

FFHClutteredGrasping: Multi-Fingered Robotic Grasping in Cluttered Environment

Lei Zhang



Universität Hamburg
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Technische Aspekte Multimodaler Systeme

April.2024



Related Works

Part 1: Robotic Grasping from Cluttered Environment

Part 2: Dataset of Multi-fingered Robotic Grasping and Human Hand-Object Manipulation

Part3: Grasp Generation of Multi-fingered Robotic Grasping in Cluttered Environment.



Grasp from Cluttered Environment with Multi-Fingered Gripper

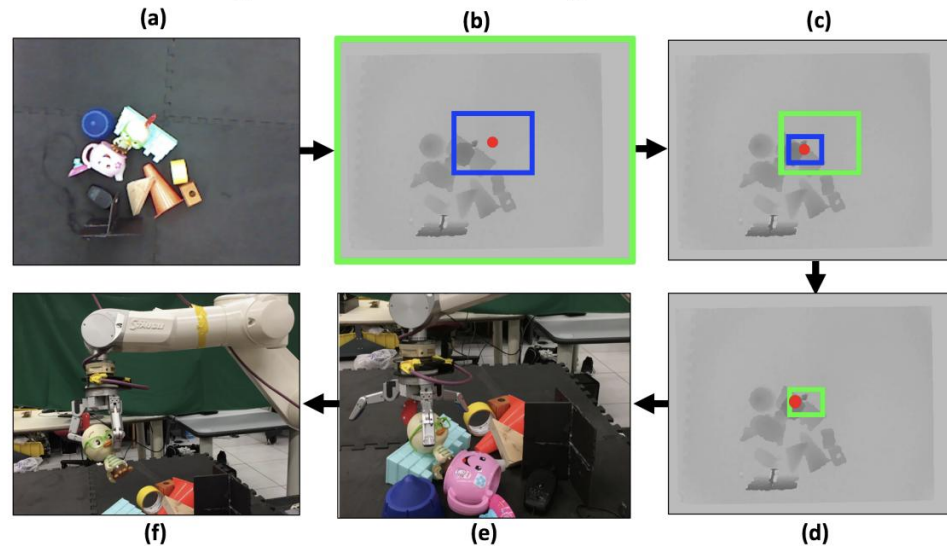
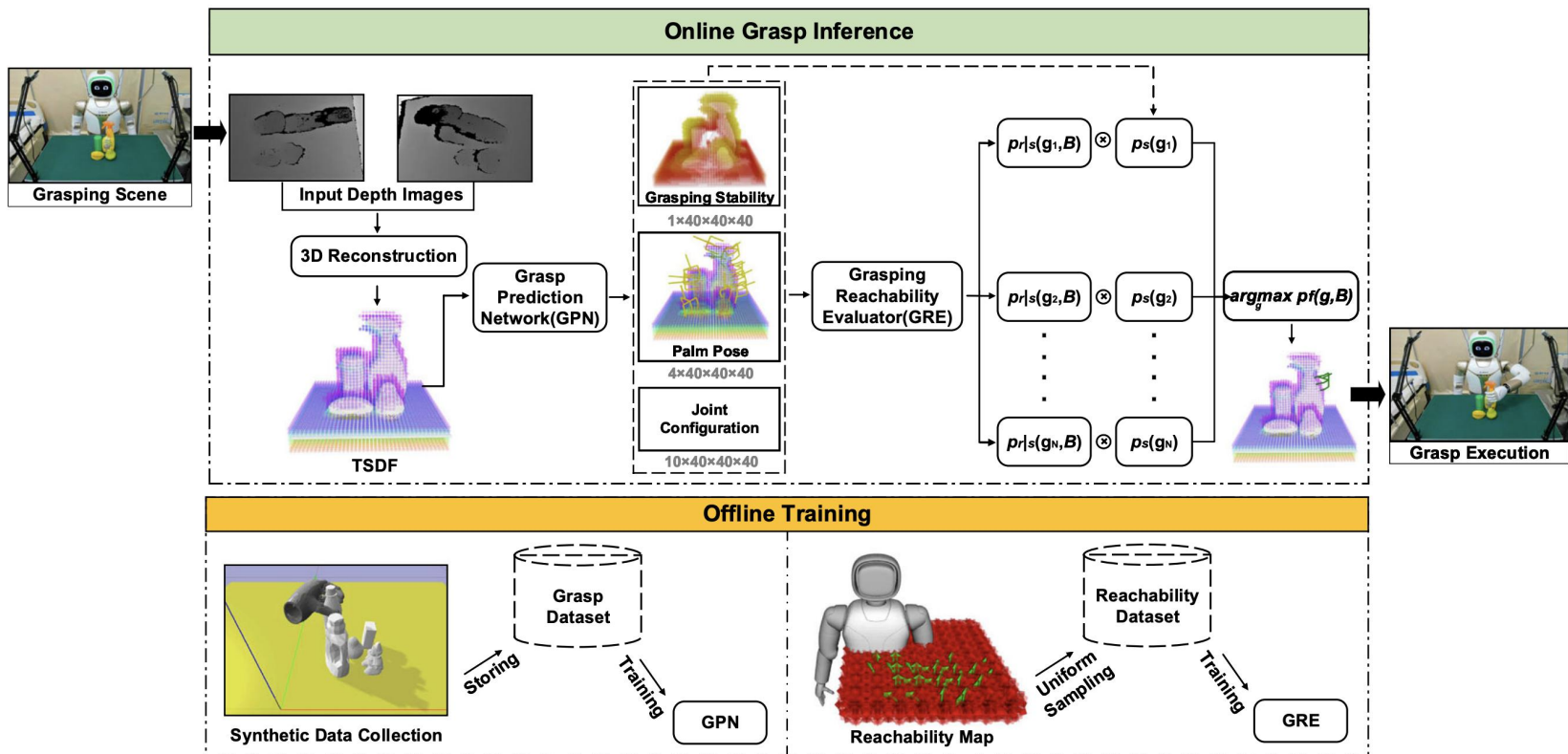


Fig. 1: Pixel-Attentive Policy Gradient Multi-Fingered Grasping. Given a scene of cluttered objects (a), our method takes in a single depth image and gradually zooms into a local region of the image to generate a good grasp. (b), (c) and (d) show the zooming process, in which the green bounding box represents the portion of the depth image the robot observes in the current timestep, and the blue bounding box represents the portion of the depth image the robot wants to observe in the next timestep. In the end, a full-DOF grasp is learned based on the final zoomed image (d) as shown in (e) and with the final pick-up shown in (f).

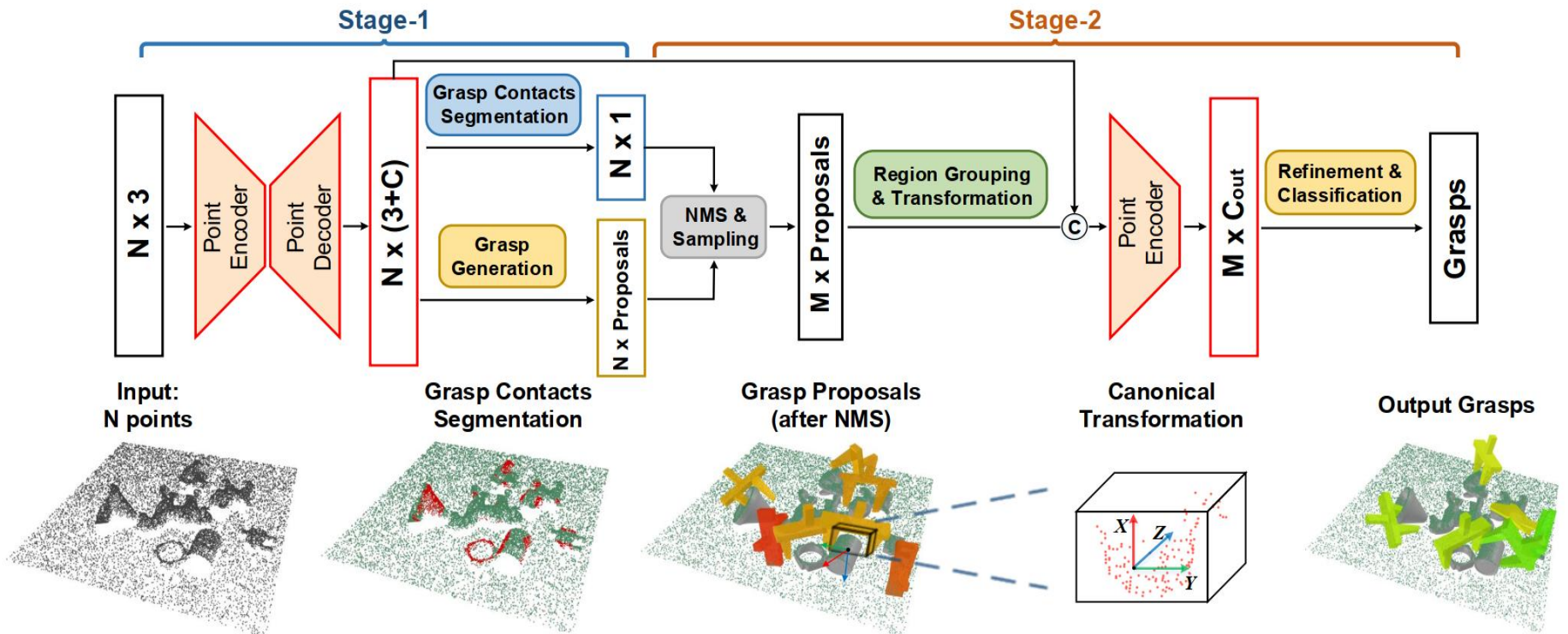
Pixel-Attentive Policy Gradient for Multi-Fingered Grasping in Cluttered Scenes, 2019 IROS

Grasp from Cluttered Environment with Multi-Fingered Gripper



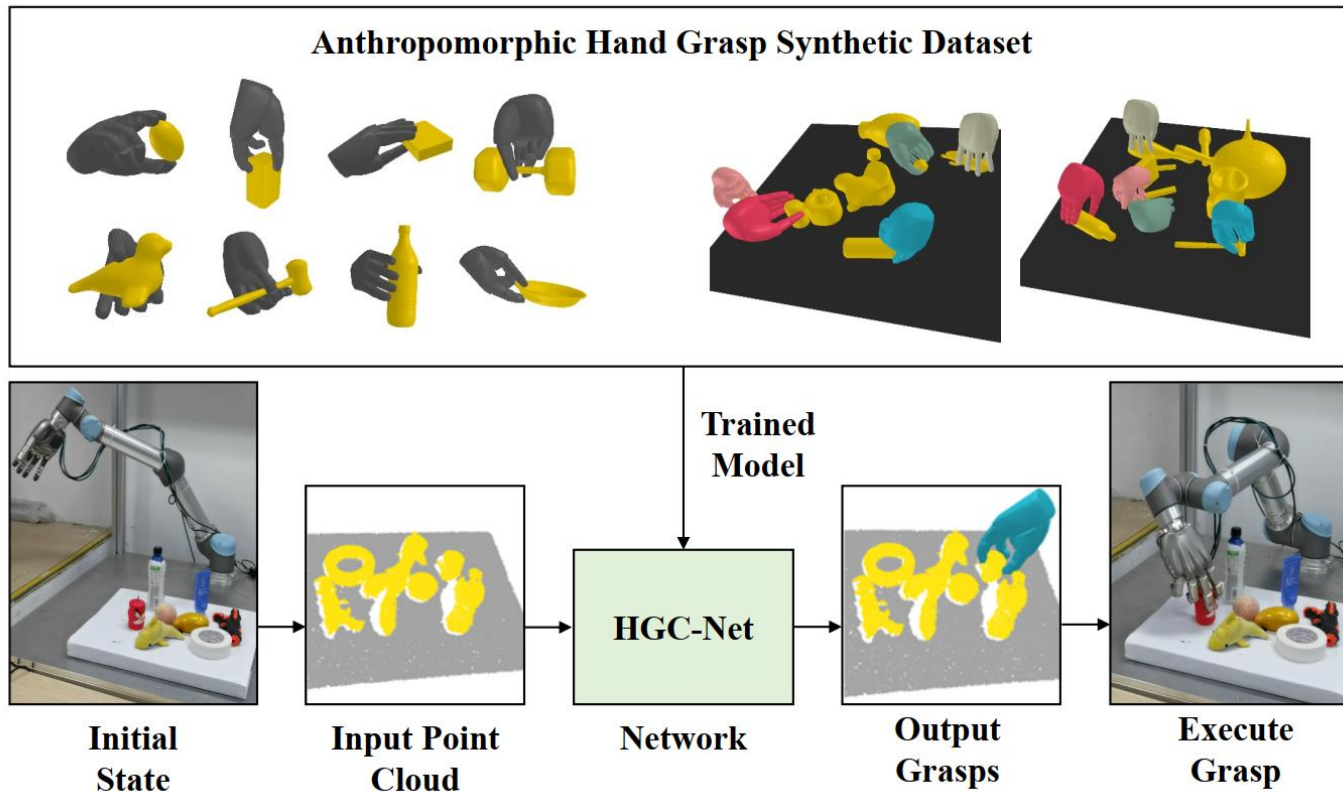
Planning multifingered grasps with reachability awareness in unrestricted workspace, 2023

Grasp from Cluttered Environment with Two-jaw Gripper



GPR: Grasp Pose Refinement Network for Cluttered Scenes, ICRA 2021

Grasp from Cluttered Environment with Multi-Fingered Gripper



HGC-Net: Deep Anthropomorphic Hand Grasping in Clutter, ICRA 2022

Grasp from Cluttered Environment with Two-jaw Gripper

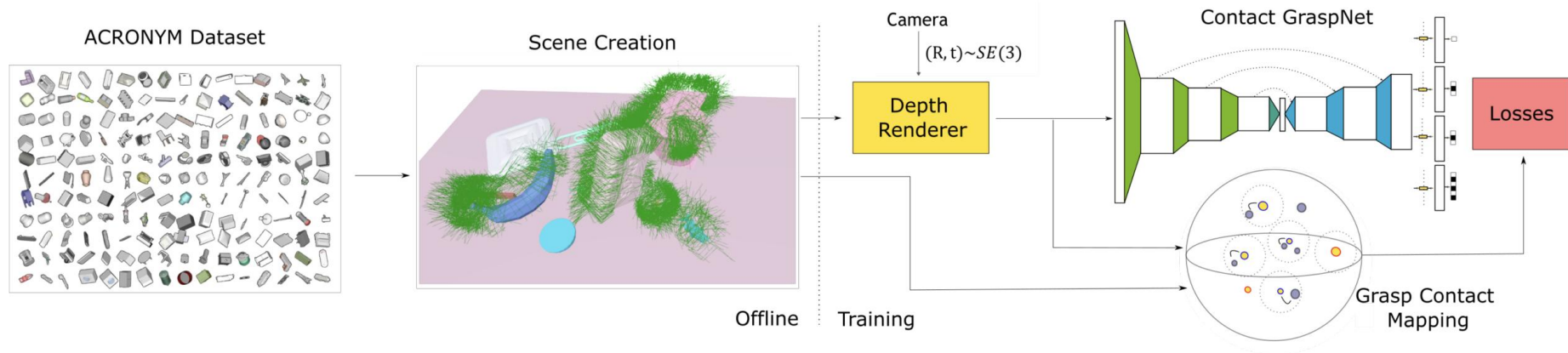


Fig. 2. Training Data Pipeline. We place object meshes with dense grasp annotations from the ACRONYM dataset [32] at random stable poses in scenes. Grasp poses that produce gripper model collisions are removed. Resulting grasps are mapped to their contacts on the mesh surface. During training, we sample virtual cameras to render point clouds from the scenes. We consider recorded points (yellow) as positive contacts if there exists a mesh contact (blue) in a 5mm radius and associate the grasp transformation belonging to the closest mesh contact to them. These per-point annotations are used to supervise the Contact Grasp Network.

Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes, ICRA 2021

GraspNet-1 Billion: Grasping dataset of two-jaw gripper

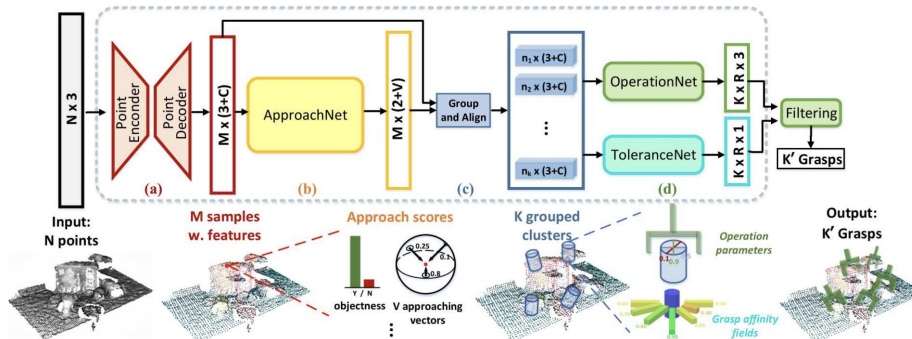
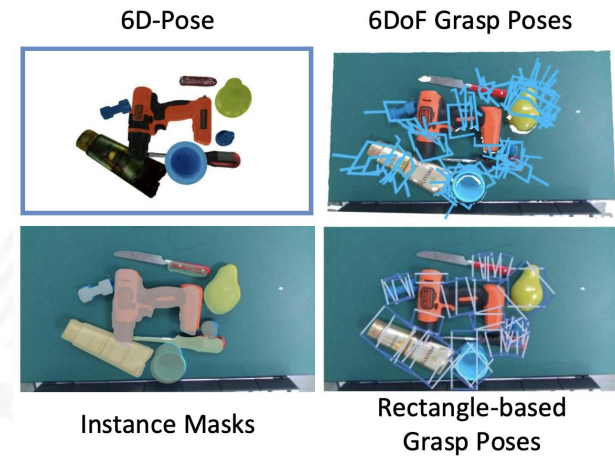


Figure 5. Overview of our end-to-end network. (a) For a scene point cloud with N point coordinates as input, a point encoder-decoder extracts cloud features and samples M points with C -dim features. (b) Approaching vectors are predicted by ApproachNet and are used to (c) grouped points in cylinder regions. (d) OperationNet predicts the operation parameters and ToleranceNet predicts the grasp robustness. See text for more details.



Dense Annotations

GraspNet-1 Billion: A Large-Scale Benchmark for General Object Grasping, CVPR 2020

Grasping dataset of human hand: DexYCB



Figure 1: Two captures (left and right) from the DexYCB dataset. In each case, the top row shows color images simultaneously captured from three views, while the bottom row shows the ground-truth 3D object and hand pose rendered on the darkened captured images.

DexYCB: A Benchmark for Capturing Hand Grasping
of Objects, CVPR 2021

Grasping dataset of human hand: Contactpose

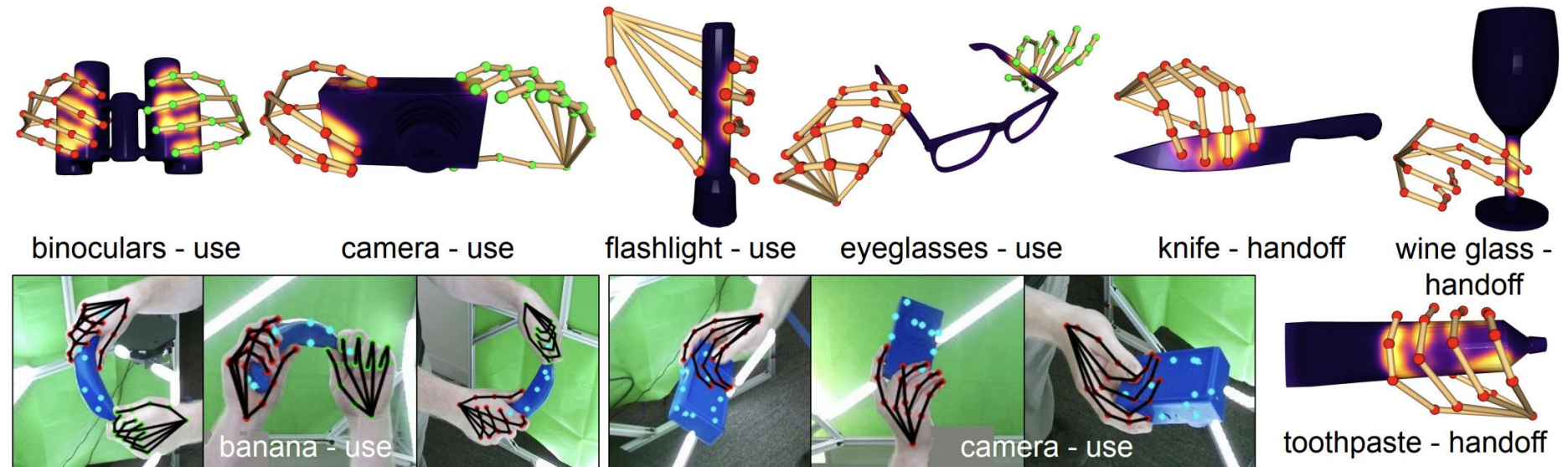


Fig. 1: Examples from ContactPose, a dataset capturing grasps of household objects. ContactPose includes high-resolution contact maps (object meshes textured with contact), 3D joints, and multi-view RGB-D videos of grasps. Left hand joints are **green**, right hand joints are **red**.

Contactpose: A dataset of grasps with object contact and hand pose, ECCV 2020

Grasping dataset of human hand: Ganhand



Figure 1: **GanHand** predicts hand shape and pose for grasping multiple objects given a single RGB image. The figure shows sample results on the YCB-Affordance dataset we propose, the largest dataset of human grasp affordances in real scenes.

Ganhand: Predicting human grasp affordances in multi-object scenes, CVPR 2020

Grasping dataset of multi-fingered robotic hand: Fast-Grasp'D

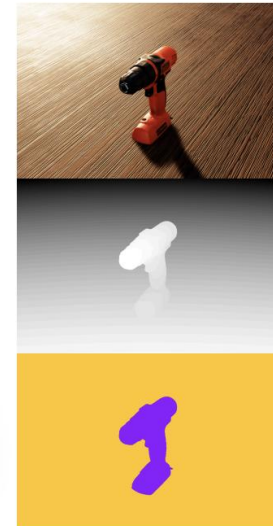


Fig. 1: The **Grasp'D-1M dataset** contains one million unique grasps, each with multi-modal visual inputs for training vision-based robotic grasping. We synthesize these grasps with a new differentiable grasping simulator, *Fast-Grasp'D*. Gradient information accelerates the grasp search, allowing us to search the full-DOF space (without eigengrasps) and simulate thousands of contacts to produce a dataset of contact-rich, stable grasps that can improve any learned grasping pipeline.

Fast-Grasp'D: Dexterous Multi-finger Grasp
Generation Through Differentiable Simulation, 2023₂

Grasping dataset of multi-fingered robotic hand: GenDexGrasp

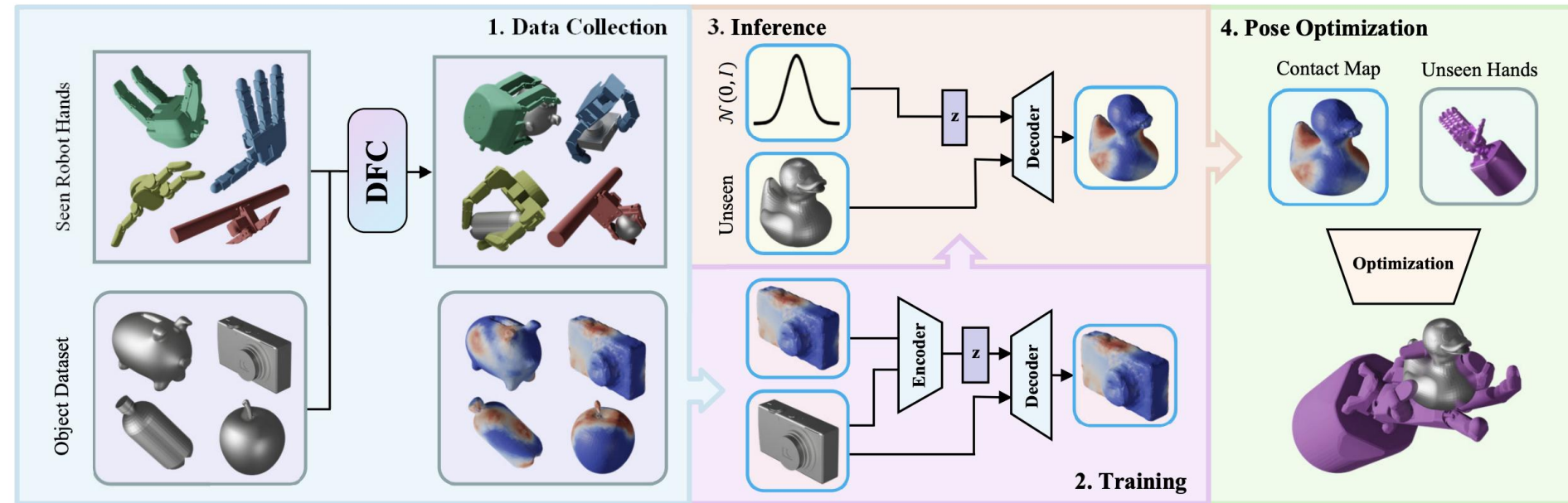


Fig. 4: An overview of the GenDexGrasp pipeline. We first collect a large-scale synthetic dataset for multiple hands with DFC. Then, we train a CVAE to generate hand-agnostic contact maps for unseen objects. We finally optimize grasping poses for unseen hands using the generated contact maps.

Grasping dataset of multi-fingered robotic hand: DexGraspNet

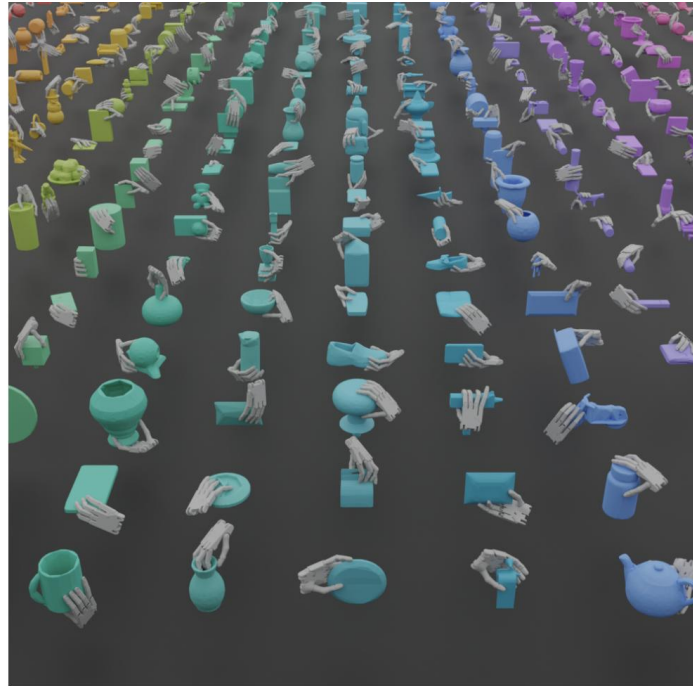
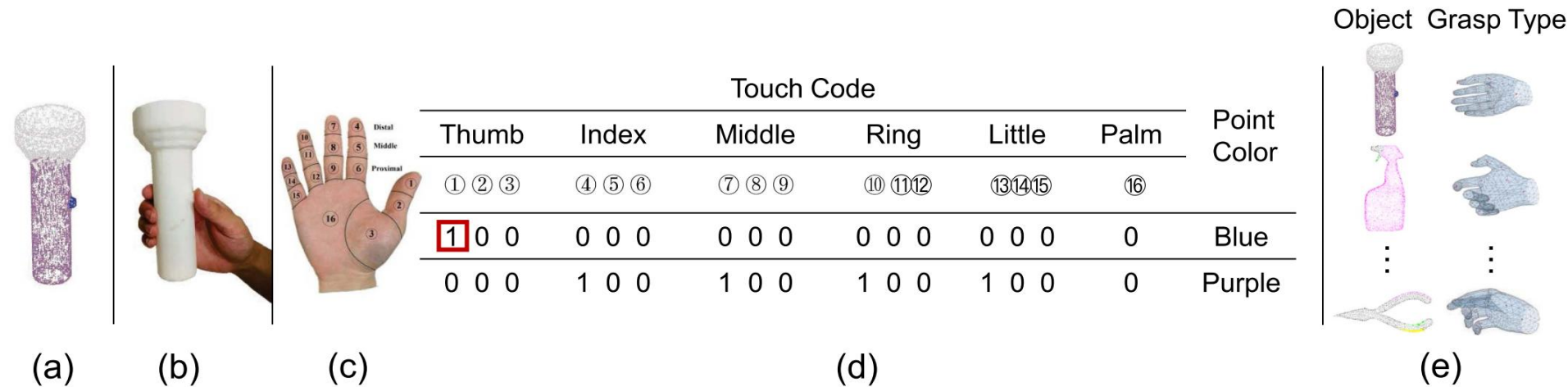


Fig. 1: A visualization of DexGraspNet. DexGraspNet contains 1.32M grasps of ShadowHand [8] on 5355 objects, which is two orders of magnitudes larger than the previous dataset from DDG [9]. It features diverse types of grasping that cannot be achieved using *GraspIt!* [10].

Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation, ICRA 2023

Different Hand-Object Representations: Touch Code



Relationship between **functional grasping** and **touch code**

Hand Object Representations: Touch Code

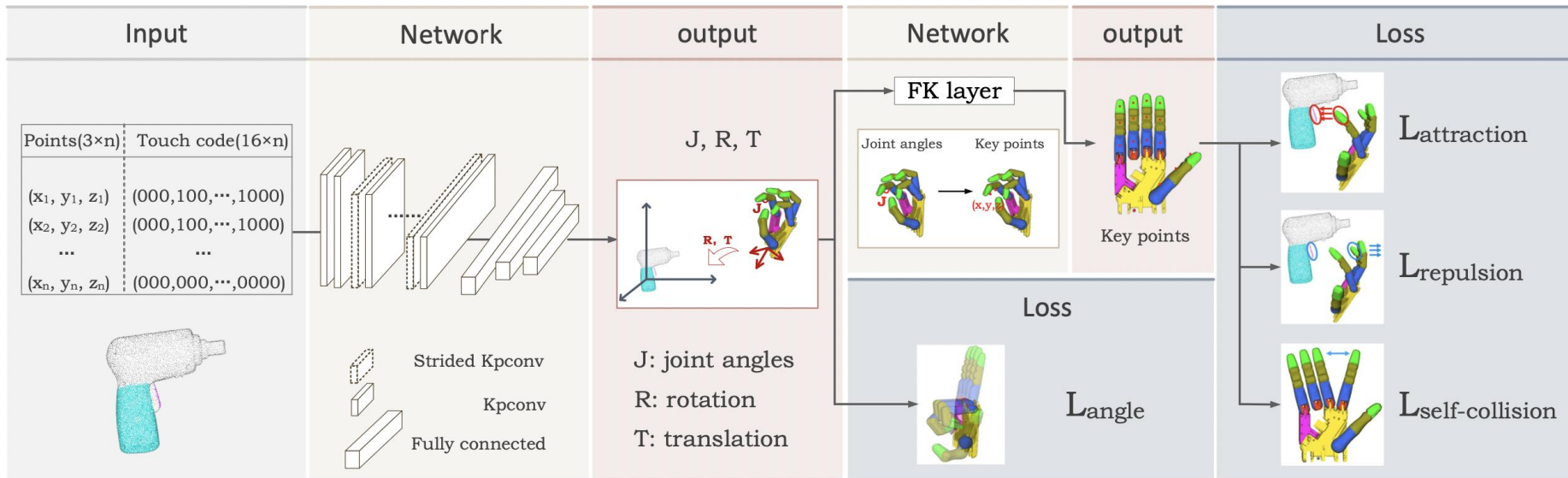
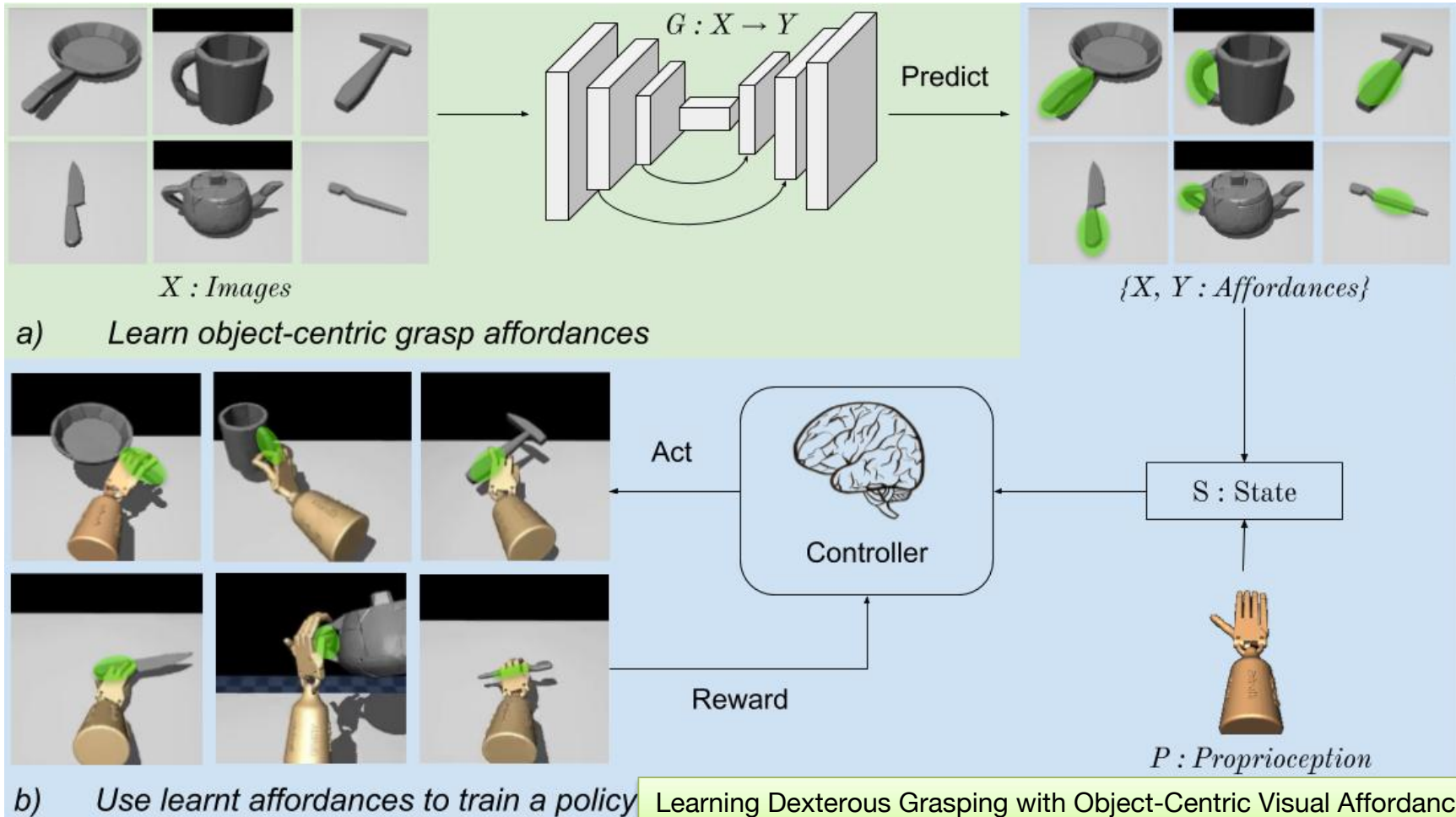


Figure 4: The overall architecture of our functional grasp synthesis framework. The original point cloud of the object with the 16-bit ‘touch code’ is fed into the network, which generates the configurations of the hand that conform to the functional grasp under the guidance of four loss functions.

Hand Object Representations: Affordance Regions



Hand Object Representations: Affordance Regions

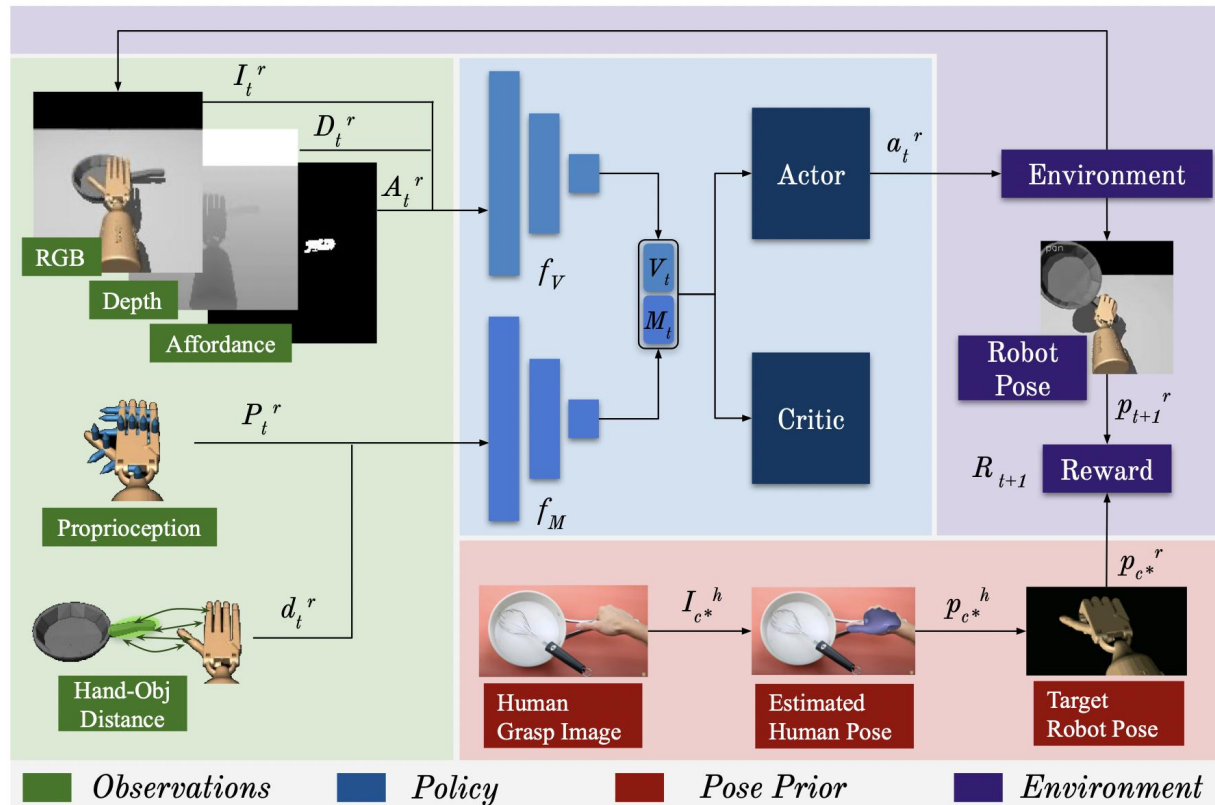


Figure 2: **Overview of DEXVIP.** We use grasp poses inferred from Internet video to train a dexterous grasping policy. An actor-critic network (blue) processes sensory observations from visual and motor streams (green) to estimate agent actions. Human hand poses derived from how-to videos (red) encourage the agent to explore worthwhile

DexVIP: Learning Dexterous Grasping with Human Hand Pose Priors from Video, University of Texas at Austin + Facebook AI

Hand Object Representations: Affordance Score

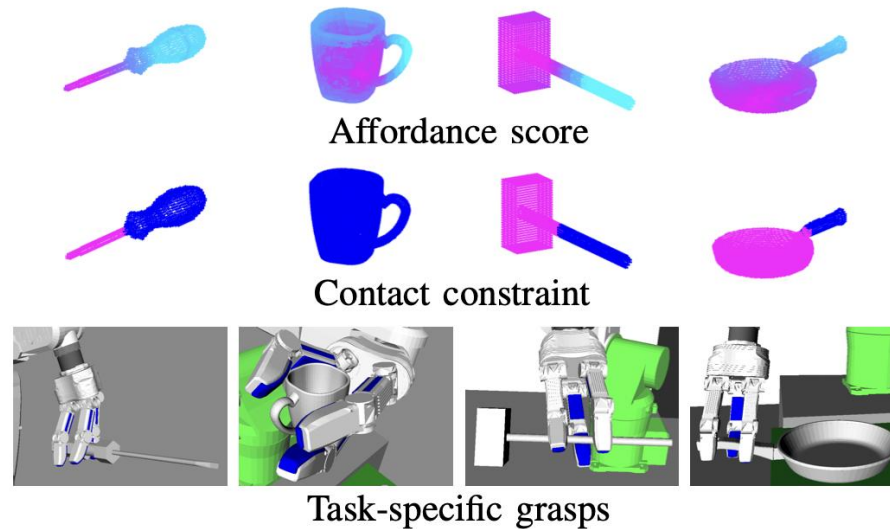


Fig. 1: Given the shape of an object and a task, we detect object part affordances. From these we formulate grasp constraints, such as a contact location constraint. These constraints are then utilized to compute task-specific grasps as shown here for example tasks poke, pour, pound and support on the objects screwdriver, mug, hammer and pan respectively. Magenta color indicates high affordance score (top) and contact avoidance constraint for grasping (middle).

Hand Object Representations: Signed Distance

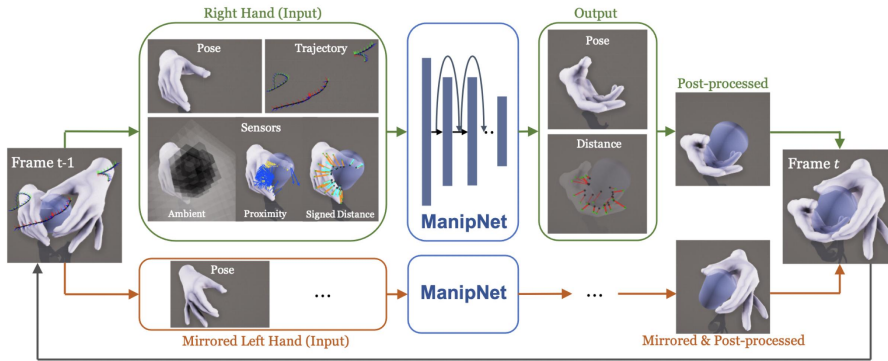


Fig. 2. The outline of our framework. Given the poses of two hands, the shapes of objects as well as the trajectories of two wrists and objects at frame $t - 1$. The inputs of the two hands will be generated separately and fed into a shared neural network. Correspondingly, the poses for the two hands at frame t will be synthesized from the outputs of the neural network.

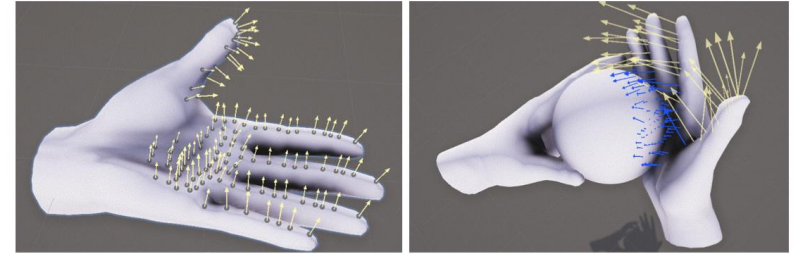


Fig. 4. Left: The 104 Proximity sensors on the hand mesh. Right: Proximity Sensors cast rays along the hand surface normal until they hit the object surface (blue arrows), or at a maximum distance (yellow arrows).

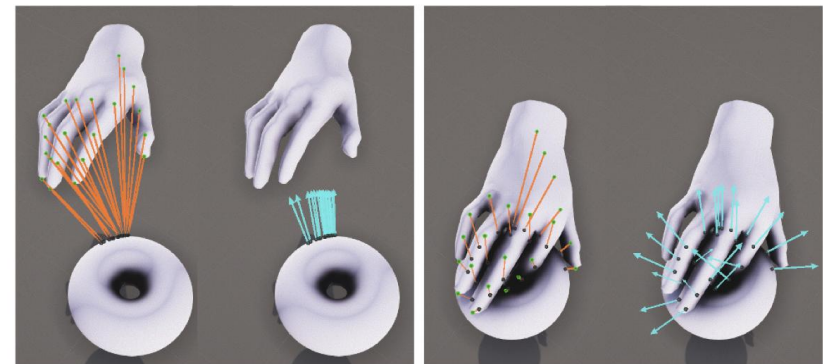


Fig. 5. Two examples of the Signed Distance sensors for the right hand. The hand joints are shown in green. Orange lines indicate the distance from the hand joints to the torus. Cyan arrows are surface normals on the torus.

ManipNet: Neural Manipulation Synthesis with a Hand-Object Spatial Representation

Hand Object Representations: Implicit Representation

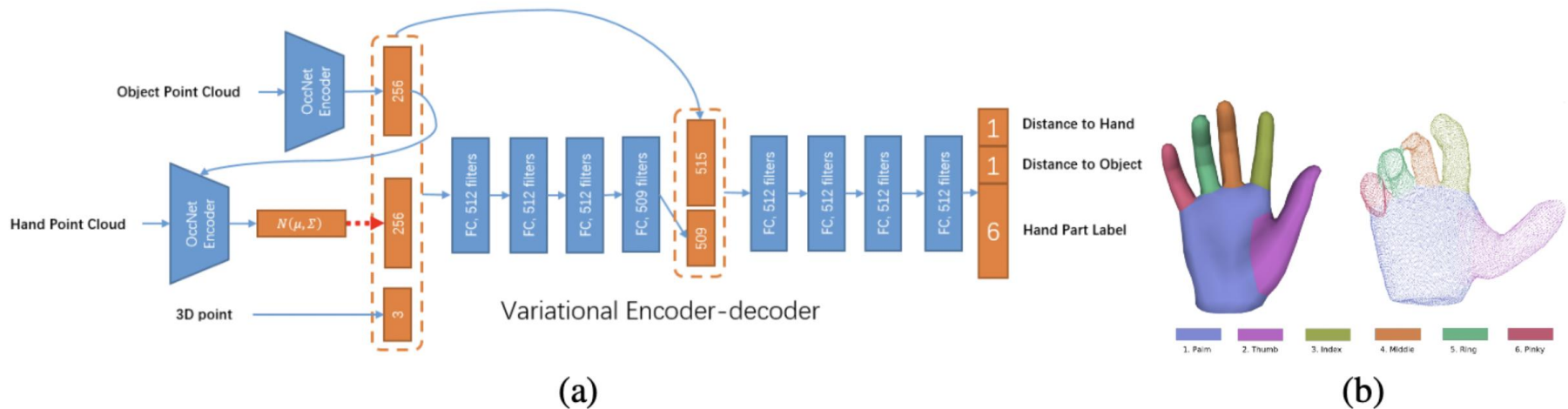


Figure 2: (a) Illustration of the generative grasping field network conditioned on the object point cloud. The red dashed arrow denotes sampling from a distribution. Architecture details are described in Appendix A. (b) Illustration of hand segmentation. Left is our hand part annotation on the MANO model. Right is an example of our *predicted* surface points with hand part labels.

Robotic and Human-Hand Grasping Dataset

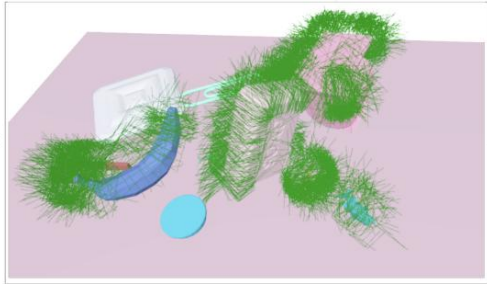
COMPARISON OF MULTI-FINGERED ROBOTIC HAND GRASPING DATASETS.

Methods	Hand Type	Cluttered Scene	Grasp Quality	Evaluation Metric	Contact Distance/ Map	Contact Semantic	Affordance
ContactPose [18], GRAB [30]	Human	✗	✗	-	✓	✓	✗
DexYCB [17]	Human	✓	✗	-	✓	✗	✗
GanHand [15]	Human	✓	✗	-	✓	✗	✓
DDGC [29]	Robot	✓	✓	GraspIt! [1]	✗	✗	✗
Columbia Grasp Database [31]	Robot	✗	✓	GraspIt! [1]	✗	✗	✗
Fast-Grasp'D [16]	Robot	✗	✓	Trial-and-Error	✗	✗	✗
DexGraspNet [4]	Human&Robot	✗	✓	Trial-and-Error	✓	✗	✗
GenDexGrasp [20]	Robot	✗	✓	Trial-and-Error	✓	✗	✗
Ours	Robot	✓	✓	Trial-and-Error	✓	✓	✓

No dataset for robotic hand, considering both cluttered scene, grasp quality. contact distance and semantic map, and affordance information.

Option for Grasp Generation from Clutter

Scene Creation



Offline

Single object/ Multiple objects

Contact-GraspNet:
Efficient 6-DoF Grasp Generation in Cluttered Scenes

Type1: Grasp Generation using Analytical Method

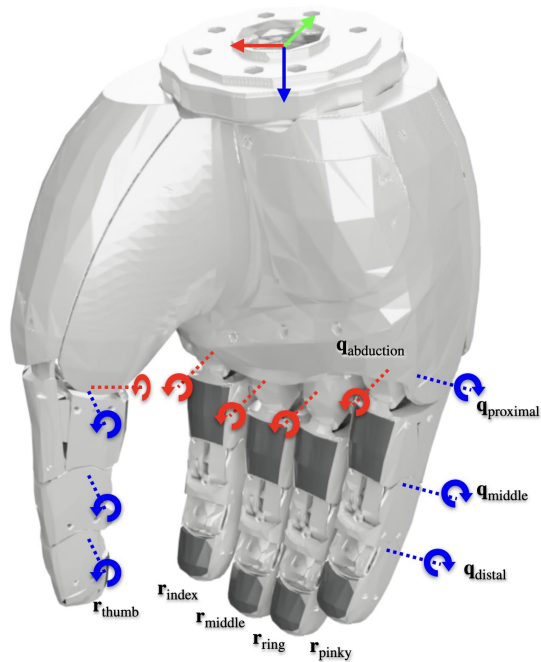


Difficult for grasping multiple-object scene using **multi-finger hand because of environment disturbance:**

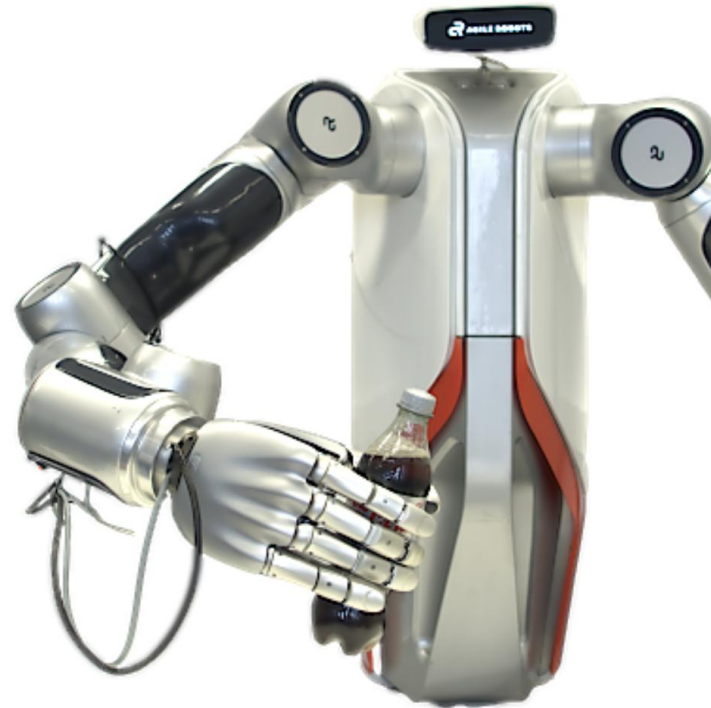
- **collision** between gripper with surrounding objects
- grasping actions lead to **unexpected collisions**

Type2: Simulation Environment

Five Finger Hand grasping



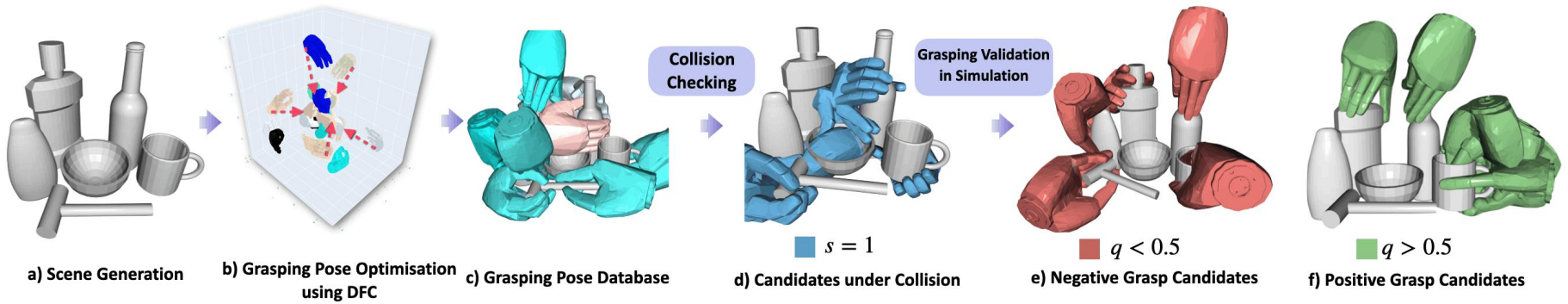
DLR-HIT II Hand



Dual-Arm Robot
(Diana 7, DLR-HIT II hand)

Grasping Generation Pipeline

Dataset Synthesis of FFH Grasping in Cluttered Environments



Data Generation: Object Dataset

over 1,700 objects from 3dNet, the Yale-CMU-Berkeley (YCB) Dataset, Princeton ModelNet, Dex-Net, and the MVTec Industrial 3D Object Detection Dataset (MVTec ITODD)



Loss: Energy Function for Grasping Optimization

$$\hat{E} = E_{\text{DFC}} + E_{\text{HO}} + E_{\text{Robot}}$$

$$E_{\text{DFC}} = \|Gc\|^2$$

$$G = \begin{bmatrix} I_3 & \cdots & I_3 \\ [\psi_1]_{\times} & \cdots & [\psi_n]_{\times} \end{bmatrix} \quad (3)$$

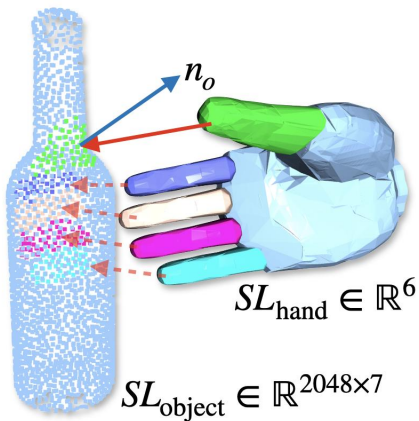
$$[\psi_k]_{\times} = \begin{bmatrix} 0 & -\psi_k^{(z)} & \psi_k^{(y)} \\ \psi_k^{(z)} & 0 & -\psi_k^{(x)} \\ -\psi_k^{(y)} & \psi_k^{(x)} & 0 \end{bmatrix}$$

where, $\Psi = \{\psi_1, \dots, \psi_n\}$ represents the set of contact point candidates, sampled from Φ_{contact} , term $c \in \mathbb{R}^{n \times 3}$ denotes the normals of object surface at the contact points in Ψ , and n indicates the number of contact points.



Loss: Energy Function for Grasping Optimization

$$\tilde{E} = E_{DFC} + E_{HO} + E_{Robot}$$



$$E_{HO} = \sum_{k=1}^n \epsilon(\psi_k, O) + \sum_{v \in \Phi} r \epsilon(v, O) \quad (4)$$

where, $\epsilon(x, O)$ represents the signed distance from each contact point ψ_k to object surface O , while $\epsilon(v, O)$ denotes the signed distance from a point v on the hand surface Φ to O . The parameter r is set to -1 when a point is inside O and 1 when outside. Penetration energy E_{pen} is calculated to prevent the hand from intruding into the object surface.

Loss: Energy Function for Grasping Optimization

$$\hat{E} = E_{\text{DFC}} + E_{\text{HO}} + E_{\text{Robot}}$$

$$E_{\text{Robot}} = E_{\text{spen}} + E_{\text{joints}}$$

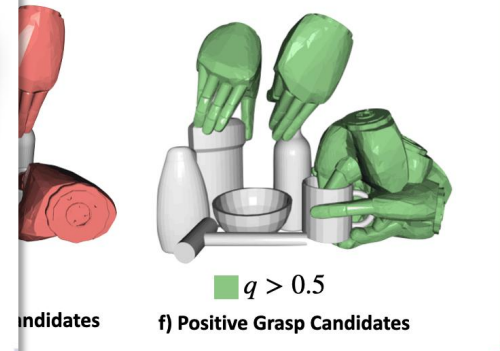
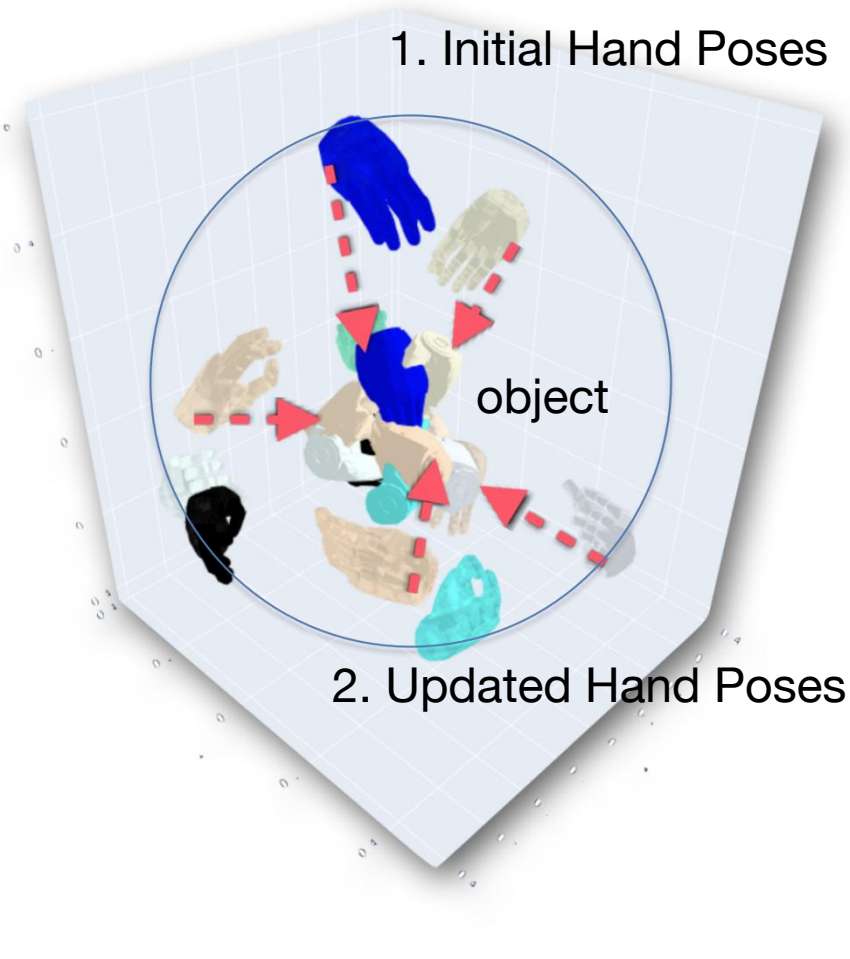
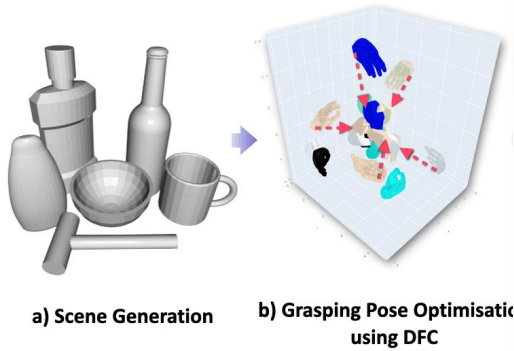
$$E_{\text{spen}} = \sum_{u \in \Phi} \sum_{v \in \Phi} [u \neq v] \max(\delta - \epsilon(u, v), 0)$$

$$E_{\text{joints}} = \sum_{i=1}^d \left(\max(\theta_i - \theta_i^{\max}, 0) + \max(\theta_i^{\min} - \theta_i, 0) \right)$$

(5)

Optimization Process

Dataset Synthesis of FFH Grasping in Clutter



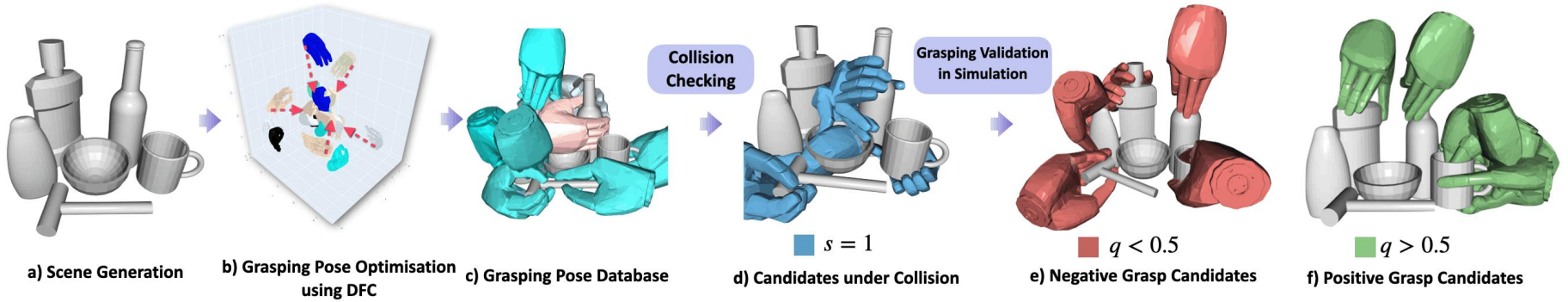
Validation Performance with Isaac Gym



Result based on DLR-HIT hand

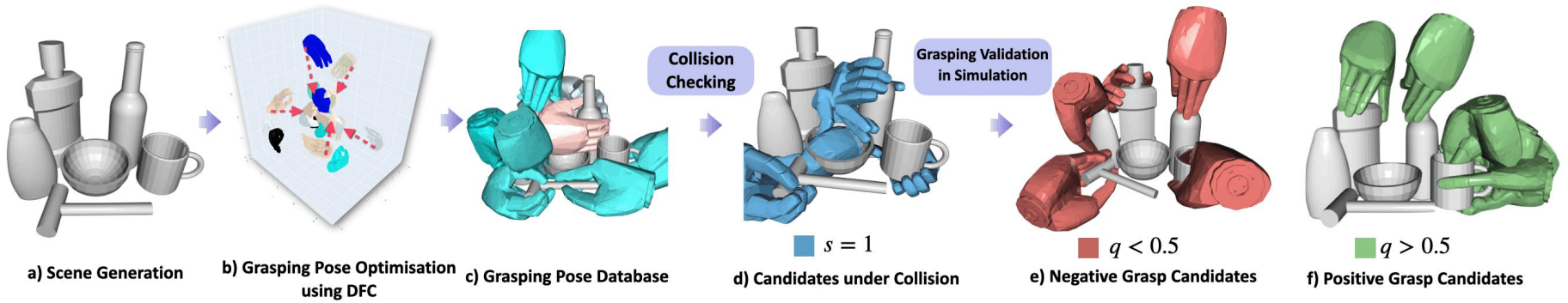
Grasping Generation Pipeline

Dataset Synthesis of FFH Grasping in Cluttered Environments

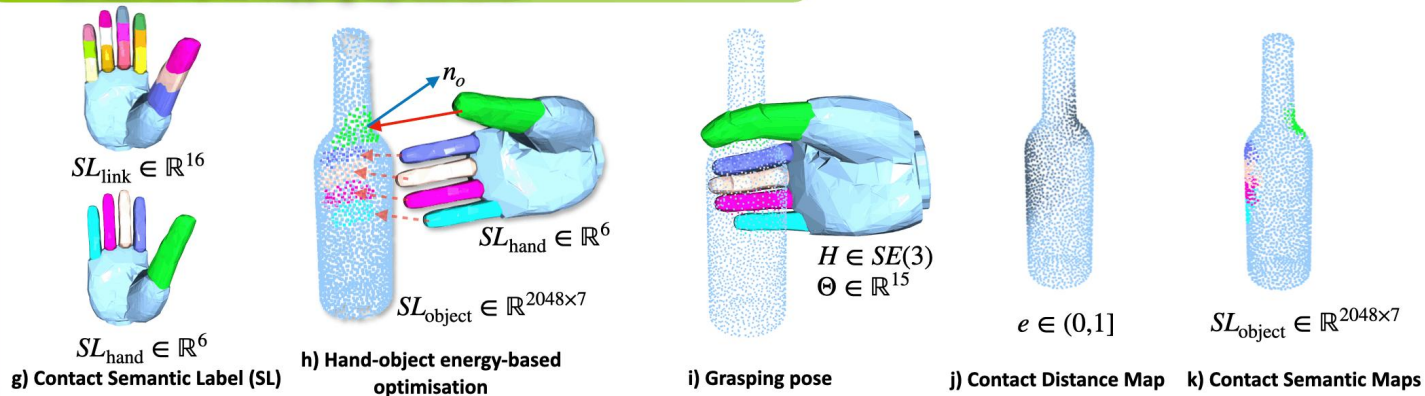


Grasping Generation Pipeline

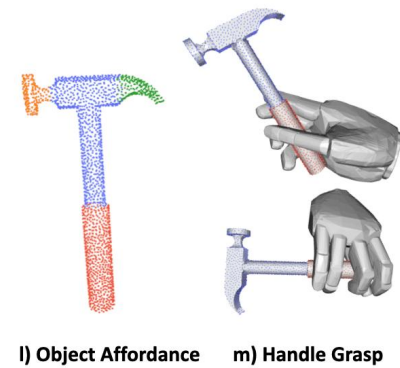
Dataset Synthesis of FFH Grasping in Cluttered Environments



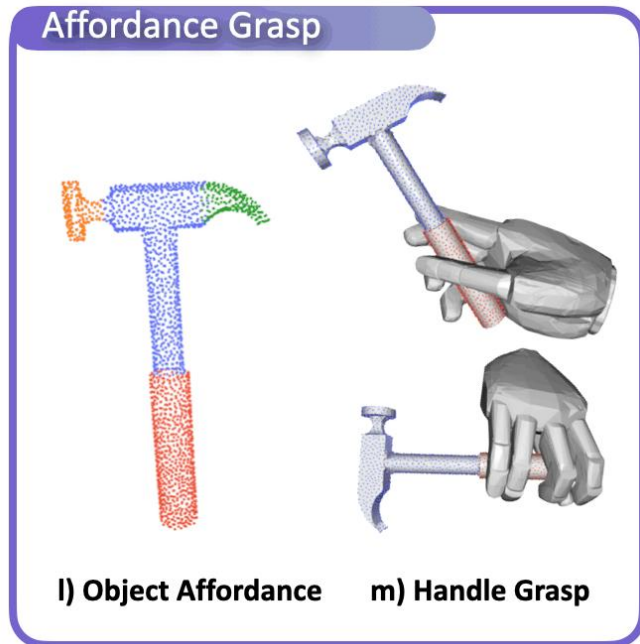
Contact Semantic Mapping Representation



Affordance Grasp



Grasping Generation Pipeline



1. object affordance annotation with Blender Plugin
2. Filter out affordance manipulation pose based on object affordance information.

```
affordance_list = ["HandleGrasp",  
"WrapGrasp", "Press", "Pour", "Cut", "Stab",  
"Pull", "Push", "Open", "Twist", "Hammer",  
"Pry"]
```

Generate Affordance Grasping

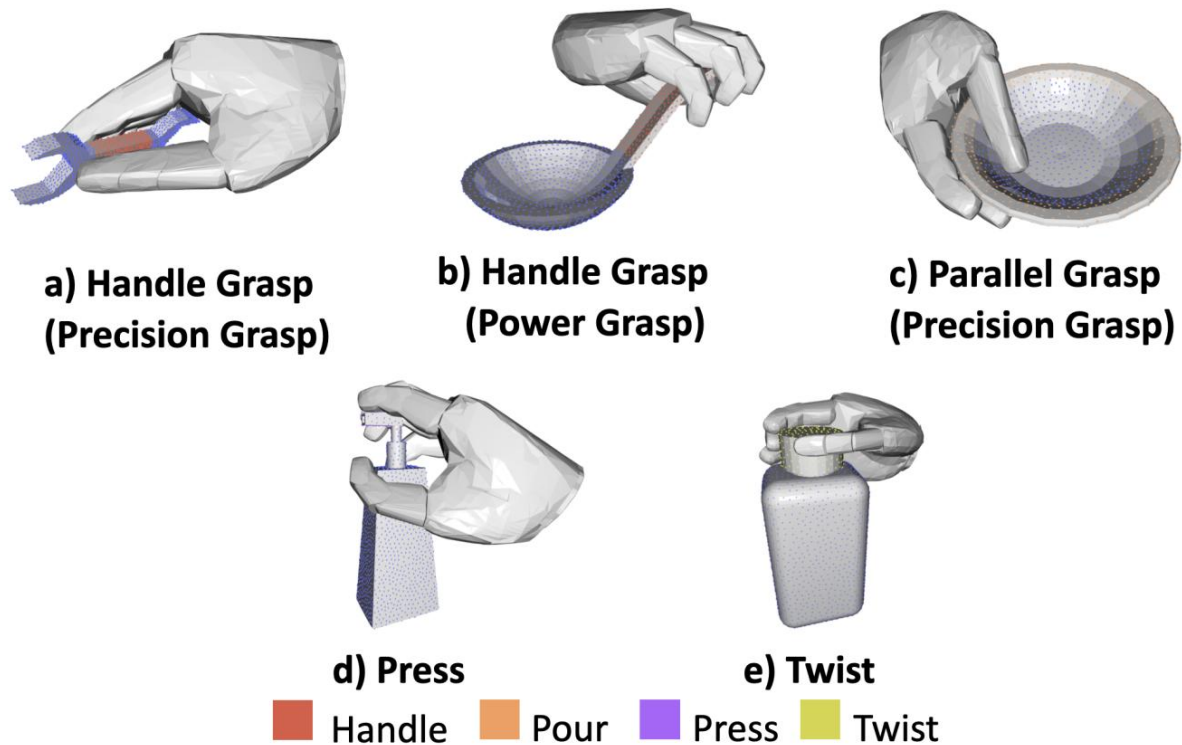


Fig. 6. Examples of grasping type considering object affordance maps and different manipulation poses.

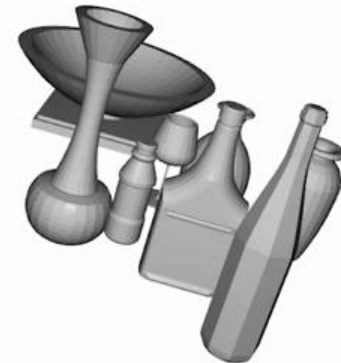
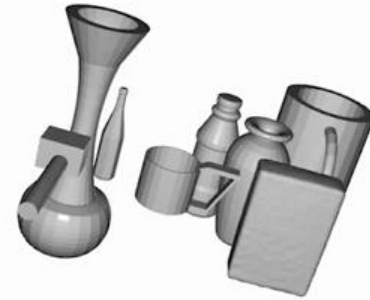
Grasping Dataset

Multi-Fingered Robotic Hand Cluttered Grasping Dataset

Grasping Dataset

Multi-Fingered Robotic Hand Cluttered Grasping Dataset

Grasping Dataset



Comparison Experiments of GraspIt!, DexGraspNet and proposed FFHClutteredGrasping

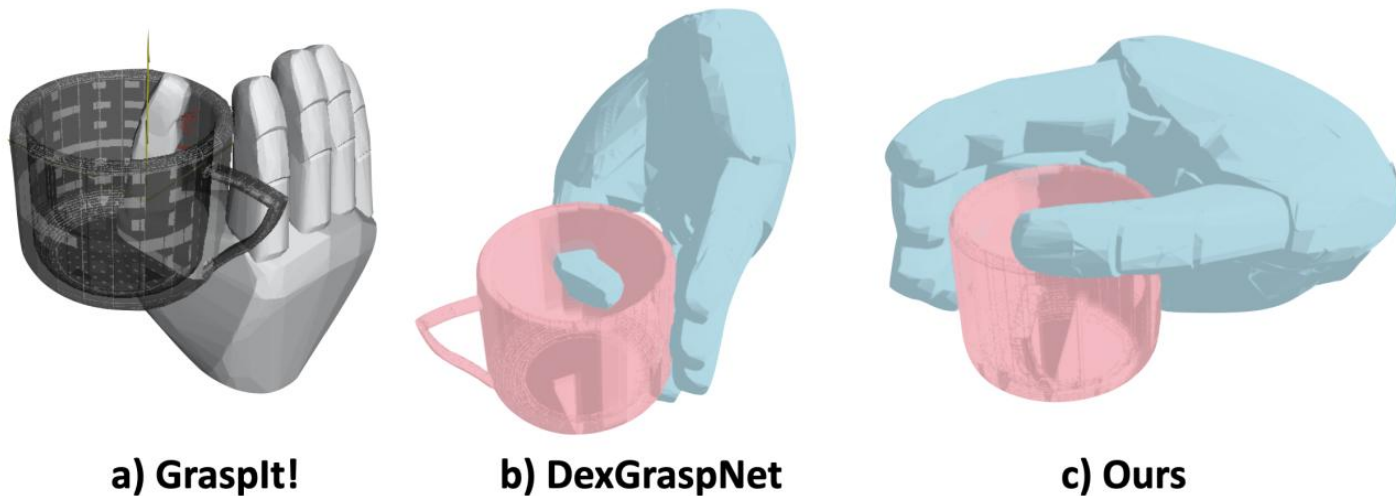


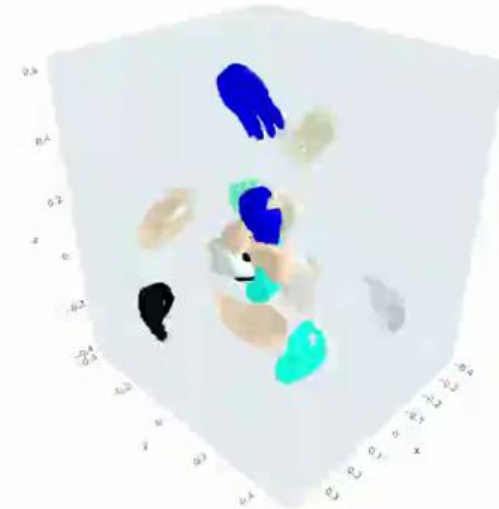
Fig. 7. Results of grasp generation using GraspIt! [1], DexGraspNet [4] and proposed method.

Grasplt! and proposed method



Grasplt!

Inefficient optimization process
local sub-optimal grasp generation

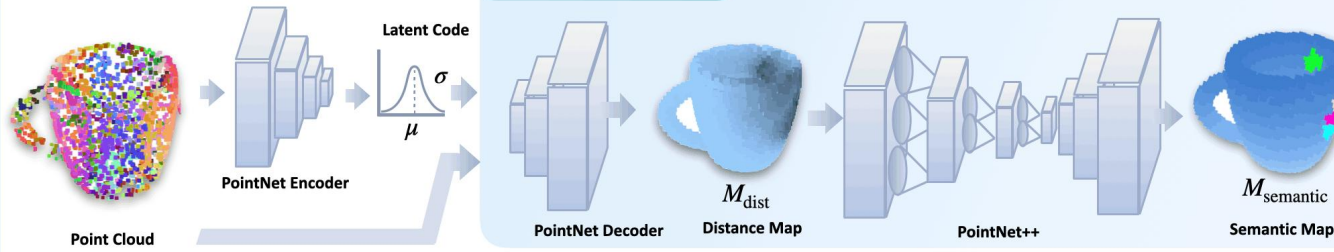


Our Methods

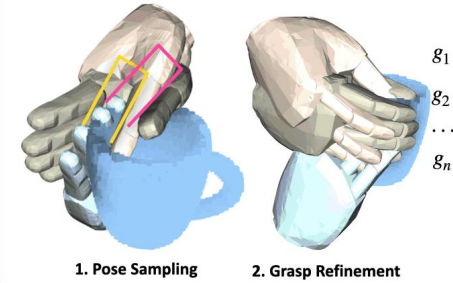
Parallel optimization of grasping poses

Contact Semantic CVAE, Grasp Detection, and Evaluation

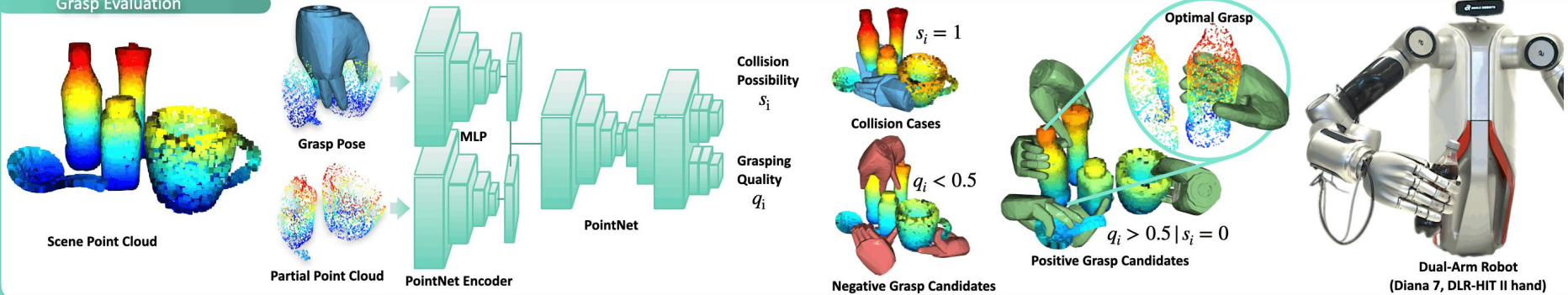
Contact Semantic CVAE



Grasp Detection



Grasp Evaluation



... Danke

Lei Zhang

lei.zhang-1@studium.uni-hamburg.de

Zhanglei.cn.de@gmail.com



Universität Hamburg
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Technische Aspekte Multimodaler Systeme

