ContactGen: Generative Contact Modeling for Grasp Generation

Shaowei Liu[1†] Yang Zhou[2] Jimei Yang[2] Saurabh Gupta[1*] Shenlong Wang[1*]
[1]University of Illinois Urbana-Champaign [2]Adobe Research
https://stevenlsw.github.io/contactgen/

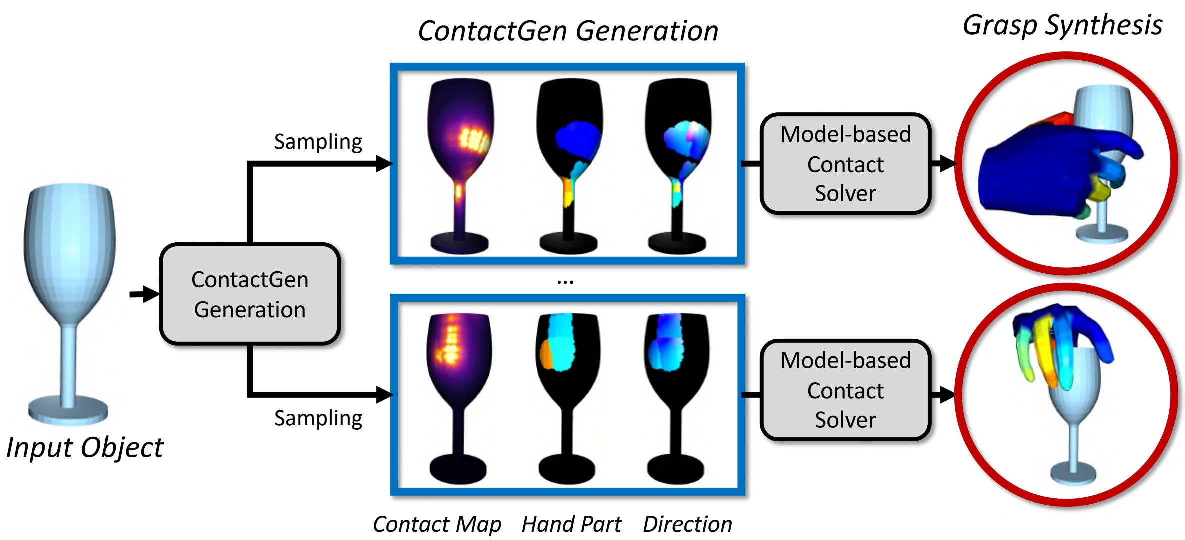

Figure 1: We introduce novel, concise and comprehensive contact representation for hand-object interaction (indicated by the dashed box). We demonstrate the process of generating contact representations from a given input object (on the left) and inferring the underlying grasp from the contact representation via model-based optimization (on the right).

Abstract

This paper presents a novel object-centric contact representation ContactGen for hand-object interaction. The ContactGen comprises 3 components: a contact map indicates the contact location, a part map represents the contact hand part, and a direction map tells the contact direction within each part. Given an input object, we propose a conditional generative model to predict ContactGen and adopt model-based optimization to predict diverse and geometrically feasible grasps. Experimental results demonstrate our method can generate high-fidelity and diverse human grasps for various objects.

1. Introduction

Modeling hand-object interaction [2, 9, 14, 21, 30, 37, 45, 57, 67] has gained substantial importance across various domains in animation, games, and augmented and virtual reality [23, 26, 62, 64]. For instance, given an object, one would like to create a computational model to reason about the different ways a human hand can interact with it, e.g., how to grasp the object using a single hand. To ensure realism and authenticity in these interactions, a precise understanding of contacts is crucial [5–7]. A thorough contact modeling should account for factors such as which regions of the object are likely to make contact, which parts of the hand will touch the object, the strength of the contact force, and the direction of the contact, among others. In contrast, the lack of thorough and precise modeling can result in unnatural and unrealistic interactions, such as insufficient contact or excessive penetration.

Previous approaches often rely on a contact map [17, 25, 27, 44, 56] applied to object point clouds, where values are bounded within the $[0, 1]$ range to indicate the status of point contacts. Nevertheless, simply modeling contact maps does not fully capture the details of contact. Specifically, even with the contact map, ambiguities remain regarding which

[†]Work started at an internship at Adobe Research.
[*]Equal advising, alphabetic order

regions of the hand are in contact and the manner of contact. Moreover, a single contact map falls short of representing the structured uncertainty inherent in hand-object interactions.

In this paper, we address the aforementioned challenges by introducing a novel contact representation called ContactGen. ContactGen provides a comprehensive presentation that encodes the specific contact parts of both the object and hand, along with the precise touch direction. Specifically, for each point on the object's surface, ContactGen models: (1) the contact region on the object's surface, represented as a contact probability; (2) the specific part of the hand making contact, be it various fingertip regions or the palm, in the form of a categorical probability; and (3) the orientation of the touch with respect to the hand part making contact, represented as a spherical coordinate. Fig. 1 depicts the three components of the presented ContactGen. Our approach significantly expands the traditional contact map, offering a precise and unambiguous representation of hand-object interactions.

Next, we introduce a novel, generative method that learns to produce diverse yet realistic ContactGen for any given object during inference. To account for uncertainties, we implement a hierarchical conditional Variational Autoencoder (CVAE) [54]. This CVAE sequentially models the contact map, the hand part map, and the contact coordinate. When provided with a 3D object as a conditional input, our CVAE initially models the probabilities of contact maps. From this, one can sample contact maps, using these samples as additional conditioning variables, to infer the distribution of the hand part map and sample from the distribution. Lastly, direction maps can be generated based on the sampled contact map and hand part map. This sequential generation strategy separates variations within the entire space into distinct components, ensuring explicit uncertainty modeling for each component.

The proposed ContactGen is applied to human grasp synthesis [28, 34, 35, 46], whose objective is to generate a diverse physically plausible human grasps for various objects. In contrast to existing work [10, 27, 29, 30, 56], which primarily addresses grasp uncertainty within the hand space, our key innovation lies in addressing this uncertainty within the object space. We achieve this by designing a novel contact solver that effectively derives hand grasp poses from ContactGen. The ContactGen is sampled from our CVAE model and linked to the specific object. As a result, our design yields more realistic and organic hand grasps, as shown by improved contact, diminished penetration, and increased stability. Additionally, the hierarchical contact modeling fosters greater diversity in the generated grasps. Our experiments validate the efficacy of our method in ensuring both diversity and fidelity in hand-object interactions, surpassing the performance of current state-of-the-art techniques.

Method	**Hand Model**	**Contact Modeling**			**Object-centric**
		location	**part**	**direction**	
Grasping Field [30]	point cloud	✓	✓	✗	✗
GraspTTA [27]	mesh	✓	✗	✗	✗
ContactOpt [17]	mesh	✓	✗	✗	✗
TOCH [73]	mesh	✓	✓	✗	✓
Ours	mesh/sdf	✓	✓	✓	✓

Table 1: Contact representation comparison between different methods of hand-object interaction. Most existing work adopt contact map, which is insufficient to recover the underlying grasp.

In summary, our contributions are as follows:

- We introduce ContactGen, a novel object-centric representation that concurrently models the contact parts of both the hand and the object, as well as the contact direction relative to each part.
- We propose a sequential CVAE model that learns to grasp the uncertainty inherent in hand-object interactions using our contact representation.
- We develop a novel human grasp synthesis algorithm, merging our suggested generative modeling with model-based optimization. This combination produces enhanced fidelity and diversity compared to current methods.

2. Related Work

Contact Modeling Various contact representations have been proposed in hand-object interaction [17, 18, 27, 34, 36, 49, 67, 68], and human-object interactions [3, 8, 16, 20, 25, 52, 66, 70, 71]. As shown in Tab. 1, existing studies [17, 27, 55, 59, 65] primarily adopt the contact map as a standard representation. Zhou *et al.* [73] proposed an object-centric TOCH field by encoding contact locations and hand correspondences on the object aimed for temporal hand pose denoising. Nonetheless, we argue the information provided by contact maps [17] or sparse hand-object correspondences on contact locations [73] falls short. Contact maps lack information about their counterparts, while sparse correspondences cannot provide detailed contact directions. In our approach, we not only infer the contact location, but also the hand part in contact and the local direction of that part. This novel representation is point-wise and object-centric, which does not require any input from the hand, but enables the comprehensive decoding of hand information.

Grasp Synthesis Grasp synthesis has gained extensive attention across both robot hand manipulation [1, 6, 24, 39, 41], animation [4, 15, 28, 49], digital human synthesis [34, 35, 51, 69], and physical motion control [22, 31, 46]. In this work, we focus on realistic human grasp synthesis [10, 27, 29, 30, 56]. The objective is to generate authentic human grasps of diverse objects. The key challenge is to achieve both physical plausibility and diversity within the generated grasps. A majority of existing approaches employ CVAE to sample hand MANO parameters [27, 55, 56, 58] or hand joints [29], which primarily model grasp variations within the hand space. These model tends to easily overfit to common grasp patterns, lacking diversity despite the use of CVAE. Karunratanakul *et al*. [30] proposed to learn an implicit grasping field. However, hand articulations are not considered, and posed hands are treated as rigid objects, the solution space spans the whole space and the generated results do not guarantee to be valid. In contrast to existing methods, we suggest learning the object-centric ContactGen within the object space. This involves breaking down the variability in hand grasping into distinct components within the ContactGen: contact location, hand part, and touching direction. This decomposition allows us to sample from the ContactGen, generating physically realistic grasps with increased diversity. Existing approaches model grasp uncertainty in the hand space often lean towards learning generalized grasp patterns with poor diversity.

Grasp Optimization Another distinct research direction focuses on analytical grasp solution [1,13,21,31,33,41,43,53,60,61,67]. The objective is to optimize grasps to minimize penetration, enhance contact, and improve overall stability. In human grasp generation, prior research [17,27,73] has formulated contact loss and penetration loss on MANO model [50] by optimizing MANO parameters. However, the optimization proves challenging. First, the objectives of promoting contact and reducing penetration inherently conflict with one another. Striking a balance between encouraging hand-object contact while preventing penetration is computationally intricate. Second, as shown in previous work [12,40], the inherent discretization and limited spatial resolution of mesh structures pose constraints. To address these challenges, we propose a hand articulation model that employs part-wise Signed Distance Function (SDF) for optimization. The SDF neatly partition the space for contact (SDF $= 0$) and penetration (SDF < 0). It captures finegraind deformation and supports contact direction within each part. The piecewise model also shares the same parameters as MANO, making it seamlessly compatible with our contact representations. By incorporating the piecewise hand model into our optimization process, we substantially enhance the grasp quality, leading to more effective and diverse grasp results.

3. Overview

Our technical sections are organized as follows: In Sec. 4, we introduce an object-centric contact representation. Our novel representation encodes contact information into three maps: the contact probability map, the hand part map, and the direction map. In Sec. 5, we use a sequential CVAE network, trained on hand-object interaction data, to infer the contact representation in a generative manner. Finally, in Sec. 6, we derive a grasp from the predicted ContactGen using our specially-developed contact solver. This incorporates our proposed piecewise hand articulation model. Fig. 1 provides a summary of our approach.

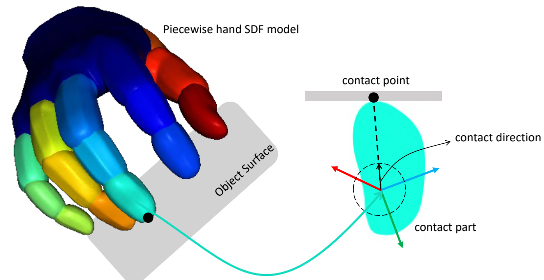

Figure 2: Illustration of the contact part and direction in ContactGen representation. The piecewise hand SDF model is partitioned into B=16 parts, represented in different color.

4. Object-Centric Contact Representation

Our object-centric contact representation $\mathbf{F} = (\mathbf{C}, \mathbf{P}, \mathbf{D})$ consists of three maps, contact map $\mathbf{C}$, part map $\mathbf{P}$ and direction map $\mathbf{D}$. All maps are defined on a set of N points $\mathbf{O} \in \mathbb{R}^{N\times 3}$ sampled from the object surface, as shown in Fig. 1.

Contact Map The contact map $\mathbf{C} \in \mathbb{R}^{N\times 1}$, each $c_i \in \mathbf{C}$ is within [0, 1], representing the contact probability of the point. This contact map closely resembles the original contact map proposed in [17]. Intuitively, the contact map illustrates *which part of the object will likely be contacted by hand*. However, relying solely on contact maps is insufficient for complex human-object interaction modeling due to ambiguities regarding *how and where the hand touches the objects*. To address this, our object-centric representation is extended by explicitly modeling the other two maps.

Part Map To locate the in-contact point on the hand surface, we use a part map $\mathbf{P} \in \mathbb{R}^{N\times B}$ (one-hot vector) to indicate the hand part label in $\{1, \cdots, B\}$ in contact with the object point $\mathbf{O}$. The hand is divided into B parts, the partition is shown in Fig. 2. Each value $\mathbf{p}_i$ in $\mathbf{P}$ is taken as the closest hand part label.

Direction Map Within each part, to describe an arbitrary point exactly on the part surface, we use its direction to the part center. The direction map $\mathbf{D} \in \mathbb{R}^{N\times 3}$, $\mathbf{d}_i \in \mathbf{D}$ records the direction of this point w.r.t. part b, as shown in Fig. 2. Imagine each part as a unit sphere, the contact direction could be any ray shooting from the part center to the sphere surface. Given the direction $\mathbf{d}_i$, the contact point location in part b could be uniquely determined by searching along the ray direction $\mathbf{d}_i$ until its part SDF $= 0$.

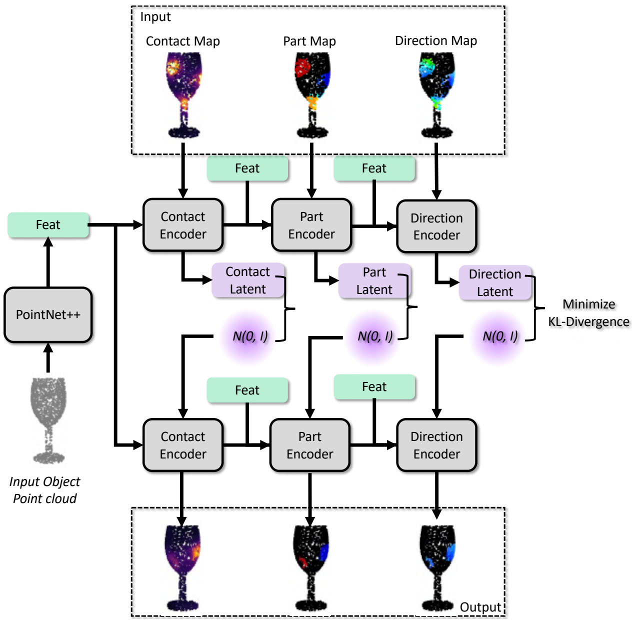

Figure 3: ContactGen CVAE model Architecture. Conditioned on the input object point cloud, we decompose the ContactGen into individual components by a sequential encoder-decoder model.

5. Generative Contact Modeling

Given an object input, we use a conditional generative model to infer possible object-centric contact representations $\mathbf{F}$. We train the model by modeling the underlying distribution $p(\mathbf{F}|\mathbf{O})$ from sampled object point clouds $\mathbf{O}$.

Sequential CVAE. As shown in Fig. 3, we model $p(\mathbf{F}|\mathbf{O})$ sequentially using a CVAE framework. We choose CVAE for its simplicity and its capability to model multi-modal uncertainty. We factorize the joint distribution of the contact feature $\mathbf{F} = (\mathbf{C}, \mathbf{P}, \mathbf{D})$ into a product of three conditional probability functions:

$$p(\mathbf{F}|\mathbf{O}) = p(\mathbf{D}|\mathbf{P}, \mathbf{O})p(\mathbf{P}|\mathbf{C}, \mathbf{O})p(\mathbf{C}|\mathbf{O}) \tag{1}$$

The contact map $\mathbf{C}$ is conditioned on object input $\mathbf{O}$; the part map $\mathbf{P}$ is additionally conditioned on contact map $\mathbf{C}$; direction map additionally conditioned on part map $\mathbf{P}$. We control each component in Eq. (1) by a latent code z sampled from Gaussian distribution. We sample latent code from posterior $\mathbf{z}_c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, $\mathbf{z}_p \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, $\mathbf{z}_d \sim \mathcal{N}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$ during training, and sample latent code from prior $\mathbf{z}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{z}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{z}_d \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at inference. $\mathcal{G}_c$, $\mathcal{G}_p$ and $\mathcal{G}_d$ denote the conditional decoder for contact map, part map and direction map. Thus, we could sample prediction $\hat{\mathbf{C}}, \hat{\mathbf{P}}, \hat{\mathbf{D}}$ using:

$$\hat{\mathbf{C}} = \mathcal{G}_c(\mathbf{z}_c; \mathbf{O}), \hat{\mathbf{P}} = \mathcal{G}_p(\mathbf{z}_p; \hat{\mathbf{C}}; \mathbf{O}), \hat{\mathbf{D}} = \mathcal{G}_d(\mathbf{z}_d; \hat{\mathbf{P}}, \mathbf{O}) \tag{2}$$

This guarantees the three generated maps are consistent with each other and decompose the complicated structured sampling of the contact representation into the conditional generation of each component. We could later recover full hand information from the sampled representation $\hat{\mathbf{C}}, \hat{\mathbf{P}}, \hat{\mathbf{D}}$.

Model architecture Our generative model is a point-based network that operates on the sampled point cloud of an input object. We extract shared object features using PointNet++ [48]. The Gaussian parameters of each component are inferred by an encoder modeled as a MLP. After sampled latent code, we have 3 sequential PointNet [47] decoders $\mathcal{G}_c$, $\mathcal{G}_p$ and $\mathcal{G}_d$ to decode each map. For part information, we employ an embedding layer to encode each part label into an embedded feature before introducing it to the network. The detailed architecture is shown in the supplementary.

Training We train the network in an end-to-end fashion. All networks are trained jointly. We use teacher forcing [63] and send GT contact map and part map as conditioning during training. The VAE loss consists of a reconstruction term and a KL regularization term. The total loss $\mathcal{L}$ is the following:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} \tag{3}$$

The reconstruction $\mathcal{L}_{rec}$ is defined as:

$$\mathcal{L}_{rec} = W_C \left(|C - \hat{C}| + \lambda_p \mathcal{L}_{CE}(P, \hat{P}) + \lambda_d \mathcal{L}_d(D, \hat{D}) \right) \tag{4}$$

$\mathcal{L}_{CE}$ is the standard cross-entropy loss between the predicted part label $\hat{P}$ and ground-truth (GT) P, $\mathcal{L}_d$ computes the cosine similarity between the GT direction $\mathbf{d}_i$ and predicted direction $\hat{\mathbf{d}}_i$ for each point. All losses are computed per-point and further weighted by $W_C = C + \delta$, where $C \in [0, 1]$ is the GT contact map value as we focus more on those contact locations and δ is a default weight for non-contacted points. $\mathcal{L}_{KL}$ regularizes the latent z-space close to normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathcal{L}_{KL}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = KL[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)||\mathcal{N}(\mathbf{0}, \mathbf{I})]$. The whole KL loss consists KL regularization from each latent space $\mathcal{L}_{KL} = \mathcal{L}_{KL}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) + \mathcal{L}_{KL}(\boldsymbol{\mu}_p.\boldsymbol{\Sigma}_p) + \mathcal{L}_{KL}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$.

6. Grasp Synthesis

Next, we will discuss how to convert the sampled object-centric contact representation into a corresponding articulated hand pose. To achieve this, we parameterize the hand using the piecewise SDF model, formulate a model-based optimization, and solve this optimization to recover the most plausible hand pose.

6.1. Piecewise hand articulation model

Following [38], we convert MANO model [50] to a piecewise SDF model. This modification enhances the compatibility of the hand model with the proposed ContactGen for grasp synthesis. The piecewise model partitions the hand into B parts and use piecewise SDF as representation $\{\text{SDF}_b\}_{b=1}^{B}$. The model is parameterized by part pose

$\mathbf{T}_b \in \mathbb{SE}(3)$ and a global shape vector β, $\mathbf{T}_b$ is the transformation from part to global coordinate frames. We use axis-angle θ_b as pose code for each part, $\mathbf{T}_b = \mathbf{T}(\theta_b)$. Given part b, the signed distance from a point $\mathbf{o}_i$ to its surface is given by $\mathrm{SDF}_b(\mathbf{T_b^{-1}o_i};\beta) = \mathrm{SDF}_\mathbf{b}(\mathbf{T}(\theta_\mathbf{b})^{-1}\mathbf{o_i};\beta)$. The direction of the point w.r.t the part is given by $\mathbf{d}_i = \frac{\mathbf{T_b^{-1}o_i}}{\|\mathbf{T_b^{-1}o_i}\|}$. The overall piecewise hand SDF model parameters are pose parameters concatenated from all parts $\theta = \oplus_b^B \theta_b \in \mathbb{R}^{B\times 3}$ and shape code β. The parameters θ and β are shared across the MANO model [50] and the piecewise SDF hand model. One can easily convert between each other. Training details of the SDF model are provided in the supplementary.

6.2. Contact solver

Given the sampled points $\mathbf{O}$ and predicted ContactGen $\hat{\mathbf{C}} \in \mathbb{R}^{N\times 1}$, $\hat{\mathbf{P}} \in \mathbb{R}^{N\times B}$ in one-hot format, $\hat{\mathbf{D}} \in \mathbb{R}^{N\times 3}$, the goal is to infer the hand model $\{\mathrm{SDF}_b\}_{b=1}^B$ parameters θ and β. The optimization objective is the following:

$$\min_{\theta,\beta} \lambda_c\mathcal{L}_c + \lambda_d\mathcal{L}_d + \lambda_p\mathcal{L}_p + \lambda_r\mathcal{L}_r \tag{5}$$

The $\mathcal{L}_c$ denotes the contact map loss, Given object points $\mathbf{O}$, For each part b we compute $\mathrm{SDF}_b(\mathbf{T}(\theta_b)^{-1}\mathbf{O};\beta) \in \mathbb{R}^{N\times 1}$, $\hat{\mathbf{P}}_b$ denotes the b-th column of $\hat{\mathbf{P}}$. The contact map loss is:

$$\mathcal{L}_c = \hat{\mathbf{C}} \sum_{b=1}^{B} \hat{\mathbf{P}}_b \cdot |\mathrm{SDF}_b| \tag{6}$$

For $\mathbf{o}_i$ with higher contact value $\hat{c}_i$ and predicted part $\arg\max \hat{p}_i = b$, we encourage the SDF of hand part b to be close to 0, driving the hand to touch the contact location. The second term $\mathcal{L}_d(\mathbf{D}, \hat{\mathbf{D}})$ encourages the direction $\mathbf{d}(\mathbf{o}_i, \theta_b)$ of hand part b to match the predicted direction $\hat{\mathbf{d}}_\mathbf{i}$.

$$\mathcal{L}_d = W_C \left(1 - \cos(\mathbf{D}, \hat{\mathbf{D}})\right) \tag{7}$$

The penetration loss $\mathcal{L}_p$ prevents object sampled points from being inside the hand:

$$\mathcal{L}_p = \sum_{b=1}^{B} -\max\left(\mathrm{SDF}_b, 0\right) \tag{8}$$

The last term $\mathcal{L}_r$ is the regularization term that prevents the model from being too complex, $\mathcal{L}_r = \lambda_r(\|\theta\|^2 + \|\beta\|^2)$. For each object and predicted ContactGen, we optimized the above objective function to get the hand grasp.

Inference We optimize θ and β from scratch. Following [65, 73], we adopted a two-stage optimization strategy. In the first stage, we only optimize the global pose of the hand. In the second stage, we freeze the hand's global pose and optimize the hand's pose and shape parameters. We use Adam [32] optimizer for both stages.

7. Experiments

7.1. Datasets

We use the GRAB dataset [56] to train the ContactGen CVAE and test grasp synthesis performance. GRAB contains real human grasps for 51 objects from 10 different subjects. We follow the official train/test split. The test set contains six unseen objects. Following [27, 29, 30], we also test on out-of-domain objects from HO3D dataset [19] test set to evaluate the generalization ability of the model.

7.2. Implementation Details

The piecewise hand SDF model was trained on the Freihand dataset [74] with 32,560 samples. The ContactGen CVAE takes $N = 2048$ points sampled from object surface as input. During training, we apply data augmentation by randomly rotating the object by $[-\frac{\pi}{6}, \frac{\pi}{6}]$ around each axis. The latent dimension of the CVAE was set to 16. We set $\lambda_p = 0.5$, $\lambda_d = 1$, and the KL weight λ_{KL} was annealed from 0 to $5e-2$ during training. We employed the standard Adam optimizer [32] with a learning rate of $1.6e-3$ and a batch size of 256. The CVAE was trained for 3000 epochs.

For grasp synthesis, we do 200 iterations with a learning rate of $5e-2$ to optimize the global hand translation and rotation, and 1000 iterations with a learning rate of $5e-3$ to optimize hand pose and shape. The regularization term $\lambda_c = 1e-1$, $\lambda_d = 1e-2$, $\lambda_r = 1e-2$, penetration weight $\lambda_p = 3.0$. We use the Adam optimizer for both stages.

7.3. Recovering Grasps from GT ContactGen

We first evaluate the effectiveness of the proposed ContactGen representation and optimization procedure at recovering hand grasps from ground truth ContactGen. We compare two alternate representations from past work.

Baselines. ContactOpt [17], utilizes contact maps on both the hand and object to refine hand pose. In our experiment, we provide it with necessary GT contact maps. **TOCH** [73] uses binary contact labels on the object's surface and hand correspondences on the MANO mesh [50] to represent contacts. For our study, we adapt TOCH to work with a single frame by removing temporal constraints. Additionally, we convert the GT contact map into binary using a threshold of 0.5 and provide the associated GT contact vertices for optimization. In all cases, the methods optimize the hand pose from scratch.

Metrics. We use the same three metrics as used in past works [19, 74]. **Mesh endpoint error (EPE)** measures the average Euclidean distance between the hand vertices of the prediction and the GT. **Mesh AUC** measures the percentage of correctly reconstructed vertices (vertices within 5cm are considered correct). **Mesh F-Score** calculates the harmonic mean of recall and precision between two meshes given a distance threshold. We report F-score at 5mm and 15mm.

Method	EPE (cm) ↓	AUC ↑	F-score ↑ @5mm	F-score ↑ @15mm
ContactOpt [17]	7.00	0.26	0.24	0.50
TOCH [73]	3.44	0.51	0.39	0.72
Ours	**1.49**	**0.77**	**0.55**	**0.91**

Table 2: Contact reconstruction comparison. We compare against ContactOpt and TOCH 's representation ability to recover GT hand grasps from contact. Due to the completeness of our representation, we are able to effectively recover the hand pose from the contact, achieving the lowest reconstruction error.

Results. The experimental results are presented in Tab. 2. We observe that both ContactOpt and TOCH are unable to accurately recover the GT hand pose due to the incompleteness of their respective representations. ContactOpt [17], despite having access to the GT hand and object contact maps, faces challenges in determining the specific hand-part that should establish contact with a given contact location on the object. TOCH shows comparatively better performance due to the richer contact information from hand-object correspondences. However, the optimization remains challenging as contact location is sparse. The hand pose isn't uniquely defined since the contact direction can also vary. Due to the completeness of the representation, our method is able to effectively recover the hand pose from the contact and achieves the lowest reconstruction error.

7.4. Grasp Synthesis Evaluation

We follow the experimental setup from [29]. We test on 6 unseen objects from the GRAB dataset [56] and out-of-domain test objects from the HO3D dataset [19].

Baselines. GrabNet [56] and GraspTTA [27] utilize CVAE to generate MANO parameters. **GraspTTA** [27] also employs test-time adaptation to enhance generated grasps. **Grasping Field** [30] (GF) uses a CVAE to predict 3D hand point clouds and fit a MANO model afterward. **HALO** [29] generates 3D keypoints using a CVAE and uses an implicit occupancy network to transform these keypoints into meshes. While all baselines predict within the hand space, our approach makes inferences in the object space.

Metrics. Following [21, 27, 29, 30, 56, 58, 65], we evaluate the generated grasps based on their a) physical plausibility and stability, b) diversity, and c) perceptual attributes.

• To assess physical plausibility, we use hand-object *Interpenetration Volume* and *Contact Ratio* following [21,27,29, 60,65,70,72]. We compute interpenetration volume by voxelizing the meshes into 1mm^3 cubes and measuring overlapping voxels. Contact ratio calculates the proportion of grasps that are in contact with objects. For grasp stability assessment, consistent with [10,21,27,30,60,61], we place the object and the predicted hand into a simulator [11], and measure the average *Simulation Displacement* of the object's center of mass under the influence of gravity.

Method	Penetration Volume ↓	Contact Ratio ↑	Simulation Displacement ↓	Entropy ↑	Cluster Size ↑
GrabNet [56]	3.65	**0.96**	**1.72**	2.72	1.93
HALO [29]	3.61	0.94	2.09	**2.88**	2.15
Ours	**2.72**	**0.96**	2.16	**2.88**	**4.11**

Table 3: Grasp result on the GRAB dataset [56]. Our method achieves the lowest penetration, highest contact, and comparable stability compared to previous approaches, while showcasing significantly larger generation diversity.

• Following [29, 72], we evaluate diversity in generated grasps by first clustering generated grasps into 20 clusters using K-means and then measuring the *Entropy* of cluster assignments and the average *Cluster Size*. Higher entropy and cluster size values indicate better diversity. Following previous work [29], we perform K-means clustering on 3D hand keypoints for all methods.

• Following [27,30,58,65], we also perform *Human Evaluation* on the naturalness and stability of generated grasps.

Results on GRAB dataset. Tab. 3 shows comparison of our method with GrabNet [56] and HALO [29] on the GRAB dataset. For each method, we randomly generate 20 grasps. Our method achieves the lowest penetration, highest contact ratio, and comparable stability compared to previous approaches. It stands out in terms of grasp variability, as indicated by the significantly larger cluster size value compared to previous methods. Qualitative results are shown in Fig. 4. Further assessment of diversity against HALO is presented in Sec. 7.4. HALO tends to generate similar grasps for a given object. Our method produces significant diversity in terms of contact locations and grasp poses.

Results on HO3D dataset. We evaluate the generalization capability of our models on the HO3D dataset [19]. Apart from the previously mentioned baselines, we extended our comparison to include Grasping Field [30] and GraspTTA [27], both trained on the ObMan dataset [21]. As indicated in Tab. 4 and shown in Fig. 6, our method achieves performance close to the best method in each metric, while keeping the highest diversity in the generated grasps. As Grasping Field [30] struggles to produce valid grasps for unseen objects, the computation of simulation displacement is infeasible. GraspTTA [27] and Grabnet [56] produce nearly identical grasps for a given object resulting in poor diversity. In comparison to HALO, our method achieves notably lower penetration and better stability.

Human evaluation. We also conduct a user study to assess the perceptual quality and stability of the generated grasps following [27, 30, 65]. We evaluate 12 objects in total from GRAB [56] and HO3D [19] dataset. The evaluation involved 10 participants. For each object, we included 3 randomly sampled GT grasps from the GRAB dataset, 3 generated grasps by HALO, and 3 generated grasps by ours. Participants were asked to rate the quality of each grasp based

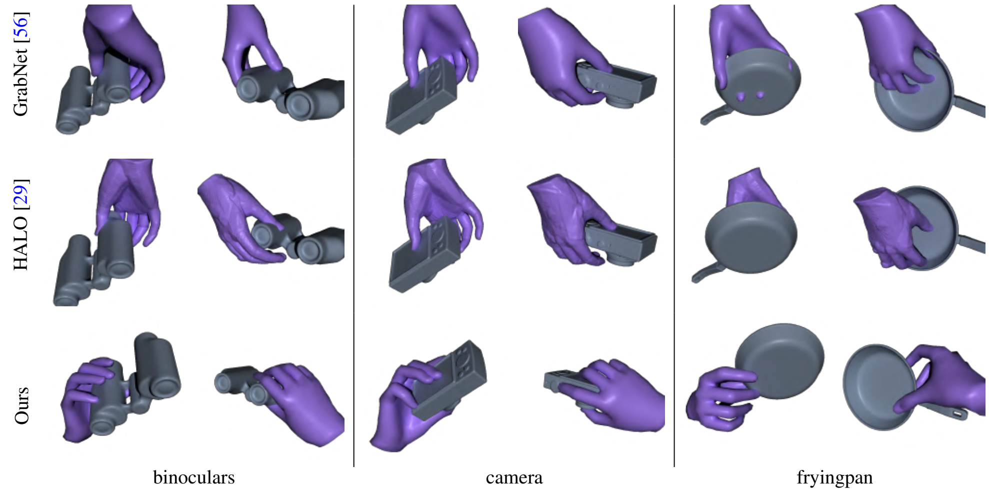

Figure 4: Qualitative comparison on GRAB dataset [56]. Each pair displays sampled grasps from dual views. Our generated grasps showcase improved object contact and reduced penetration.

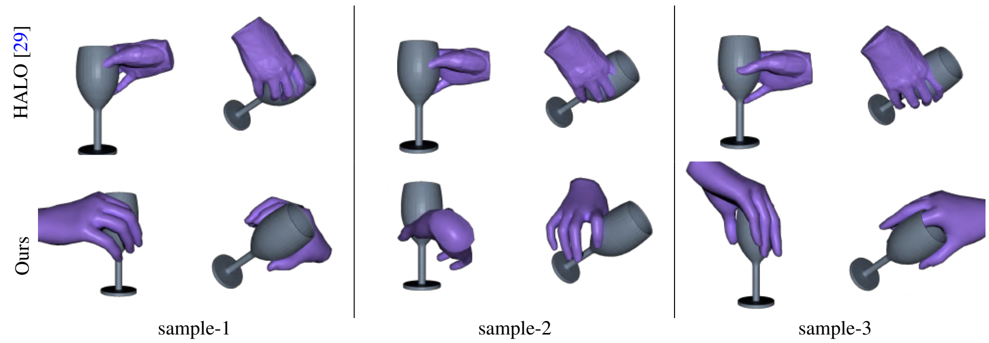

Figure 5: Generated grasp diversity comparison on GRAB dataset [56]. Each pair displays sampled grasps from two views. We observe HALO generates similar grasps for a given input object, while ours generated grasps exhibit more diverse grasp poses.

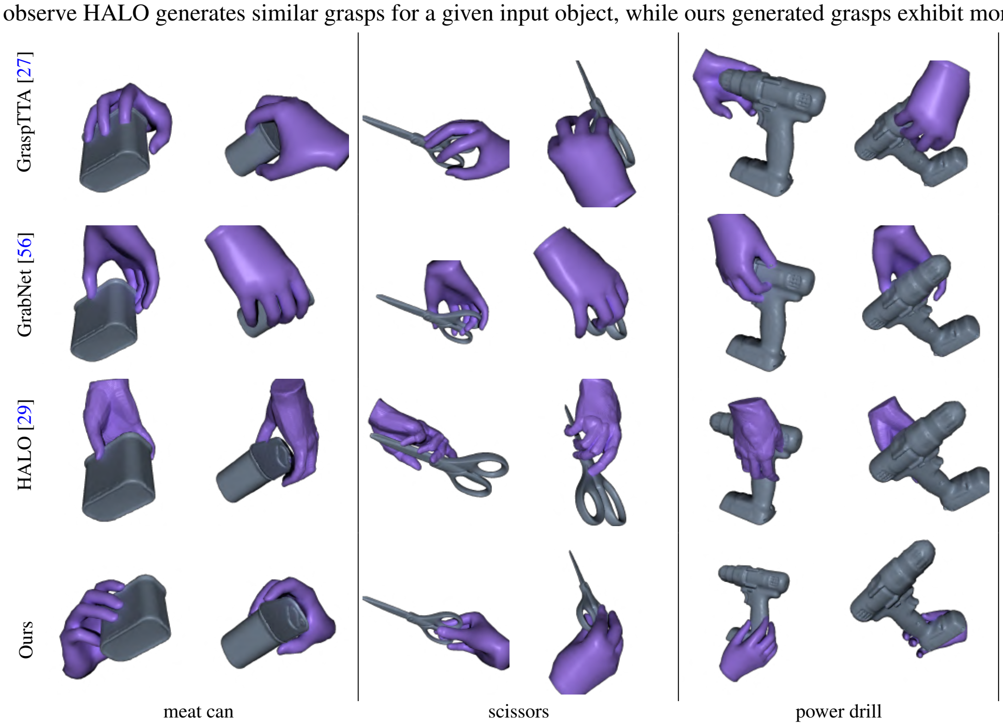

Figure 6: Qualitative comparison on out-of-domain HO3D dataset [19]. Each pair displays sampled grasps from dual views. Our method produces more plausible grasps for unseen objects.

Method	Penetration Volume ↓	Contact Ratio ↑	Simulation Displacement ↓	Entropy ↑	Cluster Size ↑
GraspTTA [27]	**7.37**	0.76	5.34	2.70	1.43
GrabNet [56]	15.50	0.99	**2.34**	2.80	2.06
GF [30]	93.01	**1.00**	-	2.75	3.44
HALO [29]	25.84	0.97	3.02	**2.81**	4.87
Ours	9.96	0.97	2.70	**2.81**	**5.04**

Table 4: Grasp result on the HO3D dataset [19]. Our method achieves performance close to the best method in each metric, while maintaining the highest diversity in the generated grasps.

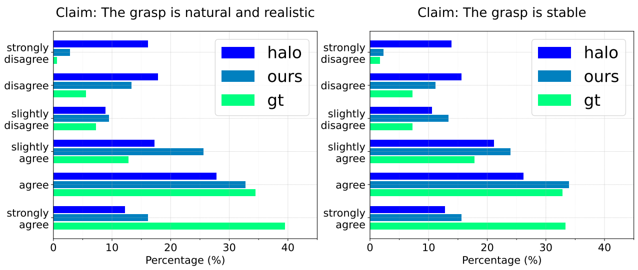

Figure 7: Grasp human studies score distribution. The distribution of scores shows that our method achieves comparable performance to the GT in both naturalness and stability.

Method	GRAB [56] Natural ↑	GRAB [56] Stable ↑	HO3D [19] Natural ↑	HO3D [19] Stable ↑
HALO	2.59 ± 1.69	2.68 ± 1.63	2.23 ± 1.41	2.09 ± 1.45
Ours	3.21 ± 1.35	3.23 ± 1.23	**3.17 ± 1.34**	**2.87 ± 1.52**
GT	**3.93 ± 1.18**	**3.73 ± 1.28**	-	-

Table 5: Grasp human study statistics. While the gap between ours and the GT exists, our method performs better than HALO [29] in terms of naturalness and stability. As HO3D dataset [19] isn't intented for grasping, GT wasn't provided.

on its naturalness and the stability of holding the object using a five-point scale ranging from strongly disagree (0) to strongly agree (5). Fig. 7 shows the score distribution, and Tab. 5 provides the mean and variance for each method's ratings. Our method outperforms HALO in terms of naturalness and stability on both datasets, but still lags behind ground truth grasps. As the HO3D dataset was not designed for grasping, appropriate ground truth hand grasps weren't available for evaluation. Further details about the human evaluation setup are available in the supplementary.

7.5. Ablations

Contact decomposition ablations. We start by comparing our hierarchical ContactGen decomposition with two more obvious choices: Joint and Separate modeling. Joint modeling utilizes a shared encoder to encode the 3 maps and a shared decoder to decode them jointly. Separate modeling encodes and decodes each ContactGen component independently, using 3 separate encoders and decoders for each map. The results are shown in Tab. 6. The separate model

Method	Penetration Volume ↓	Contact Ratio ↑	Simulation Displacement ↓	Entropy ↑	Cluster Size ↑
Joint	3.40	0.95	3.29	2.80	3.87
Separate	2.88	0.89	3.58	2.80	**4.83**
Ours	**2.72**	**0.96**	**2.16**	**2.88**	4.11

Table 6: Impact of different ContactGen decomposition on grasp quality. Two ablations model the contact latent space jointly (**Joint**) or separately (**Separate**). Both approaches fail to guarantee consistency, leading to lower contact ratios, larger penetrations, and increased simulation displacements.

Hand	C	P	D	Penetration Volume ↓	Contact Ratio ↑	Simulation Displacement ↓	Entropy ↑	Cluster Size ↑
MANO	✓			7.41	0.88	4.38	2.66	2.54
MANO	✓	✓	✓	**2.36**	**0.98**	2.85	2.60	3.56
SDF	✓			20.13	0.69	-	2.68	1.44
SDF	✓	✓		2.81	0.66	6.81	2.80	**5.51**
SDF	✓	✓	✓	2.72	0.96	**2.16**	**2.88**	4.11

Table 7: Impact of different ContactGen components on grasp quality. Contact as **C**, part as **P**, direction as **D**. Every part of the ContactGen significantly influences grasp synthesis. The proposed piecewise hand SDF model better captures intricate hand poses and enhances generation stability and diversity.

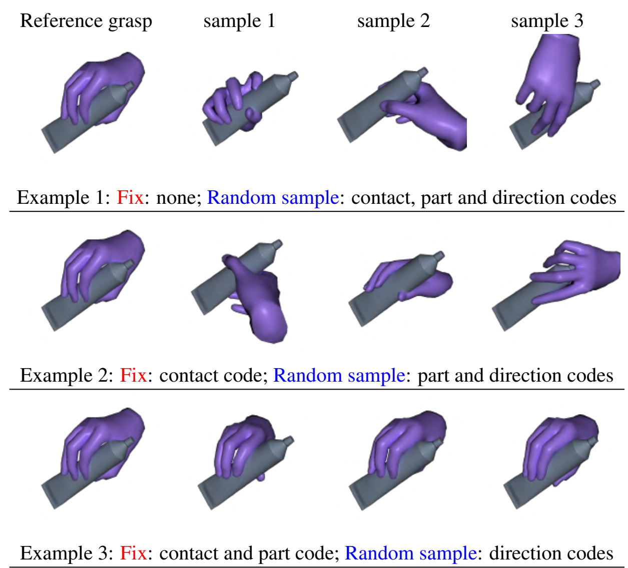

Figure 8: Visualizing grasp diversity by selectively fixing contact code, part code, and direction code.. In the top row, grasps use random contact, part, and direction codes. In the middle row, contact is fixed, only part and direction vary. In the bottom row, contact and part are fixed, only direction changes.

achieves the highest diversity, and the Joint model resulted in reduced grasp diversity. But both choices struggled to maintain consistency among the three components, failing to yield physically plausible grasps. These outcomes are characterized by either larger penetrations, decreased contact ratios, or higher simulation displacements. In contrast,

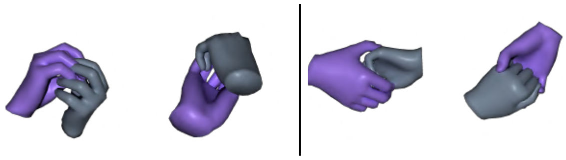

Figure 9: Hand-hand interaction synthesis on InterHand2.6M dataset [42]. Each pair displays a sample from two views.

Figure 10: Failure modes of our method. Each pair displays a sample from dual views. The **left** side shows a generated grasp appears to be more of a touch rather than a proper grasp. The **right** side shows an unsatisfactory grasp when the sampled ContactGen is infeasible on out-of-domain objects.

our proposed decomposition and the proposed sequential structure perform optimally. The generated outcomes are internally consistent and exhibit substantial diversity.

We further illustrate the decomposition of diversity in our generated grasps across different components of the ContactGen in Fig. 8. In the top row, we present generated grasps with randomly sampled ALL three latent codes, i.e., contact, part, and direction latent codes. The second row showcases generated grasps with a fixed contact code (matching the first sample in the row) and randomly sampled the remaining two latent codes. The third-row exhibits generated grasps with fixed contact and part codes (matching the first sample in the row) and randomly sampled the last direction latent codes. By conditioning on different levels of map details, we can sample and produce diverse grasp. However, as anticipated, the diversity of generated grasps diminishes when we freeze more components of the ContactGen representation, progressing from the top row to the bottom row. This trend becomes particularly evident in the bottom row, where the contact location and hand parts are fixed, yielding slight variations in contact directions.

Contact representation ablations. We conduct another ablation study to assess the contribution of different components within our ContactGen representation. The study involves evaluating the generated grasp quality by removing specific components. Additionally, we compare the performance of our proposed piecewise hand SDF model against the MANO model [50] in contact optimization. The results are shown in Tab. 7. The results highlight the critical nature of all components (contact map, part map, and direction map) for achieving optimal performance. Without the guidance of the part map, both hand model struggles to generate a coherent grasp, leading to consistently higher penetrations. The influence of the part map is more critical to the piecewise hand SDF model, as it heavily relies on part information for contact reasoning. Incorporating the direction map aids in enhancing contact and stability. Both the MANO model and the piecewise SDF model exhibit similar physical quality with the assistance of all three maps. Employing the SDF model better captures intricate hand poses, resulting in enhanced diversity and more stable outcomes.

7.6. Synthesis Beyond Grasping

Our proposed contact representation can extend its applications beyond grasping to address more complex **hand-hand interactions** scenarios. By substituting the object for another hand, our method, without any changes, can synthesize two-hand interactions. Specifically, we train our CVAE on a subset of the training set from InterHand2.6M dataset [42]. We then use the CVAE to generate ContactGen on the subset from the test split. We employ the same contact solver to decode the hand pose. By using the left hand as input, we generate corresponding right-hand poses. The qualitative results are presented in Fig. 9. This demonstrates the broad potential of our proposed ContactGen representation in addressing a wider range of interaction tasks.

8. Conclusion

In this work, we introduce ContactGen: an object-centric contact representation for hand-object interaction. The representation is compact and complete, enabling full grasp recovery from contact information. We propose a sequential CVAE to learn the ContactGen from hand-object interaction data and a model-based optimization to generate grasp from ContactGen predictions of the input object. Experiments demonstrate our method can synthesize high-fidelity and diverse grasps for various objects. The ContactGen could also be potentially used for more complex interaction scenarios synthesis beyond grasp.

Limitations and future work. We discuss two limitations of our method. First, our approach can sometimes generate touch interactions instead of grasps. This is evident on the left side of Fig. 10, where the generated grasp for a frying pan from the GRAB dataset [56] is realistic in terms of touch but not suitable for grasp. This occurs because of the inherent uncertainty in our ContactGen model. These touches exhibit strong object contact but may not align well with human grasp expectations. Second, when our method is applied to objects outside of its training domain, like those from the HO3D dataset [19], it occasionally creates unrealistic combinations of contact map, part map, and direction map, resulting in generated grasp with insufficient contact or significant penetration, as shown on the right side of Fig. 10. Addressing these limitations might involve more accurately modeling the prior of the ContactGen. Rather than using the simple Gaussian prior ($\mathcal{N}(\mathbf{0}, \mathbf{I})$) in VAE, advanced techniques like diffusion models could offer potential solutions. We leave further exploration of these possibilities for future work.

References

[1] Dafni Antotsiou, Guillermo Garcia-Hernando, and Tae-Kyun Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 3

[2] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 640–653. Springer, 2012. 1

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

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

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 1

[6] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 1, 2

[7] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 1

[8] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 2

[9] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 231–248. Springer, 2022. 1

[10] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 2, 6

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

[12] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 3

[13] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018. 3

[14] Bardia Doosti, Shujon Naha, M. Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. *CVPR*, pages 6607–6616, 2020. 1

[15] George ElKoura and Karan Singh. Handrix: animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 110–119, 2003. 2

[16] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 2

[17] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 1, 2, 3, 5, 6

[18] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010. 2

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent 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. 5, 6, 7, 8, 9

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 2

[21] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1, 3, 6

[22] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[23] Markus Höll, Markus Oberweger, C. Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 175–182, 2018. 1

[24] Kaijen Hsiao and Tomas Lozano-Perez. Imitation learning of whole-body grasps. In *2006 IEEE/RSJ international con-*

ference on intelligent robots and systems, pages 5657–5662. IEEE, 2006. 2

[25] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 1, 2

[26] Wolfgang Hürst and Casper Van Wezel. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 62(1):233–258, 2013. 1

[27] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 1, 2, 3, 5, 6, 7, 8

[28] Maciej Kalisiak and Michiel Van de Panne. A grasp-based motion planning algorithm for character animation. *The Journal of Visualization and Computer Animation*, 12(3):117–129, 2001. 2

[29] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *2021 International Conference on 3D Vision (3DV)*, pages 11–21. IEEE, 2021. 2, 3, 5, 6, 7, 8

[30] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *arXiv preprint arXiv:2008.04451*, 2020. 1, 2, 3, 5, 6, 8

[31] Jun-Sik Kim and Jung-Min Park. Physics-based hand interaction with virtual objects. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3814–3819. IEEE, 2015. 2, 3

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[33] Robert Krug, Dimitar Dimitrov, Krzysztof Charusta, and Boyko Iliev. On the efficient computation of independent contact regions for force closure grasps. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 586–591. IEEE, 2010. 3

[34] Paul G Kry and Dinesh K Pai. Interaction capture and synthesis. *ACM Transactions on Graphics (TOG)*, 25(3):872–880, 2006. 2

[35] Ying Li, Jiaxin L Fu, and Nancy 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

[36] C Karen Liu. Dextrous manipulation from a grasping pose. In *ACM SIGGRAPH 2009 papers*, pages 1–6. 2009. 2

[37] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 1

[38] Sandro Lombardi, Bangbang Yang, Tianxing Fan, Hujun Bao, Guofeng Zhang, Marc Pollefeys, and Zhaopeng Cui. Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In *2021 International Conference on 3D Vision (3DV)*, pages 278–288. IEEE, 2021. 4

[39] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019. 2

[40] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 3

[41] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 2, 3

[42] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 9

[43] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910, 2019. 3

[44] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 33:7841–7851, 2020. 1

[45] Paschalis Panteleris and Antonis Argyros. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 575–584, 2017. 1

[46] Nancy S Pollard and Victor Brian Zordan. Physically based grasping control from example. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 311–318, 2005. 2

[47] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas 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

[48] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4

[49] Hans Rijpkema and Michael Girard. Computer animation of knowledge-based human grasping. *ACM Siggraph Computer Graphics*, 25(4):339–348, 1991. 2

[50] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017. 3, 4, 5, 9

[51] Anis Sahbani, Sahar El-Khoury, and Philippe Bidaud. An overview of 3d object grasp synthesis algorithms. *Robotics and Autonomous Systems*, 60(3):326–336, 2012. 2

[52] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2

[53] Jungwon Seo, Soonkyum Kim, and Vijay Kumar. Planar, bimanual, whole-arm grasping. In *2012 IEEE International Conference on Robotics and Automation*, pages 3271–3277. IEEE, 2012. 3

[54] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 2

[55] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022. 2, 3

[56] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 1, 2, 3, 5, 6, 7, 8, 9

[57] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019. 1

[58] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 3, 6

[59] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 568–584. Springer, 2022. 2

[60] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 201–221. Springer, 2022. 3, 6

[61] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016. 3, 6

[62] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. A hand-pose estimation for vision-based human interfaces. *IEEE Trans. Ind. Electron.*, 50:676–684, 2003. 1

[63] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989. 4

[64] Min-Yu Wu, Pai-Wen Ting, Yahui Tang, En-Te Chou, and L. Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *J. Vis. Commun. Image Represent.*, 70:102802, 2020. 1

[65] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 257–274. Springer, 2022. 2, 5, 6

[66] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 2

[67] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 1, 2, 3

[68] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 2

[69] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2

[70] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651. IEEE, 2020. 2, 6

[71] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 2

[72] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6194–6204, 2020. 6

[73] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 1–19. Springer, 2022. 2, 3, 5, 6

[74] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 5