HOPE Generator: human-object interaction data synthesis

Mirali Ahmadli, Antonino Emanuele Scurria, Bartlomiej Borzyszkowski

ML4Science Project in the Laboratory of Computational Neuroscience & AI
EPFL, Swiss Federal Institute of Technology in Lausanne

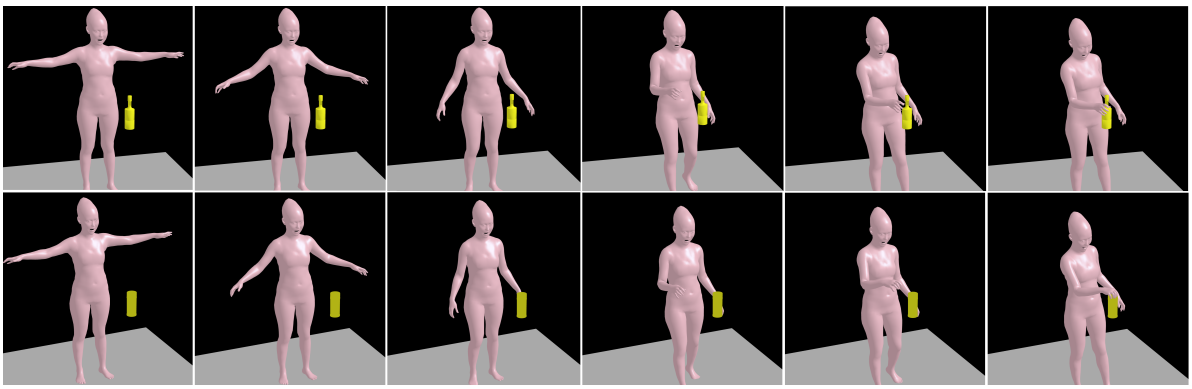

Fig. 1: We introduce the HOPE Generator, a method that leverages and extends GOAL [1] to generate naturalistic, dynamic 4D whole-body human-object interactions with over 1800 objects from the OakInk dataset [2]. For each object, the method generates realistic final grasp (most right) and a sequence of motion that leads to it from the initial pose (most left).

***Abstract*—Hands are the primary means by which humans manipulate objects in the real-world, and measuring hand-object interactions (HOI) and hand-object pose estimation (HOPE) hold great potential for understanding human behavior. However, existing datasets are too small and lack comprehensive awareness of the object's affordance and the hand's interactions with it. For this reason, they are insufficient to elucidate fundamental principles of human movement. We propose to address this problem by exploiting knowledge from three datasets: (large-scale) OakInk [2], GRAB [3], and (large-scale) HOI4D [4] that together contain a wide variety of objects and hand grasps. We propose to adapt GOAL [1] in order to construct a method that generates dynamic whole-body grasps for the collected objects. As a result, our method, HOPE Generator, includes dynamic human-object interactions for over 2600 object instances from the connected datasets. We show quantitatively and qualitatively that this method generalizes well to the unseen objects, thus providing a promising opportunities for motion analysis. Our code is available for reproduction at: github.com/CS-433/ml-project-2-hope-generator**

I. Introduction and Related work

Human object interaction is a rapidly growing field in computer vision. It involves the development of algorithms and techniques for detecting, tracking, and analyzing the interactions between humans and objects in the real-world environments. These techniques have a wide range of applications, including robotics, virtual reality, and human-computer interaction. Moreover, HOI offers numerous opportunities for research in computational neuroscience, e.g. to study human behavior. Our work is hosted by the Laboratory of Computational Neuroscience & AI in order to develop datasets and methods that would help to gain better insights into the sensorimotor control in the long term.

Some of the key challenges in recording HOI with RGB cameras include dealing with occlusions, handling changes in pose and appearance, and variations in lighting and background conditions. On the other hand, using motion capture (MoCap) technology is also limited due to cost, setup requirements, range of motion, and unrealistic environments. For these reasons, available datasets of HOI recordings are relatively small-scale and lack diversity, as they typically contain a limited number of objects [2, 4].

In order to better understand and model complex interactions that occur between people and objects it is essential to generate large-scale data repository. In recent years, intense research efforts have been made in this direction and only in 2022 several SOTA works addressed this problem [11, 12]. However, existing methods usually simplify the task and only focus on the hands, ignoring the rest of the body. We argue that what is really needed to understand the human behavior is to generate the motion of whole-body avatars grasping objects, by jointly considering the body, hands, and objects.

We propose to leverage GOAL [1], a novel approach for generating 4D whole-body motion for hand-object grasping. It uses two neural networks: first, GNet generates a target grasp with a realistic body and hand-object contact; second, MNet generates the motion between the initial and target pose. Although GOAL achieves remarkable results, the authors show its operation only on a small number of objects. Our aim is to reproduce the original results and further adapt the method to a wide range of objects and diverse interaction types (Fig. 1).

We propose to exploit knowledge from the three real HOI datasets and extend them using our synthetic data. Firstly, we use GRAB, a dataset of whole-body humans that manipulate objects. As GRAB was recorded using MoCap, it is characterized by high precision at a cost of scale. It contains 51 objects and a total of 1334 dynamic interactions that are used for training and evaluation of our models. Secondly, we adapt OakInk, a large-scale dataset

that contains 3D object meshes with virtual, single-hand, static grasps. We extend them by applying our generator to synthesize dynamic whole-body grasps for all 1800 objects from OakInk. Finally, we include HOI4D, a large-scale 4D egocentric dataset that includes another 800 different articulated and rigid-body object. Altogether, our method scales to dynamic interactions with over 2600 object instances from the three collected datasets.

Our contributions are concluded in three-fold:

1) We combine three large-scale HOI datasets: OakInk[2], GRAB[3], and HOI4D[4] that together contain over 2600 objects with a wide variety of grasps;
2) We introduce HOPE Generator, a method that leverages and extends GOAL [1] to generate naturalistic, dynamic 4D whole-body human-object interactions for over 1800 objects from the OakInk dataset;
3) We show quantitatively and qualitatively that our method generalizes well to the unseen objects.

II. Datasets

In this section we are going to analyze more in detail the datasets previously introduced. It is important to underline that we chose these 3 datasets because they complete each other and are based on the same body models (SMPL-X and MANO[5, 10]).

A. GRAB

GRAB dataset is focused on obtaining accurate motions without RGB images. Motion capture is used to obtain the dataset with a Vicon system with 54 infrared cameras that capture 16 MP at 120 fps. The markers and the framework used can be seen in [3]. To capture precisely the surface geometry of the objects and therefore the interactions a 3D CAD of the object is used: each object is then represented by a mesh. The link between the MoCap and the mesh of the modelization of the human subject is the body model SMPL-X that will be later presented. Contact is estimated through 3D proximity between the 3D human and the 3D mesh of the objects: namely if the distance between the human model and the object is less than a set tolerance we achieve contact (it helps with measurement errors and the fact that human soft tissue deforms whereas SMPL-X body model can't deform). The dataset contains 1334 sequences and over a 1.6M frames of MoCap with different objects and tasks (the following tables provides a synthetic overview of the dataset). The important limitations of GRAB dataset compared to OakInk and HOI4D is that we are not dealing with a large-scale dataset: indeed the dataset is made up of intercations of 10 different subjects with 51 everyday objects. Thus we only have a very small number of available objects (and thus a small number of interactions) compared to the following datasets.

B. OakInk

OakInk is a large-scale dataset of static grasps and is made up of 2 different related bases: the first is associated with the object-centric affordance knowledge (Oak base); the second is associated with the human-centric interaction knowledge (Ink base). The Oak base deals with 1800 household objects that are designed for one-hand manipulations. The objects obtained by different sources (see [2]) are then collected in a graph structure that specifies the functionality and other features. The Ink base, namely the human-centric part of the dataset, is collected through a platform of a multi-camera system and an infrared motion capture system. The dataset is made up of 230K image frames of 12 human subjects performing up to 5 intent-oriented interactions, namely 'use', 'hold', 'lift-up', 'hand-out', and 'receive', with 100 real world object of 32 different categories; then the interactions with these first 100 objects are transferred to the 1700 virtual counterparts that share the same category. The procedure doesn't simply consist in copying the hand pose since the shape of objects of the same category (for example cameras) may vary; instead the latter procedure is based on different steps: the first step is to perform a continuous shape interpolation from the source object to the target object (DeepSFD [13], a neural generative model, is used in this step); then the second step is to map the contact regions from the source object to the target one and to refine the pose (this second step is formulated as an iterative optimization problem; to have more details about the framework used see [2] section 3.3). Finally the OakInk dataset has 50.000 different hand objects interactions.

C. HOI4D

HOI4D is a large-scale dataset for hand object interactions that consists of 2.4M RGB-D images egocentric video frames over 4000 sequences of 9 human subjects interacting with 800 objects. HOI4D is built using a Kinetic v2 RGB sensor and an Intel RealSense D455 RGB-D sensor. The objects are divided into 16 categories that include both rigid and articulated objects: each category consists of 50 objects instances obtained through CAD models built from high resolution RGB images. Moreover, for all the categories, 54 tasks are defined. The procedure to obtain the data annotation consists, at first, in splitting the dynamic RGB-D sequence in moving content and static content to ease panoptic labeling by annotating framewise 2D motion segmentation; then the 3D static scene is reconstructed via a SLAM algorithm (see [7, 8] for more details); finally a 3D static panoptic segmentation generated manually, annotating the reconstructed scene and the the 3D static panoptic segmentation, and the 2D motion segmentation are merged to obtain a 4D dynamic scene panoptic segmentation. The hand pose annotation is instead estimated by the 2D annotations through an optimization problem related to a specific loss function (to read more details please see [4]).

TABLE I: Comparison of the datasets.

Dataset parameters	GRAB	OakInk	HOI4D
# of subjects	10	12	9
# of objects	51	1800	800
# of interactions	1334	50000	4000
full-body	yes	no	no
hand(s)	R+L	R	R+L
body / hand model	SMPL-X	MANO	MANO
egocentric	no	no	yes

III. Methodology

This section discusses methodology, including the machine learning pipeline and data synthesis of the HOPE

Generator. We use zero-shot learning, specifically interactions are generated without retraining the models and the original weights from GOAL are used.

A. Body model

We use the SMPL-X statistical 3D whole-body model, which combines shape and pose parameters and allows to represent a wider range of body types, including gender and ethnicity. It jointly represents the body, head, face and hands. As input, it takes shape β, pose θ, and expression ψ, parameters and then outputs a 3D mesh M, with vertices given as V and triangles F.

B. Architecture overview

Our architecture leverages GOAL [1], which consists of two neural networks:

1) GNet: GNet is a conditional Variational Auto Encoder (cVAE) that generates a static whole-body grasp, conditioned on the given object and its location (Fig. 2). Firstly, it encodes whole-body grasps into an embedding space. Then, the decoder takes a sample from this space and outputs SMPL-X parameters $\hat{\theta}$, the head direction vector $\hat{q}$, and hand offset vectors $\hat{d}^h$. Additionally, the interaction features $\hat{q}$ and $\hat{d}^h$ are used to refine the predicted SMPL-X parameters $\hat{\theta}$ to get a more realistic whole-body grasp. The loss function used for GNet is define as :

$$\mathcal{L}_{\text{GNet}} = \lambda_{\boldsymbol{v}}\mathcal{L}_{\boldsymbol{v}} + \lambda_{\boldsymbol{v}}^{h}\mathcal{L}_{\boldsymbol{v}}^{h} + \lambda_{\boldsymbol{\Theta}}\mathcal{L}_{\boldsymbol{\Theta}} + \lambda_{\text{q}}\mathcal{L}_{\text{q}} + \lambda_{d}^{h\to\circ}\mathcal{L}_{d}^{h\to\circ} + \lambda_{KL}\mathcal{L}_{KL},$$

where $\mathcal{L}_{\boldsymbol{v}} = \|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_1, \mathcal{L}_{\boldsymbol{v}}^{h} = \left\|\boldsymbol{v}^h - \hat{\boldsymbol{v}}^h\right\|_1, \mathcal{L}_{\boldsymbol{\Theta}} = \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2, \mathcal{L}_{\text{q}} = \|\boldsymbol{q} - \hat{\boldsymbol{q}}\|_2, \mathcal{L}_{d}^{h\to\circ} = \left\|\boldsymbol{d}^{h\to\circ} - \hat{\boldsymbol{d}}^{h\to\circ}\right\|_1$ $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, λ are weights and $\boldsymbol{v} \in \mathbb{R}^{N_b\times 3}$ are the 3D coordinates of the N_b smapled SMPL-X vertices. Hat variables are inferred; non-hat ones are ground truth.

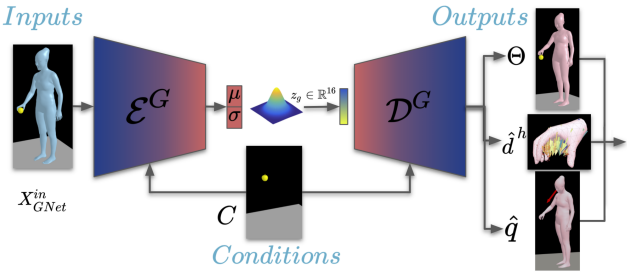

Fig. 2: Overview of the cVAE network architecture in GNet.

2) MNet: Auto-regressive network that generates motion by taking in each iteration 5 past frames, X_p, and generating the next 10 frames, X_f (Fig. 3). It uses linear interpolation as the optimization process that refines the motion when hand vertices get closer than 10cm to the target location generated by GNet. The loss function associated to MNet is defined as follows:

$$\mathcal{L}_{\text{MNet}} = \lambda_v\mathcal{L}_v + \lambda_v^h\mathcal{L}_v^h + \lambda_\Theta\mathcal{L}_\Theta + \lambda_{\boldsymbol{d}}^{h\circ o}\mathcal{L}_{\boldsymbol{d}}^{h\to 0} + \lambda_v^f\mathcal{L}_{\boldsymbol{v}}^{\text{f}}$$

where all the first loss terms are defined as in the GNet loss function while the last term is a loss function related to the position of the feet, namely $\mathcal{L}_v^f = \left\|\boldsymbol{v}^{\text{f}} - \hat{\boldsymbol{v}}^{\text{f}}\right\|$.

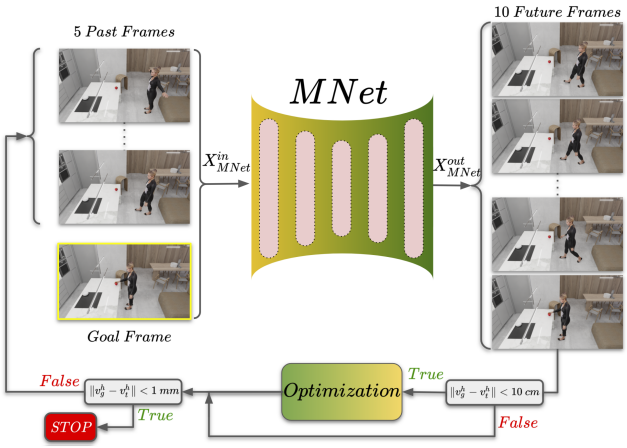

Fig. 3: Overview of the MNet network together with the linear interpolation as the optimization post-processing step.

C. Data preparation

1) Data preprocessing for GNet: From our datasets (GRAB and OakInk) we collect all the frames with stable right-hand grasps of the objects: to do so we follow a selection criteria used for the training data of [3]. Then the object is placed at the origin (during this step the height of the object is preserved because a change of the height of the object would cause a change of the body pose related to the grasp).

2) Data preprocessing for MNet: MNET generates the motion from a standard 'T pose' of the body model to the goal grasps: all the frames (starting from the first and ending with a stable grasp frame) are gathered (the same selection criteria as GNET is used for the stability of the grasp). Then several sub-sequences are generated with a length of 21 frames with a stride of 1 frame: the central frame of the sub-sequence is considered as the 'current frame', while the last 10 and the first 10 frames are respectively considered as 'future frames' and 'past frames'. Then following [9] all past and future frames are made relative to the body coordinate system of the current frame (with the gravity direction upwardly oriented).

IV. Results

A. Qualitative evaluation

For generated grasps, we use the refinement procedure in the GOAL [1]. This optimization refines hands for more realistic and physically plausible grasps. This is shown in Figure 4 where pink grasp and green grasp are respectively before and after the refinement step.

While method generalizes well for most objects, it sometimes fails to grasp thin objects (such as pens) and depending on the orientation of the given object, it does not grasp the object from instinctive handles for human (Figure 5).

Figure 6 shows the generated motion from initial T-pose to the target grasping pose.

B. Quantitative evaluation

We report our quantitative metrics on both generated grasps and generated motions.

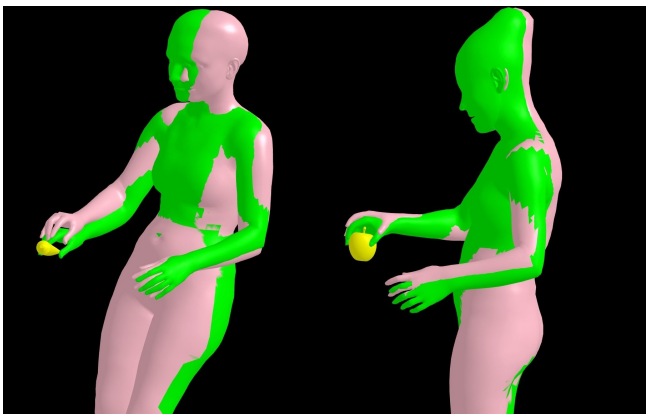

Fig. 4: Optimization process of the grasping pose with two different objects.

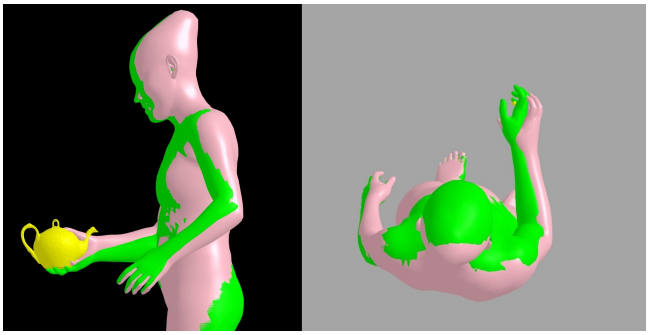

Fig. 5: Examples of penetration for a teapot (left) and lack of contact for a pen (right).

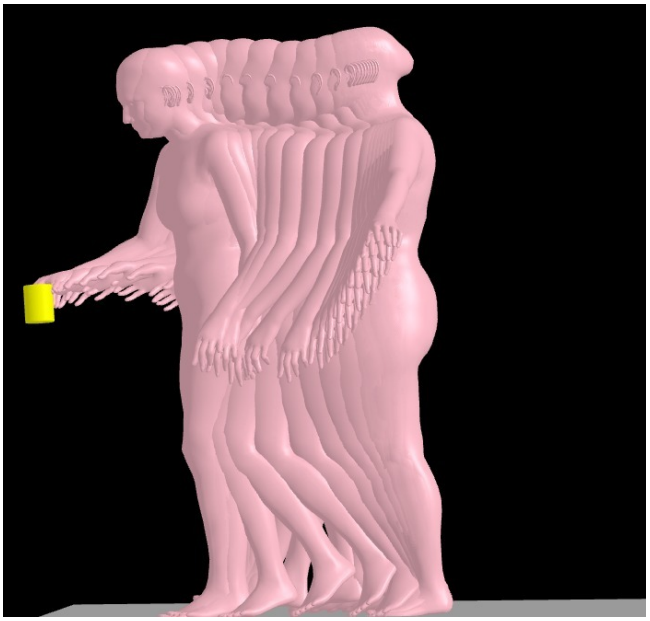

Fig. 6: Motion of the subject towards the goal grasping pose.

For generated grasps, we calculate penetration volume and contact ratio. For penetration volume, we find the intersection mesh between human and object meshes, and calculate its volume. For contact ratio, we calculate how many grasps are in contact and get the ratio. To check if they are in contact, we check if penetration volume is nonzero or if they are in contact on the surface of the object.

For generated motions, we report vector-to-vector loss, which is L2 norm / Euclidean distance between target grasp and last generated frame.

For objects in OakInk dataset, GNet fails to generate grasp that is in contact with thin objects even after the optimization (Figure 5) and generates grasp with unrealistic interaction such as tea pot where the human mesh penetrates the object a lot (Figure 5).

For generated motions, MNet performs better on hands for OakInk dataset than GRAB, meanwhile body vertices have $6cm$ shift with target grasps.

TABLE II: Penetration Volume and Contact Ratio of generated grasps GNet.

Grasp Synthesis	Penetration Vol. (cm^3)	Contact ration
GNet - GRAB	2.22	1.00
GNet - OakInk	2.46	0.97

TABLE III: V2V loss between target grasp and generated motion MNet.

Motion	Body	Hand	Feet
MNet - GRAB	$19.7mm$	$28.0mm$	$9.9mm$
MNet - OakInk	$62.07mm$	$12.0mm$	$9.89mm$

V. Conclusion

In our work we introduce HOPE-Generator, a method to generate a large-scale dataset on HOI that generalizes well to the unseen objects. During our work we were able to merge the contributions of different advanced papers in computer vision and to face and overcome the limits of the available datasets on HOI. The qualitative and quantitative evaluation shows that our work is able to generate natural and physically admissible grasping motions.

A. Future Work

HOPE-Generator provides a method to generate large-scale dataset on HOI and so new possibilities to study the grasping motions. Even though the motion generated is physically realistic we can sometimes observe grasps that are not 'humanly realistic' (for example grasping a teapot from the upper part and not from the handle): one of our future plans is to train the model to generate grasps from the handles (and the most natural parts to grasp) of the objects. Another possible future development is to consider long range interactions and focus more on the walking movement (the model isn't currently able to generate long-distance interactions and also the walking movement has to be optimized not to obtain 'sliding' phenomena).

References

[1] Taheri, Omid and Choutas, Vasileios and Black, Michael J. and Tzionas, Dimitrios "GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping", Conference on Computer Vision and Pattern Recognition (CVPR),2022

[2] Yang, Lixin and Li, Kailin and Zhan, Xinyu and Wu, Fei and Xu, Anran and Liu, Liu and Lu, Cewu, "OakInk: A Large-Scale Knowledge Repository for Understanding Hand-Object Interaction", IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2022

[3] Taheri, Omid and Ghorbani, Nima and Black, Michael J. and Tzionas, Dimitrios, "GRAB: A Dataset of Whole-Body Human Grasping of Objects", European Conference on Computer Vision (ECCV), 2021

[4] Liu, Yunze and Liu, Yun and Jiang, Che and Lyu, Kangbo and Wan, Weikang and Shen, Hao and Liang, Boqiang and Fu, Zhoujie and Wang, He and Yi, Li, "HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022

[5] Pavlakos, Georgios and Choutas, Vasileios and Ghorbani, Nima and Bolkart, Timo and Osman, Ahmed A. A. and Tzionas, Dimitrios and Black, Michael J., "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image", Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019

[6] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation", Conference on Computer Vision and Pattern Recognition (CVPR),2019

[7] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun, "Robust reconstruction of indoor scenes", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015

[8] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun., " Open3d: A modern library for 3d data processing", arXiv preprint, 2018

[9] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito, "Neural state machine for character-scene interactions", Transactions on Graphics (TOG), 2019

[10] Romero, Javier and Tzionas, Dimitrios and Black, Michael J., "Embodied Hands: Modeling and Capturing Hands and Bodies Together", ACM Transactions on Graphics, 2017

[11] Turpin, Dylan, et al. "Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands." European Conference on Computer Vision. Springer, Cham, 2022.

[12] Christen, Sammy, et al. "D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[13] Park, Jeong Joon and Florence, Peter and Straub, Julian and Newcombe, Richard and Lovegrove, Steven, "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019