Hand-Object Contact Consistency Reasoning for Human Grasps Generation

Hanwen Jiang* Shaowei Liu* Jiashun Wang Xiaolong Wang
UC San Diego

Accepted at ICCV'2021 (oral presentation)

2021 ICCV OCTOBER 11-17 VIRTUAL

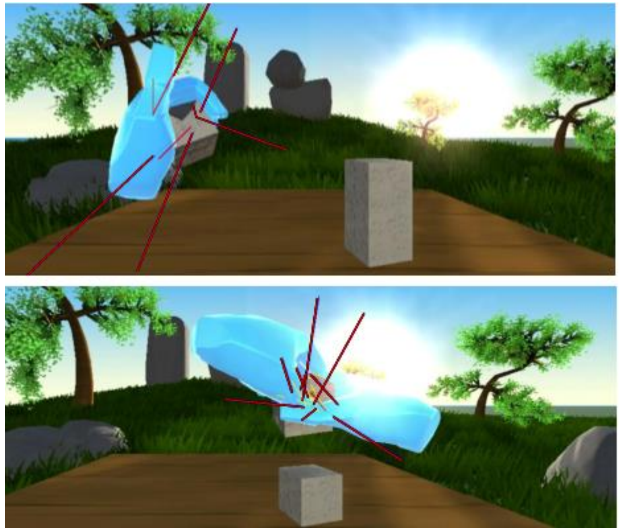
Motivation

- Understanding how to grasp 3D objects

Motivation

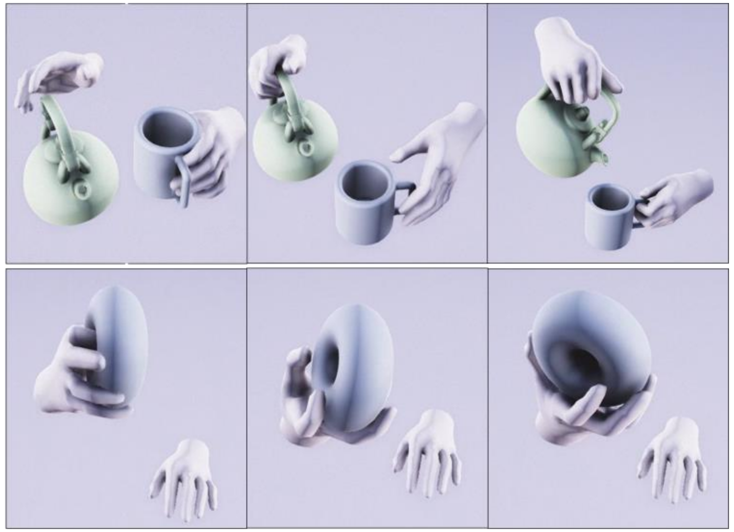
- Possible applications

Realistic VR Experiences



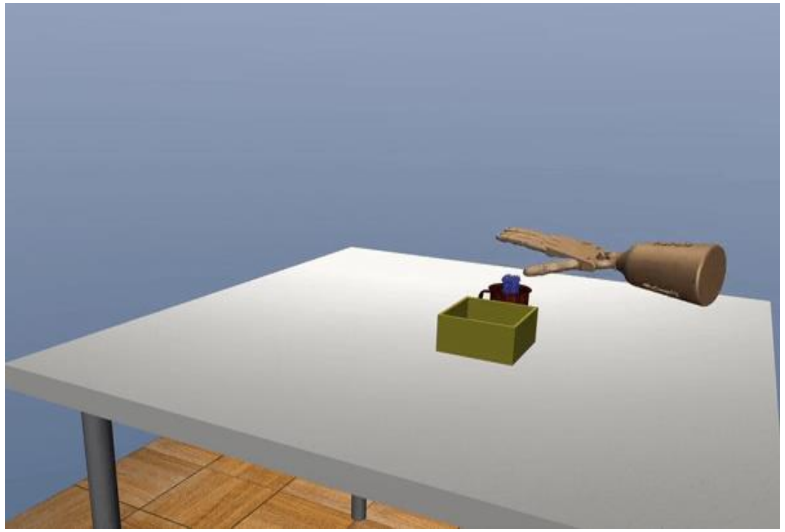
Holl et al, VR '18

Realistic Animation



Zhang et al, SIGGRAPH '21

Robot Imitation Learning



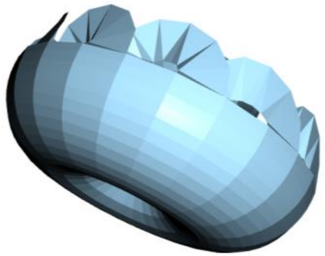
Qin et al, Arxiv '21

Task Definition

- Human grasp generation

Input

Object point clouds
(Sampled on mesh)



Output

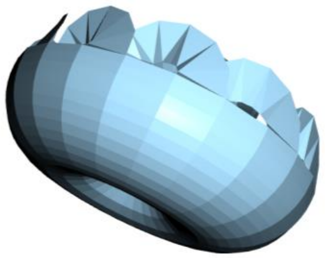
Human grasp
(Hand Mesh)

Task Definition

- Human grasp generation

Input

Object point clouds
(Sampled on mesh)



Output

Human grasp
(Hand Mesh)

Input + Output

* One object can have multiple different 6-Dof poses in the world coordinate.

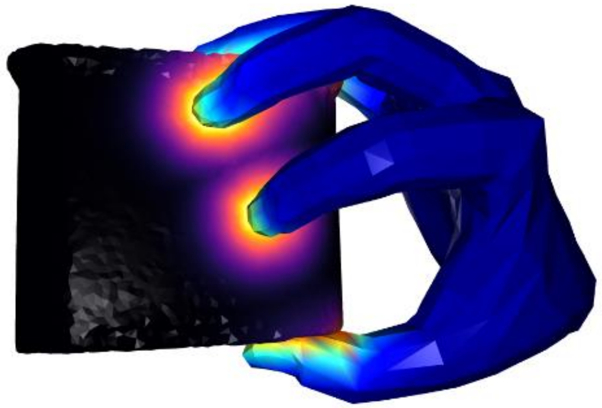
Targets, problems and solution

- Targets:
  - Physical plausibility
  - Natural hand poses
- Problem:
  - Human hand has higher degree-of-freedom than grippers
- Solution:
  - Reason contact consistency between hand-object

Contact consistency

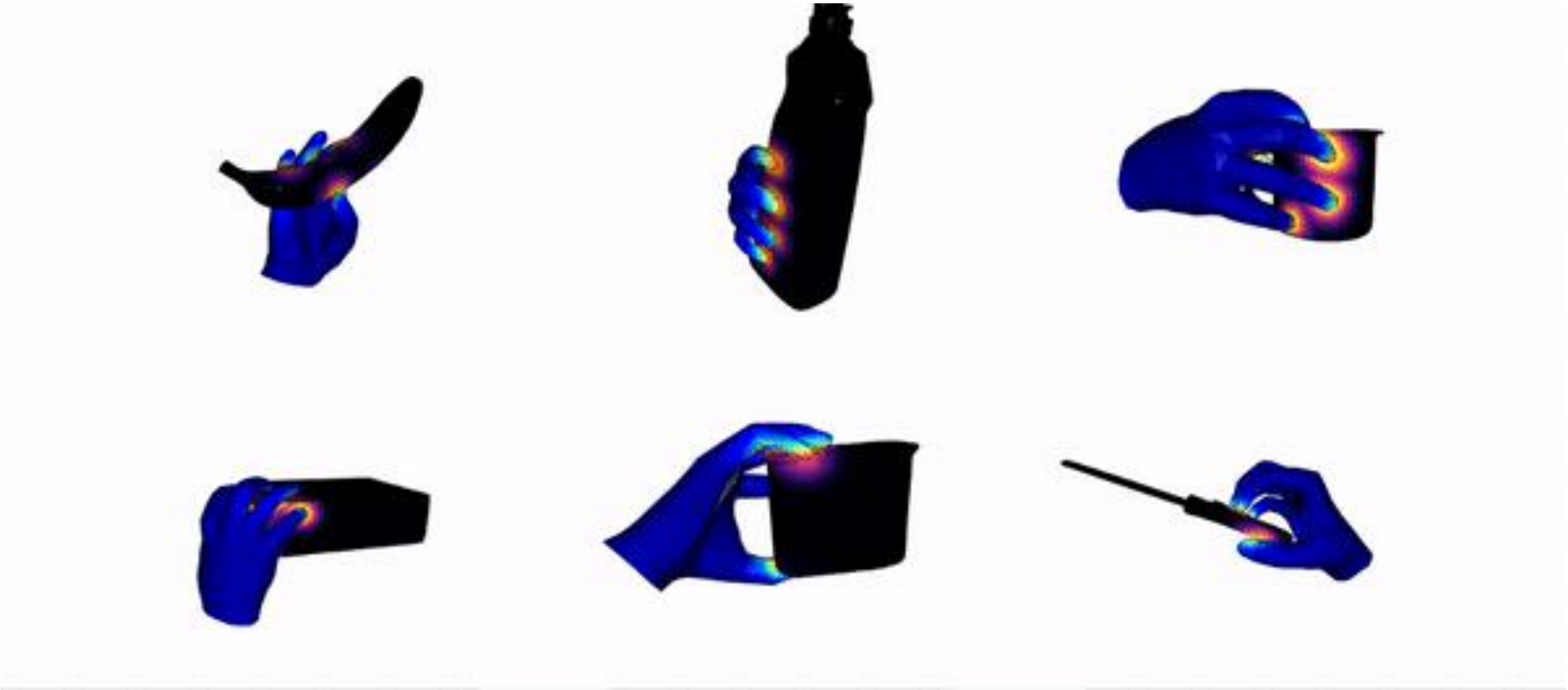
- Hand-centric:
  - Hand should contact object surface

- Object-centric:
  - Object possible contact regions should be touched



Contact consistency

- For out-of-domain objects
  - Contact consistency reasoning can improve generalization ability



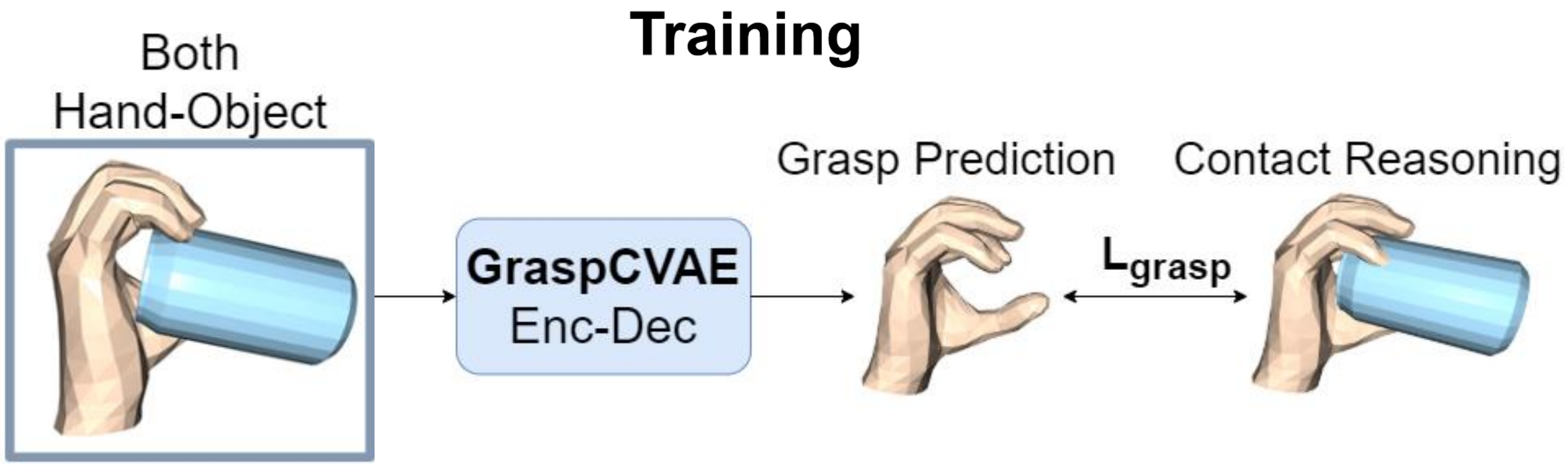
We reason contact consistency between hand-object from two aspects

- Framework design
- Loss function design

Framework

- We propose two networks

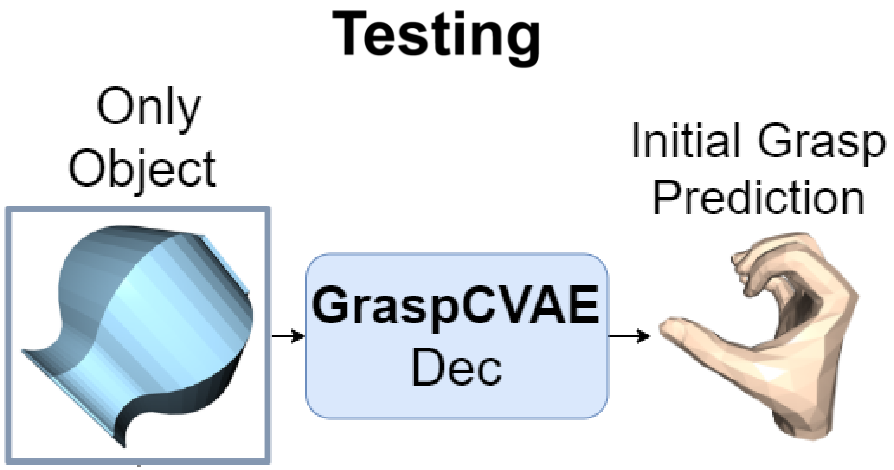
- GraspCVAE
  - Target: Generating grasps
  - In training, it learns to reconstruct hands with both hand-objects as inputs



Framework

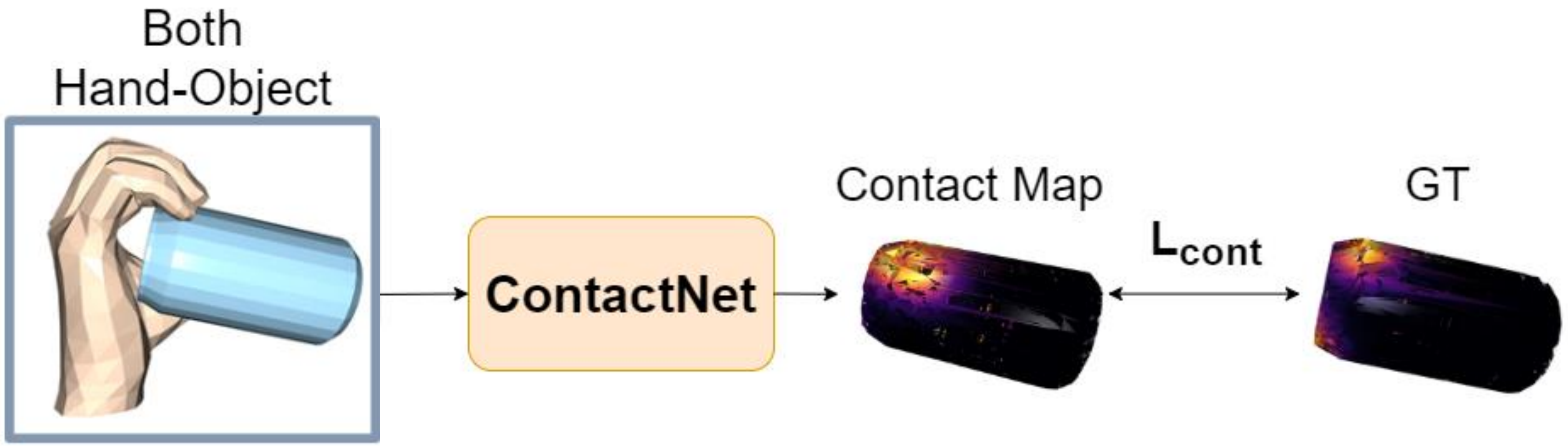
- We propose two networks

- GraspCVAE
  - Target: Generating grasps
  - In testing, it generates grasps given only the object as input



Framework

- We propose two networks
- ContactNet
  - Target: Predicting object contact map
  - With both hand-object as inputs



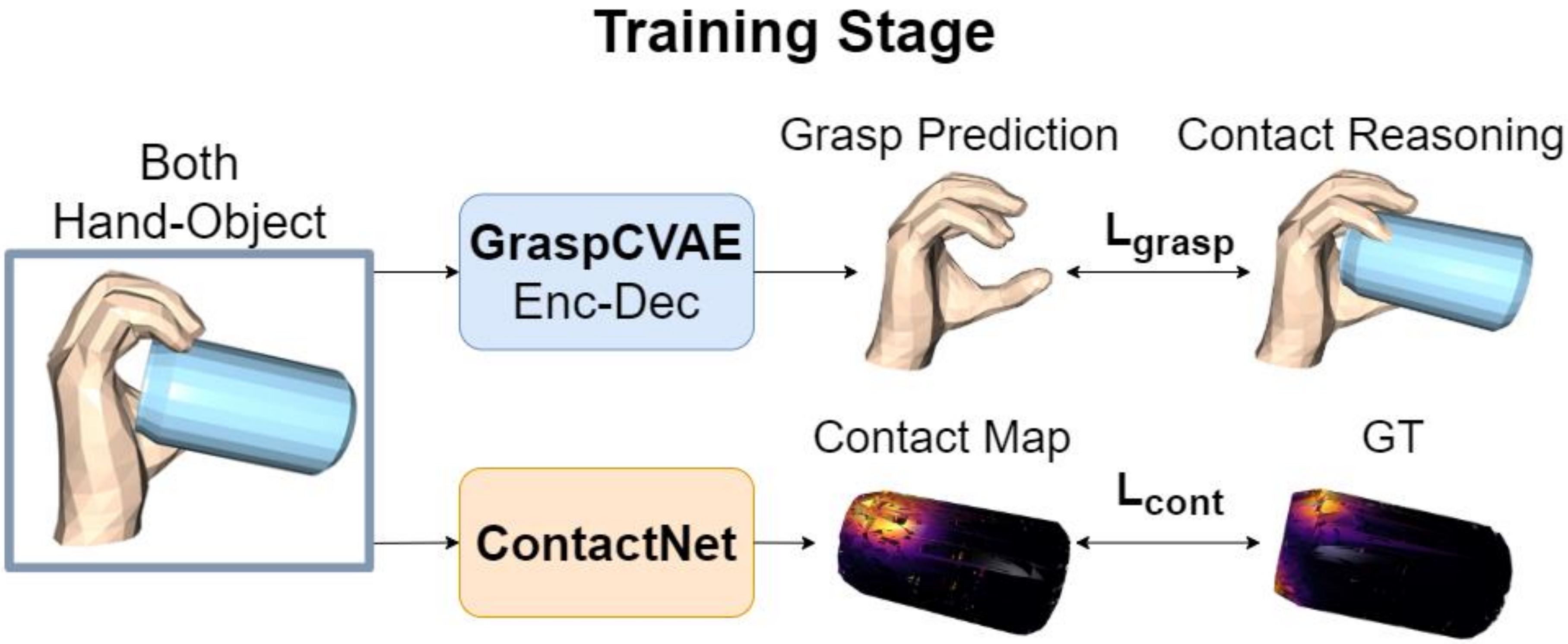
* Contact Map: Brighter regions should be more close to hand.

Framework: (1) Training

- Train the two networks separately on ground-truth data

Framework: (1) Training

- Train the two networks separately on ground-truth data



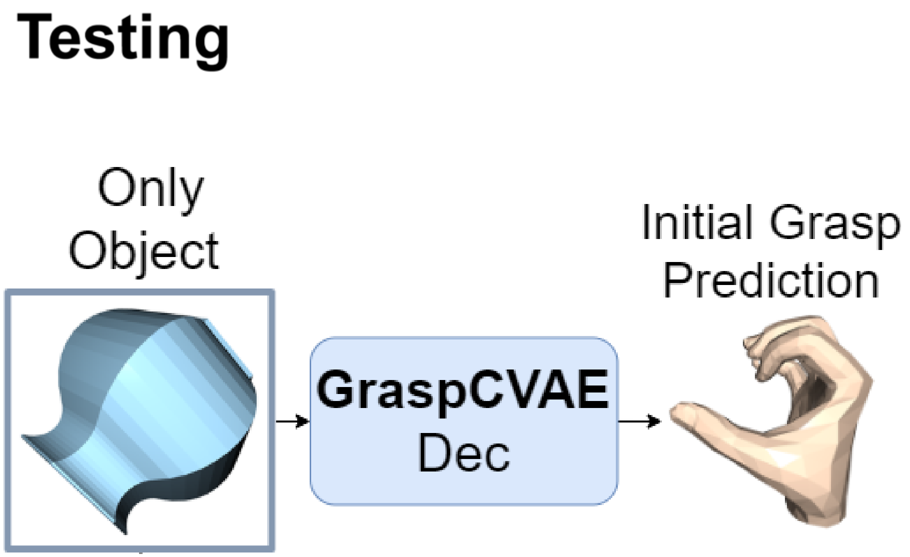
* Contact Map: Brighter regions should be more close to hand.

Framework: (2) Testing with Test-Time Adaptation (TTA)

- Unify the two networks in a cascade manner
- ContactNet: Provides self-supervision signals

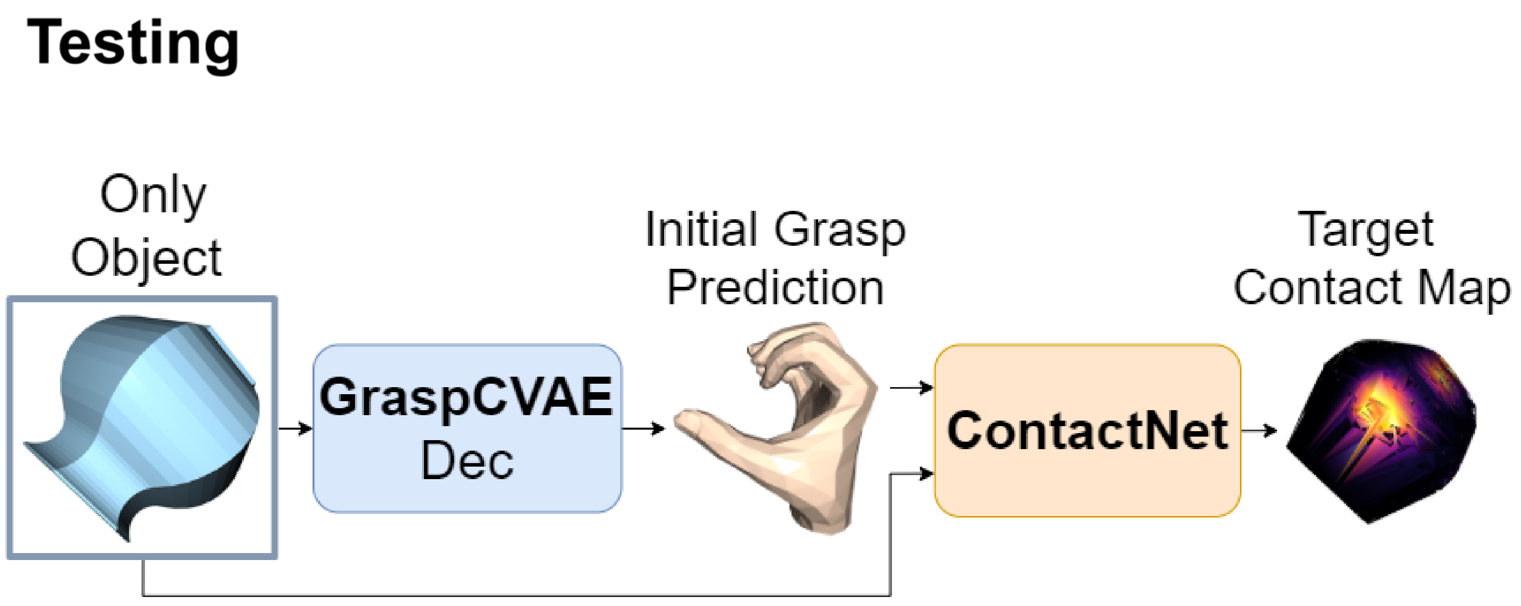
Framework: (2) Test-Time Adaptation (TTA)

- Step 1: Generate an initial grasp by GraspCVAE



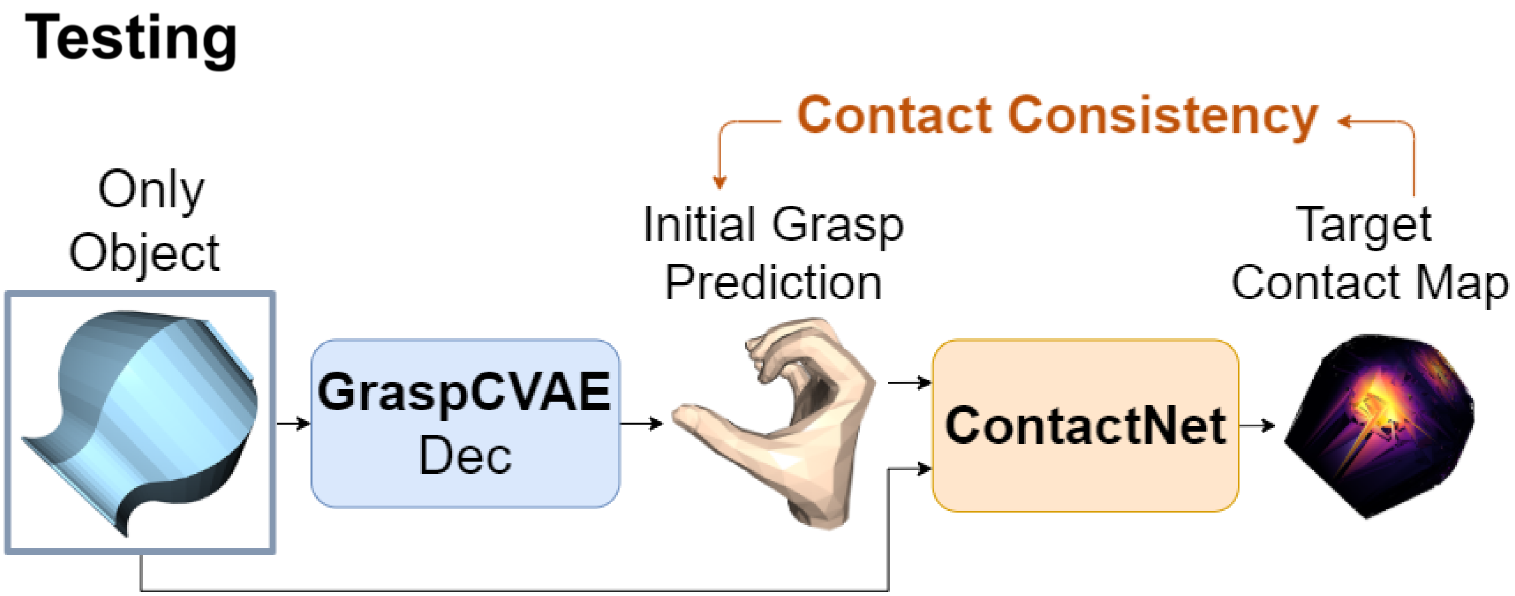
Framework: (2) Test-Time Adaptation (TTA)

- Step 2: Generate a target contact map by ContactNet



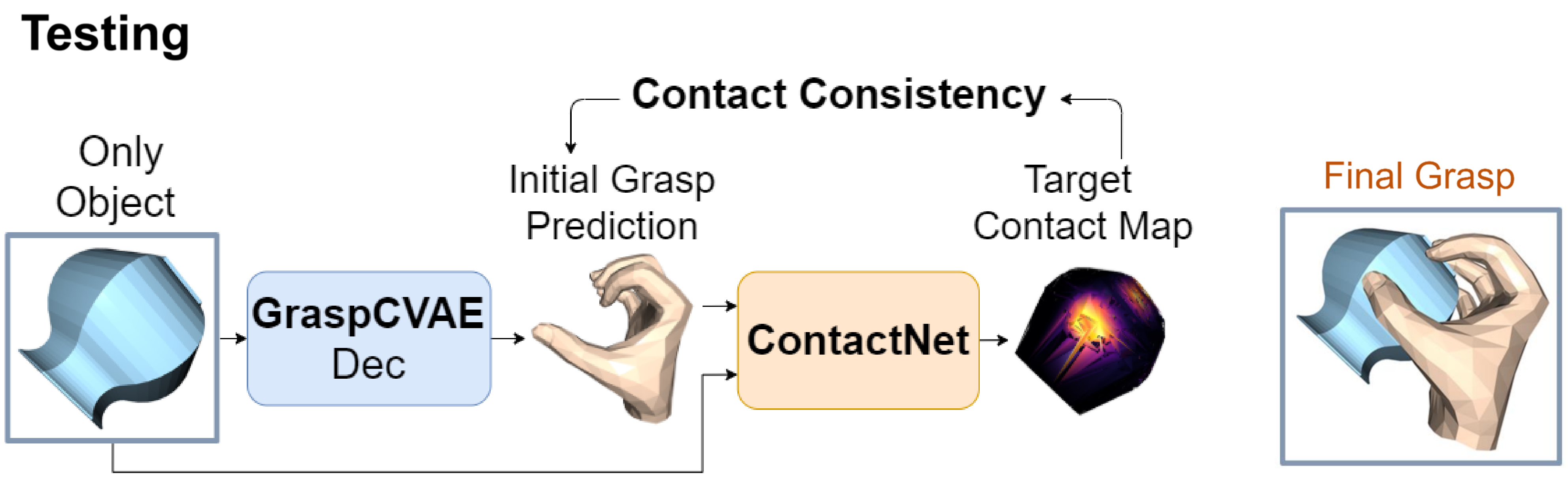
Framework: (2) Test-Time Adaptation (TTA)

- Step 3: Leverage the consistency between two outputs as a self-supervised loss



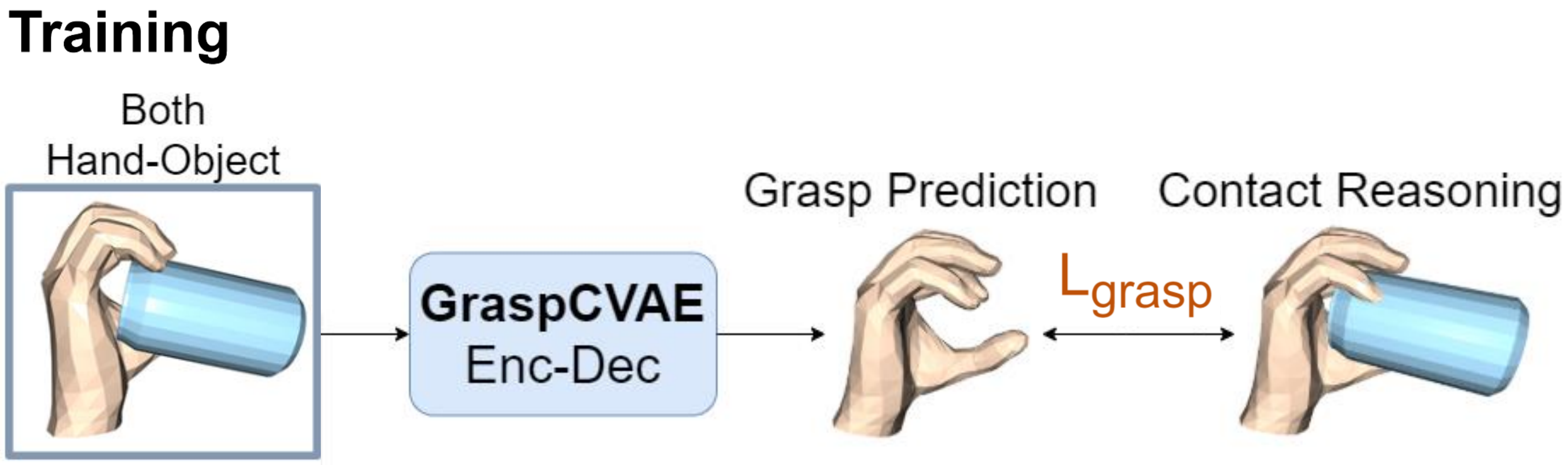
Framework: (2) Test-Time Adaptation (TTA)

- Step 4: Learn from the consistency loss for 10 iterations



Loss functions

- An object-centric loss: ensures object common contact regions to be touched by hand
- A hand-centric loss: encourages hand touching object surface



Datasets

- Train with ObMan [1] training set

- Test on
  - ObMan [1] testset (in-domain)
  - HO-3D [2] and FPHA [3] datasets (out-of-domain)

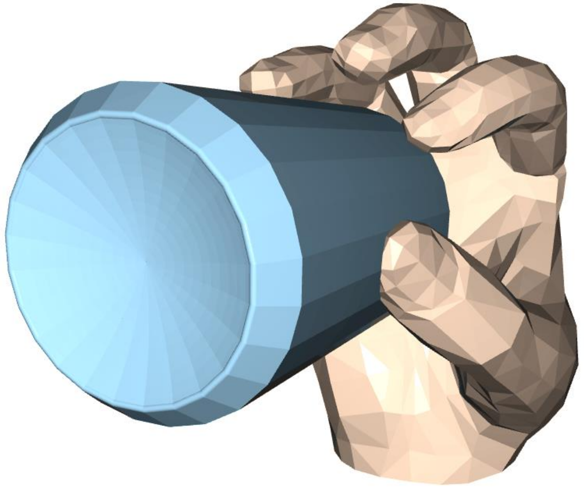
[1] Hasson, Yana et al. “Learning Joint Reconstruction of Hands and Manipulated Objects.” CVPR (2019).
[2] Hampali, Shreyas et al. “HOnnotate: A Method for 3D Annotation of Hand and Object Poses.” CVPR (2020).
[3] Garcia-Hernando, Guillermo et al. “First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations.” CVPR (2018).

First, we show examples of generated grasps for both in-domain and out-of-domain objects.

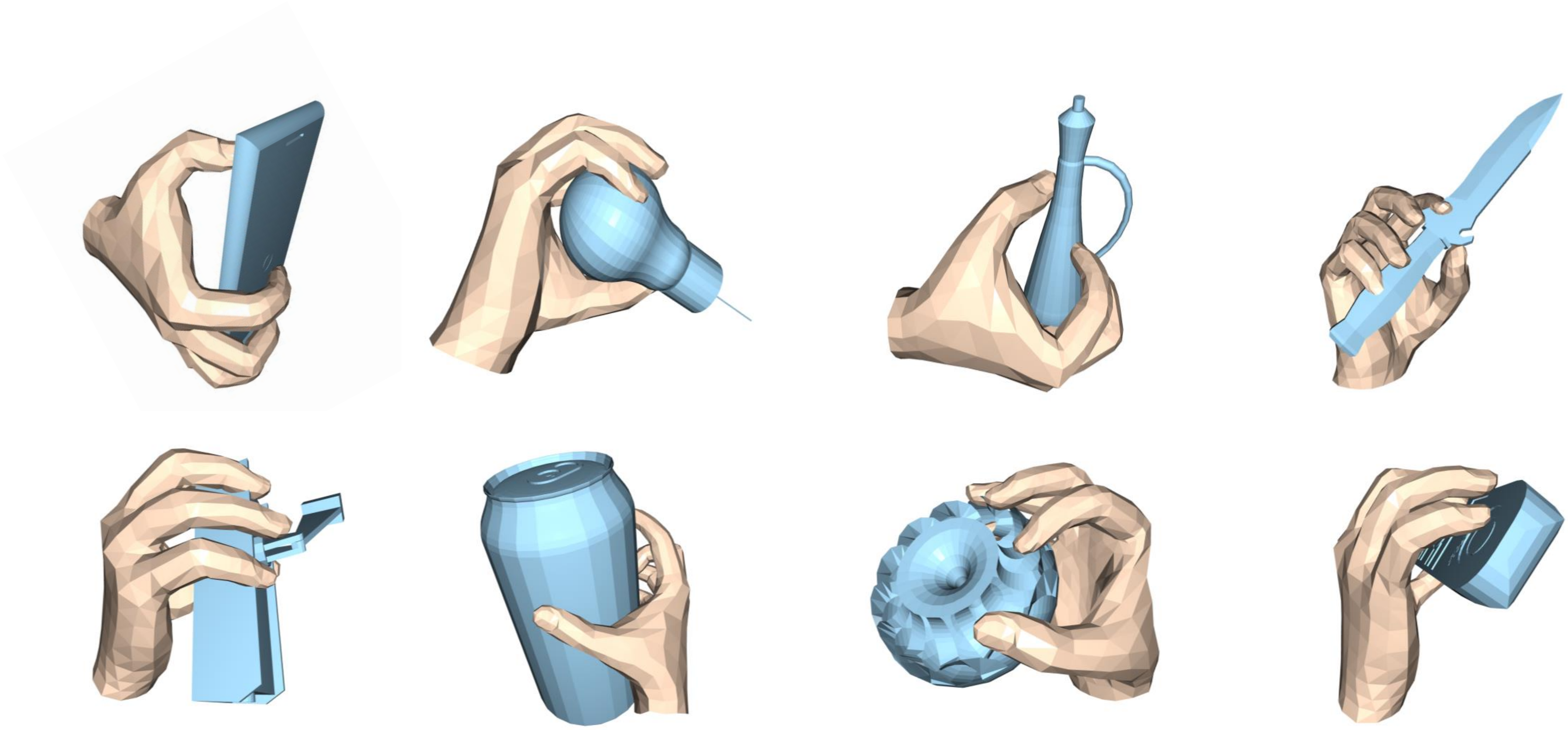
Generated grasps given in-domain objects

Example 1

Example 2

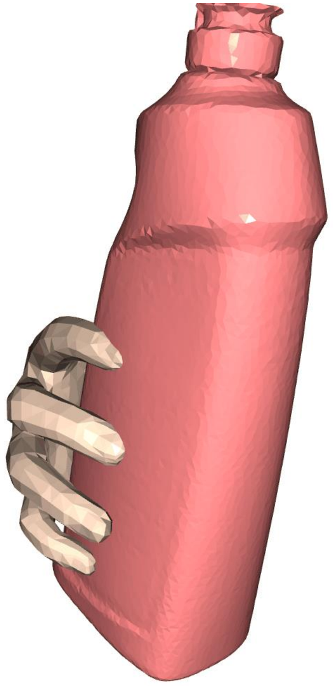


Generated grasps given in-domain objects

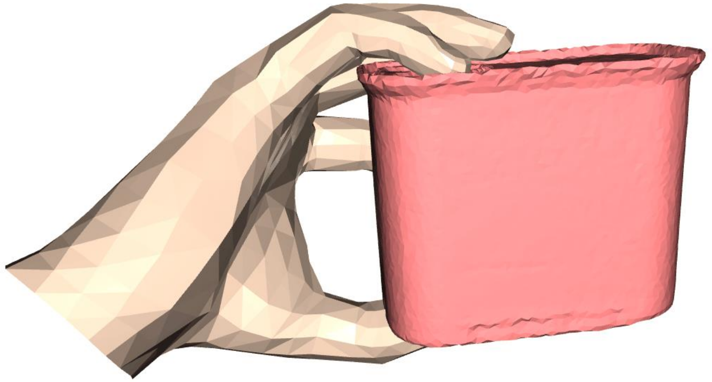


Generated grasps given out-of-domain objects

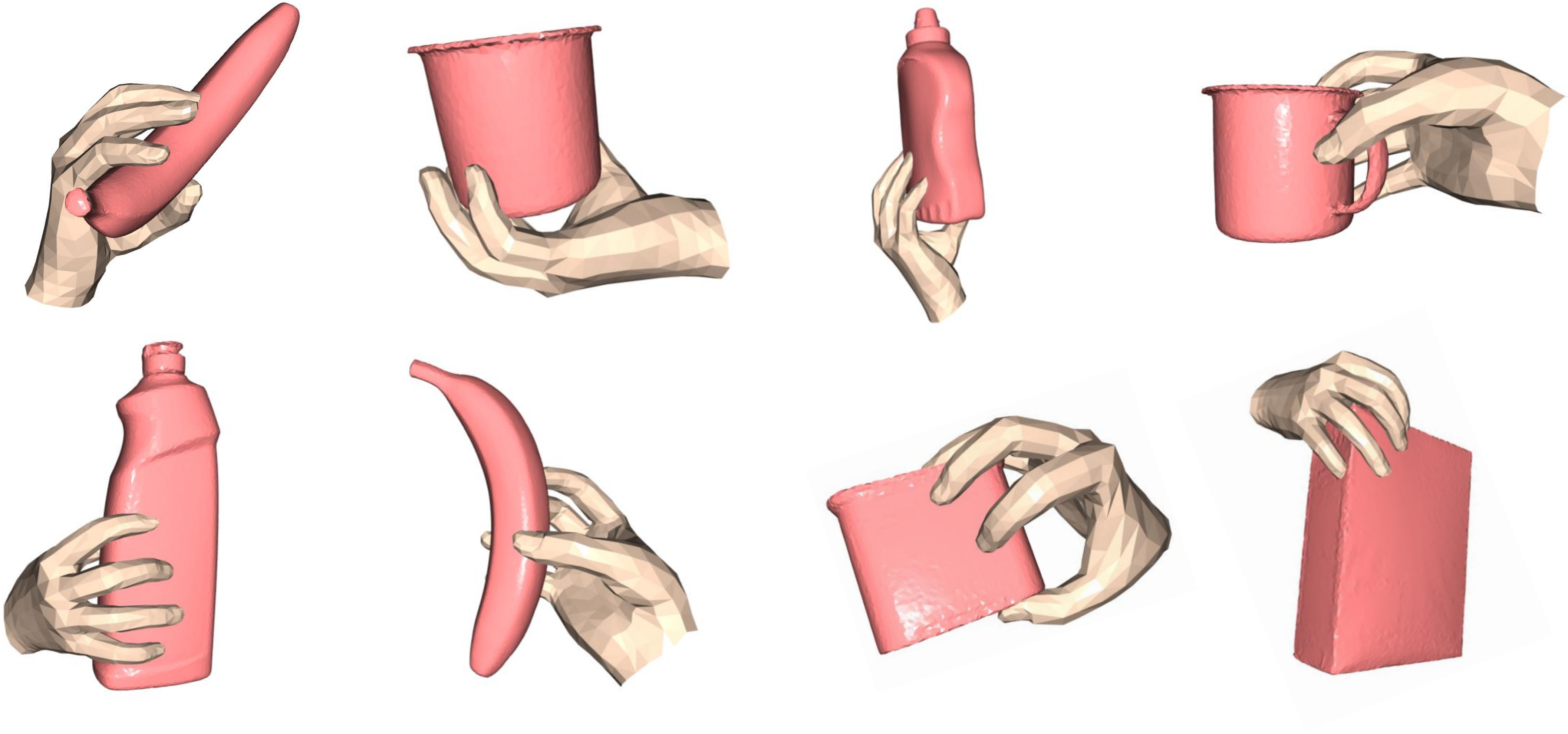
Example 1



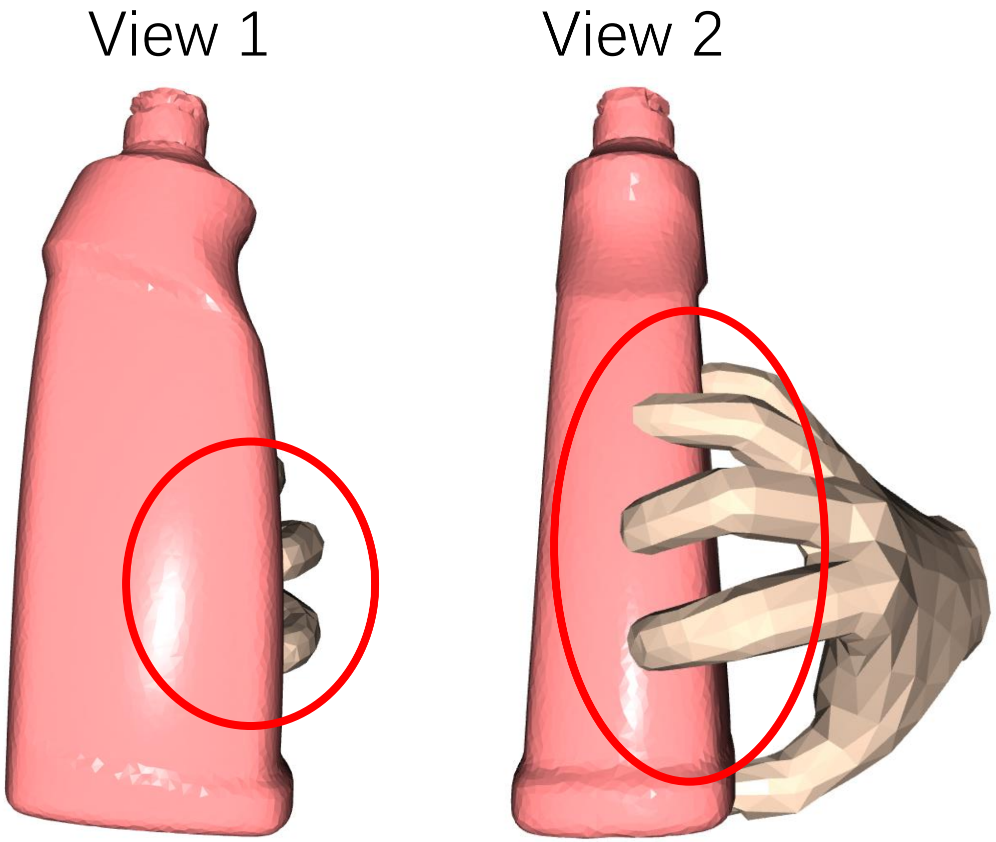
Example 2



Generated grasps given out-of-domain objects

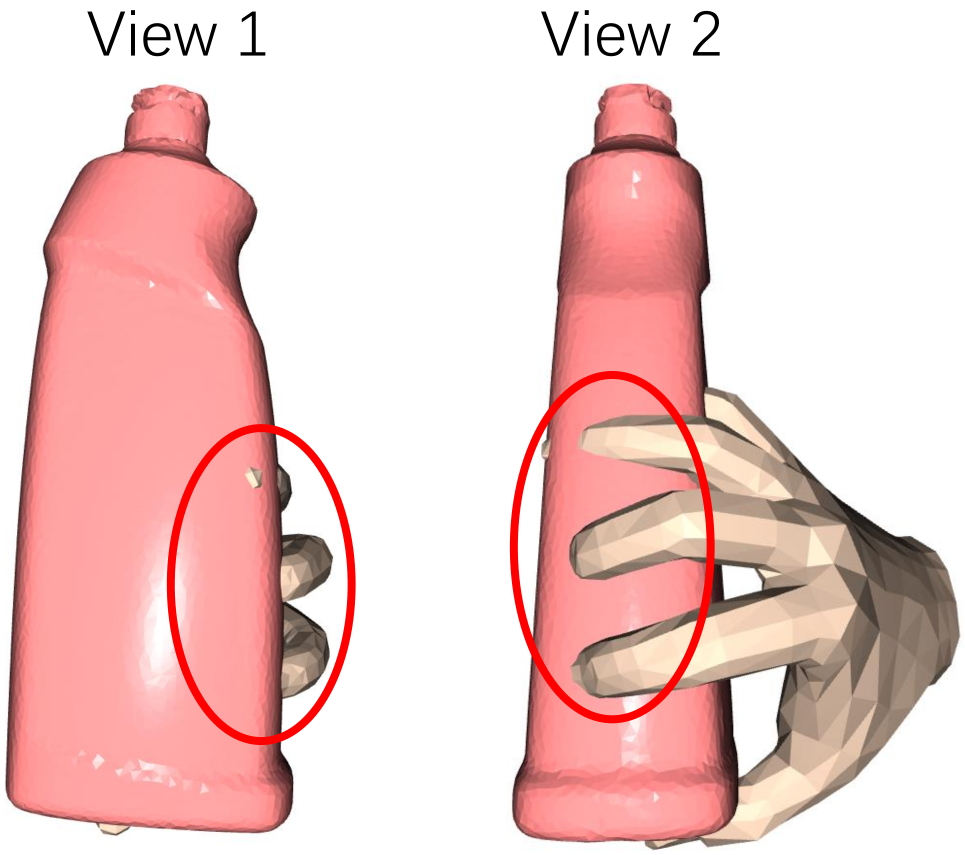


TTA visualization



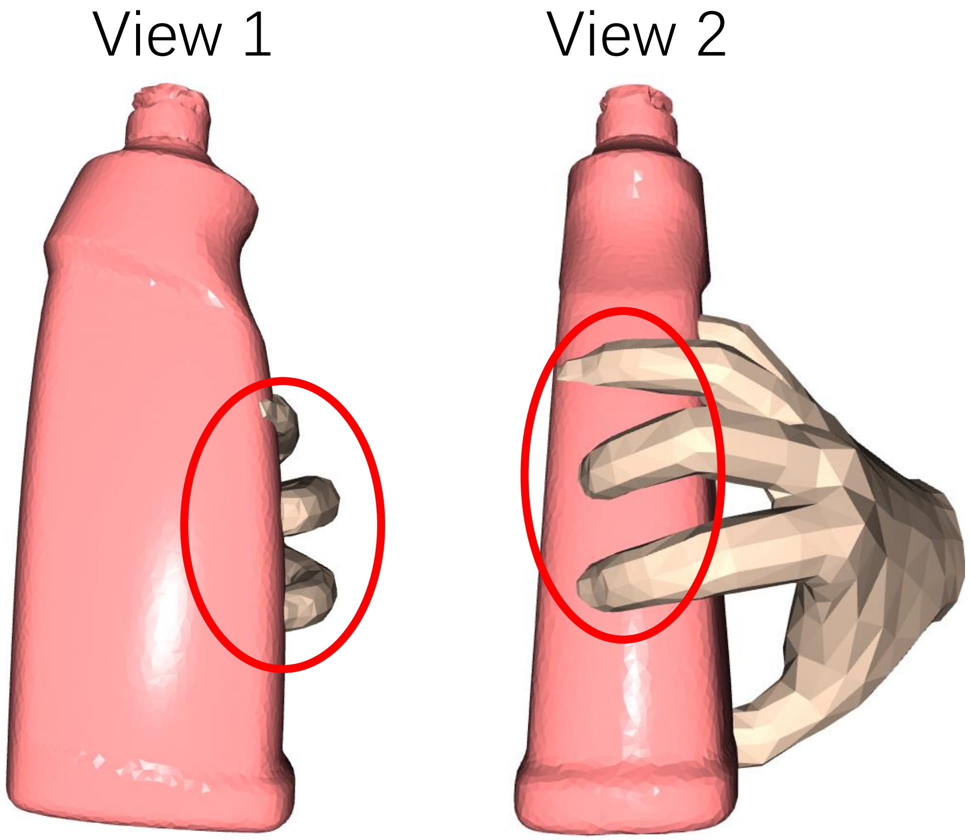
Before: hand penetrates into the object.

TTA visualization



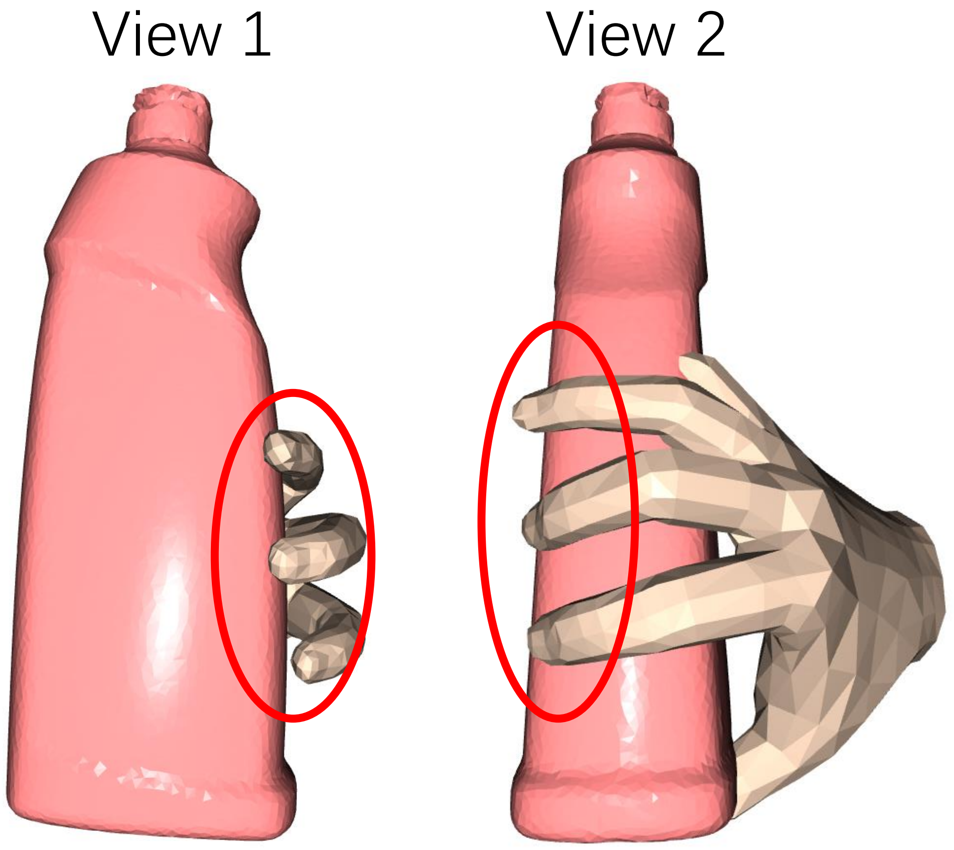
Iteration 2: penetration decreases.

TTA visualization



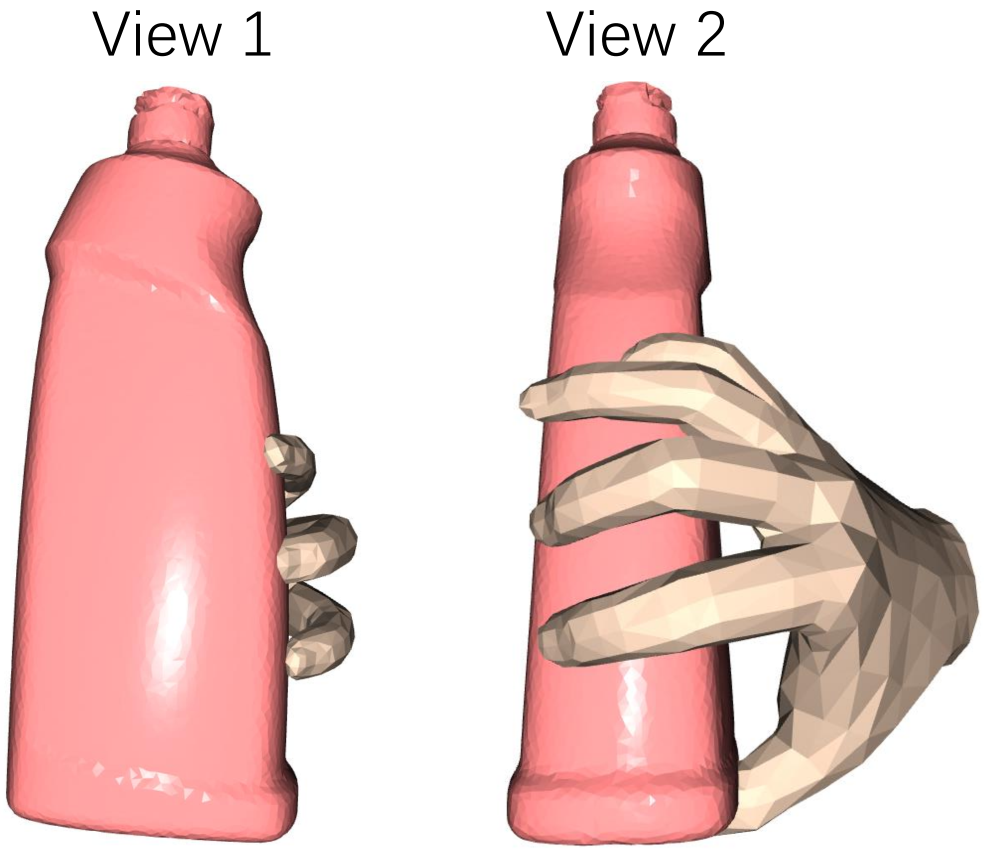
Iteration 5: penetration decreases.

TTA visualization



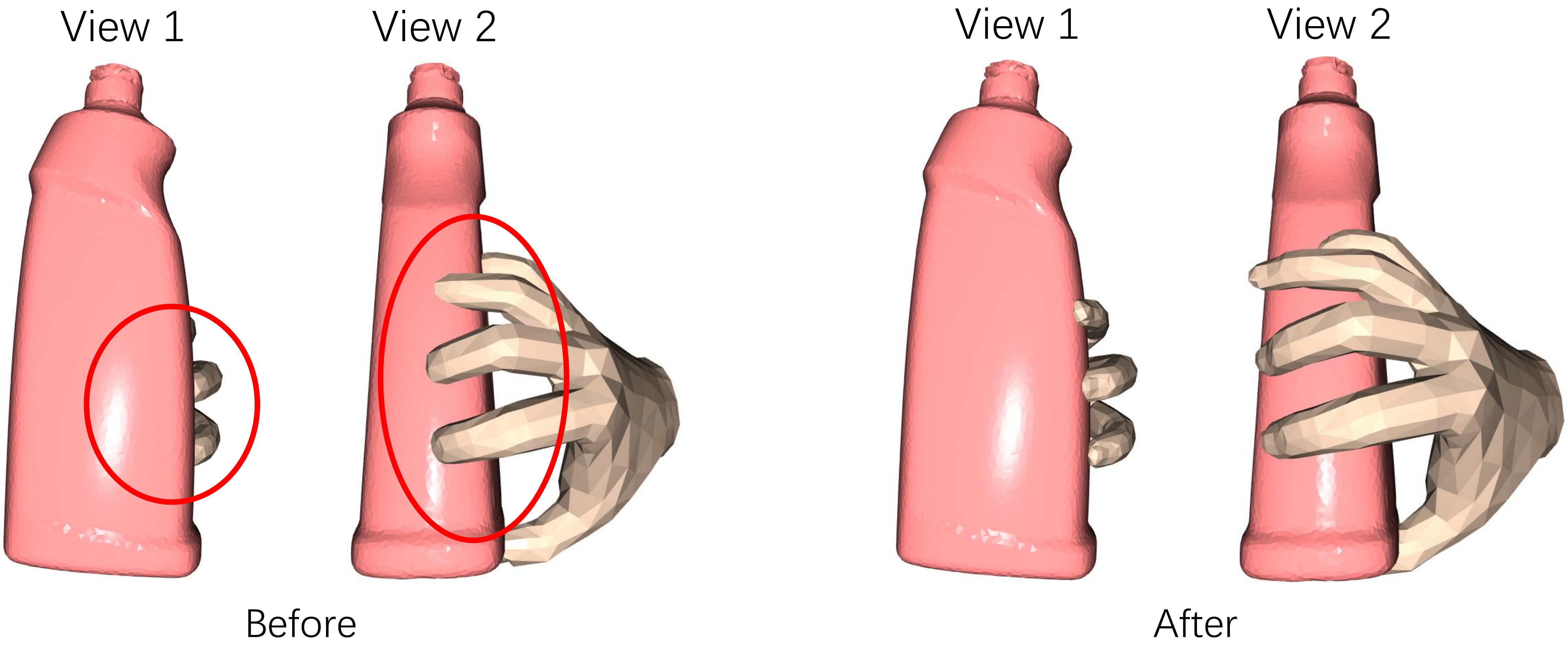
Iteration 8: hand is repelled out of the object.

TTA visualization

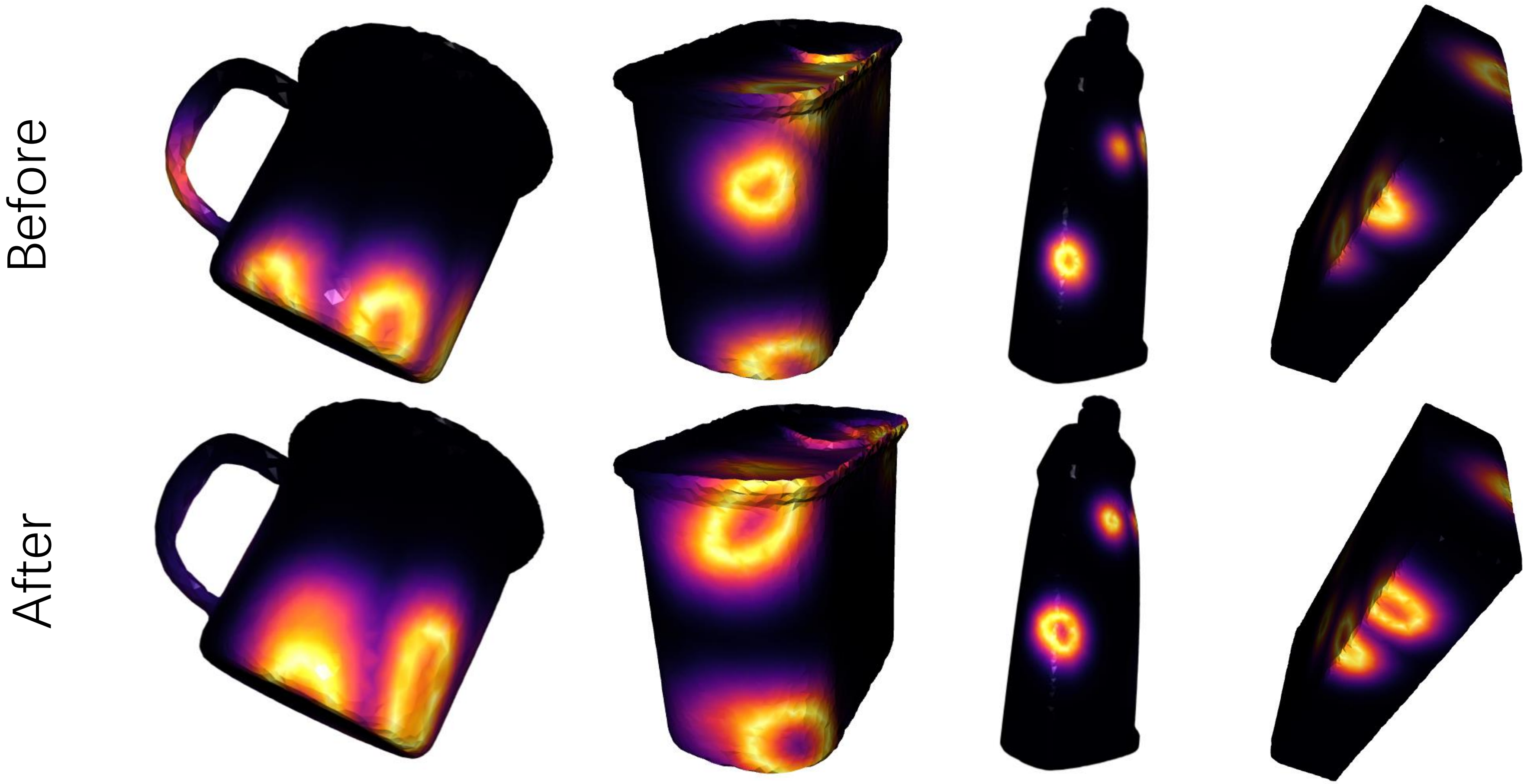


Iteration 10: hand become closer to object surface.

TTA visualization



TTA visualization



Object contact regions become larger (grasps more stable).

* Contact Map: Brighter regions should be more close to hand.

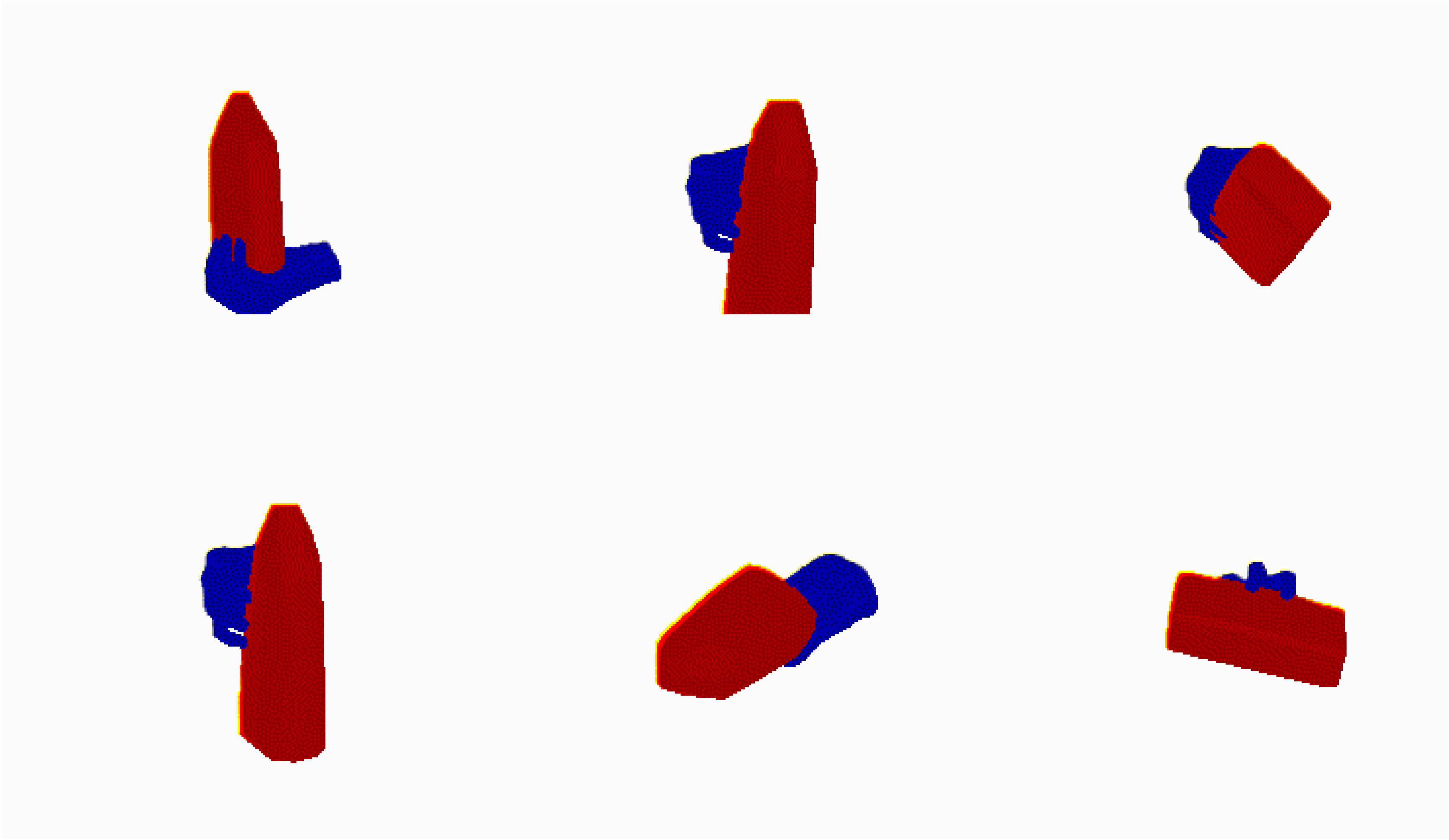
Generated diverse grasps

Quantitative evaluation metrics

- Penetration (↓)
- Grasp displacement in the simulation (↓)
- Perceptual score (↑)
- Contact measurements (↑)
  - Contact ratio, number of fingers in contact, etc.

Grasp simulation

(measures grasp stability)



Quantitative results

Overall performance (in-domain objects)

		Obman		
		GT	GF [25]	Ours
Penetration	Depth (cm) ↓	**0.01**	0.56	0.46
	Volume (cm^3) ↓	**1.70**	6.05	5.12
Grasp Displace.	Mean (cm) ↓	1.66	2.07	**1.52**
	Variance (cm) ↓	± 2.44	± 2.81	± **2.29**
Perceptual Score	{1, ..., 5} ↑	3.24	3.02	**3.54**
Contact	Ratio (%) ↑	**100**	89.40	99.97

Karunratanakul, Korrawe et al. “Grasping Field: Learning Implicit Representations for Human Grasps.” *3DV* (2020).

Quantitative results

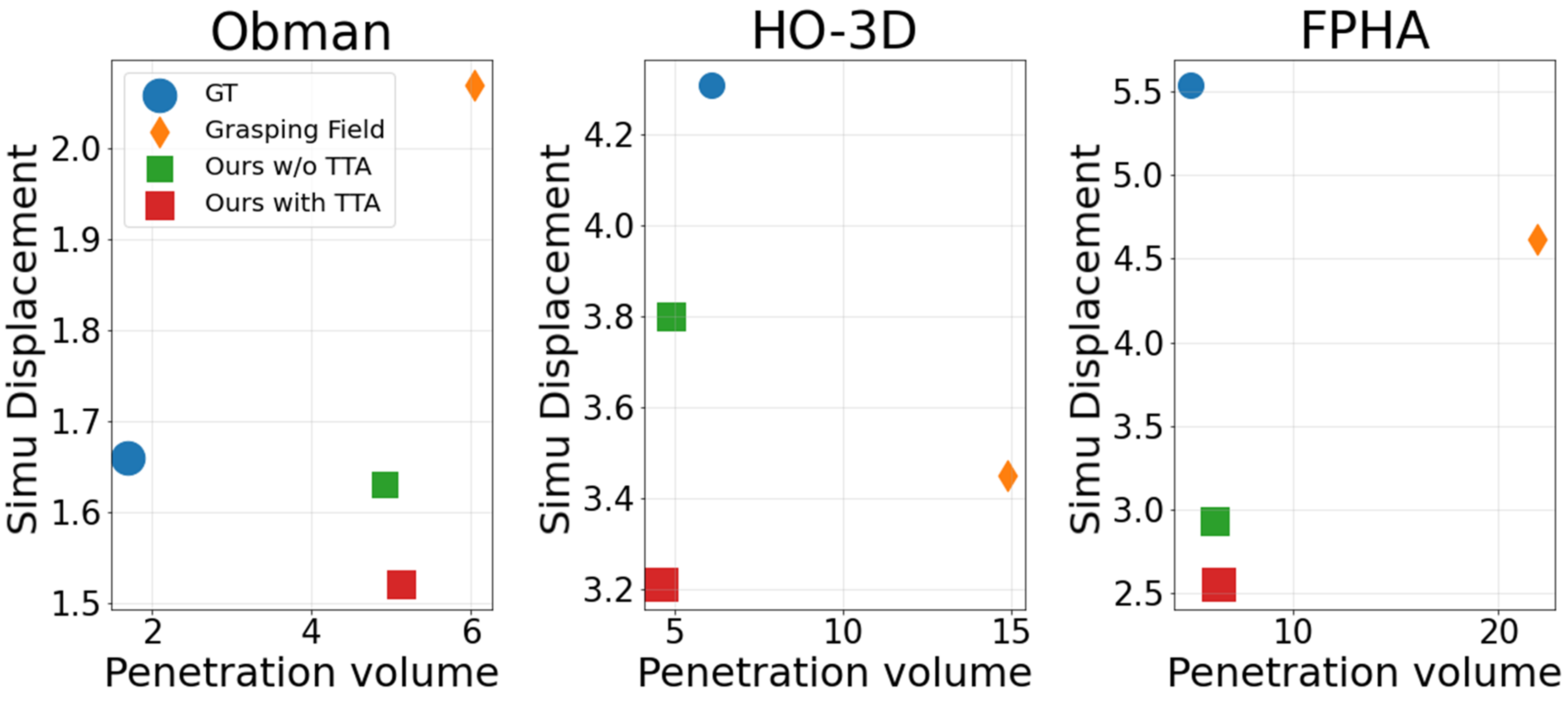
Overall performance (out-of-domain objects)

		HO-3D GT	GF [25]	Ours
Penetration	Depth $(cm)\downarrow$	2.94	1.46	**1.05**
	Volume $(cm^3)\downarrow$	6.08	14.90	**4.58**
Grasp Displace.	Mean $(cm)\downarrow$	4.31	3.45	**3.21**
	Variance $(cm)\downarrow$	$\pm$ 4.42	$\pm$ 3.92	**±3.79**
Perceptual Score	$\{1, ..., 5\}\uparrow$	3.18	3.29	**3.50**
Contact	Ratio (%) $\uparrow$	91.60	90.10	**99.61**

Karunratanakul, Korrawe et al. "Grasping Field: Learning Implicit Representations for Human Grasps." *3DV* (2020).

Quantitative results

Ablation on TTA



Quantitative results

Ablation on TTA

- Optimization
  - Fix network parameters, directly optimize hand parameters

- TTA:
  - Optimize the network parameters
  - Offline: Re-initiate the network weights for each sample
  - Online: Re-initiate the network after many samples

Quantitative results

Ablation on TTA (out-of-domain objects)

	Penetration ↓		Grasp Displace. ↓	Contact ↑
	Depth	Volume	Mean ± Variance	Ratio (%)
w/o TTA	**0.94**	**4.21**	4.98 ± 4.48	86.63
Optimization	1.07	4.59	4.14 ± 4.31	91.45
TTA-offline	1.09	4.88	3.80 ± 4.20	92.31
TTA-online	1.05	4.58	**3.21 ± 3.79**	**99.61**

Summary

- We reason contact consistency between hand-object
  - In both training and testing
  - For more realistic and generalizable human grasp generation

Thanks for listening!