Grasping Field: Learning Implicit Representations for Human Grasps

Korrawe Karunratanakul[1] Jinlong Yang[2] Yan Zhang[1]
Michael J. Black[2] Krikamol Muandet[2] Siyu Tang[1]
[1]ETH Zurich [2]Max Planck Institute for Intelligent Systems
{korrawe.karunratanakul,yan.zhang,siyu.tang}@inf.ethz.ch {jyang,krikamol,black}@tue.mpg.de

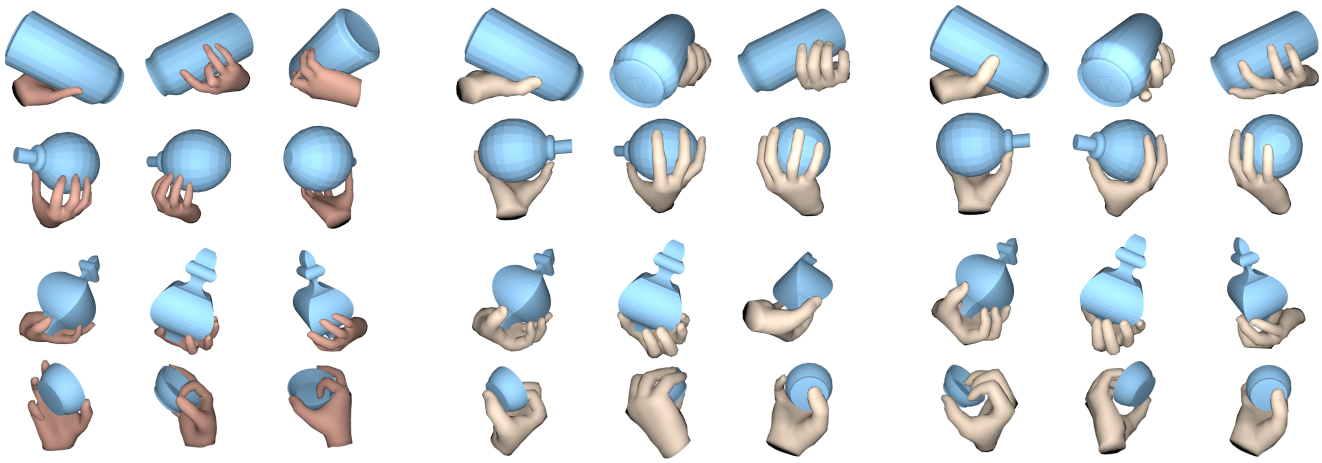


Figure 1: Ground truth grasps and generated grasps. Each row corresponds to one object. Left three columns show the ground truth grasps, each from three different viewpoints. The middle three columns show one generated example, and the right three columns show another generated example. Note that these objects are never seen during training. See Appendix E (Fig. E.4) for more examples.

Abstract

Robotic grasping of house-hold objects has made remarkable progress in recent years. Yet, human grasps are still difficult to synthesize realistically. There are several key reasons: (1) the human hand has many degrees of freedom (more than robotic manipulators); (2) the synthesized hand should conform to the surface of the object; and (3) it should interact with the object in a semantically and physically plausible manner. To make progress in this direction, we draw inspiration from the recent progress on learning-based implicit representations for 3D object reconstruction. Specifically, we propose an expressive representation for human grasp modelling that is efficient and easy to integrate with deep neural networks. Our insight is that every point in a three-dimensional space can be characterized by the signed distances to the surface of the hand and the object, respectively. Consequently, the hand, the object, and the contact area can be represented by implicit surfaces in a common space, in which the proximity between the hand and the object can be modelled explicitly. We name this 3D to 2D mapping as Grasping Field, parameterize it with a deep neural network, and learn it from data. We demonstrate that the proposed grasping field is an effective and expressive representation for human grasp generation. Specifically, our generative model is able to synthesize high-quality human grasps, given only on a 3D object point cloud. The extensive experiments demonstrate that our generative model compares favorably with a strong baseline and approaches the level of natural human grasps. Furthermore, based on the grasping field representation, we propose a deep network for the challenging task of 3D hand-object interaction reconstruction from a single RGB image. Our method improves the physical plausibility of the hand-object contact reconstruction and achieves comparable performance for 3D hand reconstruction compared to state-of-the-art methods. Our model and code are available for research purpose at https://github.com/korrawe/grasping_field.

1. Introduction

Capturing and synthesizing hand-object interaction is essential for understanding human behaviours, and is key to a number of applications including augmented and virtual reality, robotics and human-computer interaction. Despite substantial progress, fully automatic synthesis of highly realistic human grasps remains an unsolved problem. The anatomical complexity of the human hand and the variety of manufactured and natural objects make it extremely challenging to pose the hand such that it interacts with the object in a natural and physically plausible way. Recent data-driven approaches explore deep learning technology to learn and leverage powerful object representations, yet they are mainly limited to simple robotic end effectors, such as parallel jaw grippers [69]. In this work, we seek to understand: 1) what is an efficient and expressive representation for modeling hand-object interaction, that can facilitate realistic human grasp synthesis given an unseen 3D object; and 2) how can we learn such a representation from data.

Our key observation is that human grasping is rooted in physical hand-object *contact*. Through this contact, humans are able to grasp and manipulate objects naturally. To better model hand-object interaction, we must find a way to effectively represent the contact between hands and objects. To this end, we propose a novel interaction representation that is based on regressing a continuous function that we call the *Grasping Field*. The grasping field maps any 3D point to a 2D space, where each dimension of the 2D space indicates the signed distance to the surface of the hand and the object respectively (see Sec. 3.1 for a formal definition). Inspired by [51, 58, 61], we further utilize a deep neural network to parameterize the grasping field and learn it from data. As a result, the learned grasping field serves as a powerful representation to facilitate hand-object interaction modelling.

Based on the grasping field representation, we propose a generative model, in which we generate plausible hand grasps given an object point cloud. We show that our model can produce physically and semantically plausible synthetic grasps, which are similar to the ground truth. Generated grasps on unseen objects are shown in Fig. 1 and Fig. E.4.

We further demonstrate the effectiveness of the grasping field representation by considering the task of 3D hand and object reconstruction from a single RGB image. In recent work, Hasson et al. [29] introduce an end-to-end learnable model to reconstruct 3D meshes of the hand and object simultaneously, producing the state-of-the-art results on several datasets. Physical constraints, such as no interpenetration and proper contact, are enforced during the training. However, there are several drawbacks of their mesh-based representation for hand-object interaction modeling. First, they heuristically pre-define regions of the hand that can be in contact with objects. Second, their object representation is limited to objects of genus zero. Third, the resolution of their contact inference is limited by the resolutions of the hand and object meshes. In contrast, with the grasping field representation, it is not necessary to first compute the hand and object meshes, and then compute the contact region. Instead, one can easily infer the contact region by querying the signed distances of input 3D points. Furthermore, the physical constraints, such as no inter-penetration and proper contact, can be efficiently computed and enforced. As demonstrated in our experiments, our model considerably reduces the interpenetration between the reconstructed hand and the object, and improves the quality of 3D hand reconstruction, compared with [29].

In summary, our contributions are: (1) We propose the grasping field, a simple and effective representation for hand-object interaction; (2) Based on the grasping field, we present a generative model to yield semantically and physically plausible human grasps given a 3D object point cloud; (3) We further propose deep neutral networks to reconstruct the 3D hand and object given an RGB input in a single pass; (4) We perform extensive experiments to show that our method outperforms the baseline [29] on 3D hand reconstruction and on synthesizing grasps that appear natural.

2. Related work

Human grasp and contact. There is a large body of work on capturing and recognizing human grasps [13, 22, 31, 56, 64, 92]. Recently, [22] introduced a stretch-sensing soft glove to capture accurate hand pose without extra optical sensors. Puhlmann et al. [67] utilized a touch screen to facilitate the capturing process of human grasping. As physical contact is fundamental to hand-object interaction, researchers have proposed methods to capture and modeling contact from diverse modalities [5, 42], but these often interfere with natural movement. Concurrent to this work, [7] proposed a new dataset of hand-object contact paired with RGB-D images. Our work differs in that our focus is on learning an interaction representation, which is efficient and easy to interface with deep neural networks.

Grasp synthesis. Grasp synthesis is a longstanding problem in robotics and graphics, resulting in an extensive literature [1, 4, 6, 16, 36, 40, 43, 47, 59, 65, 70, 76, 77, 81, 95]. As early as 1991, Rijpkema and Girard [70] proposed a knowledge-based approach to incorporate the role of the human hand, object, environment and animator for the task of computer-animated grasping. More recent works can be categorized into three types of approaches: analytic, data-driven and hybrid approaches. For the analytic approaches [39, 77], the grasps are often synthesized by formulating the problem as a constrained optimization problem that satisfies a set of criteria measuring the stability or other properties of the grasps. The data-driven approaches [63, 69] often employ machine learning methods to learn representations for synthesizing grasps. An excel-

lent survey of data-driven grasp generation is presented in [3]. Recent hybrid approaches [47, 46] combine analytic models and deep learning tools to synthesize grasps for various end effectors. Finally, the most related approaches to our work was presented in [11] and [84], where neural networks are used to predict hand parameters of the MANO hand model [74] given object information. In [11], the model learns to predict the best grasp type from the grasp taxonomy [18] according to the RGB images of the objects. Then, the predicted hands are optimized together with the object meshes to refine the contact points. While in [84], the parameters are generated directly from the given Basis Point Set [66] of the objects. Our work differs from the previous works in that, by also considering the object distance field, we propose a learnable representation for modelling hand-object interaction that can be used without contact post-processing. Empowered by deep neural networks, the learned representation enables us to synthesize realistic human hand grasping a given object naturally.

Hand pose estimation. Hand pose estimation is a long-standing problem, and various input modalities have been considered, e.g., RGB images [2, 15, 32, 54, 60, 79, 80, 96] or RGB-D and depth sensors [24, 35, 48, 55, 57]. Due to the lack of large scale 3D ground truth data, synthetic data has often been used for training [14, 29, 49]. Recently, instead of estimating the hand skeleton, recovering the pose and the surface of the hand has become popular using statistical hand models, e.g., the MANO model [74], that can represent a variety of hand shapes and poses [25, 93, 97]. Using the template derived from MANO, [41] show that it is also possible to regress hand meshes directly using mesh convolution. In this work, we represent the 3D hand by a signed distance field, instead of a parametric hand model, due to the difficulty of incorporating object interaction into the model parameter space. For fair comparison with the parametric hand model representation, we fit the MANO model [74] into our resultant signed distance fields. The experimental results indicate the advantage of our new interaction representation.

Object model representation. Learning 3D object models using various types of representations has also been explored [9, 23, 33, 44, 50, 52, 82, 90, 91]. Recently, the community has focused on using the implicit functions such as the Signed Distance Function (SDF) [61], Occupancy Networks [51], Implicit Field [10], and their derivations [20, 21], as these can model arbitrary object topology with adjustable resolution. Due to these advantages, we also adopt implicit functions to capture hand-object interaction.

Hand-object interaction. Reconstructing hand and object jointly has been studied with both RGB input and RGB-D input [62, 71, 72, 73, 78, 83, 86, 87, 88, 89]. Recently, Hasson et al. [27, 29] achieved promising results on explicitly modeling the contact by combining a parametric hand model MANO [74], with the mesh based representation for the object. As data for hand-object interaction is limited, we opt to use their synthetic dataset, the ObMan dataset [29], which is sufficiently large for training a neural network. Our work differs from previous hand-object reconstruction work mainly by focusing on the novel representation of contact and learning both hand and object in the signed-distance space, which allows arbitrary shape modelling and easier distance field manipulation. Furthermore, we go beyond the reconstruction task by proposing generative models to synthesize realistic human grasps given a 3D object.

3. Method

3.1. Grasping field

The *grasping field* (GF) is based on the signed distance fields of the object and the hand, formally defined as a function $f_{GF} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, mapping a 3D point to the signed distances to the hand surface and the object surface, respectively. In this way, the contact and inter-penetration relations between the hand and the object can be explicitly and efficiently represented. Specifically, the hand-object contact manifold is given by $\mathcal{C} = \{\boldsymbol{x} \mid f_{GF}(\boldsymbol{x}) = \mathbf{0} \text{ for } \boldsymbol{x} \in \mathbb{R}^3\}$. The volume of hand-object inter-penetration is given by $\mathcal{I} = \{\boldsymbol{x} \mid f_{GF}(\boldsymbol{x}) < \mathbf{0} \text{ for } \boldsymbol{x} \in \mathbb{R}^3\}$.

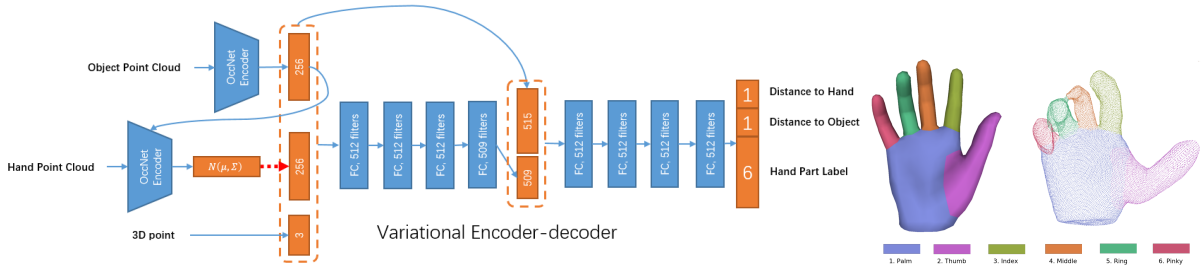
Inspired by [61] and [51], we propose to model f_{GF} using a deep neural network, and learn it from data. Therefore, one can infer hand-object interaction in 3D space without the explicit hand and object surfaces. The learned GF can be considered as an interaction prior, which enables us to infer various grasping poses of the hand, only based on the 3D object. Furthermore, in contrast to previous works, e.g. [26, 29, 94], which can only evaluate body-object interactions after obtaining the body and the object meshes, when using GF as the representation in hand-object reconstruction from images, we model the hand, the object, and the contact area by the implicit surfaces in a common space, largely improving the physical plausibility of the reconstruction.

According to the aforementioned merits of the GF, we use it to address two tasks in this paper; i.e. hand grasp generation given 3D objects and hand-object reconstruction from RGB images. Different GF networks are designed specifically for different tasks.

3.2. Grasping field for human grasp synthesis

In this section, we show how to use GF to synthesize human grasps. Given an object point cloud, the goal is to generate diverse hand grasps that interact with the object in a natural manner.

Network architecture. The network architecture is shown in Fig. 2a; we adopt the encoder-decoder framework.



(a) (b)

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

To extract features from point clouds, we use the PointNet encoder [68] with residual connection. The encoder is trained jointly with other network layers from scratch. The encoder-decoder network takes a query 3D point, and two point clouds of the hand and the object as input, and produces the signed distances of the query point to the hand and the object surfaces. In addition, the encoded object point cloud feature is fed into the hand point cloud encoder, leading to a hand distribution conditioned on the object. Note that this variational encoder-decoder network only requires both hand and object point clouds during the training. During inference, only the conditioning object point cloud and the query point are required. The hand features are sampled from the learned latent space, as in a standard VAE [38]. The training loss consists of the following terms:

The reconstruction loss $\mathcal{L}_{rec}$: For each query point $\boldsymbol{x}$, the input object point cloud p^o and the input hand point cloud p^h, the reconstruction loss is designed for the hand and object individually:

$$\mathcal{L}_{rec} = |c(f_{CGF}(\boldsymbol{x}, p^o), \delta) - c(\text{SDF}_{p^o}(\boldsymbol{x}), \delta)| + |c(f_{CGF}(\boldsymbol{x}, p^h), \delta) - c(\text{SDF}_{p^h}(\boldsymbol{x}), \delta)|, \quad (1)$$

where f_{CGF} is the grasping field network (Fig. 2a). $\text{SDF}_{p^h}(\cdot)$ and $\text{SDF}_{p^o}(\cdot)$ are the ground truth SDF for the hand and object, respectively. In addition, $c(s, \delta) := \min(\delta, max(-\delta, s))$ is a function to constrain the distance s within $[-\delta, \delta]$. δ is set to 1cm in all experiments.

KL-Divergence $\mathcal{L}_{kl}$: In order to generate new hand grasps, we use a KL-divergence loss to regularize the distribution of hand latent vector h, obtained from the hand point cloud encoder $h = E_h(p^h|p^o)$, to be a normal distribution. The loss is given by

$$\mathcal{L}_{kl} = \textbf{KL-div}\left(N(\mu(h), \sigma(h))||N(0, \boldsymbol{I})\right), \quad (2)$$

where $N(0, \boldsymbol{I})$ denotes a standard high-dimensional normal distribution, μ and σ denotes mean and standard deviation. For generation, the hand latent vector h is sampled from a standard normal distribution.

Classification loss $\mathcal{L}_{cls}$: Besides predicting the signed distances of a query point, we also train the network to produce the hand part label of a query point to parse the hand semantically. To achieve this, we introduce a classification loss, which is given by a standard cross-entropy loss. The hand part annotation is based on the MANO model [74] as illustrated in Fig. 2b.

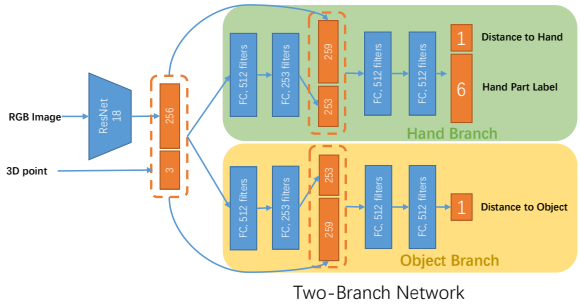
3.3. Grasping field for 3D hand-object reconstruction from a single RGB image

Our proposed grasping field is an expressive representation for modelling hand-object interactions in 3D. Here we address the challenging task of 3D hand-object reconstruction from a single RGB image, i.e. $f_{CGF} : \mathbb{R}^3 \times \mathcal{I} \rightarrow \mathbb{R}^2$, in which $I \in \mathcal{I}$ is a 2D image. We model such a conditional GF by two types of deep neural networks and learn their parameters from data.

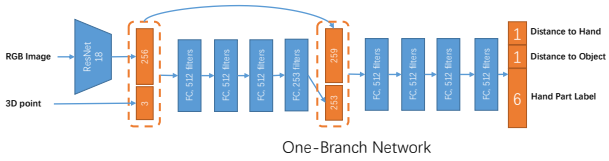
Network architecture. The network architectures are illustrated in Fig. 3, which are designed to recover both hand and object in a single pass. To enable a direct comparison with [29], the two-branch network is employed (Fig. 3a), which addresses hand and object individually. Similar to [29], we introduce contact and inter-penetration losses during the training to facilitate a better 3D reconstruction on the contact regions of the hand and the object. To introduce hand-object interactions in early stages, we propose a one-branch network (Fig. 3b), which uses the same image encoder and has the same number of layers with the two-branch model. See Appendix A for architecture details.
The training loss consists of the following terms:

The reconstruction loss $\mathcal{L}_{rec}$: For each query point $\boldsymbol{x}$ and the input image I, the reconstruction loss is designed for the hand and object individually, and is given by

$$\mathcal{L}_{rec} = \sum_{p \in \{p^h, p^o\}} |c(f_{CGF}(\boldsymbol{x}, I), \delta) - c(\text{SDF}_p(\boldsymbol{x}), \delta)|, \quad (3)$$



(a)



(b)

Figure 3: Two different network architectures of the GF conditioned on the image. The blue blocks denote network modules and layers. A ReLU layer and a dropout layer (dropout ratio 0.2) are between every two consecutive fully-connected (FC) layers. The orange blocks denote feature vectors, and the feature dimensions are presented inside of these feature blocks. The orange dashed boxes denote feature vector concatenation.

in which f_{CGF} is our conditinal grasping field network, and $\text{SDF}_p(\cdot)$ is the ground truth SDF for the component p (hand or object). $c(s,\delta) := \min(\delta, max(-\delta, s))$ is the thresholding function to constrain the distance s within $[-\delta, \delta]$ as with the generative model proposed in Sec. 3.2.

The inter-penetration loss $\mathcal{L}_{ip}$: To avoid surface inter-penetration between the reconstructed hand and object, we define the inter-penetration loss as

$$\mathcal{L}_{ip} = \sum_{\boldsymbol{x}} \max(-\langle \mathbf{1}, f_{CGF}(\boldsymbol{x}, I)\rangle, 0), \tag{4}$$

where $\mathbf{1}$ is a 2D one-vector, and $\langle\cdot,\cdot\rangle$ denotes a dot product. This loss function actually penalizes the negative sum of predicted signed distances to the object and to the hand. If the hand and the object are separate and have no contact, the signed distance sum of every point in 3D space is always positive, and hence is ignored by our inter-penetration loss. On the other hand, if the hand and the object have inter-penetration, then this inter-penetration loss does not only penalize the points in the intersection volume, but also all 3D points in the space, indicating that the predicted hand and object are incorrect. Compared to the inter-penetration methods in [26, 94], which only penalize the intersection volume, our loss applies stronger constraints.

Contact loss $\mathcal{L}_{cont}$: Our proposed contact loss encourages hand-object contact, and is given by

$$\mathcal{L}_C = \sum_{\boldsymbol{x}} \min(\alpha |f_{CGF}(\boldsymbol{x}, I)|^2, 1), \tag{5}$$

where α is a hyper-parameter. We can see that $f_{CGF}(\boldsymbol{x}, I) = 0$ corresponds to the hand-object contact surface. Therefore, it ignores points with predicted grasping field $|f_{CGF}(\boldsymbol{x}, I)|^2 \geq \frac{1}{\alpha}$, and only encourages points with $|f_{CGF}(\boldsymbol{x}, I)|^2 < \frac{1}{\alpha}$ to be the contact points. In our study, we empirically set $\alpha = 0.005$ based on the hand-object interactions in the training data. Finally, we employ the same **Classification loss $\mathcal{L}_{cls}$** as the one proposed in Sec. 3.2.

3.4. From grasping field to mesh

With the trained grasping field conditioned on images or point clouds, one can compute the signed distances to the hand and object of a query 3D point. To recover the hand, object and their interactions, we first randomly sample a large number of points, and evaluate their signed distances. The point clouds belonging to the hand and the object can be selected, according to point-object signed distances close to zero. Then, the hand mesh and the object mesh are obtain by marching cubes [45].

In addition, the hand mesh can be recovered by fitting the MANO [74] model to the hand point cloud. In this case, we can obtain hand segmentation, hand joint positions, and a compact representation of the hand simultaneously, according to the pre-defined topology in MANO.

Denoting the MANO model by $\mathcal{M}(\beta, \theta)$ with the parameter β and θ representing hand shape and pose respectively, we minimize $\sum_{l=1}^{6} d(\tilde{p}^h, \mathcal{M}(\beta, \theta)^l)$ to recover the hand configuration, in which l denotes the 6 parts of hand, i.e. the palm and the 5 fingers, $\tilde{p}^h$ denotes the hand point cloud produced by our model, $\mathcal{M}(\beta, \theta)^l$ denotes the MANO hand mesh belonging to the hand part l, and $d(\cdot,\cdot)$ denotes the Chamfer distance [17, 61]. The hand segmentation is shown in Fig. 2b.

The implementation details are thoroughly presented in Appendix A.

4. Experiments

We demonstrate the effectiveness of the grasping field representation on two challenging tasks: human grasp generation given a 3D object and 3D hand-object reconstruction from a single image.

Dataset. To train the generative model for human grasp synthesis, we need ground truth 3D meshes of interacting hands and objects. Unfortunately, existing datasets often lack the desired properties.The limitations include small dataset size and lack of 3D ground truth hand pose or shape.

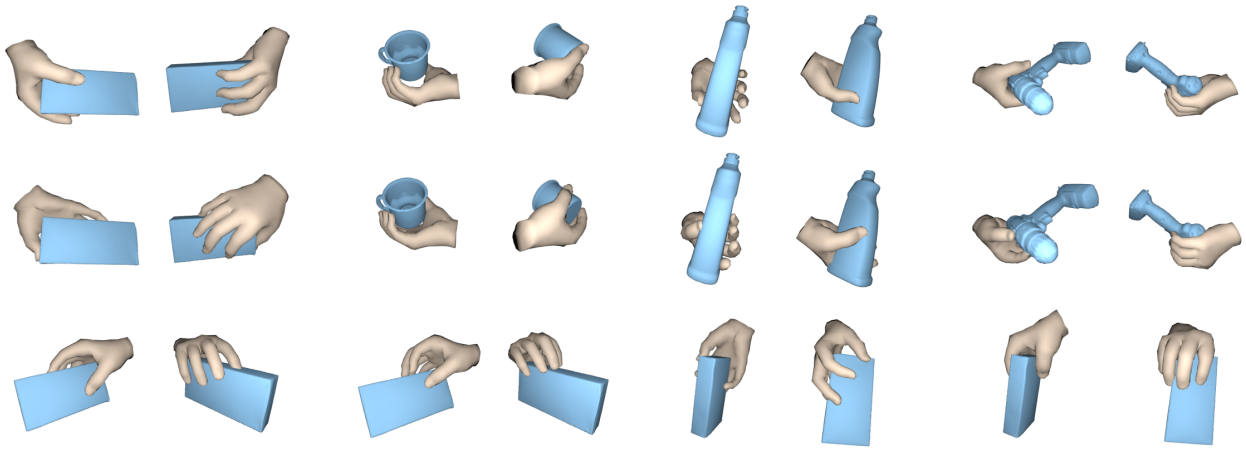


Figure 4: Generated grasps conditioned on objects from the HO3D dataset. Each pair shows the sampled grasp from two viewpoints. The model is trained only on the ObMan [29] dataset.

Consequently we use the synthetic **ObMan dataset** [29] to train our model. The data is generated from a statistical hand model, MANO [74], and 2772 object meshes covering 8 classes of everyday objects from the ShapeNet dataset [8]. Hand-object interaction is generated using a physics simulator, GraspIt [53], resulting in high-quality hand-object interaction. Due to the limited number of grasp types in the **FHB dataset** [19] and the **HO-3D dataset** [25], they are not suitable for training the generative model (see Appendix B). Instead, we use them to test the generalization ability of the generative model trained on the **ObMan** grasps.

For the 3D reconstruction task, we also mainly use the **ObMan dataset** for training and testing. To test the effectiveness of our network on real-world images, however, we follow the same approach as [28] to train and test on the **FHB dataset**.

Evaluation metrics. Our goal is to generate physically plausible and semantically meaningful 3D human hand given an object. Therefore, we quantitatively evaluate the generated samples according to physics-based metrics and use large-scale perceptual studies to measure the visual realism of the grasps. For the 3D reconstruction task, we use Chamfer distance and hand joint error. Details of the evaluation metrics are in Appendix C.

(1) Physical metrics: A valid human grasp implies stable hand-object contact without interpenetration. Consequently, we use the following evaluation metrics: ***a) Intersection volume and depth***. The hand and object mesh are voxelized and the interpenetration depth is the maximum distance from all the intersected voxels to the surface of another mesh. ***b) Ratio of samples with contact***. We define a *contact* between the object and the hand when any point on the surface of the hand is on or inside the surface of the object. We calculate the ratio of samples over the entire dataset that have interpenetration depth more than zero. ***c) Grasp stability***. Using physics simulation [12], we hold the hand constant, apply gravity, and measure the average displacement of the object's center of mass during a fixed time period.

(2) Semantic metric: We perform perceptual studies using Amazon Mechanical Turk to evaluate the naturalness of our generated grasps. Details of the study are presented in Appendix C and D.

(3) 3D reconstruction quality: We use the Chamfer distance between reconstructed and ground truth hand surfaces to evaluate the hand reconstruction quality as in [61]. Hand joint distance is computed following [29, 96].

4.1. Evaluation: Human grasps generation

Baseline. To our knowledge, there is no previous model that learns to synthesize natural human grasps given a 3D object. Rather than randomly placing the hand around the object, we trained a strong baseline model for grasp generation. Specifically, we replace the decoder (i.e. the grasping field) of our conditional VAE model (Fig. 2a) with fully connected layers to regress MANO hand parameters. Then given a 3D object point cloud and a random sample, our baseline model generates MANO parameters directly. Generated grasps from the baseline are shown in Appendix E.

Results. We show the systematic quantitative evaluation of the generative method in Tab. 1 and qualitative results in Fig. 1, 4 and Appendix E (Fig. E.3, E.4). The baseline and the GF model are only trained on the **ObMan** training set,

Table 1: Evaluation of the grasp synthesis on the objects from the ObMan test set, FHB and HO3D. GT* indicates that the ground truth grasps are obtained by fitting the MANO model to the data. Best results except the ground truth are shown in boldface.

	ObMan			**FHB**			**HO3D**		
	GT	Baseline	GF	GT*	Baseline	GF	GT	Baseline	GF
Contact ratio (%) ↑	-	66.89	**89.4**	92.2	48.8	**97.0**	93	44.3	**90.1**
Intersection vol. (cm^3) ↓	-	14.46	**6.05**	16.6	**9.65**	21.9	10.5	**5.86**	14.9
Intersection depth (cm) ↓	-	0.94	**0.56**	1.99	**1.77**	2.37	1.47	**1.01**	1.46
Physics simulation (cm) ↓	1.66	4.56	**2.07**	6.69	8.59	**4.62**	4.31	8.25	**3.45**
	± 2.42	± 4.57	± 2.81	± 5.48	± 3.67	± 4.48	± 4.42	± 4.18	± 3.92
Perceptual score {1...5} ↑	3.24	2.40	**3.02**	3.49	2.43	**3.33**	3.18	2.03	**3.29**

and tested extensively on the objects from the **ObMan** test set, **FHB** and **HO3D**. Our proposed GF performs substantially better than the baseline on **ObMan** and achieves comparable quality as the ground truth grasps. When the model is tested on the **FHB** objects, which are never seen during training, it achieves a comparable perceptual score compared to the ground truth grasps. Surprisingly, on **HO3D**, our synthesized grasps are judged more realistic than the ground truth grasps of real humans (3.29 vs 3.18). These perceptual studies suggest that our method makes an important step towards the fully automatic synthesis of realistic human grasps.

Regarding the physical plausibility, we observe that our model achieves a better contact ratio and grasp stability (physics simulation) than the ground truth grasps on **FHB** and **HO3D**. This is likely due to the GF results having a larger intersection volume. One reason is that there are a very limited number of objects in these two datasets. Some of the test objects are very different from the training objects, resulting in more inter-penetration for the generated grasps. Overall, the combination of visual realism and grasp stability suggests that our results are approaching the level of natural human grasps.

4.2. Evaluation: 3D hand-object reconstruction

Apart from serving as a powerful representation for the synthesis task, the proposed GF also facilitates the 3D reconstruction task. In the following, we analyze the different network architectures and training losses proposed in Sec. 3.3. We compare with the baseline method [29] on the **ObMan** dataset and [28] on the **FHB** dataset. The results are summarized in Tab. 2. Due to many limiting factors of the real-world datasets such as **FHB** and **HO3D** (see Appendix B for detailed data analysis), learning a reasonable model for joint object and hand reconstruction is extremely challenging. Instead, to evaluate the effectiveness of the GF representation for 3D reconstruction on real-world images, we follow the setting of the latest work [28], where the object 3D model is given as input. Note that this is a commonly used setting in previous works (e.g. [28, 34, 85]).

Network design. We first analyze the two different network architectures presented in Fig. 3 denoted with and without '2De' respectively. Both architectures achieve comparable performance for hand reconstruction, however differ significantly for intersection error, where the one decoder model achieves considerably better performance, due to the efficient joint modeling of hand-object interaction. Compared with the baseline [29], the intersection volume and depth are reduced from 6.25 and 1.20 to 0.65 and 0.32, respectively. The contact ratio are comparable among two architectures and baseline model. All our models considerably improve the quality of hand reconstruction, compared with [29]. The object reconstruction quality is behind hand quality for all model variations including the baseline model. Note that the **ObMan** dataset contains more than 1600 objects from 8 different classes. The object reconstruction performance is decreased as it remains unclear how to learn the implicit representations to reconstruct a large variety of object classes with a single model [51, 61] and such a task is beyond the focus of this work. Please see Appendix E (Fig. E.1) for visualization.

Training losses. The effect of the contact and interpenetration loss (+L) is shown in Tab. 2 (a), when the loss is imposed during the training of the two-decoder network, the intersection volume and depth are reduced and the overall quality of the interaction is considerably improved. In contrast, for the one-decoder model, our observation is that, for a large portion of 3D points, the signed distances to the object and to the hand are highly correlated, the model that jointly predicts both signed distance values does not need to enforce this auxiliary training loss.

MANO fitting. As shown in Tab. 2 (a), MANO fitting (indicated by GF-MANO) does not have a substantial influence on the reconstruction quality. This implies that on the one hand, the reconstructed hand of our GF model is realistic enough without a statistical model to regularize it, and that on the other hand, the output hand part labels are accurate

Table 2: 3D reconstruction results on **ObMan** (a) and **FHB** (b). 2De refers to the 2 decoder model; L indicates the corresponding model is trained with the contact and interpenetration loss; GF-MANO refers to the MANO hand obtained by fitting the MANO model to the SDF.

Models	Hand error		Joints	Intersection		Contact	Object Error	
	Mean	Med	(cm)	Vol	Depth	(%)	Mean	Med
GF	0.419	0.283	-	0.65	0.32	90.8	12.8	6.4
GF+L	0.400	0.261	-	0.00	0.00	5.63	14.2	6.8
GF-2De	0.408	0.262	-	8.56	1.01	99.6	11.1	5.7
GF-2De+L	0.384	0.237	-	0.23	0.20	69.6	11.7	5.9
GF-MANO	0.405	0.272	1.13	0.59	0.27	83.3	-	-
GF-2De-MANO	0.419	0.276	1.14	5.75	0.87	98.9	-	-
Hasson et al. [29]	0.533	0.415	1.13	6.25	1.20	94.8	6.7	3.6

(a)

Models	Hand Joints (cm)
GF-MANO (MANO joints)	2.60
GF-MANO (FHB marker)	2.94
Hasson et al.* [28]	2.74

(b)

enough for us to fit the MANO model and retrieve hand joints or shape parameters for applications that need these without undermining the shape and contact estimation. A qualitative illustration is presented in Fig. E.2.

Hand reconstruction on real-world images. To analyse the effectiveness of the proposed grasping field representation for the 3D reconstruction task on real-world images, we compare our method with the latest work [28] on the **FHB** dataset. Compared to [29], the key difference in [28] is that the object is given as part of the input. We explore the same network architecture as [28] and only replace the decoder part with the grasping field. The implementation details are presented in Appendix A (Fig. A.3).

As stated in [28], definition of hand joint locations vary between datasets. Without hand surface annotation in the FHB dataset, it is difficult to train an accurate regressor that maps between the FHB markers and the MANO joints. Assuming that the joints are identically defined, we fit the MANO model to the FHB markers by minimizing the distance between the MANO joints and the FHB markers. Then the MANO joints obtained in such way are considered as our pseudo ground truth joints, and the obtained MANO surface is used to supervise the training.

We compare the predicted MANO joints with the pseudo ground truth joints as well as the original FHB markers assuming identical joints. As our model is not trained to optimise for the FHB marker locations, the reconstruction error is larger than [28] as shown in Tab. 2 (b). When we evaluate our prediction on the pseudo ground truth MANO joints, the reconstruction error decreases from $2.94cm$ to $2.6cm$. This suggests that the proposed grasping field representation is effective for the task of 3D hand reconstruction from a single image, achieving comparable performance with respect to the start-of-the-art.

5. Conclusion and Discussion

In this work, we propose a novel representation for hand-object interaction, namely the grasping field. Learning from data, the GF captures the critical interactions between hand and object by modeling the joint distribution of hand and object shape in a common framework. To verify the effectiveness, we address two challenging tasks: human grasp generation given a 3D object and shape reconstruction given a single RGB image. The experiments show that the generated hand grasps appear natural and are physically plausible while the hand reconstruction achieves comparable performance as the state-of-the-art.

A limitation of our work is that there is no explicit modeling of the object functionality and human action in the current grasping field representation. In reality, a person holds an object differently based on different intentions. For instance, using a knife or passing it to someone else result in completely different human grasps. One promising future research direction is to incorporate human intention and object affordances into the grasping field for action specific grasps generation. Furthermore, we believe the proposed grasping field representation opens up avenues for several other future research directions. For instance, 3D human hand generation given only an object image and synthesizing the motion of hand-object interactions.

Acknowledgement. We sincerely acknowledge Lars Mescheder and Michael Niemeyer for the detailed discussions on implicit function, Dimitrios Tzionas, Omid Taheri, and Yana Hasson for insightful discussions on hand interaction, Partha Ghosh and Qianli Ma for the help with VAE. **Disclosure.** MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at MPI. He is an investor in Meshcapde GmbH.

References

[1] D. Antotsiou, G. Garcia-Hernando, and T. Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[2] S. Baek, K. I. Kim, and T.-K. Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2020. 3

[3] J. Bohg, A. Morales, T. Asfour, and D. Kragic. Data-driven grasp synthesis—a survey. *Trans. Rob.*, 30(2):289–309, Apr. 2014. 3

[4] C. W. Borst and A. P. Indugula. Realistic virtual grasping. In *IEEE Proceedings. VR 2005. Virtual Reality, 2005.*, pages 91–98, 2005. 2

[5] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 2

[6] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox. ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019. 2

[7] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, August 2020. 2

[8] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[9] Z. Chen, A. Tagliasacchi, and H. Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[10] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[11] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 3

[12] E. Coumans et al. Bullet physics library. *Open source: bulletphysics. org*, 15(49):5, 2013. 6, 5

[13] De-An Huang, Minghuang Ma, Wei-Chiu Ma, and K. M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *CVPR*, 2015. 2

[14] E. Dibra, S. Melchior, T. Wolf, A. Balkis, A. C. Öztireli, and M. H. Gross. Monocular RGB hand pose inference from unsupervised refinable nets. In *CVPR Workshops*, 2018. 3

[15] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[16] G. ElKoura and K. Singh. Handrix: Animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, page 110–119, Goslar, DEU, 2003. Eurographics Association. 2

[17] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 5, 1

[18] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015. 3

[19] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 6, 2

[20] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[21] K. Genova, F. Cole, D. Vlasic, A. Sarna, W. T. Freeman, and T. Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7154–7164, 2019. 3

[22] O. Glauser, S. Wu, D. Panozzo, O. Hilliges, and O. Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 38(4), 2019. 2

[23] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry. AtlasNet: A papier-mâché approach to learning 3D surface generation. In *CVPR*, 2018. 3

[24] H. Hamer, K. Schindler, E. Koller-Meier, and L. V. Gool. Tracking a hand manipulating an object. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1475–1482, Sep. 2009. 3

[25] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3196–3206, 2020. 3, 6, 2

[26] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, Oct. 2019. 3, 5

[27] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 571–580, 2020. 3

[28] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 6, 7, 8

[29] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 4, 6, 7, 8, 5

[30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[31] G. Heumer, H. B. Amor, M. Weber, and B. Jung. Grasp recognition with uncalibrated data gloves - a comparison of classification methods. In *2007 IEEE Virtual Reality Conference*, 2007. 2

[32] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, and J. Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[33] H. Kato, Y. Ushiku, and T. Harada. Neural 3D mesh renderer. In *CVPR*, 2018. 3

[34] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017. 7

[35] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, 2012. 3

[36] J. Kim and J. Park. Physics-based hand interaction with virtual objects. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3814–3819, 2015. 2

[37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2014. 1

[38] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[39] R. Krug, D. Dimitrov, K. Charusta, and B. Iliev. On the efficient computation of independent contact regions for force closure grasps. pages 586 – 591, 11 2010. 2

[40] P. G. Kry and D. K. Pai. Interaction capture and synthesis. *ACM Trans. Graph.*, 25(3):872–880, July 2006. 2

[41] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4990–5000, 2020. 3

[42] M. Lau, K. Dev, W. Shi, J. Dorsey, and H. Rushmeier. Tactile mesh saliency. *ACM Trans. Graph.*, 35(4), 2016. 2

[43] Y. Li, J. L. Fu, and N. S. Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):732–747, 2007. 2

[44] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas. Fpnn: Field probing neural networks for 3d data. In *Advances in Neural Information Processing Systems*, pages 307–315, 2016. 3

[45] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, page 163–169, New York, NY, USA, 1987. Association for Computing Machinery. 5

[46] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg. Dex-net 3.0: Computing robust robot suction grasp targets in point clouds using a new analytic model and deep learning. 09 2017. 3

[47] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26), 2019. 2, 3

[48] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker. Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[49] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Héloir, and D. Stricker. DeepHPS: End-to-end estimation of 3D hand pose and shape by learning from synthetic depth. In *3DV*, 2018. 3

[50] D. Maturana and S. Scherer. VoxNet: A 3D convolutional neural network for real-time object recognition. In *IROS*, 2015. 3

[51] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2, 3, 7, 1

[52] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson. Deep Level Sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802*, 2019. 3

[53] A. T. Miller and P. K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 6

[54] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[55] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *ICCV*, 2017. 3

[56] Y. Nakamura, D. Troniak, A. Rodriguez, M. Mason, and N. Pollard. The complexities of grasping in the wild. pages 233–240, 11 2017. 2

[57] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[58] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger. Texture fields: Learning texture representations in function space. In *Proceedings IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[59] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, page 2088–2095, USA, 2011. IEEE Computer Society. 2

[60] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445, March 2018. 3

[61] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2, 3, 5, 6, 7, 1

[62] T. Pham, N. Kyriazis, A. A. Argyros, and A. Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 3

[63] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In D. Kragic, A. Bicchi, and A. D. Luca, editors, *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 3406–3413. IEEE, 2016. 2

[64] S. Pirk, V. Krs, K. Hu, S. D. Rajasekaran, H. Kang, Y. Yoshiyasu, B. Benes, and L. J. Guibas. Understanding and exploiting object interaction landscapes. *ACM Transactions on Graphics (TOG)*, 36(3):1–14, 2017. 2

[65] N. S. Pollard and V. B. Zordan. Physically based grasping control from example. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '05, page 311–318, New York, NY, USA, 2005. Association for Computing Machinery. 2

[66] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[67] S. Puhlmann, F. Heinemann, O. Brock, and M. Maertens. A compact representation of human single-object grasping. In *IROS*, 2016. 2

[68] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[69] J. Redmon and A. Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1316–1322, 2015. 2

[70] H. Rijpkema and M. Girard. Computer animation of knowledge-based human grasping. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '91, page 339–348, New York, NY, USA, 1991. Association for Computing Machinery. 2

[71] G. Rogez, M. Khademi, J. S. Supančič III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *ECCV Workshop on Consumer Depth Cameras for Computer Vision*, 2014. 3

[72] G. Rogez, J. S. Supančič III, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015. 3

[73] J. Romero, H. Kjellström, and D. Kragic. Monocular real-time 3D articulated hand pose estimation. In *IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS)*, pages 87–92, 2009. 3

[74] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3, 4, 5, 6

[75] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[76] T. Schmidt, K. Hertkorn, R. Newcombe, Z. Marton, M. Suppa, and D. Fox. Depth-based tracking with physical constraints for robot manipulation. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 119–126, 2015. 2

[77] J. Seo, S. Kim, and V. Kumar. Planar, bimanual, whole-arm grasping. In *2012 IEEE International Conference on Robotics and Automation*, pages 3271–3277, 2012. 2

[78] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[79] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[80] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 3

[81] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *ECCV*, 2016. 2

[82] H. Su, H. Fan, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 3

[83] J. S. Supančič III, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *ICCV*, 2015. 3

[84] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[85] B. Tekin, F. Bogo, and M. Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[86] A. Tsoli and A. Argyros. Joint 3D tracking of a deformable object in interaction with a hand. In *ECCV*, 2018. 3

[87] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 3

[88] D. Tzionas and J. Gall. 3d object reconstruction from hand-object interactions. In *ICCV*, 2015. 3

[89] H. Wang, S. Pirk, E. Yumer, V. G. Kim, O. Sener, S. Sridhar, and L. J. Guibas. Learning a generative model for multi-step human-object interactions from videos. In *Computer*

Graphics Forum, volume 38, pages 367–378. Wiley Online Library, 2019. 3

[90] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *ECCV*, 2018. 3

[91] J. Wu, Y. Wang, T. Xue, X. Sun, W. T. Freeman, and J. B. Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *NIPS*, 2017. 3

[92] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *CVPR*, 2015. 2

[93] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, October 2019. 3

[94] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. Tang. Generating 3d people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 5

[95] W. Zhao, J. Zhang, J. Min, and J. Chai. Robust realtime physics-based motion control for human grasping. *ACM Trans. Graph.*, 32(6), Nov. 2013. 2

[96] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4903–4911, 2017. 3, 6, 5

[97] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. 3

Grasping Field:
Learning Implicit Representations for Human Grasps
Appendix

A. Implementation Details

In Sec. 3.2 and Sec. 3.3, we present the neural networks that are used for human grasps generation and reconstruction, respectively. Here we discuss the implementation details.

A.1. Architecture

In this section, we explain the network architectures used in our experiments. The same decoder architecture is used in our image reconstruction and the hand generation tasks. We change the encoder architectures according to the input type. In our experiments, both the encoder and decoder are jointly trained end-to-end. Figure A.1 illustrates the one-branch decoder with 8 fully-connected layers used in all tasks.

For image reconstruction, we use the ResNet18 [30] model pretrained on the ImageNet dataset [75] as an encoder. We change the last layer of the encoder to produce a latent vector of size 256 for the decoder.

For the point cloud input, we use two separated PointNet encoders [17] with additional pooling and expansion layers presented in [51]. In each encoder, 3D points are first mapped to 512-dimension feature vectors followed by 5 ResNet-blocks, producing a latent vector of size 256. The latent codes for hand and object are then concatenated to make a 512-dimension latent code.

For the hand generation task, we change the first layer in the hand encoder to produce a 256-dimension vector for each point then concatenate it with the 256-dimension object latent vector. Figure A.2 shows the details of the point cloud model.

For image reconstruction with known objects, we assume that the object mesh in the normalized pose is given. We sample surface points from the given object and use a PointNet encoder to compute object latent vector of size 128. The object latent vector is then concatenated with a hand latent vector of size 128 from ResNet18 encoder, producing a latent code of size 256 for the decoder. The overview of the network is shown in Figure A.3.

A.2. Data preparation

To prepare the sampled 3D points and their distances to the hand and object surfaces for training, we follow the point sampling method provided by [61]: For each pair of the hand and object meshes, we translate both meshes such that the hand root joint is at the origin then scale them to fit in a unit cube. The scaling factor is the same for the entire dataset to ensure the hands are normalized across dataset. After that, 40,000 points are sampled in a unit cube. Following [61], 95% of the total points were sampled near the surface to capture the details of both meshes. For the Chamfer distance calculation, we sample 30,000 points from the surface of the ground truth mesh and reconstructed mesh following [61]. In case the reconstructed mesh contains more that one connected component, only the largest watertight connected component is retained.

A.3. Training

The contact loss is disabled in the beginning. When computing the reconstruction loss $\mathcal{L}_{rec}$, hand points to the object surface, and object points to the hand surface, are not considered until the contact loss is enabled. In our trials, we observe dramatic degradation when such a mask is not used or when the contact loss is enabled in the beginning.

For the generative GF network conditioned on an object point cloud, the KL loss, $\mathcal{L}_{kl}$, is employed in an annealing scheme; the loss weight is kept at 0 in the first 200 epochs and then linearly increased to 0.1 over the next 200 epochs. We find that such a annealing scheme is essential in our trials. Applying the KL loss in the beginning causes our generative network posterior to collapse.

In all experiments, we use Adam optimizer [37] with learning rate of 10^{-4} and decay it to 5×10^{-5} at after 600 epochs. We train the models for 1,200 epochs without hand-part classification loss and another 100 epochs with the classification loss. Weight decay is used in all layers in the decoder.

A.4. Inference

During inference, we use Marching Cube with resolution 128 to obtain hand and object meshes. As the object can vary in size, we use a two-stage approach to dynamically scale the cube size in the Marching Cube algorithm. First, to find the boundary of the reconstructed meshes, we query equally space points in a unit cube centered at the origin point to locate the

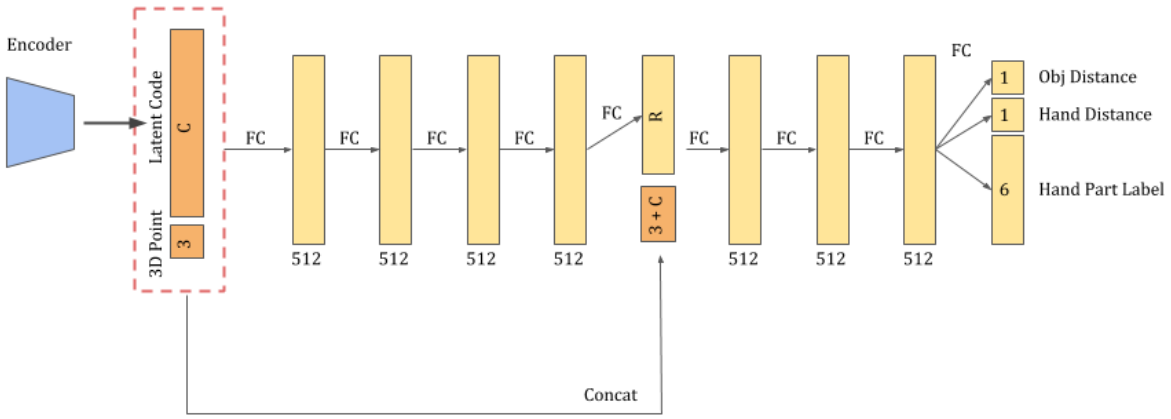


Figure A.1: Decoder architecture. The fully-connected layers are denoted as "FC" in the diagram. The latent vector C has 256 and 512 dimensions in the image reconstruction and conditional hand mesh generation task, respectively. The latent code from the encoder is concatenated with 3D point query then given to the decoder. The same latent vector is concatenated again at the middle of the decoder following [61], with the concatenated vector R having size 253 and 509, respectively. Every "FC" layer except the last layer is followed by a ReLu activation and a dropout layer with drop rate 0.2. The last layer produces the distance to object surface and the distance to hand surface along with hand part classification scores

negative signed-distance values which indicate the inside of the mesh. Then, we query again with a cube that covers every negative-value point. Using this approach, no mesh is produced if no negative point is found in the first stage.

B. Dataset Analysis

In this section, we provide detailed analyses on the **FHB** [19] and the **HO-3D** [25] datasets. Although these datasets considerably contribute to the studies of hand-object interactions with detailed 3D annotation, our analyse shows that they might not be suitable for learning human grasps and modelling the accurate contact relation between hand and object. First, the number of objects and the types of grasps are limited. As shown in Tab. 5, the number of object is 3 in the **FHB** dataset and 10 in the **HO3D** dataset. Second, the (pseudo) ground truth meshes of the interacting hand and object exhibit frequent interpenetration. Sampled ground truth meshes from the HO-3D dataset are illustrated in Fig. B.1.

We evaluate the interpenetration between hand and object meshes quantitatively. We use the same evaluation metric as the one presented in the main paper, namely, the intersection volume (cm^3) and depth (cm) (Sec. 4). The results are shown in Table 5.

For the HO-3D dataset, 91.94% of the training examples exhibit hand-object contact. However, among these training examples, the average intersection volume and depth are 10.91 cm^3 and 1,56 cm respectively.

For the FHB dataset, we use the similar subset as the previous work [29], namely, we exclude the milk bottle related examples and the examples where the distance from hand joints to the object mesh is more than 1cm. We refer to this dataset as FHB_c. We further fit the MANO hand model with the provided joint location. For FHB_c, 97.1% of the training examples have hand-object contact. However, similar level of intersection between hand and object meshes can be observed in Table. 5.

Overall, the evaluation shows considerable intersection volume and depth of the training data. Therefore we use the ObMan dataset as our main training dataset, where the ground truth quality of the contact regions is more suitable for

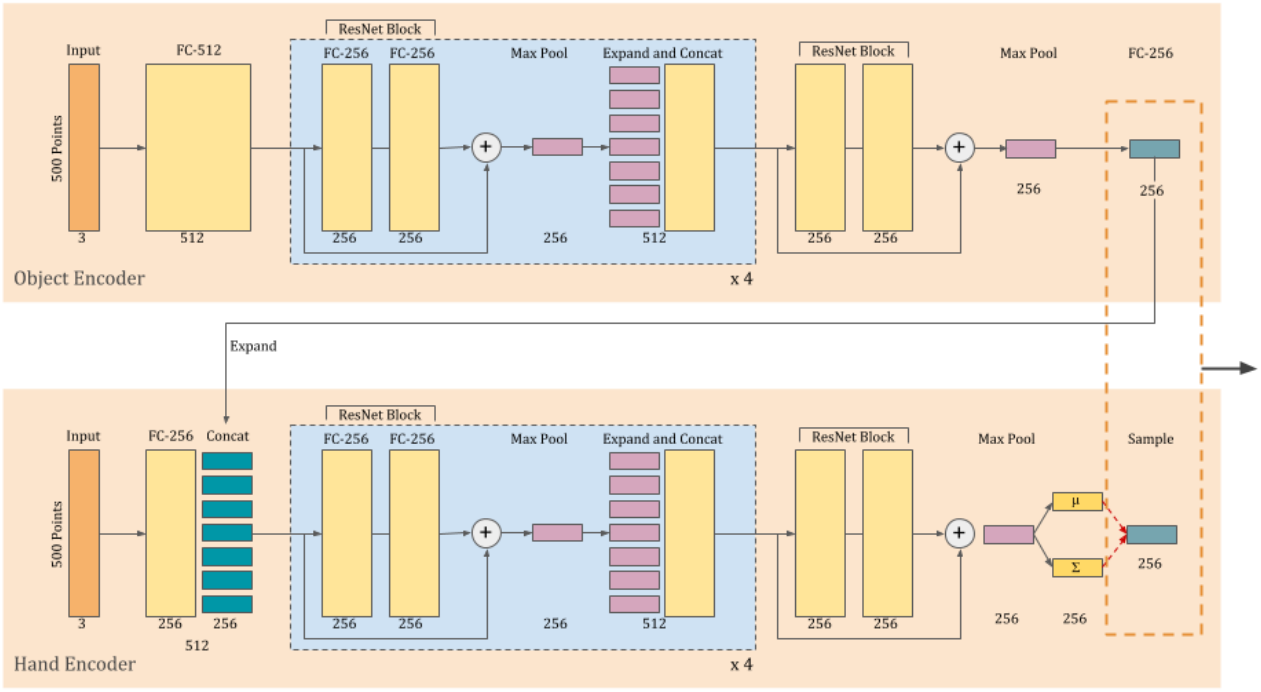


Figure A.2: The encoder architecture for the conditional VAE. Each box represents a vector obtained from applying the layer written above. For the object encoder, we use the same architecture as the model for point cloud completion used in [51]. The hand encoder is conditioned on the latent code from the object decoder. The combined latent codes of hand and object are then concatenated with a 3D point and passed to the decoder

learning physically plausible human grasps.

Furthermore, as shown in the experiment section (Tab. 1), our generated grasps that are learned from the ObMan dataset obtain a higher perceptual score than the ground truth grasps from the HO3D data in the perceptual study, suggesting that the physical plausibility, i.e. no interpenetration and proper contact, plays an important role on the naturalness of human grasps.

C. Details of the evaluation metrics

Evaluation Metrics. For human grasps synthesis, our goal is to generate physically plausible and semantically meaningful 3D human hand given an object. Therefore, we propose to quantitatively evaluate the generated samples using physics metrics and a large-scale perceptual study to measure the perceptual fidelity. In addition, for the quantitative evaluation of our reconstruction networks, we use Chamfer distance and hand joint error.

(1) Physical metric: A valid human grasp implies hand-object contact without interpenetration. Naturally we propose

Table 5: Characteristics of the FHB_c and HO-3D dataset. The intersection volume and depth are calculated from the training sets and are considered when there is contact between hand mesh and object mesh. For the FHB_c dataset, the evaluation is done on the pseudo-ground truth meshes.

Dataset	# of frames (train/test)	# of objects	Intersection	
			Vol(cm^3)	Depth(cm)
FHB_c	5082 / 5658	3	10.59	2.34
HO-3D	66034 / 11524	10	10.91	1.56

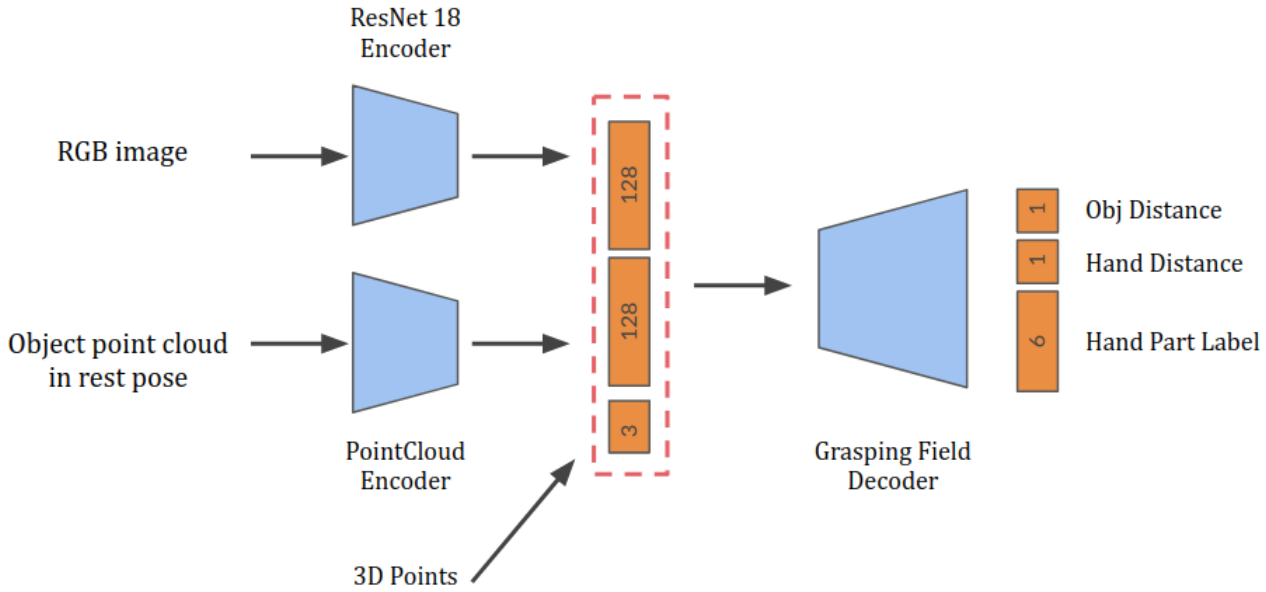


Figure A.3: The architecture of the GF conditioned on the image, given a known object.

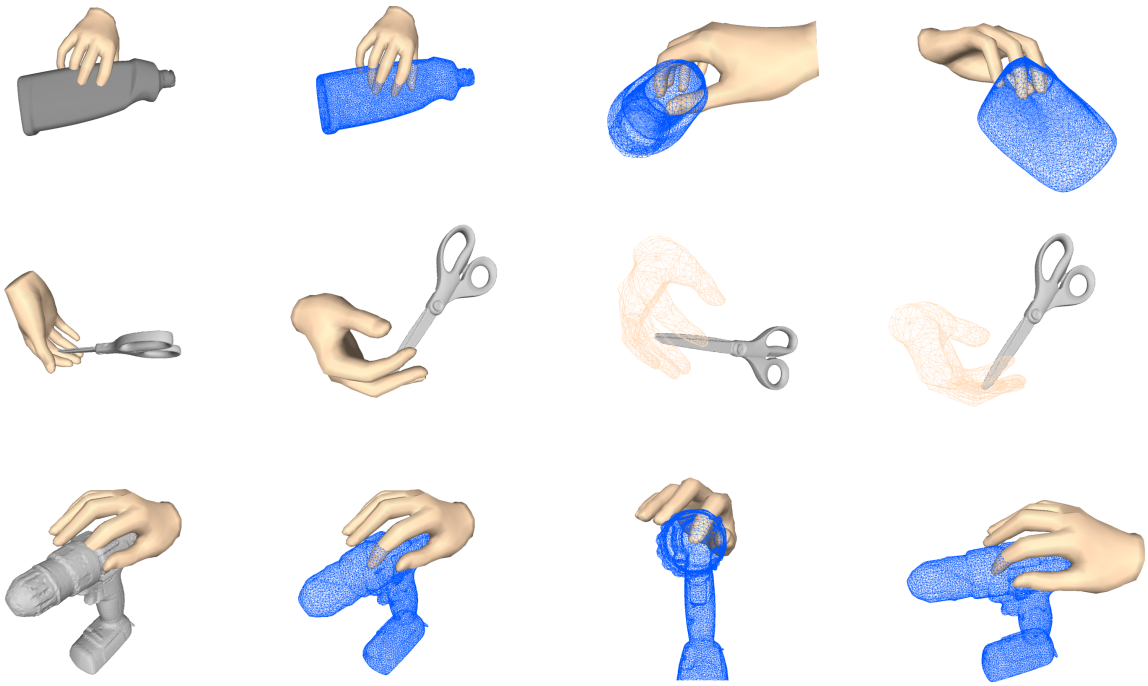


Figure B.1: Ground truth meshes from the HO-3D dataset

with the following evluation metrics:

Intersection volume and depth. We follow [29] to report intersection volume and depth. The hand and object mesh are voxelized using a voxel with edge length of 0.5cm. The interpenetration depth is the maximum distance from all points on

the interpenetrated surface to another surface. If the meshes do not overlap, the interpenetration depth is defined as 0.

Ratio of samples with contact. We define a *contact* between object and hand when any point on the surface of hand is on or inside the surface of the object. To measure the performance of models on hand-object contact quality, we calculate the ratio of samples over the entire dataset that have interpenetration depth more than zero. As all of the samples in the dataset should have contact between hand and object, the best ratio of frames with contact is 100%. A good hand and object reconstruction model should have high ratio of contact and small interpenetration volume and depth.

Simulation displacement. Following [29], we use physics simulation to evaluate the stability of the grasps. In the simulated environment [12], we fix the hand and measure the average displacement of the mass center of the object in a give time period. Small displacement suggests a stable grasp.

(2) Semantic metric: We perform perceptual studies on Amazon Mechanical Turk to evaluate the authenticity of our generated grasps. For each randomly generated sample, we render images from 6 different views, and request participants to score from 1 (low fidelity) to 5 (high fidelity).

(3) 3D reconstruction quality: We use the Chamfer distance between reconstructed and ground truth hand surfaces to evaluate the hand reconstruction quality. Surface distance is approximated by mean square point cloud Chamfer distance (cm^2) as implemented in [61]. The MANO wrist is sealed to form a watertight mesh for fair comparison. Joins distance is computed following [29, 96]. After MANO parameters are recovered from the predicted hand mesh as described in Sec. 3.4, we compute mean Euclidean distance over 21 joints following [29, 96]. Note, since scale and global translation can not be determined by a single image, for each predicted hand, we optimize the scale and global translation to match the ground truth by minimizing the Chamfer distance between them. Similarly to hand, we also use Chamfer distance as measurement of object surface quality. The predicted object mesh is transformed according to the corresponding predicted hand transformation estimated from the above to align with the ground truth object mesh.

D. Details of the perceptual study

Figure D.1 shows the user interface for evaluating the generated grasps on the Amazon Mechanical Turk (AMT). Users are asked to rate the plausibility of the hand-object interactions individually. Each entry consists of images from six different views and is rated by three different users.

E. Qualitative results

Figure E.5 shows the generated grasps from our baseline VAE model conditioned on the object surface point cloud. The MANO parameters are directly predicted by the decoder.

Figure E.1 shows the reconstruction results on the test images of the ObMan dataset [29]. We observe that our model can recover hand meshes with proper interaction with the object.

Figure E.2 shows the comparison between reconstructed mesh before and after MANO fitting. The hand meshes also come from the single image reconstruction task. We observe that the MANO fitted meshes match the inferred meshes, even in the case where the rasterized hand mesh has merged fingers.

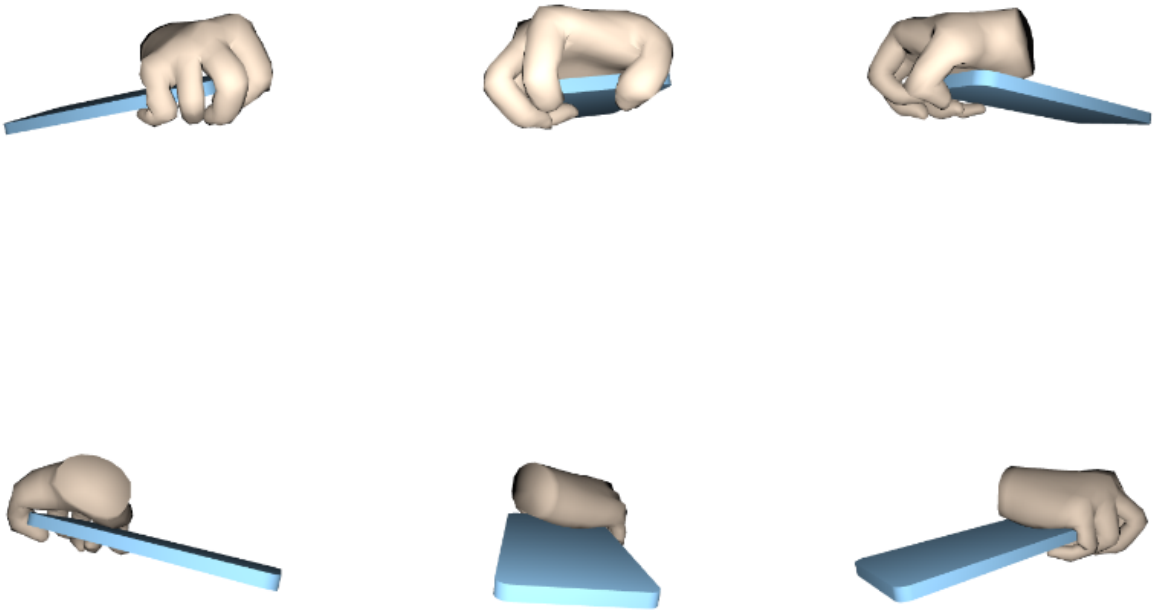
Figure E.3 shows **randomly sampled** grasps from our VAE model conditioned on the object surface point cloud. We observe that our model can generate a variety of grasps given an object. Figure E.4 shows the generated results of the same model conditioned on the objects from HO3D datset. It should be note that this model is only trained on the ObMan dataset and have never seen these objects before.

Hand-Object Interaction

Claim: The human hand is interacting very naturally with the object. What is your opinion?

`1.Strongly disagree  2.Disagree  3.Neither agree nor disagree  4.Agree  5. Strongly agree`

The same grasping is visualized from 6 different views. Use the slider at the bottom to give your score and submit.



Your Rating:

You must ACCEPT the HIT before you can submit the results.

Figure D.1: Amazon Mechanical Turk user interface for grasping evaluation

Figure E.1: Reconstruction results from RGB images. From left to right: input images, recovered mesh from two different view points, MANO fitted hand prediction, ground truth hand and object.

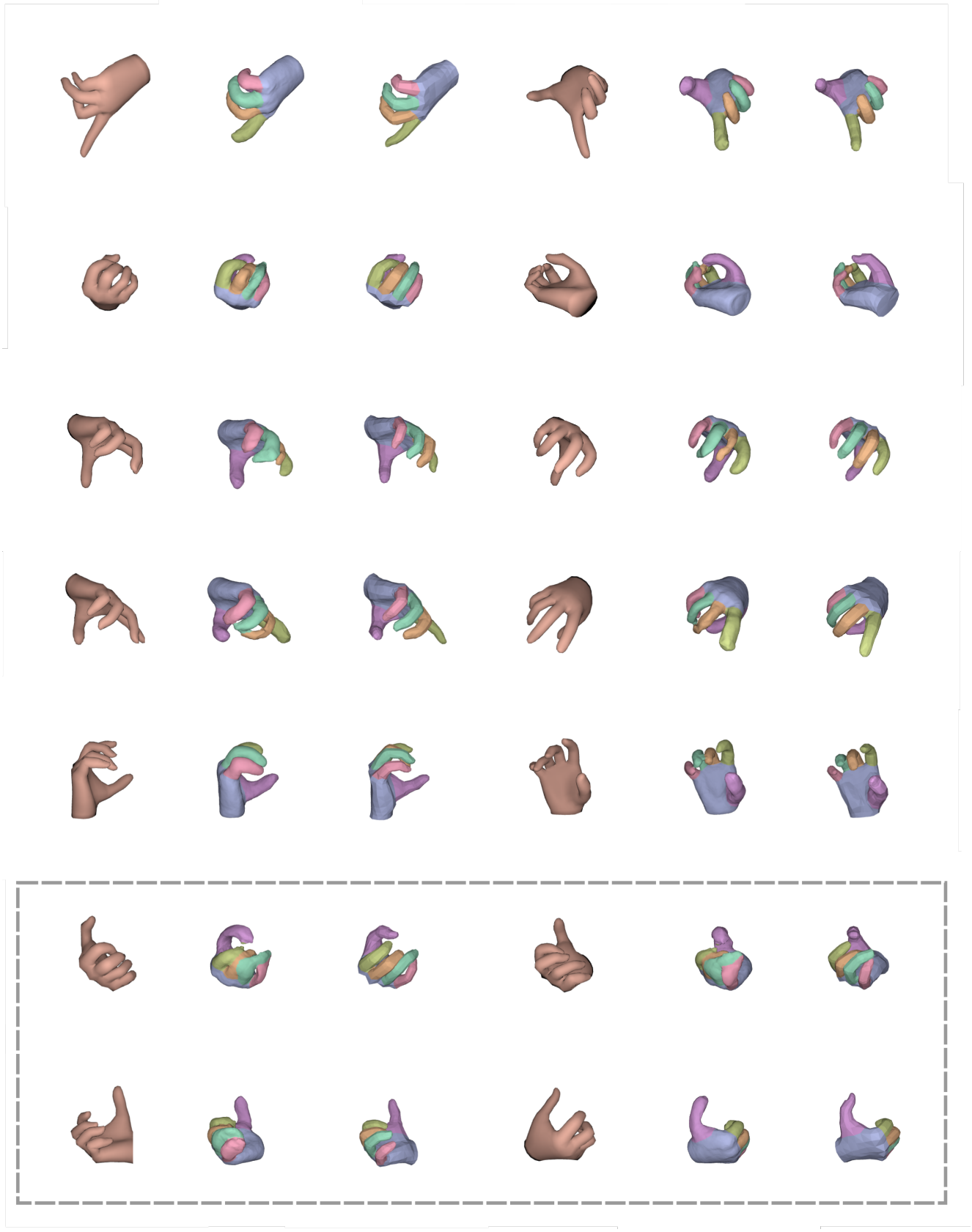


Figure E.2: MANO fitting results on the reconstructions from RGB images. Each row shows a set of ground truth hand mesh, reconstructed hand, and MANO fitted prediction, from two different views. The last two rows demonstrate the robustness of our fitting method where the reconstructed meshes from the estimated SDF values are less satisfactory. However, even with merged fingers, we can still recover reasonable hand mesh.

Figure E.3: **Randomly selected** hands generated from the conditional VAE. Each row shows hand and object ground truth followed by three sets of sampled hand meshes, before and after MANO fitting, all from the same view. The samples presented on the two bottom rows are less satisfactory as the generated SDFs have artifacts and we can observe interpenetration between the fitted MANO hand meshes and the object meshes.

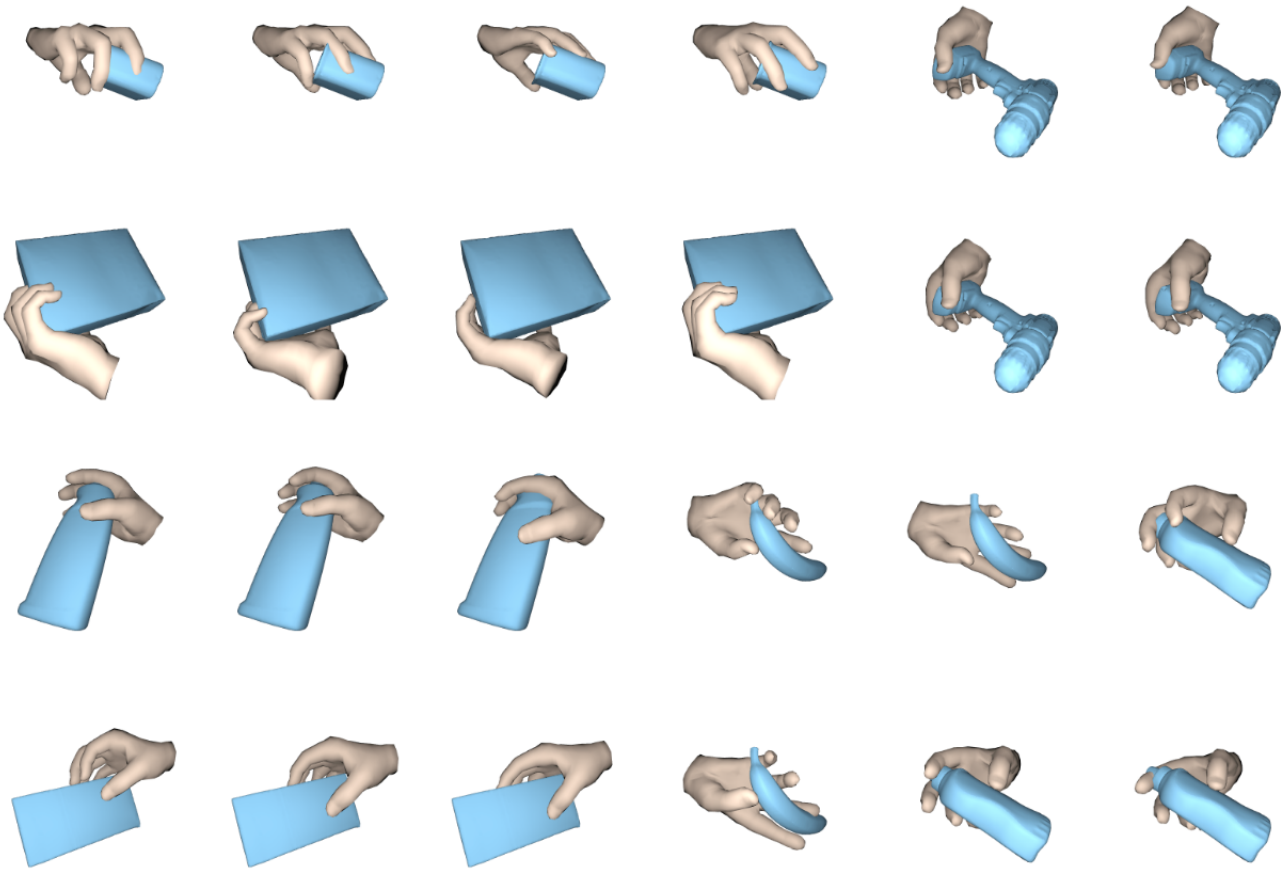


Figure E.4: Hands generated from the VAE conditioned on objects from HO3D dataset. The model is trained only on the ObMan dataset.

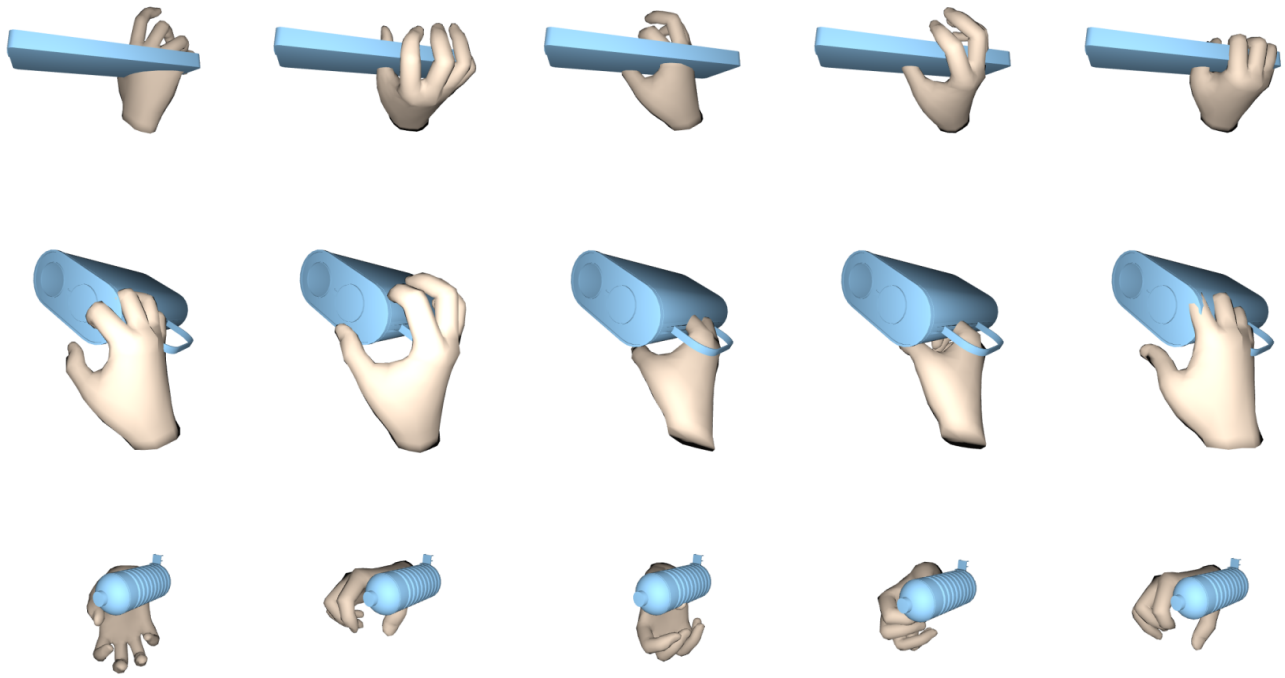


Figure E.5: Each row shows five randomly sampled hands given an object from the baseline conditional VAE. We can observe interpenetration between the hand meshes and the object meshes. In some cases, the hands are not in contact with the objects.