Local Neural Descriptor Fields:
Locally Conditioned Object Representations for Manipulation

Ethan Chun [1] Yilun Du [1] Anthony Simeonov [1] Tomas Lozano-Perez [1] Leslie Kaelbling [1]

[1]Computer Science and Artificial Intelligence Laboratory, MIT, USA

Abstract

- **Context:** A robot operating in a household environment will see a wide range of unique and unfamiliar objects. Given this variety, it is infeasible to train an agent on all objects it will encounter.
- **Problem:** When given a small set (5 - 10) of manipulation demonstrations on a single category of objects, can we successfully execute this skill on novel objects types in arbitrary SE(3) orientations?
- **Solution:** We present Local Neural Descriptor Fields (L-NDF) a method to generalize object manipulation skills acquired from a limited number of demonstrations, to novel objects from unseen shape categories.

Introduction

We design an imitation learning paradigm where we:

1. Encode a set of task demonstrations into a local neural descriptor field using a **SE(3) equivariant convolutional occupancy network** encoder.
2. Encode test objects using the same encoder.
3. Perform **local pose optimization** to match the demonstration pose to a pose on the test objects.

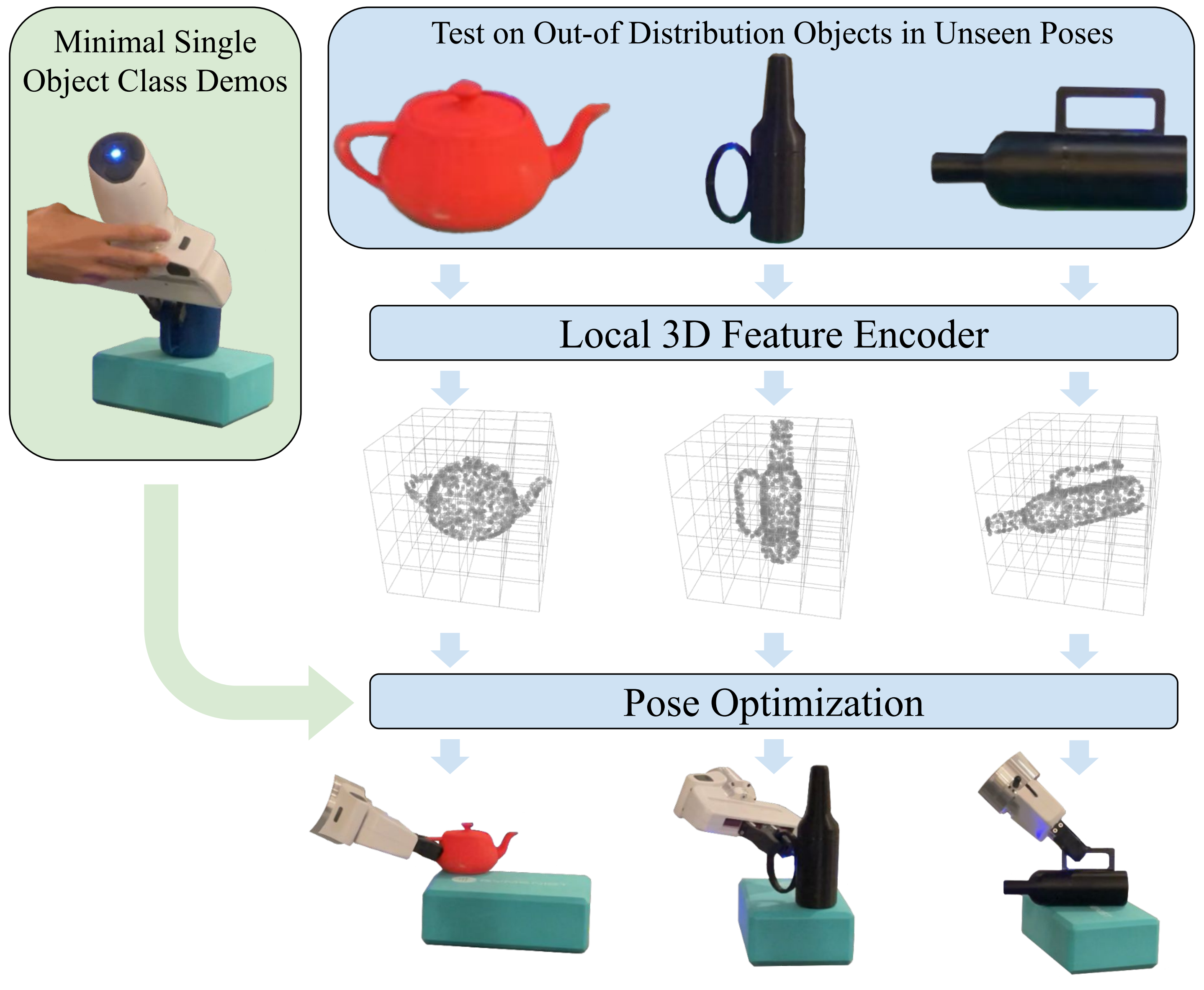
Minimal Single Object Class Demos
Test on Out-of Distribution Objects in Unseen Poses
Local 3D Feature Encoder
Pose Optimization

Figure 1. **L-NDF System Overview** – We encode all objects using a fine-tuned local encoder, then perform pose optimization to best match our demonstration pose.

Proposed Method

Architecture

We use a convolutional occupancy network encoder to encode a partial point cloud into a voxel grid of latents (illustrated in Fig. 2). These are queried to produce the final latent code.

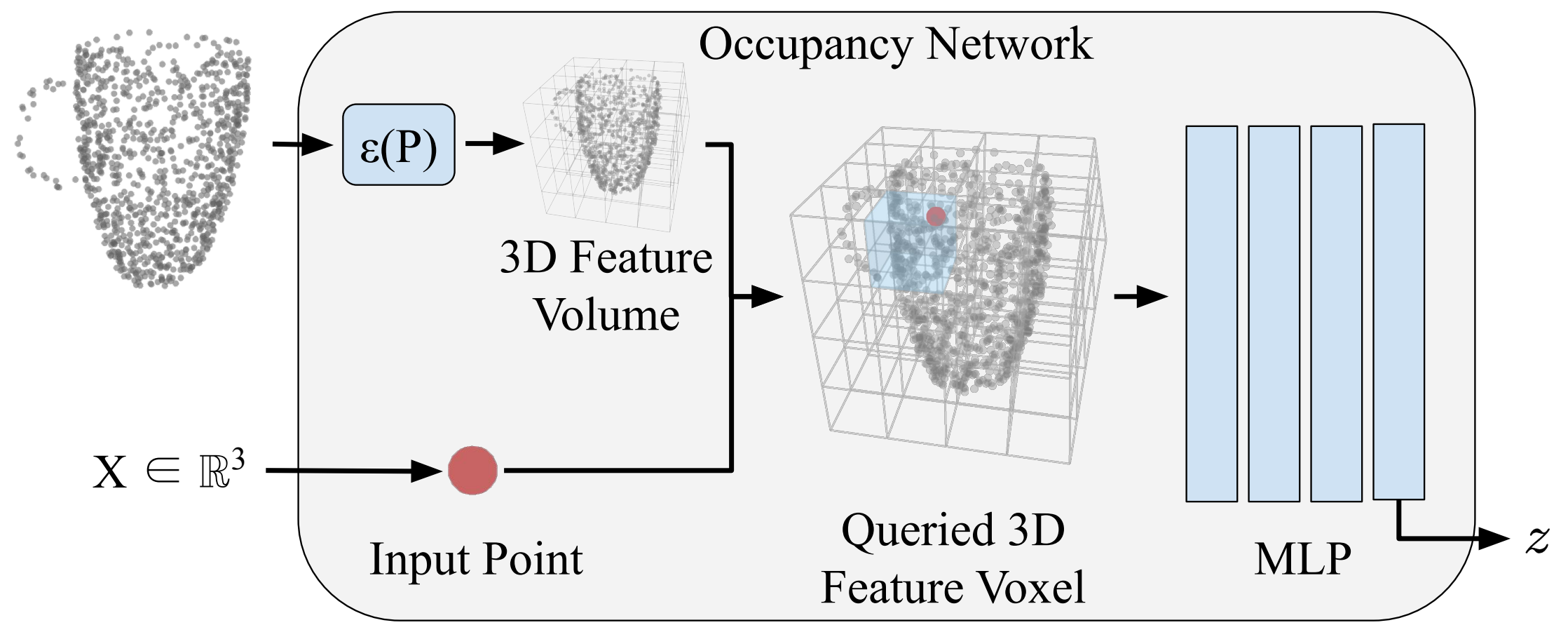

Figure 2. **Local Neural Descriptor Field Architecture** – A L-NDF takes any coordinate in 3D space, x, and a conditioning point cloud $\mathbf{P}$. It then uses an encoder $\epsilon(\mathbf{P})$ to encode $\mathbf{P}$ into a 3D feature volume from which the voxel containing $\mathbf{x}$ is queried. These feature are passed into an MLP decoder where the activations of the decoder's final layer are extracted to create the spatial descriptor, z.

Enforcing SE(3) Equivariance

We use a contrastive loss term to shape the network activations towards SE(3) equivariance.

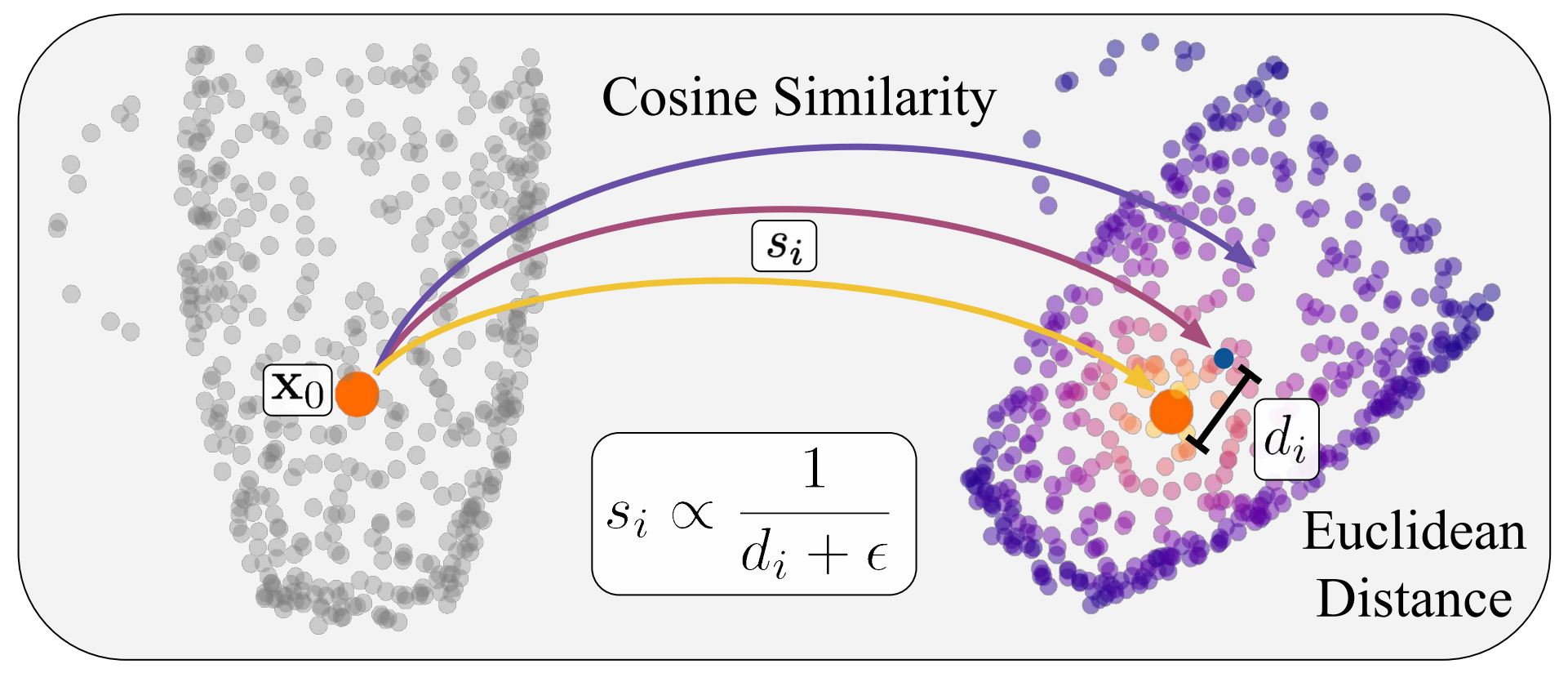

Figure 3. **Contrastive Loss Term for L-NDF** – The spatial descriptor of a 3D coordinate, $\mathbf{x}$, with respect to an observed point cloud, $\mathbf{P}$, is similar across any transform, $\mathbf{T} \in \mathrm{SE}(3)$. Additionally, geometrically farther points have decreasingly similar descriptors.

Pose Optimization

Query points are a proxy for the demonstration and target poses. We initialize query points at random rotations and translations within the bounding box of the observed point cloud. Additionally, we provide heuristics for selecting task-specific query points.

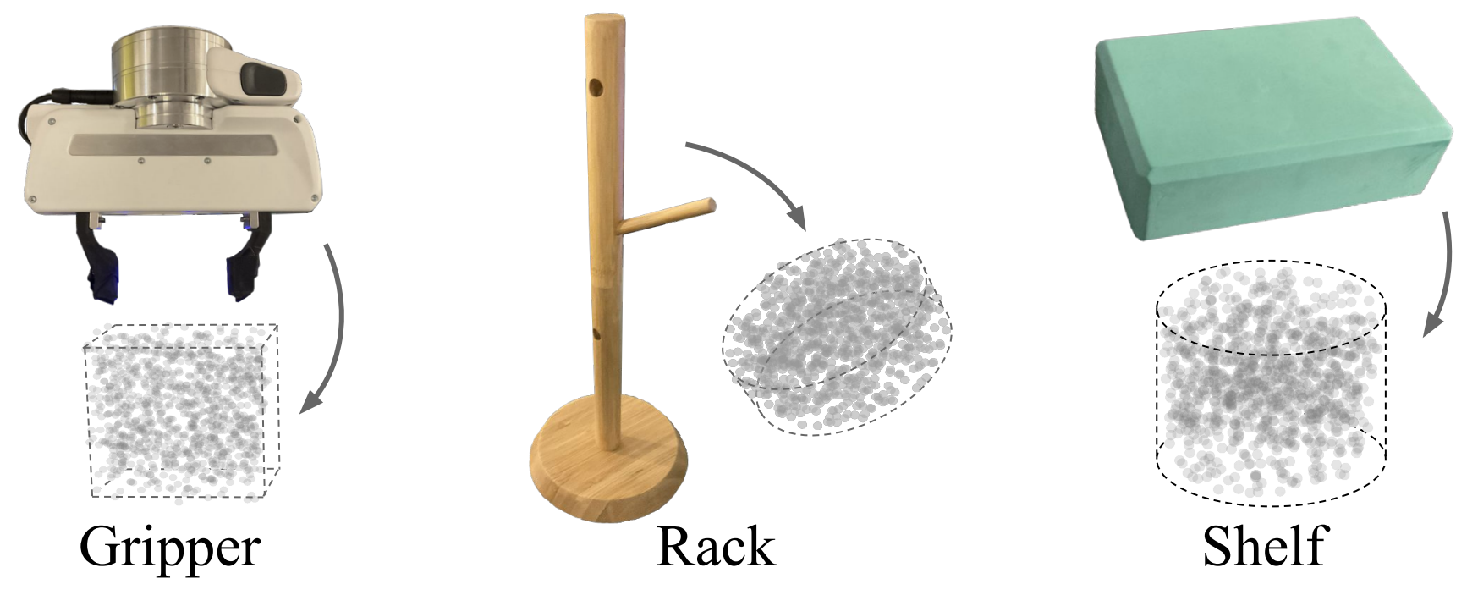

Figure 4. **Selecting Query Points** – For grasp and rack placement tasks, we use query points similar in size to contact geometry of the known object (gripper and peg size). For placement surfaces, we find larger query point selections performs well.

Experiments

Experimental Design

We evaluate L-NDF on a set of tasks involving mug-like objects, bowl-like objects, and bottle-like objects. We utilize both simulation environments and real-world testing to evaluate the performance of L-NDF.

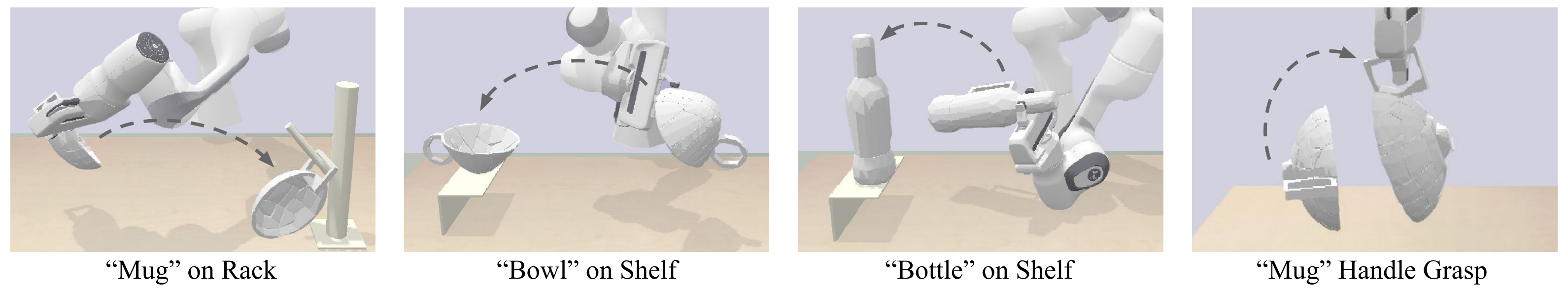

Figure 5. **Experimental Setup** – We provide ten simulated demonstrations of each task, then execute each on a set of 200 unseen objects. We measure grasp success, place success, and overall success. Grasp and place success check that the simulated object is in a stable configuration. Overall success checks if both grasp and place success occurred.

Experimental Results

We find that L-NDFs successfully generalize demonstrations to test objects in both simulation and the real world. Additionally, we demonstrate the viability of L-NDFs in a cluttered environment.

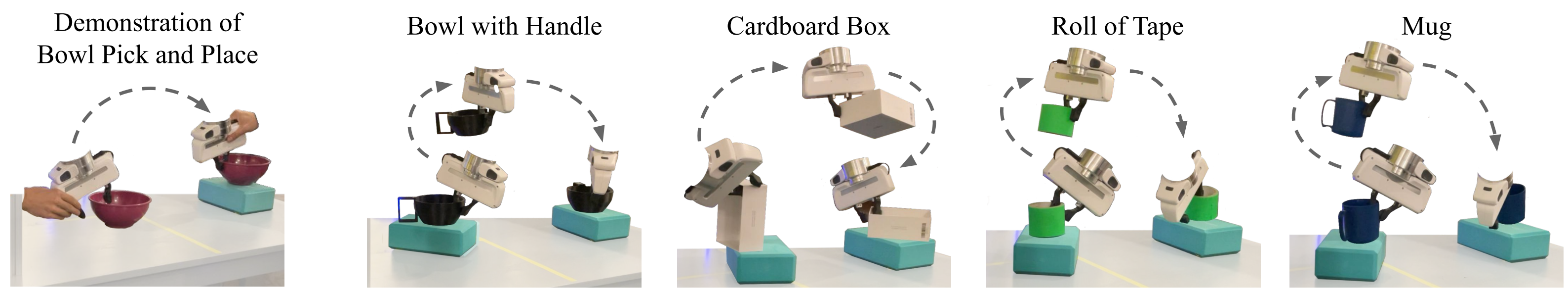

Figure 6. **Real world Execution** – We provide four real world demonstrations of grasping and placing two different bowls. We then successfully grasp and place a variety of unseen objects using a Franka Panda arm.

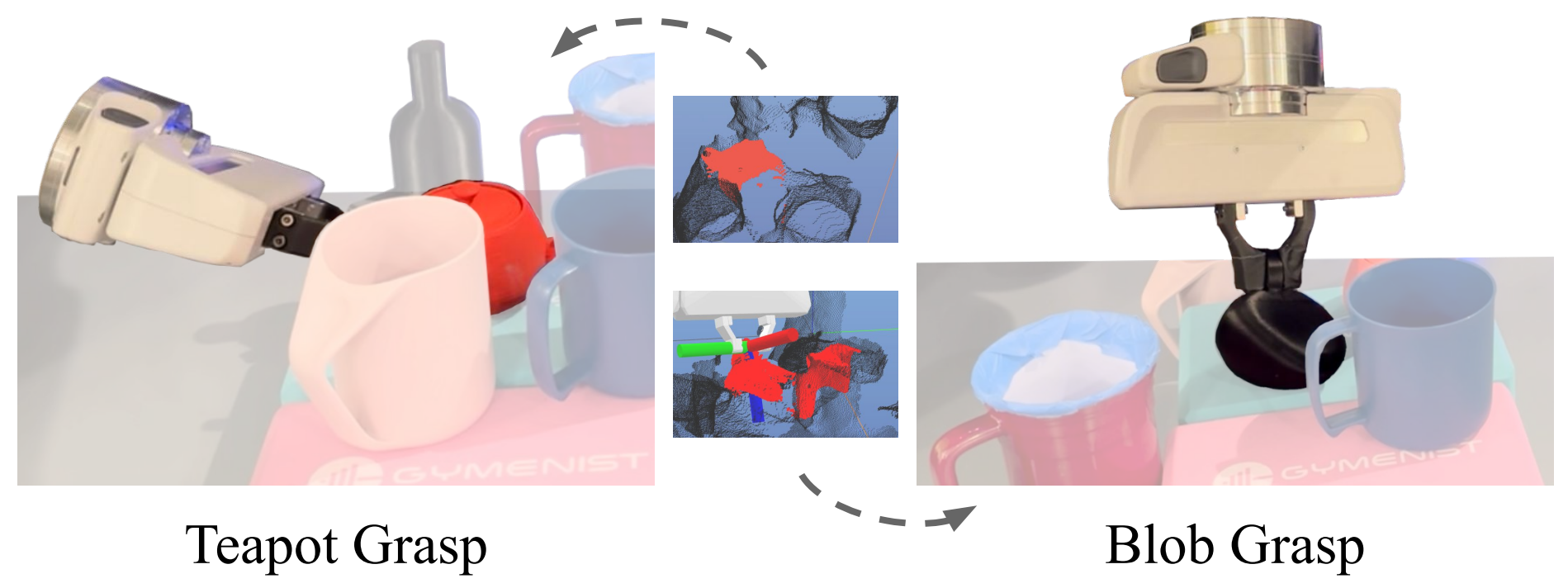

Figure 7. **Operating in Clutter** – We provide four real world demos of grasping a mug in an uncluttered scene. We then grasp a variety of objects from a cluttered environment using partial point clouds. We used Mask R-CNN for scene segmentation.

Conclusion

- We introduce Local Neural Descriptor Fields, an object representation that allow few-shot imitation learning of manipulation tasks on potentially novel categories of objects in both simulation and the real world.
- Additional information and code can be found at https://elchun.github.io/lndf/.