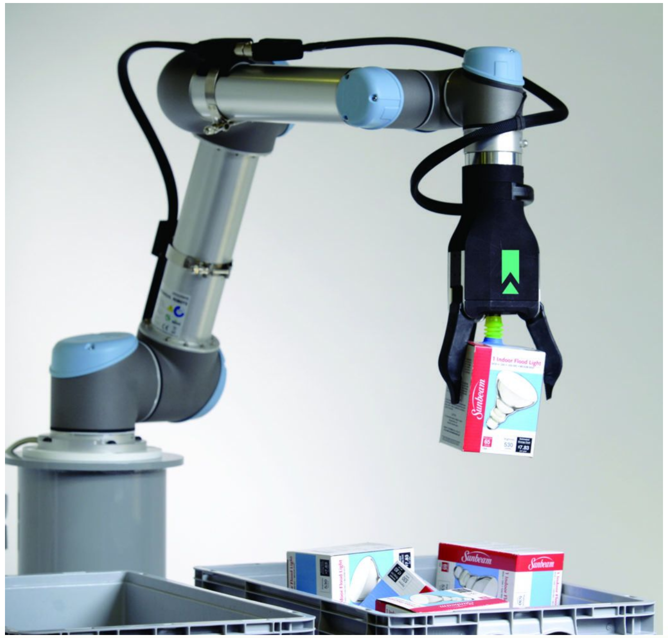
source



source

Key terms

- **4-DoF grasp**: top-down; position of end effector in x-y-z, rotation about z
- **6-DoF grasp**: position of end effector in SE(3) (3D position, 3-axis rotation)
- **approach vector**: line along which the end effector approaches the target
- **antipodal points**: pairs of points with collinear and opposite normal vectors



[41]

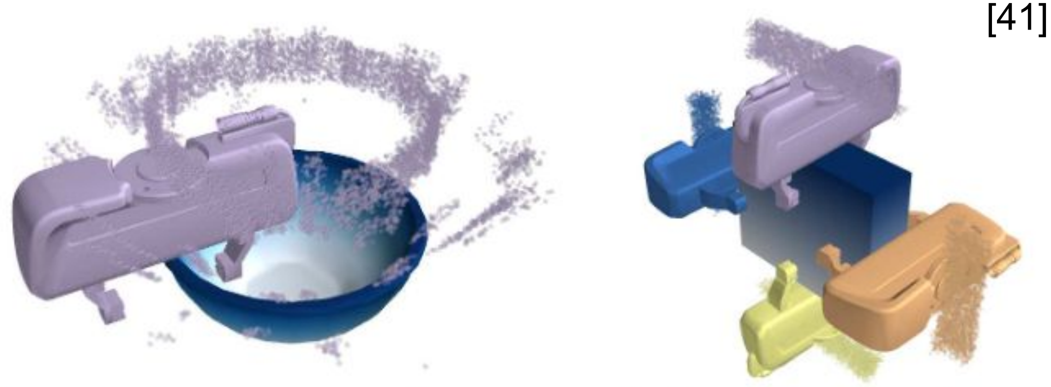
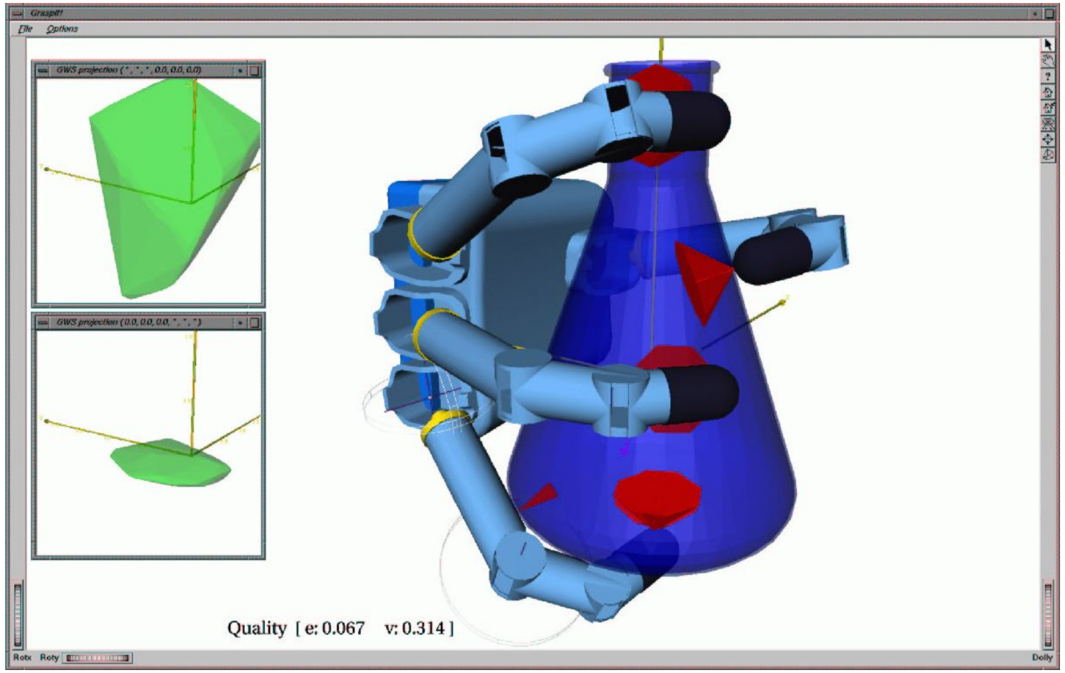


Figure 6. We use training data generated with a physics simulator. The colored dots around the objects depict successful grasps for a bowl (left) and a box (right). For each continuous grasp subspace an exemplary gripper pose is shown.

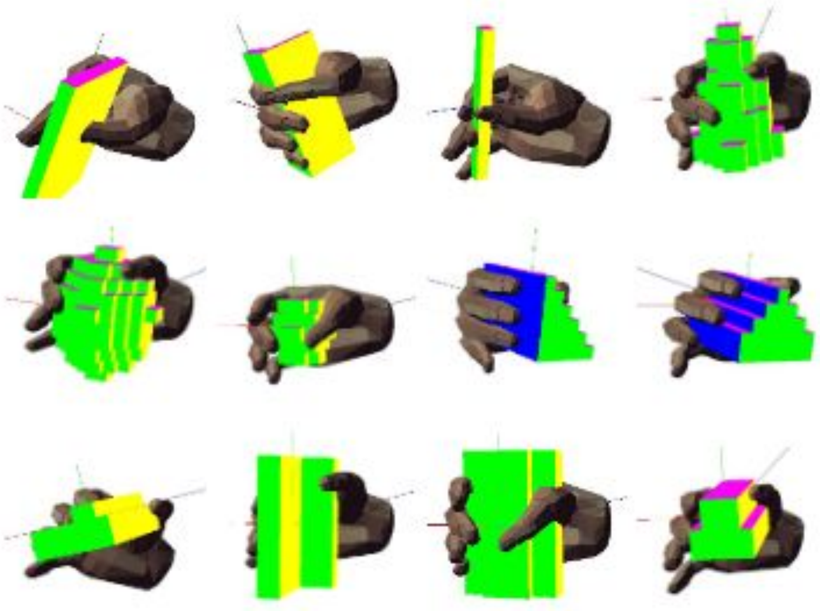
Data-driven approaches

- grasping simulators (Graspit!, Simox)
- hand-designed features
- grasping with only RGB or RGB-D
- supervised learning: where to grasp

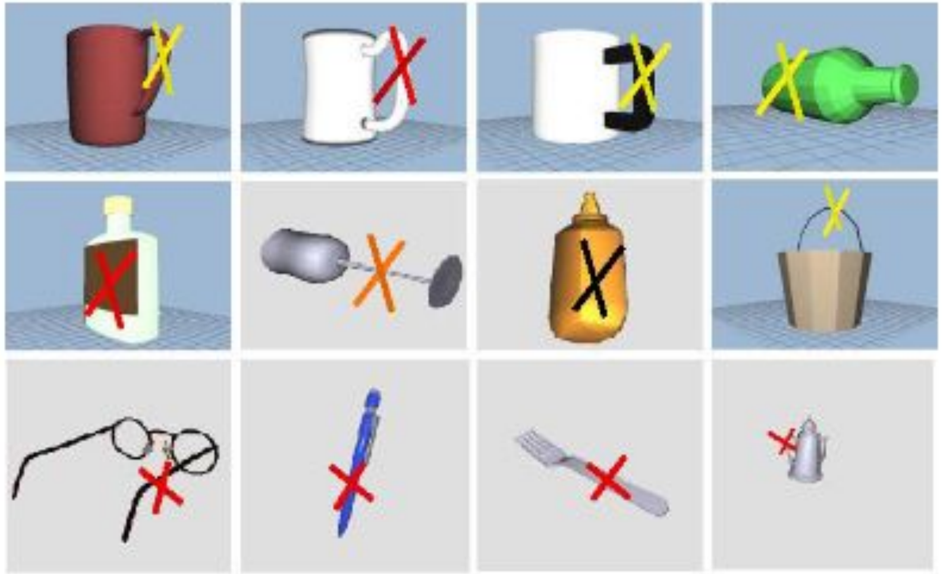


Quality [e: 0.067 v: 0.314]

[22]



[24]

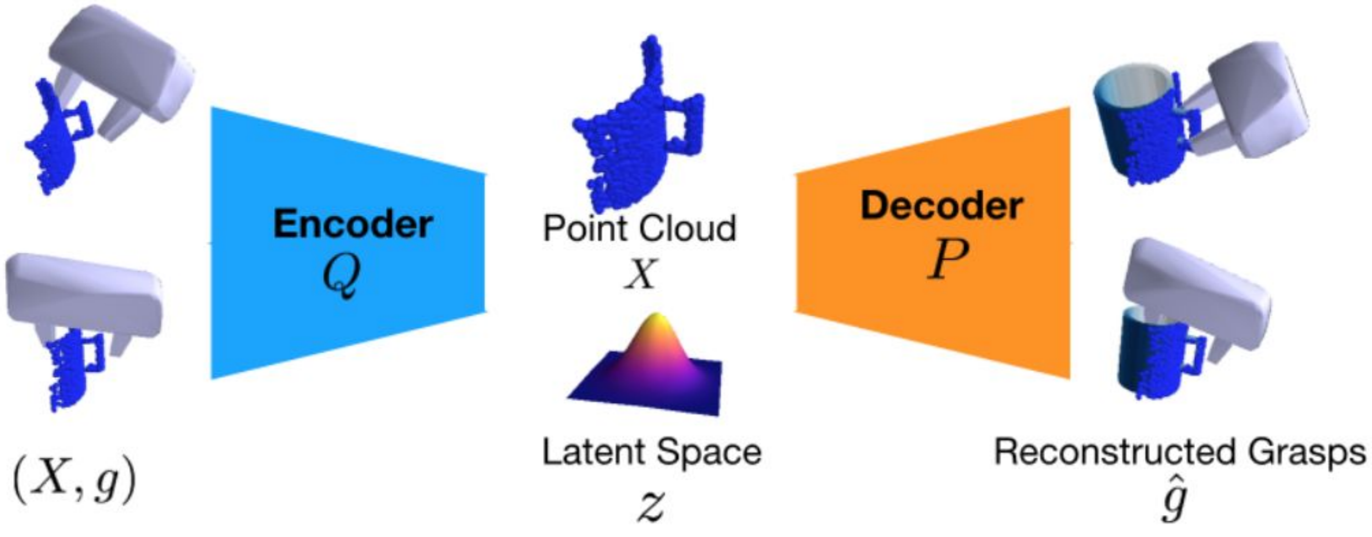


[23]

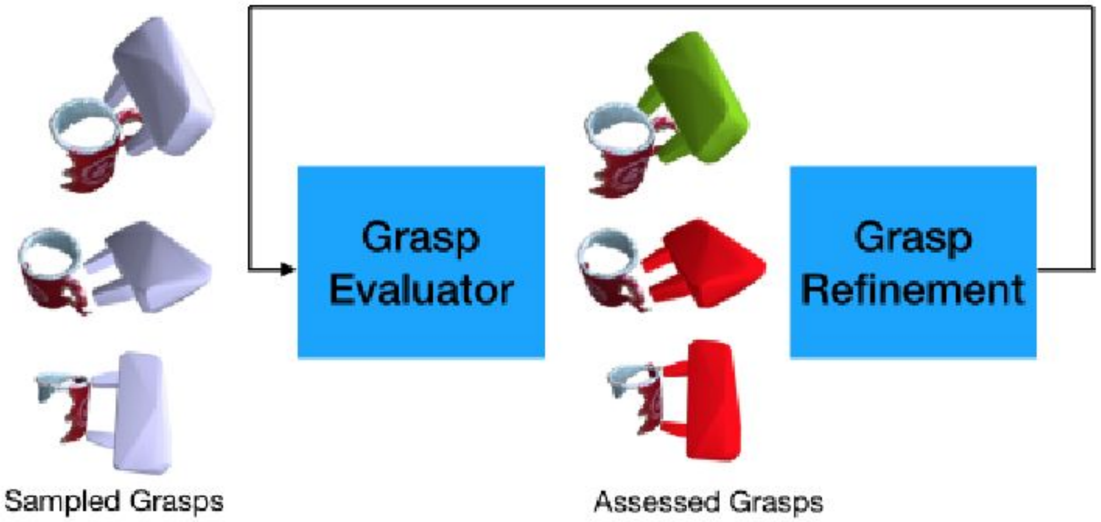
Sampling-based deep learning approaches

1. sample information
   a. randomly or systematically sample grasp pose in Euclidean or latent space
   b. remove infeasible grasps (collisions, empty grasps)
   c. generative models for learning distribution: VAEs, GMMs, GANs
2. evaluate sample according to (learned) quality function
3. (optionally) refine sample using optimization (grad. descent on quality function)

VAE training [203]



Test-time architecture [203]



Exemplar methods

- Key idea: maintain database of successful grasps, find most applicable example
- Patten [131]: metric learning to encode objects with similar geometry
- Mahler [34]: CNN to provide similarity metric, then sample from known grasps

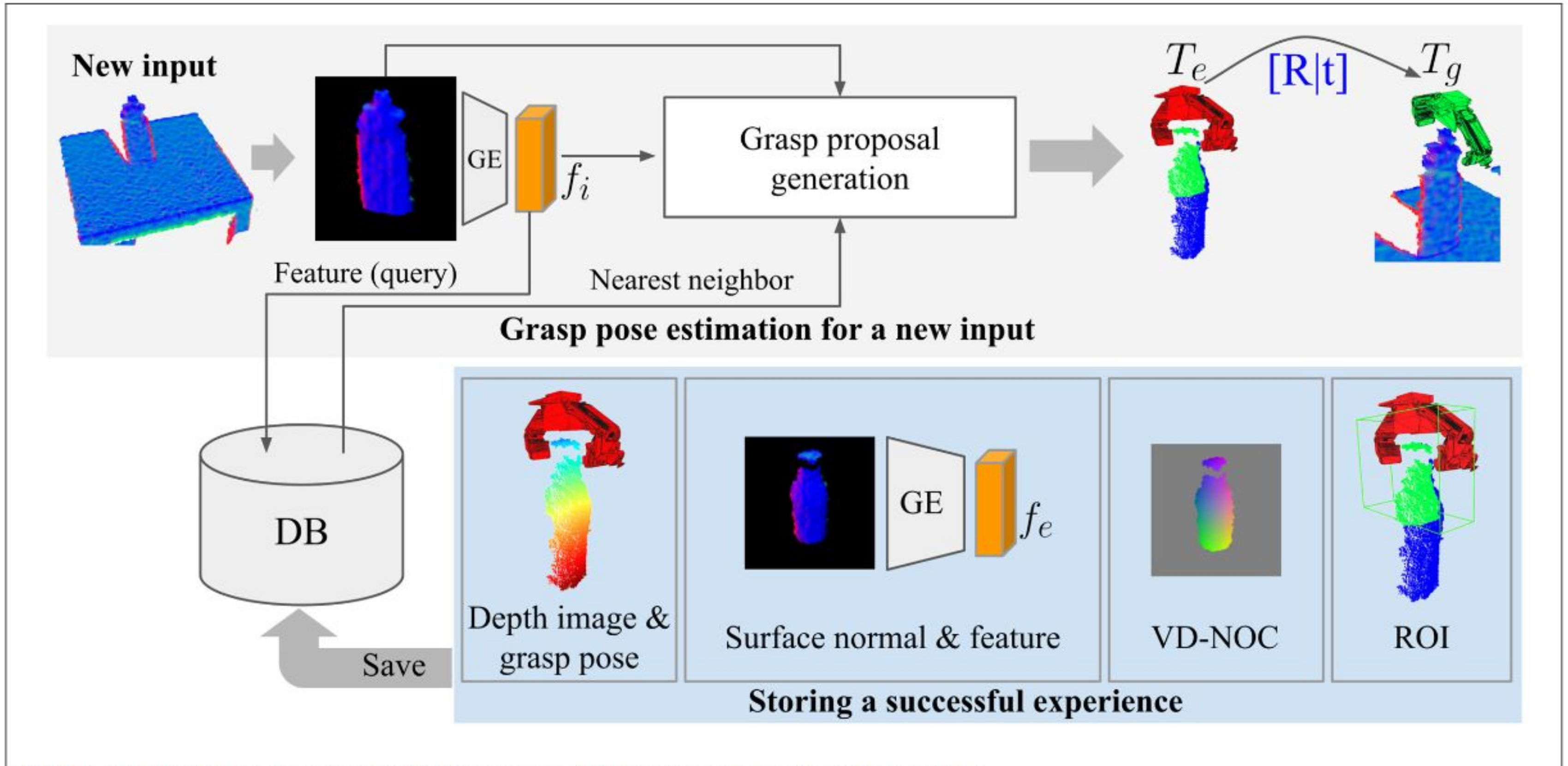
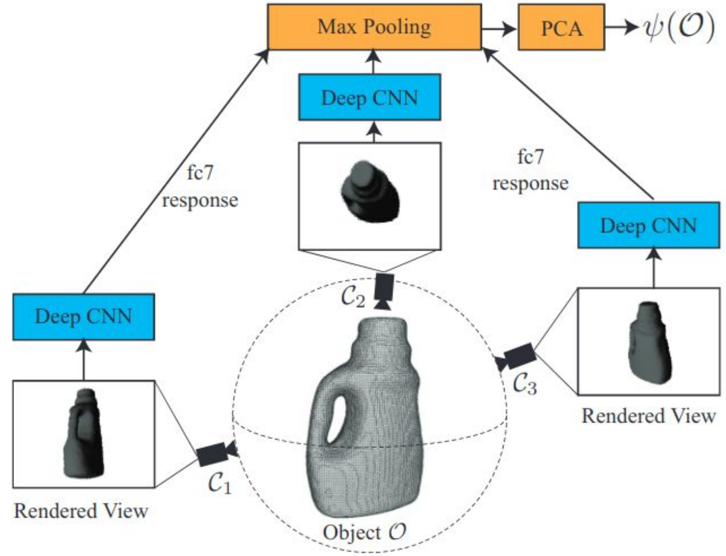


FIGURE 1 | Overview of storing and retrieving experience with the incremental grasp learning framework.

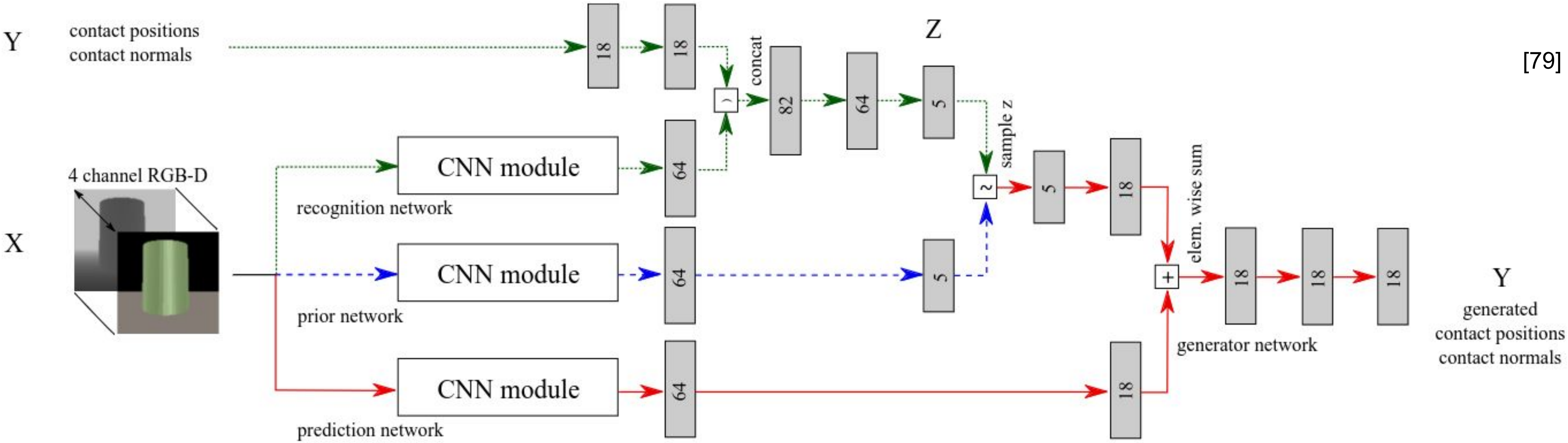
[131]

[34]



Regression

- Process entire sample space simultaneously (end-to-end)
  - predict grasp parameters, quality from single network
  - full 6-DoF pose output



[79]

(a) CVAE architecture we use in our experiments. Dotted arrows denote components used during training, dashed arrows are components used during testing, and solid arrows are used for both training and testing. The CNN Module is expanded in (b).

Regression: simplifying techniques

- difficult to regress in 6DoF
- can used reduced-dimensional representations
- solve for remaining DoF based on regressed grasp
- discretize sample space [81]
- assume grasping centroid [86]
- predict contact point, conditional grasp [100]

[94]

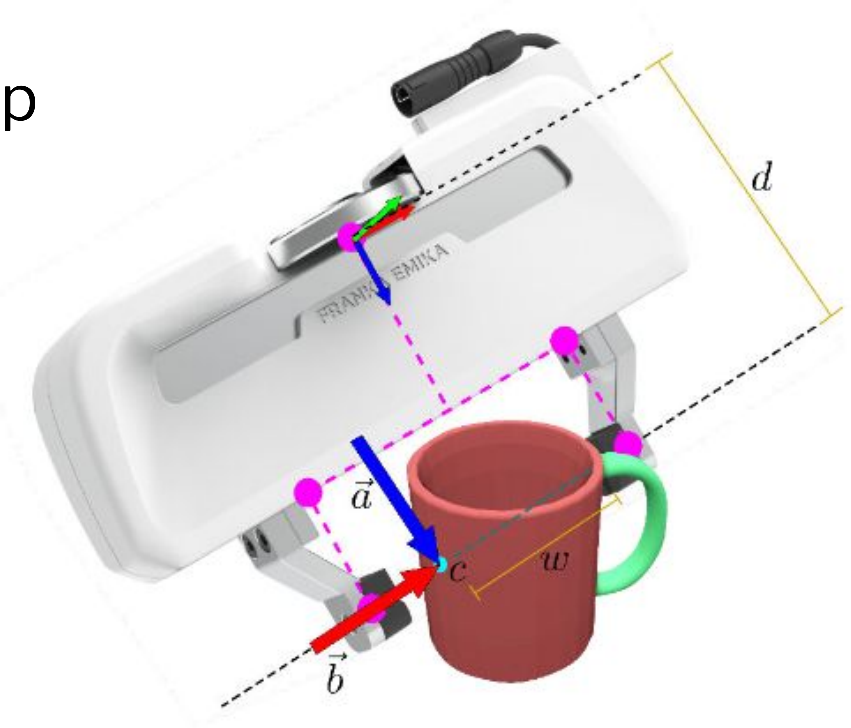


Fig. 3. Our grasp representation: c depicts an observed contact point. $\mathbf{a}$ and $\mathbf{b}$ constitute the 3-DoF rotation, w is the predicted grasp width, d the distance from baseline to base frame. In pink we show the five gripper points $\mathbf{v}$ that we used in the l_{add-s} loss.

Off-policy RL: learning from demonstrations

- Song [122]: Q-learning from human demonstrations with hardware
  - deterministic policy on learned Q-function
- Wang [127]: DDPG from demonstrations, transfers from PyBullet sim to real
  - demos from optimization-based motion and grasp planner (“expert”)

[122]

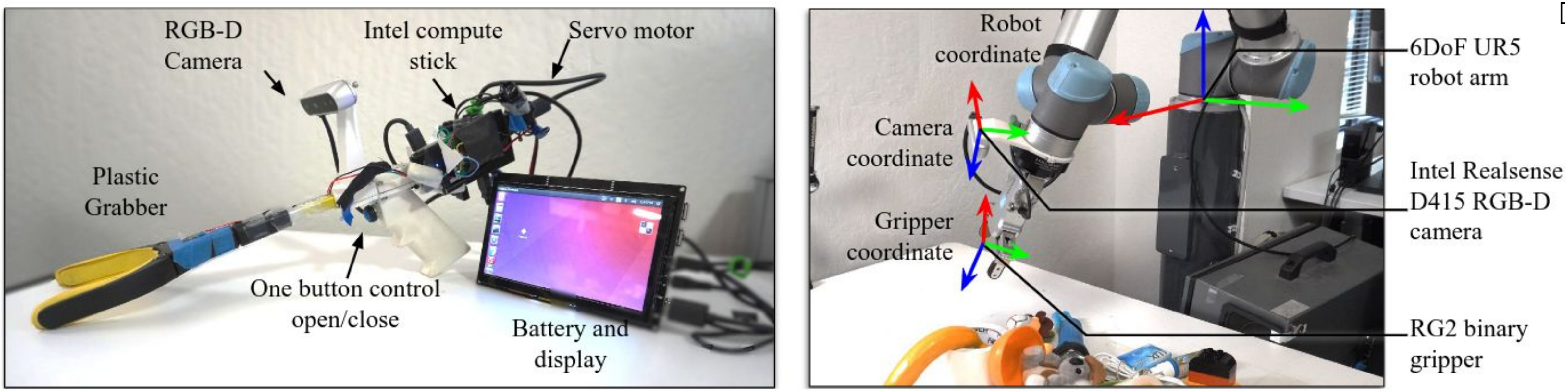


Fig. 2. **Hardware setup.** Our low-cost handheld device (left) consists of a plastic grabber tool equipped with an RGB-D camera and a servo that controls the binary opening of the grabber fingers. This device was designed to be analogous to the real robot’s end effector setup (right), while providing a low-user-friction interface that enables untrained people to collect grasping data in almost any environment.

On-policy RL: learning from demonstrations

- training with experiences from most recent policy (PPO, DQN, A2C, TRPO)
- Kawakami [125]: separate {orienting, approaching, closing} into separate tasks
  - start with imitation learning (collected with VR), then PPO for each task
- Mandikal and Grauman [121]: actor-critic reward based on CNN for affordances

[125]

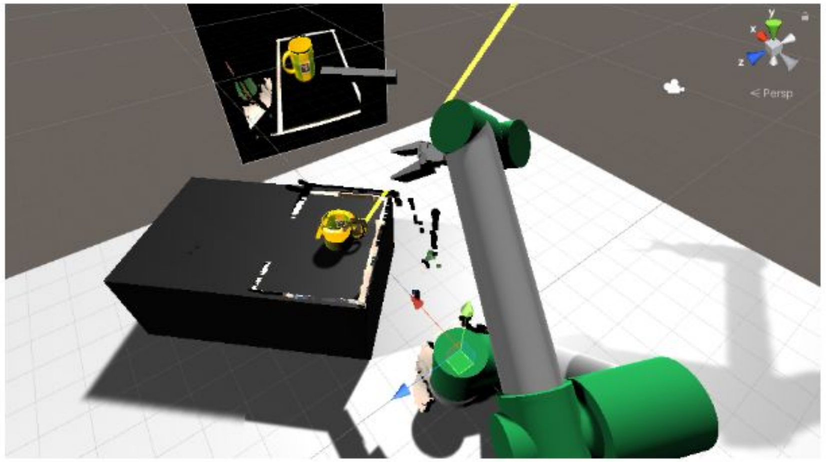
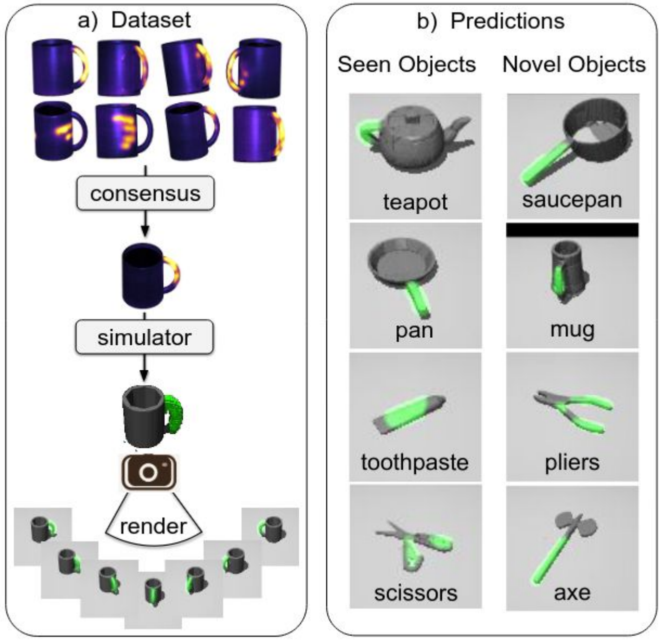
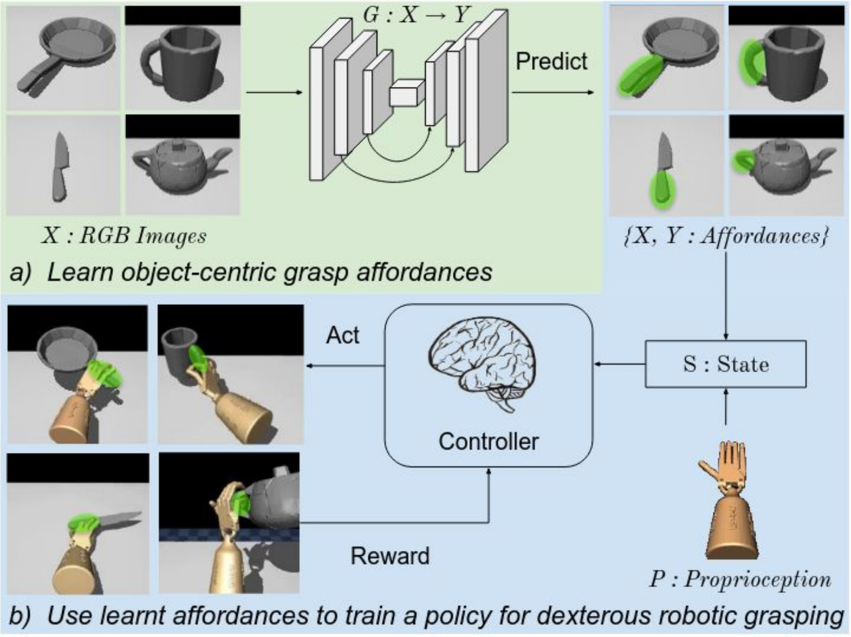


Fig. 1: VR interface for robot control



[121]

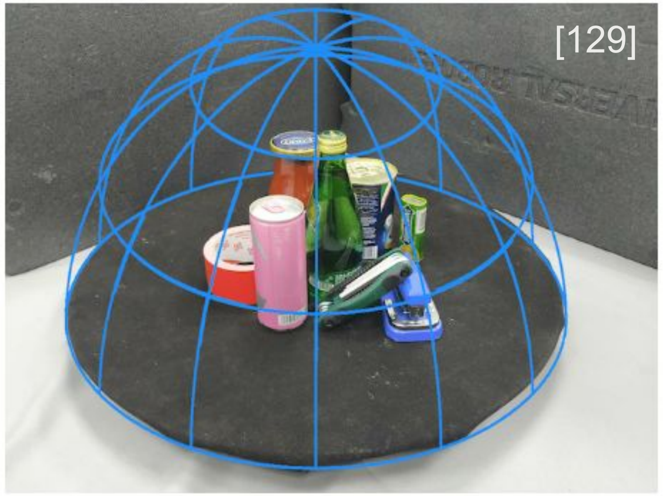
Fig. 2: Affordance anticipation. a) Training images generated from 3D thermal maps from ContactDB. Green denotes label masks overlaid on images. b) Sample predictions for seen and novel objects from ContactDB and 3DNet, respectively. Our anticipation model predicts functional affordances for novel objects and viewpoints (e.g., graspable handles and rings).


G : X → Y
Predict
X : RGB Images
{X, Y : Affordances}
a) Learn object-centric grasp affordances
Act
S : State
Controller
Reward
P : Proprioception
b) Use learnt affordances to train a policy for dexterous robotic grasping

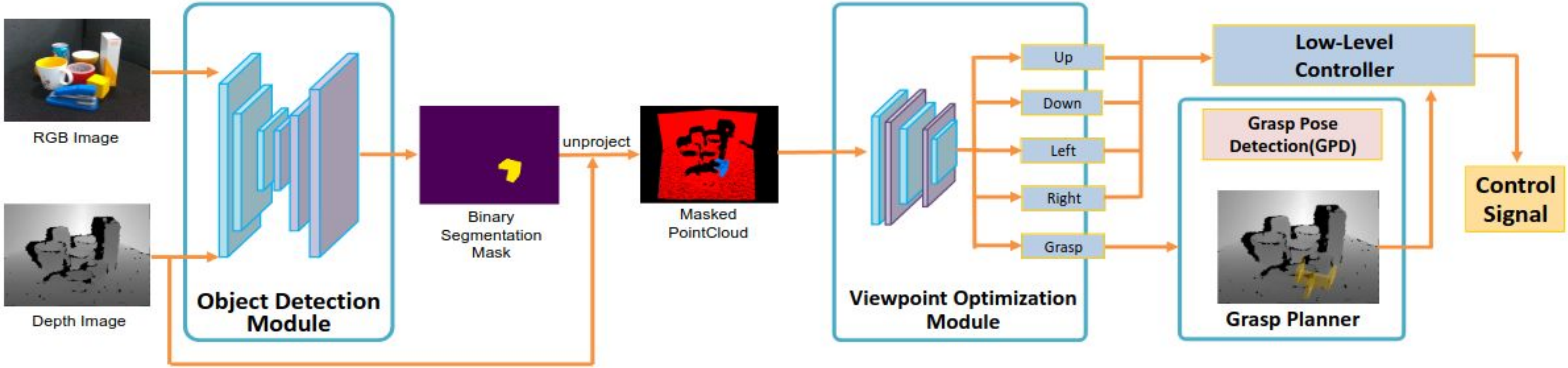
[121]

Fig. 3: Overview of our GRAFF model. a) In Stage I, we train an affordance prediction model that predicts object-centric grasp affordances given an image. b) In Stage II, we train an RL policy that leverages these affordances along with other visuomotor sensory inputs (RGB-D image + hand joint variables) to learn a stable grasping policy.

On-policy RL: viewpoint search


[129]

- Chen [129]: A2C to optimize viewpoint first
  - CNN to predict 6-DoF grasp pose (GPD)
  - dense reward for increasing visible portion of object of interest
  - sparse reward for grasping
  - real viewpoint data collected using a turntable
  - this "real embodied simulator" improves sim-to-real transfer



[129]

On-policy RL: contact force input

- Merzic [119]: TRPO using contact feedback as input
  - entirely in simulation environment: Gazebo
  - simulated contact force measurement + proprioception
  - perfect or noisy knowledge of object pose
  - reward based on weighted combination:
    - change in links in contact with object
    - change in distance from gripper to object
    - joint torques, object linear velocity
    - drop test
  - results: contact force feedback improves grasp success, especially with sensing noise

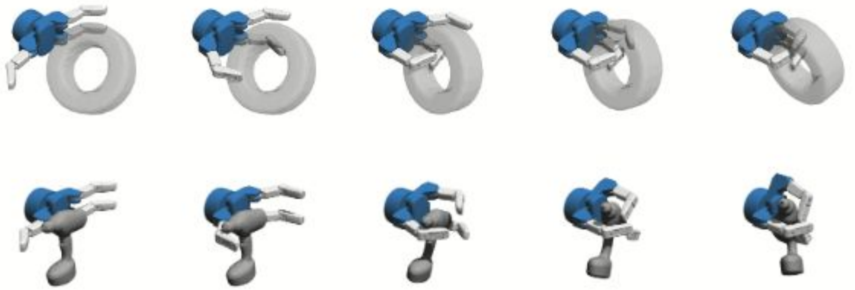


Fig. 1: Examples of learned grasp strategies using a multi-fingered hand and contact feedback.

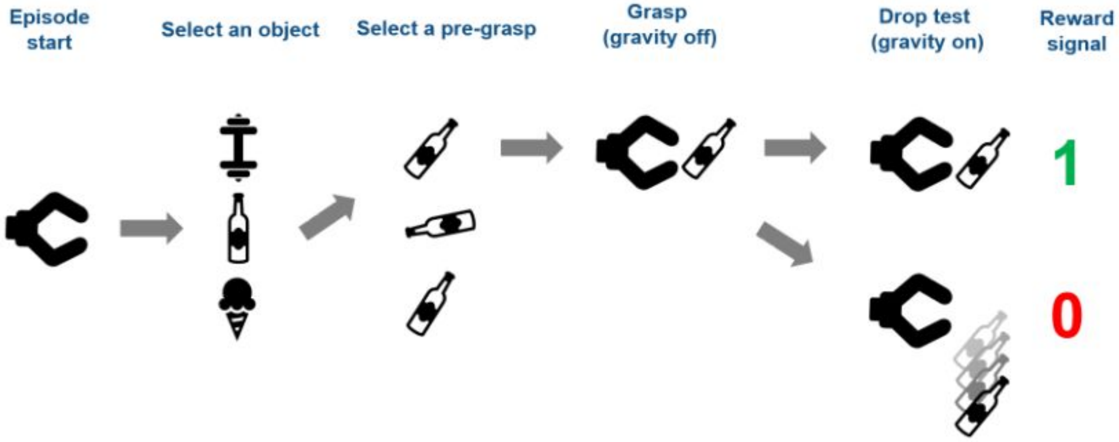


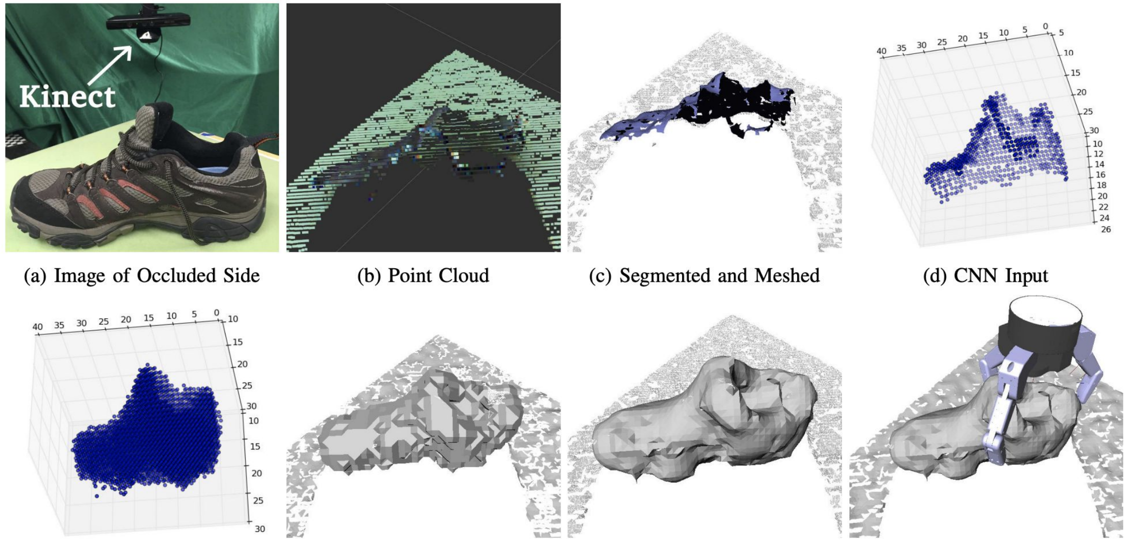
Fig. 4: Breakdown of a single grasping episode.

Supporting methods based on deep learning

- Deep learning can also be used in certain part of the grasping pipeline to improve the success rate of a grasping task
  - Shape approximation
  - Affordance

Supporting methods | Shape approximation

- Shape completion:
  - Estimate the full object model from partial view
  - Sample grasps around completed shape
  - Better capture the geometry & Uncertainty

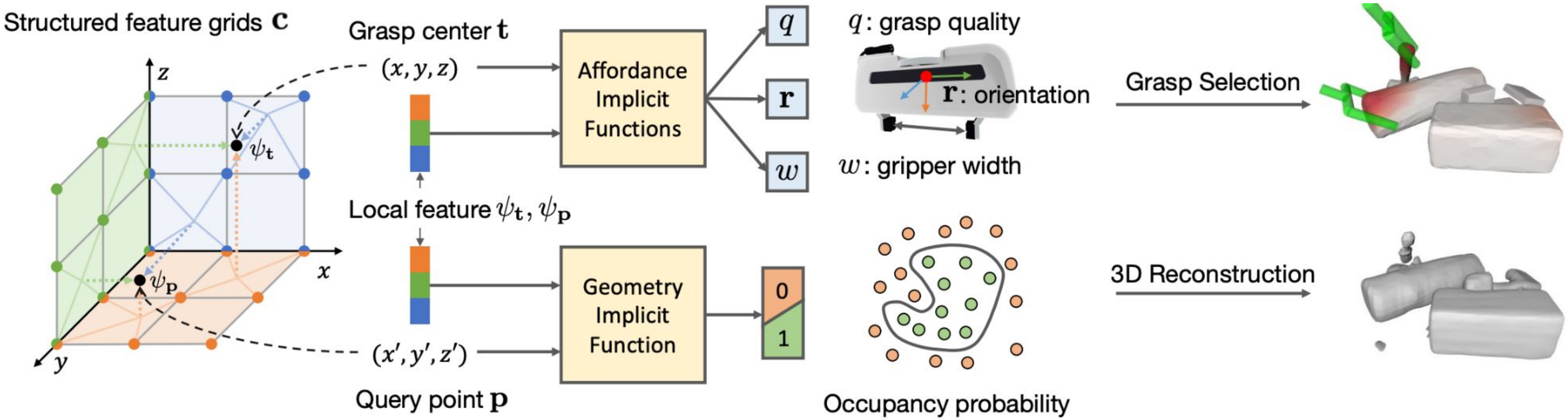


(a) Image of Occluded Side (b) Point Cloud (c) Segmented and Meshed (d) CNN Input

(e) CNN Output (f) Fast Mesh (g) Detailed Mesh (h) Grasp Planning

Supporting methods | Shape approximation

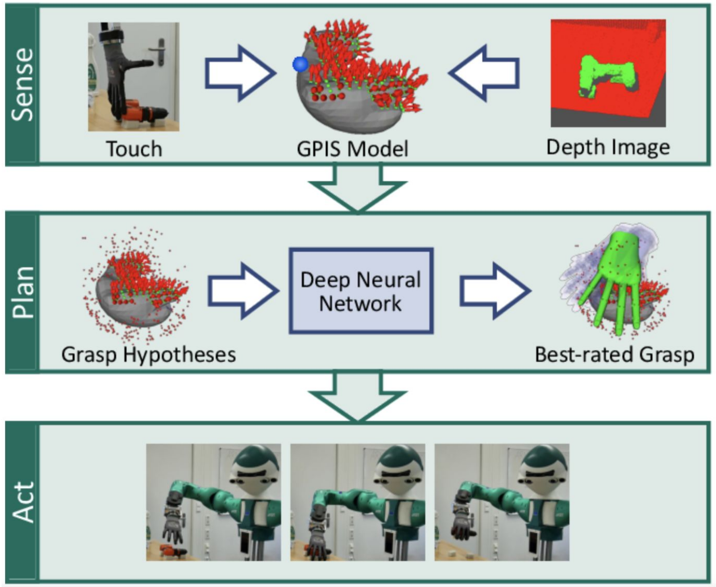
- Shape completion as auxiliary task:
    - Exploit the synergy between grasping and shape completion
    - Obtain more informed quality function or regression model

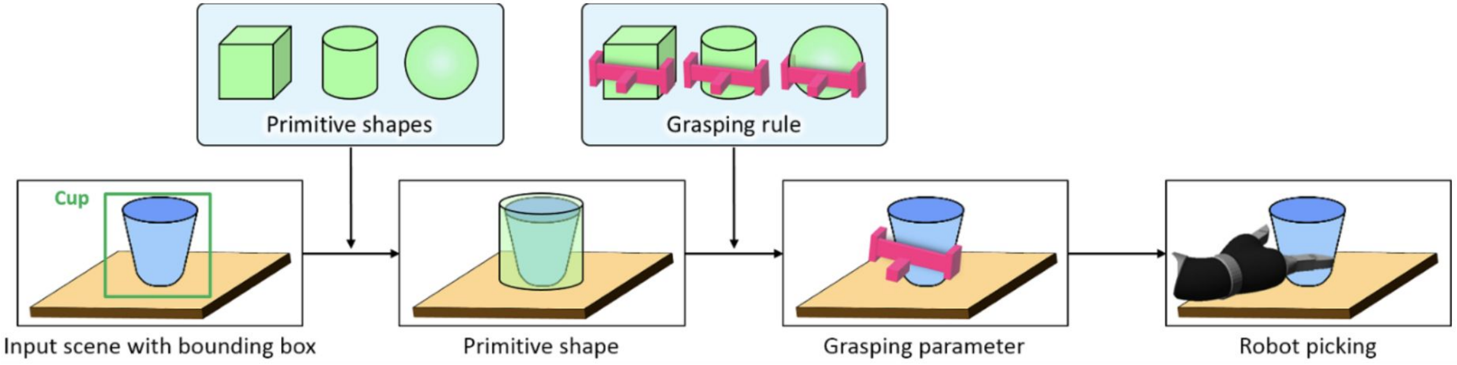


Synergies between affordance and geometry: 6-dof grasp detection via implicit representation

Supporting methods | Shape approximation

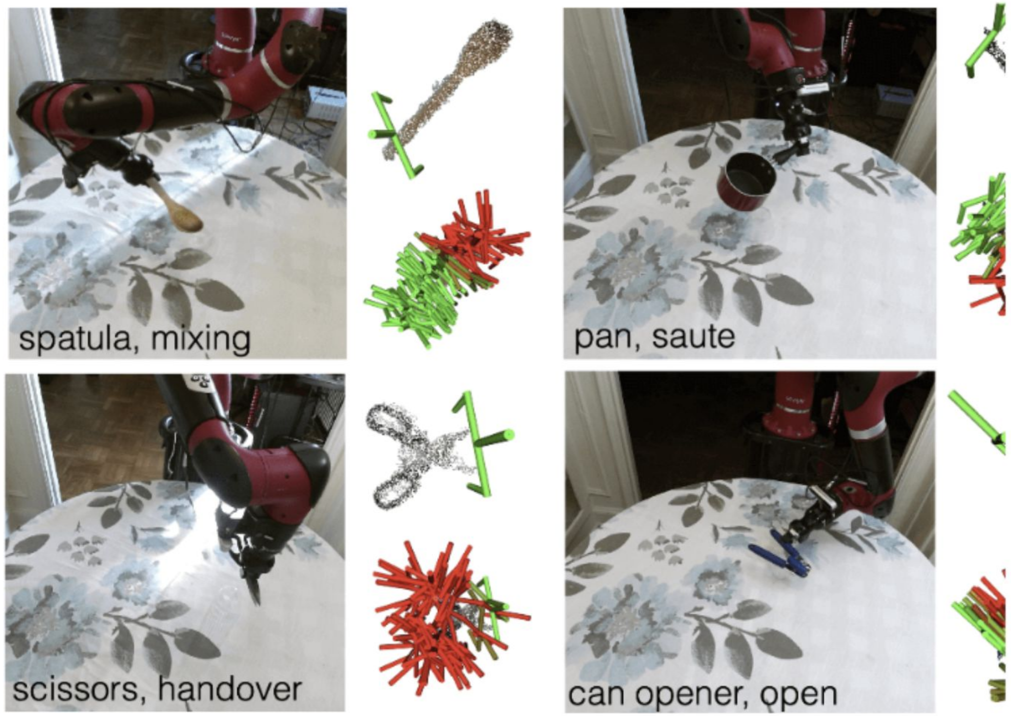
- Other
  - Visual-tactile grasping:
    - [44] gather tactile info to complete the shape during grasping
  - Approximates the object using shape primitives [147]





Supporting methods | Affordance

- Success of grasping -> additional considerations for what kind of task it is used for
- Geometry -> higher-level reasoning (functional)
- Methods
  - Segmentation + analytical
  - Affordance-aware quality function
  - Keypoint-based approach



Dataset design

- Object sets
- Household items such as food, toys and tools.
    - YCB, BigBIRD, KIT
- Large-scale object model repositories
    - ShapeNet [157], 3DNet [158], Grasp [36],
    - PSB [159], ModelNet [160]

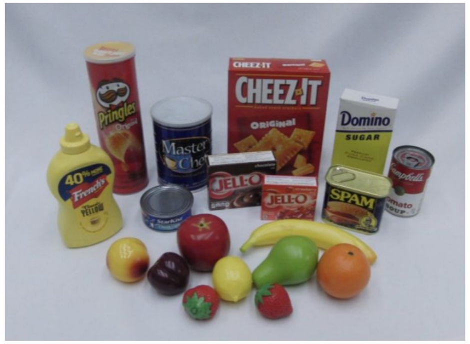


Fig. 1: Food items in the YCB Object Set: back row: chips can, coffee can, cracker box, box of sugar, tomato soup can; middle row: mustard container, tuna fish can, chocolate pudding box, gelatin box, potted meat can; front: plastic fruit (lemon, apple, pear, orange, banana, peach, strawberries, plum).

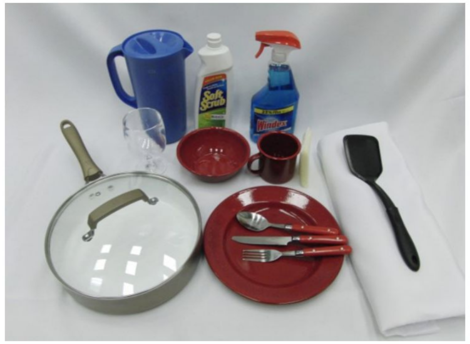


Fig. 2: Kitchen items in the YCB Object Set: back row: pitcher, bleach cleanser, glass cleaner; middle row: plastic wine glass, enamel-coated metal bowl, metal mug, abrasive sponge; front: cooking skillet with glass lid, metal plate, eating utensils (knife, spoon, fork), spatula, white table cloth

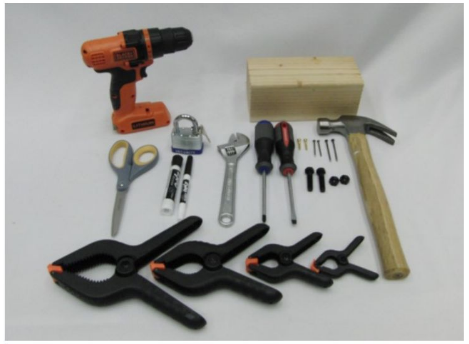


Fig. 3: Tool items in the YCB Object Set: back: power drill, wood block; middle row: scissors, padlock and keys, markers (two sizes), adjustable wrench, phillips and flat screwdrivers, wood screws, nails (two sizes), plastic bolt and nut, hammer; front: spring clamps (four sizes).

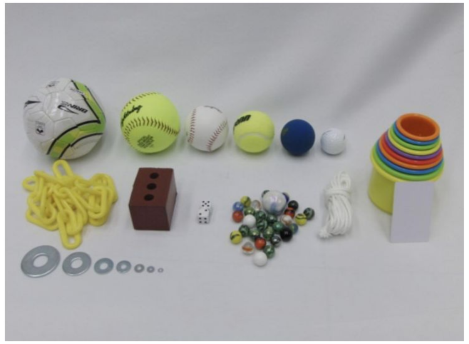
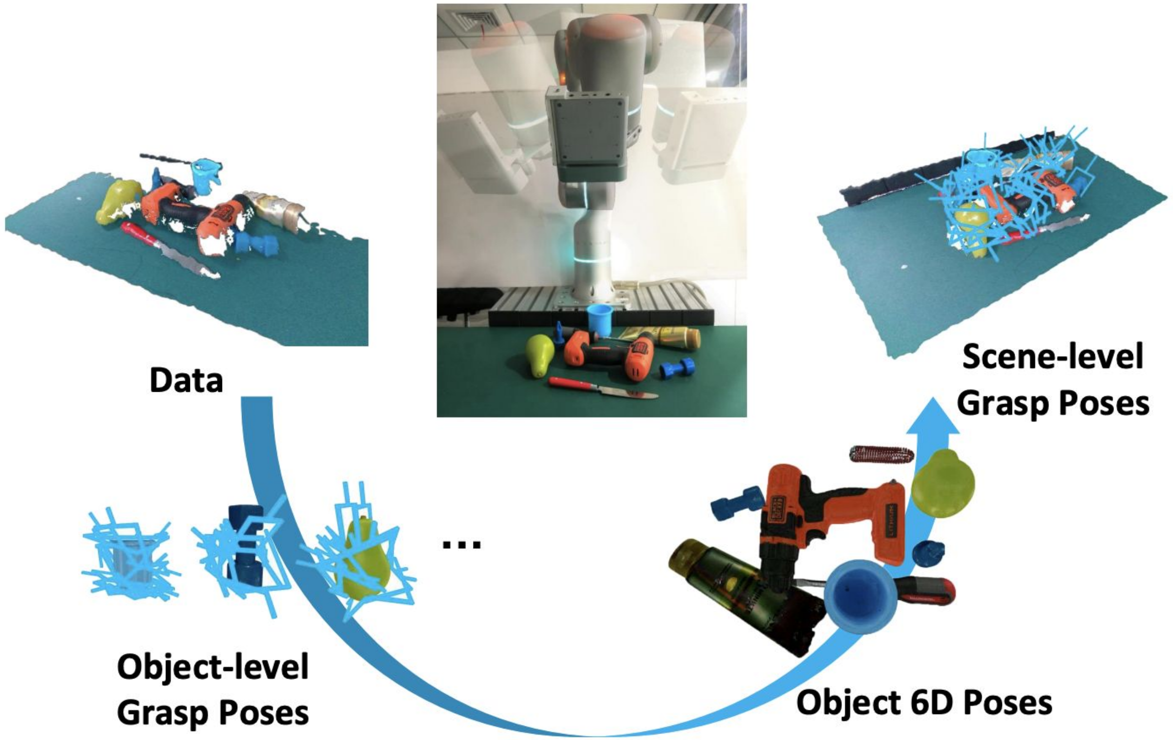


Fig. 4: Shape items in the YCB Object Set: back: Mini soccer ball, softball, baseball, tennis ball, racquetball, golf ball, front: plastic chain, washers (seven sizes), foam brick, dice, marbles, rope, stacking blocks (set of 10), credit card blank.

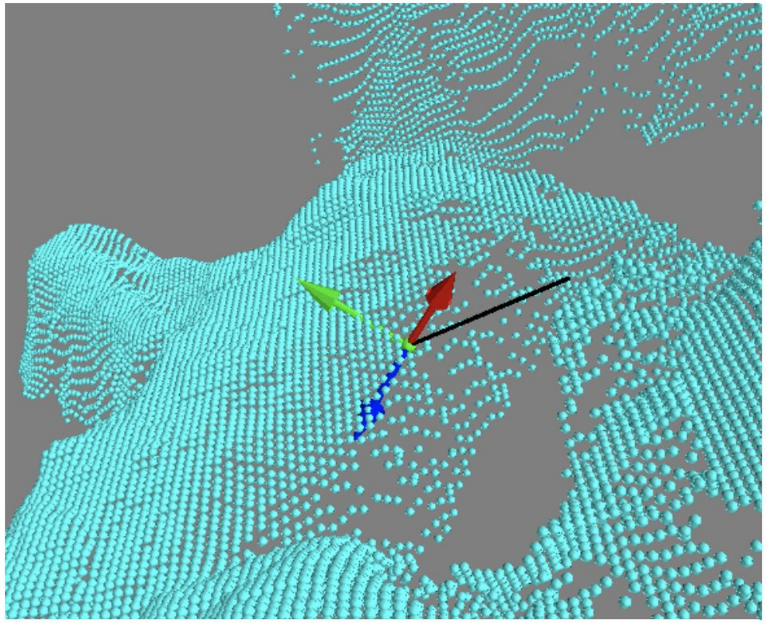
Dataset design

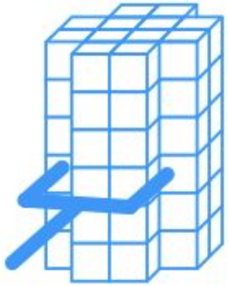
- Procedurally Generated Datasets
  - GraspNet-1Billion



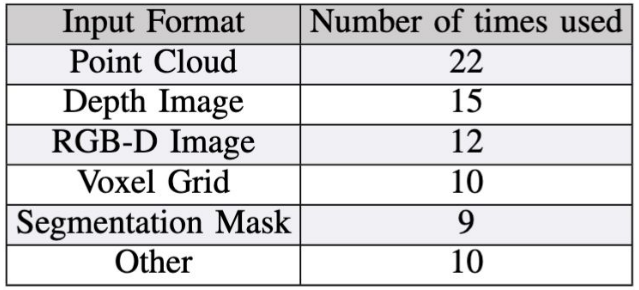
Dataset design | Data representation

- Point cloud
- Image
- Voxel Grid





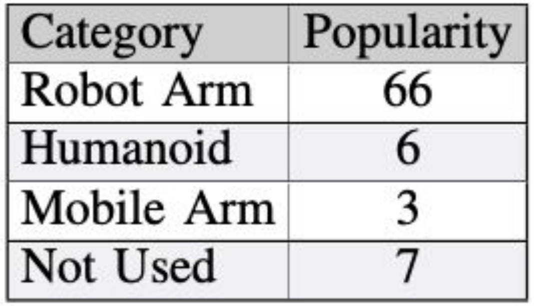
grasp pose $\mathbf{X}_g(\hat{\delta}, \hat{\omega})$



Input Format	Number of times used
Point Cloud	22
Depth Image	15
RGB-D Image	12
Voxel Grid	10
Segmentation Mask	9
Other	10

Benchmark | Experimental Evaluation

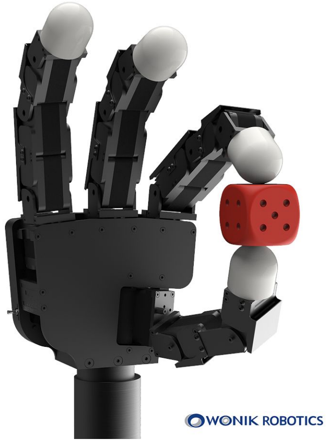
- Usually evaluated in real world
- Real world evaluation carries more weights
- Most works study robot arm, some use mobile arm or humanoid



Category	Popularity
Robot Arm	66
Humanoid	6
Mobile Arm	3
Not Used	7

Franka Panda

Franka gripper



Dexterous hand

Benchmark | Object Configurations

- Singulated
- Piled clutter
- Structured clutter

Singulated

Piled clutter

Structured clutter

Benchmark | Performance metrics

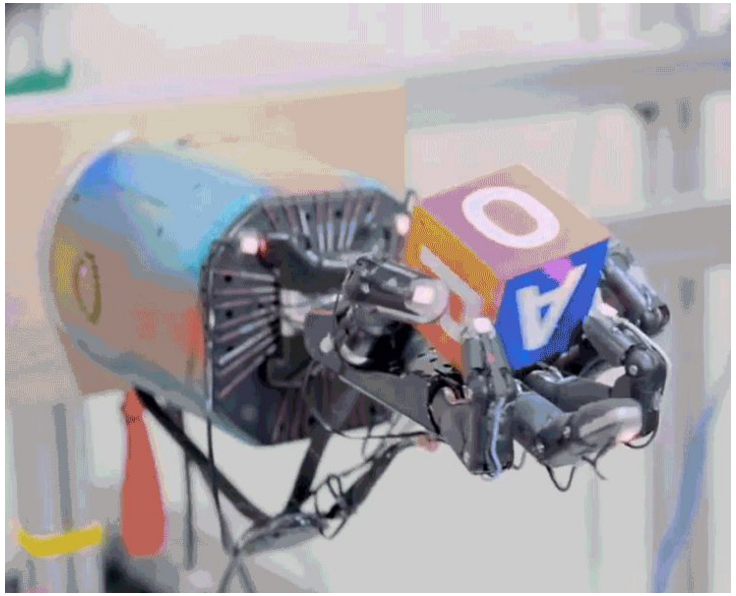
- **Grasp Success Rate**: The percentage of successful grasps
- **Completion / Clearance Rate**: The percentage of objects that are removed from the clutter (No. of Objects Grasped / Total No. of Objects in Clutter).
- **Coverage**: The percentage of sampled ground truth grasps that are within a threshold distance of any of the generated grasps.
- **Computation Time**: Time required to compute grasp hypothesis generation.

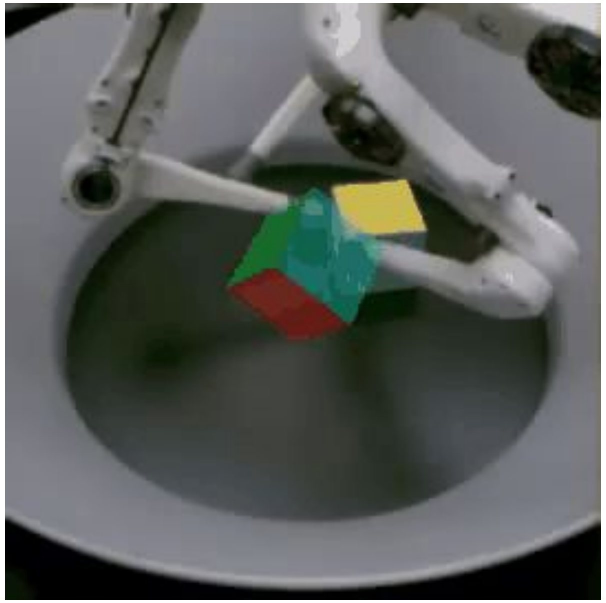
Discussion

Discussion and future directions

Many papers do not consider the semantics of the scene, focusing mostly on geometry. They also do not generally predict motion; manipulation is often handled separately, with the object assumed to be held firmly.

What level of understanding of the scene is required for (meaningful) grasping tasks? How about coupled planning and prediction?



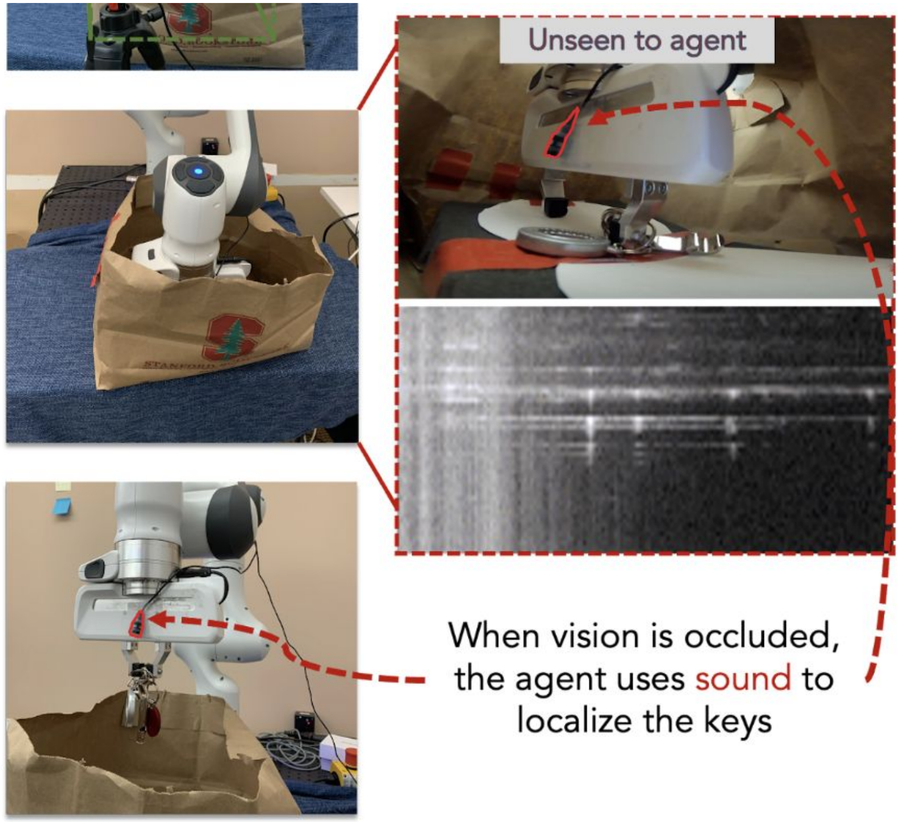


Discussion and future directions

Most papers use vision as the sole modality of perception.

What other modalities do you think are helpful for grasping, and why?

- Tactile information:
    - predict if the grasp is robust
    - slip detection
    - account for uncertainty of object pose
    - reconstruction of shape
- Sound



When vision is occluded, the agent uses sound to localize the keys

Discussion and future directions

- Most works try to find a collision-free path to the grasp pose. However, it’s not always possible to find a collision-free path
  - when the scene is densely cluttered
  - when the grasp pose is occluded
- **Instead of avoiding contact, how to leverage contact for better grasping?**