Learning Joint Reconstruction of Hands and Manipulated Objects

Yana Hasson[1,2] Gül Varol[1,2] Dimitrios Tzionas[3] Igor Kalevatykh[1,2] Michael J. Black[3] Ivan Laptev[1,2] Cordelia Schmid[1,4]

[1] Inria [2] Département d'informatique de l'ENS, CNRS, PSL Research University

[3] MPI for Intelligent Systems, Tübingen [4] Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK

MAX-PLANCK-GESELLSCHAFT

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

Goal & Contributions

- Reconstruct **hand and object meshes** given a single RGB image as input.
- Generate a **synthetic** dataset of hands interacting with objects.
- Enforce **physics constraints** in an end-to-end learning framework.

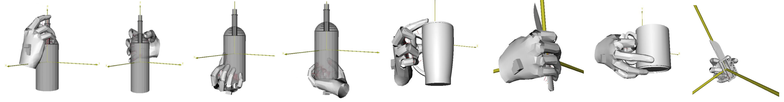
Motivation

- Understanding **object manipulation** is critical to teach robots how to perform useful tasks.
- Ground truth shapes are difficult to obtain for hands and objects in real images.
- Synthetic data is cheap to generate and comes with 3D ground truth, allowing to train CNNs in a fully supervised frameworks.

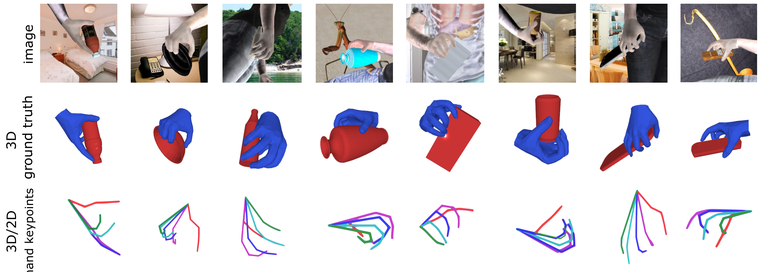
ObMan dataset: Synthetic OBject MANipulation

Generation of diverse grasps leveraging robotics and graphics

- 8 object categories (bowls, botles, cans, ...), 2.7K instances of objects from ShapeNet [1]
- Automatically generated realistic object grasps using GraspIT [8], following [3]



Samples from the generated ObMan dataset



A diverse dataset of hand-object configurations

- Large variety of body and hand poses
- Realistic textures from hand scans
- Object textures selected randomly from ShapeNet
- Randomized lighting

split	train	val	test
#object instances	4K	0.4K	0.4K
#grasp instances	15K	3K	3K
#frames	141K	6K	6K

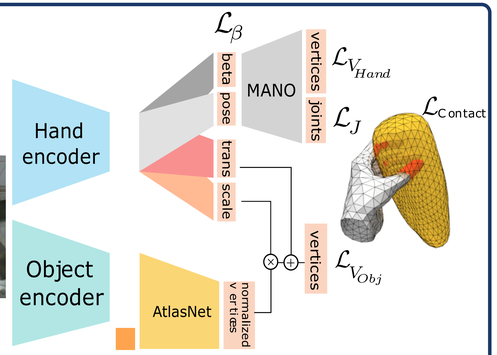
ObMan dataset statistics

Hand-Object Reconstruction

Hand reconstruction: regressing MANO [6] hand model parameters.

Object reconstruction: deforming a sphere using AtlasNet [4].

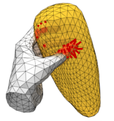
Relative position: regressing object scale and translation relative to the hand



Contact losses: enforcing physically plausible graps

$$\mathcal{L}_{\text{Contact}} = \lambda_R \mathcal{L}_R + (1-\lambda_R)\mathcal{L}_A$$

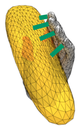
Repulsion loss: penalizes interpenetration between hand and object



V_{Obj}: object vertices
V_{Hand}: hand vertices
$\text{Int}(Obj)$: object interior
$\text{Ext}(Obj)$: object exterior
d: usual distance
C_i: Contact regions

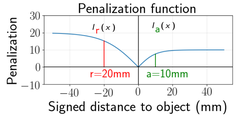
$$\mathcal{L}_R(V_{Obj}, V_{Hand}) = \sum_{v \in V_{Hand}} \mathbb{1}_{v \in \text{Int}(V_{Obj})} l_r(d(v, V_{Obj}))$$

Attraction loss: encourages contact regions of the hand to be close to the object



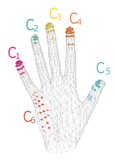
$$\mathcal{L}_A(V_{Obj}, V_{Hand}) = \sum_{i=1}^{6} l_a(d(C_i \cap \text{Ext}(Obj), V_{Obj}))$$

Contact function: soft penalization of vertex distances

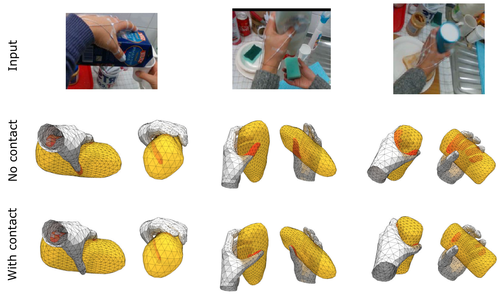


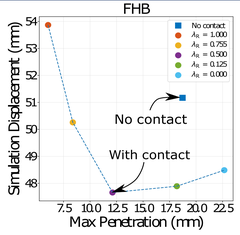
$$l_\alpha(x) = \alpha \tanh\left(\frac{x}{\alpha}\right)$$

Contact regions: Vertices on MANO [6] that are often in contact with objects



Results: First Hand Action Benchmark [2]





Simulation Displacement: quantifies the **stability** of the grasp
Max Penetration: quantifies the **interpenetration** between the hand and the object

Attraction $\mathcal{L}_a$ encourages contacts, but induces interpenetration when used independently

Repulsion $\mathcal{L}_r$ prevents interpenetration, and balances the effect of the attraction term

Used in combination, $\mathcal{L}_a$ and $\mathcal{L}_r$ reduce interpenetration, while preserving hand-object reconstruction accuracy

	ObMan Dataset					FHB Dataset				
	Hand Error	Object Error	Maximum Penetration	Simulation Displacement	Intersection Volume	Hand Error	Object Error	Maximum Penetration	Simulation Displacement	Intersection Volume
No contact loss	11.6	641	9.5	31.3	12.3	28.1	1579	18.7	51.2	26.9
Only attraction ($\lambda_R = 0$)	11.9	637	11.8	26.8	17.4	28.4	1587	22.7	48.5	41.2
Only repulsion ($\lambda_R = 1$)	12.0	639	6.4	38.1	8.1	28.6	1604	6.0	53.9	7.1
Attraction+ Repulsion ($\lambda_R = 0.5$)	11.6	638	9.2	30.9	12.2	28.8	1565	12.1	47.7	17.6

Results: CORe50 dataset [5]

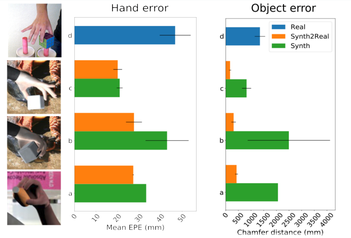
Synthetic to real qualitative results

Transfer: synthetic-to-real

To investigate the domain gap between our synthetic renderings and real datasets, we increasingly match the statistics (hand pose, object shape) of the small Hands in aCtion (HIC) [7] dataset.

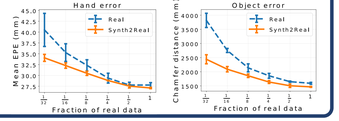
dataset	object shape	hand pose	image domain
(d)	HIC	HIC	real
(c)	HIC	HIC	synthetic
(b)	HIC	GraspIt	synthetic
(a)	ShapeNet	GraspIt	synthetic

Matching the target hand pose and object shape distributions in the synthetic dataset is crucial for good performances on the real dataset.



Pretraining on ObMan before fine-tuning on FHB improves both hand and object reconstructions in low-data regimes

Real: Encoders initialized with ImageNet weights, hand and object decoders initialized randomly, trained on real data
Synth: Trained on synthetic dataset from ImageNet weights
Synth2Real: All weights initialized from hand-object reconstruction task trained on ObMan, fine-tuned on real dataset



Experiment: effect of occlusions

Hands-only images (H-image) and object-only images (O-image) in our dataset enable to systematically study the effect of mutual occlusions on hand pose estimation and object reconstruction.

H O -image O -image H -image

Training images	Evaluation images: H-image	Evaluation images: HO-image
H-image	10.3	14.1
HO-image	11.7	11.6

mepe (mm)

Training images	Evaluation images: O-image	Evaluation images: HO-image
O-image	0.0242	0.0722
HO-image	0.0319	0.0302

chamfer distance

Training with occlusions is crucial when targeting images of hand-object interactions.

References

[1] Chang et al., ShapeNet: An information-rich 3D model repository, 2015
[2] Garcia-Hernando et al., First-person hand action benchmark with RGB-D videos and 3D hand pose annotations, CVPR 2018.
[3] Goldfeder et al, The Columbia grasp database, ICRA 2009.
[4] Groueix et al., AtlasNet: A papier-mâché approach to learning 3D surface generation, CVPR 2018.
[5] Lomonaco et al., Core50: a new dataset and benchmark for continuous object recognition, CoRL 2017.
[6] Romero et al., Embodied hands: Modeling and capturing hands and bodies together, SIGGRAPH Asia 2017
[7] Tzionas et al., Capturing hands in action using discriminative salient points and physics simulation, IJCV 2016
[8] Miller et al., Graspit! A versatile simulatorfor robotic grasping, Robotics Automation Magazine 2004

Models Data Code

https://hassony2.github.io/obman
yana.hasson@inria.fr