GEARS: Local Geometry-aware Hand-object Interaction Synthesis

Keyang Zhou[1,2] Bharat Lal Bhatnagar[1,2] Jan Eric Lenssen[2] Gerard Pons-Moll[1,2]
[1]University of Tübingen, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

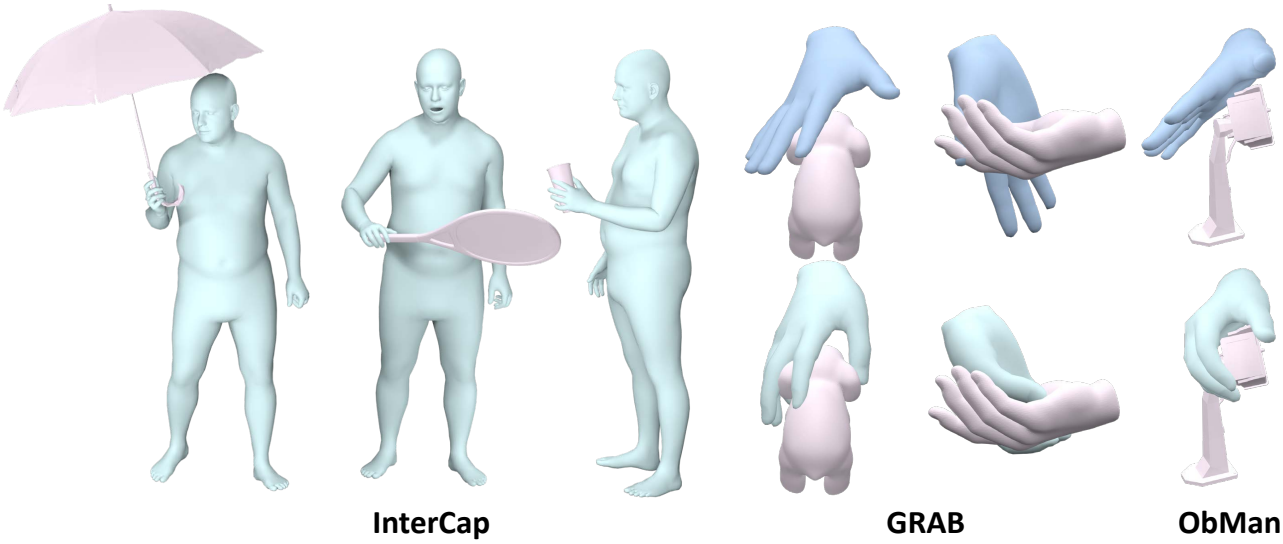

Figure 1. We propose GEARS, a method to synthesize sequence of hand poses during interaction with an object. GEARS takes hand and object trajectory as input. It generates realistic hand poses that are well-adapted to object surface, irrespective of object category and size. We show sample results on different datasets. Hands colored in blue are inputs while hands colored in cyan are our predictions.

Abstract

Generating realistic hand motion sequences in interaction with objects has gained increasing attention with the growing interest in digital humans. Prior work has illustrated the effectiveness of employing occupancy-based or distance-based virtual sensors to extract hand-object interaction features. Nonetheless, these methods show limited generalizability across object categories, shapes and sizes. We hypothesize that this is due to two reasons: 1) the limited expressiveness of employed virtual sensors, and 2) scarcity of available training data. To tackle this challenge, we introduce a novel joint-centered sensor designed to reason about local object geometry near potential interaction regions. The sensor queries for object surface points in the neighbourhood of each hand joint. As an important step towards mitigating the learning complexity, we transform the points from global frame to hand template frame and use a shared module to process sensor features of each individual joint. This is followed by a spatio-temporal transformer network aimed at capturing correlation among the joints in different dimensions. Moreover, we devise simple heuristic rules to augment the limited training sequences with vast static hand grasping samples. This leads to a broader spectrum of grasping types observed during training, in turn enhancing our model's generalization capability. We evaluate on two public datasets, GRAB and InterCap, where our method shows superiority over baselines both quantitatively and perceptually.

1. Introduction

We humans mostly rely on hands to interact with different objects in the surrounding environment. Learning the high-dimensional space of plausible hand-object interactions is an important and challenging task that needs to be solved in many applications. These include modeling digital humans in Augmented and Virtual Reality, or reasoning about potential grasps in robotics.

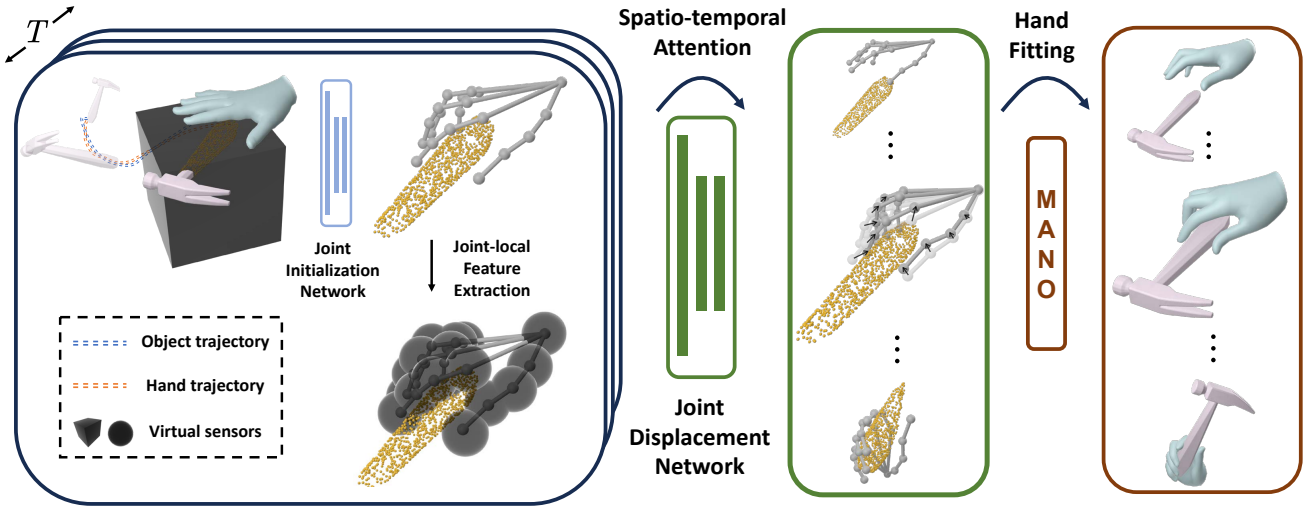

Figure 2. An overview of our method. The input consists of the hand trajectory, object trajectory and object template mesh. For each time frame, the object mesh is cropped with a cube-shaped virtual sensor positioned and oriented based on the wrist. The cropped object points together with the hand trajectory are fed to the Joint Initialization Network to predict coarse joints locations. We then place more fine-grained geometry sensors at each joint to extract joint-local object features. The features are subsequently processed by the Joint Displacement Network to refine the initialized joints. Finally, we fit MANO hand model [27] to the joints to get the hand mesh sequence.

Real world objects can largely differ in size, topology and geometry. Learning a model which can adapt to object surface is a particularly demanding task, especially when existing dynamic hand-object interaction data is very scarce. One crucial factor determining generalization capability lies in how the object is encoded relative to the hand. Previous work [32, 40] proposed to use occupancy-based or distance-based virtual sensors to represent local surface geometry. However, these features have two limitations. First, they are inherently constrained by their expressiveness. Occupancy-based sensor attaches an occupancy grid to the hand. Occupancy grids with a low resolution can only detect coarse object geometry. On the other hand, increasing the grid resolution would result in an exponential increase in feature size. Distance-based sensor measures the distance from a fixed set of basis points rigidly attached to the hand to the closest points on the object surface. It gives more fine-grained features and it is also less computationally expensive. However, a discrete collection of hand-to-object distance cannot faithfully describe local object geometry properties such as normal directions and curvature. Moreover, features computed by both of the aforementioned sensors are global with respect to the hand, which means it is difficult to model the intricate correlation between the movement of each finger. As the result, these methods exhibit limited generalization capabilities to unseen objects of different sizes.

The ability of humans to perform dexterous object manipulations is attributed to the dense tactile sensory receptors in the skin. Thus, we hypothesize that the ability to reason about local geometry is key to generalization to arbitrary surfaces. Inspired by this, we propose a novel hand-object interaction sensor which is local to every hand joint. Specifically, we establish a canonical frame at each joint, and use a shared module to process local object points within a small radius of the joint. This way, the module learns joint-agnostic local features, which are highly generalizable from limited training data. We further fuse together features at each joint by self-attention operations, enabling the model to learn the compositional relationship between different joints in forming the hand pose.

Due to the limited availability of dynamic human-object interaction data, we present a simple yet effective method for generating dynamic hand sequences from static grasps. Static hand grasping data is easily accessible and exhibits a diverse range of object geometry and grasping type. With our data augmentation procedure, we can turn them into artifical grasping sequences. We show that adding them to our training dataset can further improve the results.

Our contributions are as follows:

- We propose a learning-based method to synthesize diverse hand motion sequences interacting with objects. Though trained only on small hand-held objects, we show that our model naturally generalizes to objects of larger sizes (see Figure. 1).
- We introduce a novel hand-object interaction sensor, which detects local object surface geometry relative to hand joints. This is proven essential to our model's generalization capabilities.
- With a simple yet effective data augmentation trick, we

are able to utilizing the vast amount of existing static hand grasp data to train our model.
- Our code and pre-trained model will be released to enhance further research in this direction.

2. Related Work

Static Grasp Synthesis. Synthesizing stable hand grasps given target objects has been extensively studied in computer graphics [18] and robotics [28, 29]. Conventional analytical approaches assume a simplified contact model and solve a constrained optimization problem to satisfy the force-closure condition [1, 8, 26, 46]. In contrast, data-driven approaches generate grasp hypotheses from human demonstrations or annotated 3D data, and rank them according to certain quality metrics [2, 20]. Modern robotic grasping simulators usually combine the merits of both [19, 23]. Recently, there has been an increasing interest in training neural network-based models for hand grasp generation [3, 7, 12, 13, 17, 22, 30, 49]. For example, [15, 16] modeled the hand-object proximity as an implicit function.

Dynamic Grasp Synthesis. In comparison with static grasp synthesis, generating dynamic manipulation of objects is more challenging since it additionally requires dynamic hand and object interaction to be modeled. This task is usually approached by optimizing hand poses to satisfy a range of contact force constraints [21, 25, 39, 44]. With the advent of deep reinforcement learning, a number of work explored training hand grasping control policies in physics simulation [6, 37, 41]. Hand motions generated by these works are physically plausible but lack natural variations. Zheng *et al*. [45] modeled hand poses in a canonicalized object-centric space, achieving category-level generalization for both rigid and articulated objects. More similar to our work are ManipNet [40] and GRIP [32], which utilized occupancy-based and distance-based sensors to extract local object features near the hand and then directly regressed hand poses from the features. We argue that these features are limited by resolution and they are global with respect to the hand, hence hindering generalization capability. In contrast, we adopt a novel joint-centered point-based sensor which captures local object geometry in finer details while enabling modeling the correlation among hand joints.

Full-body Human-object Interaction Synthesis. Generating realistic human motion sequences in 3D scenes has received considerable attention in recent years [11, 14, 24, 34, 35, 42, 43]. However, these work usually models coarse body motion only and ignore fine-grained finger articulations. Another line of work focused on generating full-body motion for grasping [9]. A typical solution to this problem is first generating the final static grasping pose and then using a motion infilling network to generate the intermediate poses [31, 36]. [33] and [47] leveraged the existing body pose prior and hand-only grasping prior to circumvent the limited diversity in available full-body grasping data. Braun *et al*. [4] adopted a physics-based approach, training separate low-level policies for the body and fingers, and then integrating them with a high-level policy which operates in latent space.

Grasp Refinement. As consumer-level hand tracking devices including RGB/depth cameras, data gloves and IMUs become more accessible, it is relatively easy to acquire hands that are approximately correct but may contain noise and artifacts. Refining hand poses in accordance with hand-object interaction emerges as a practical research problem [30]. [38] and [10] proposed to identify the potential contact area on the object surface and subsequently adjust the hand to align with the predicted contact points. Limited by their contact representations, they can only handle static hand grasp. Zhou *et al*. [48] improved upon them and extended the binary contact map representation to a spatio-temporal correspondence map, enabling the refinement of a hand motion sequence. We deviate from these work in our assumptions, as we only require hand and object trajectories as input.

3. Method

Given the trajectories of a hand and an object in interaction, we aim to generate hand poses that align with the object motion. The object shape is assumed to be known. We tackle this problem in three steps. First, we estimate a coarse initial hand pose for each frame individually. We place virtual sensors on the initialized hand joints, detecting nearby object surface points and extracting hand-object interaction features based on these points. Local to each joint, the features are fed to a spatio-temporal attention network, which learns the correlation among hand joints and generates displacements to the initialized hand joints. Lastly, we solve an optimization problem and fit a parametric hand model to the predicted joints. See Figure. 2 for an overview of our method.

Specifically, our input consists of the hand trajectory $\{\boldsymbol{w}^t, \boldsymbol{R}_H^t\}_{t=1}^T$, the object trajectory $\{\boldsymbol{o}^t, \boldsymbol{R}_O^t\}_{t=1}^T$ and the object template mesh $\boldsymbol{M}_O = \{\boldsymbol{V}_O, \boldsymbol{F}_O\}$, with $\boldsymbol{w}^t, \boldsymbol{o}^t \in \mathbb{R}^3$ denoting hand and object translations and $\boldsymbol{R}_H^t, \boldsymbol{R}_O^t \in \mathrm{SO}^3$ denoting hand and object global orientations respectively. We use the MANO model [27] as our hand representation, which is parameterized by shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$. Hence the hand trajectory is composed of the wrist joint coordinates and the global orientation of the target MANO hand at each frame.

3.1. Joint Initialization Network

Given the object position and orientation at frame t, we first obtain the object mesh at that frame by $\boldsymbol{V}_O^t = \boldsymbol{R}_O^t \boldsymbol{V}_O + \boldsymbol{o}^t$. In order to predict an initial hand pose, only the part of the

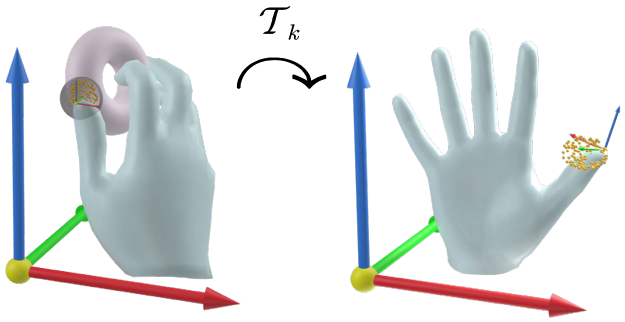

Figure 3. Visualization of our joint-local geometry sensor. (Left) Given the joints positions and the object mesh, we sample points on the object surface within a specified radius centered at each joint. The object points are represented in a joint-local frame. (Right) We transform the sampled object points from global frame to the canonical frame defined by the MANO template hand.

object which is close to the wrist matters. Hence we crop the object by a cube-shaped virtual sensor $\boldsymbol{S}^t$ rigidly attached to the wrist. The resulting partial object mesh is denoted by $\boldsymbol{M}_O^t{}' = \left\{\boldsymbol{V}_O^t{}', \boldsymbol{F}_O^t{}'\right\}$, where $\boldsymbol{V}_O^t{}' = \{\boldsymbol{v}_i \in \boldsymbol{V}_O^t : \boldsymbol{v}_i \notin \boldsymbol{S}^t\}$ and $\boldsymbol{F}_O^t{}' \subseteq \boldsymbol{F}_O^t$.

Let $\boldsymbol{P}^t \in \mathbb{R}^{N\times 3}$ denote the point cloud sampled on $\boldsymbol{M}_O^t{}'$. We subsequently express the point cloud relative to the wrist:

$$\tilde{\boldsymbol{P}}^t = \boldsymbol{R}_H^t{}^T \left(\boldsymbol{P}^t - \boldsymbol{w}^t\right). \tag{1}$$

We additionally sample the hand trajectory centered on the current frame. We sample k frames both in the past and in the future, and express them relative to the wrist in a similar fashion as 1. The inputs to the hand pose initialization module are $\left[\tilde{\boldsymbol{w}}^{t-k:t+k}, \tilde{\boldsymbol{R}}_H^{t-k:t+k}, \tilde{\boldsymbol{P}}^t\right]$, where $\tilde{\boldsymbol{w}}^{t-k:t+k}$ and $\tilde{\boldsymbol{R}}_H^{t-k:t+k}$ are canonicalized sampled wrist positions and orientations respectively. In particular, we first use PointNet to extract a global feature vector from the partial point cloud $\boldsymbol{P}$. This feature vector is then concatenated with the trajectory and fed to a three-layer fully-connected network. The output of the network is denoted by $\boldsymbol{j}_{\text{init}}^t$, which represents the initialized coordinates relative to the wrist. The training loss for this module is defined by

$$L_{\text{init}} = \left\|\boldsymbol{j}_{\text{init}} - \boldsymbol{j}_{\text{gt}}\right\|_2^2, \tag{2}$$

where $\boldsymbol{j}_{\text{gt}}$ denotes groundtruth joint coordinates.

3.2. Local Geometry Sensor

Although coarse and inaccurate, the initialized joints offer an indication of where the hand could potentially interact with the object. To refine the initial joint positions, we need to sense local geometry properties of the object near the interaction regions. We introduce a novel joint-centered point-based local geometry sensor to overcome these limitations. Specifically, given the predicted joints $\{\boldsymbol{j}_i\}_{i=1}^J$, we can utilize inverse kinematics to analytically derive the joint rotations which satisfy:

$$\boldsymbol{j}_k - \boldsymbol{j}_{\text{pa}(k)} = \boldsymbol{R}_{k,\text{pa}(k)} \left(\bar{\boldsymbol{j}}_k - \bar{\boldsymbol{j}}_{\text{pa}(k)}\right), \tag{3}$$

where $\boldsymbol{R}_{k,\text{pa}(k)}$ is the relative angle between the k-th joint and its parent, and $\{\bar{\boldsymbol{j}}_k\}$ are joints in the rest pose. In this manner, we can define the template frame of the k-th joint by

$$\mathcal{T}_k = \prod_{i\in \text{A}(k)} \left[\begin{array}{c|c} \boldsymbol{R}_{i,\text{pa}(i)} & \boldsymbol{j}_i \\ \hline \boldsymbol{1} & \boldsymbol{0} \end{array}\right], \tag{4}$$

where $\text{A}(k)$ denotes the list of ancestors of joint k and $\mathcal{T}_k$ is the transformation which brings joint k from the template frame to the global frame.

By sampling object surface points within a given radius r of the k-th joint along with their normal vectors, we get $\boldsymbol{F}_k = \{\boldsymbol{P}_k, \boldsymbol{N}_k\}$, where $\boldsymbol{P}_k = \{\boldsymbol{v}_i \in \boldsymbol{V} : \|\boldsymbol{v}_i - \boldsymbol{j}_k\|_2^2 < r\}$. We then transform the sampled points to the template frame, by

$$\bar{\boldsymbol{F}}_k = \{\bar{\boldsymbol{P}}_k, \bar{\boldsymbol{N}}_k\} \tag{5}$$

$$= \{\mathcal{T}_k^{-1}(\boldsymbol{P}_k - \boldsymbol{j}_k), \mathcal{T}_k^{-1}\boldsymbol{N}_k\}. \tag{6}$$

Since we now have the sampled object points in a joint-centered canonical frame, we apply a learnable module f_{feat} to process the transformed points. Note that this module is shared between joints, which greatly reduces the learning complexity. We hence arrive at a hand-object interaction feature $\boldsymbol{f}_k = f_{\text{feat}}(\bar{\boldsymbol{F}}_k)$ for each joint k. We implement f_{feat} with a three-layer PointNet architecture.

3.3. Joint Displacement Network

With the local object features aggregated at each joint, we propose to use a transformer architecture to predict displacement vectors to the initialized joints. Achieving a visually plausible and smooth hand sequence requires modeling spatio-temporal inter-joint dependencies. Hence we apply the self-attention operation in both spatial and temporal dimensions. Concretely, we first project the initialized joint coordinates to per-joint embedding vectors with a fully-connected network g_{embed}:

$$\boldsymbol{e}_k = g_{\text{embed}}(\mathcal{T}_k^{-1}\boldsymbol{j}_k), \tag{7}$$

where k is the joint index. We concatenate the joint-local sensor features with the joint embeddings and obtain $\boldsymbol{X} =$

concat($\boldsymbol{f}, \boldsymbol{e}$), which is the input feature tensor for our transformer. Note that X contains features for all the joints in all time frames. We divide $\boldsymbol{X}$ along its spatial and temporal dimension, and apply a self-attention function to them separately.

Spatial self-attention. The spatial self-attention module divides $\boldsymbol{X}$ into batches of frames, and each batch contains joint features in a single frame, denoted by $\boldsymbol{X}_S$. This module takes the hands in different frames as static identities, and focuses on learning the correlations between different fingers. Following conventional self-attention operations, we linearly project $\boldsymbol{X}_S$ to queries $\boldsymbol{Q}_S$, keys $\boldsymbol{K}_S$ and values $\boldsymbol{V}_S$. The output feature is hence obtained by

$$\tilde{\boldsymbol{X}}_S = \text{sa}(\boldsymbol{Q}_S, \boldsymbol{K}_S, \boldsymbol{V}_S) \tag{8}$$

$$= \text{softmax}\left(\frac{\boldsymbol{Q}_S \boldsymbol{K}_S^T}{\sqrt{l}}\right)\boldsymbol{V}_S, \tag{9}$$

where l is the length of key, query and value vectors. See Fig. 4 (Left) for an illustration.

Temporal self-attention. On the other hand, the temporal self-attention module divides $\boldsymbol{X}$ into batches of joints, and each batch contains features of a specific joint across the whole sequence, denoted by $\boldsymbol{X}_T$. This module models the trajectory of each individual joint, ensuring that all joints move in a temporally smooth and consistent manner. We similarly project $\boldsymbol{X}_T$ to queries $\boldsymbol{K}_T$, keys $\boldsymbol{Q}_T$ and values $\boldsymbol{V}_T$ respectively. The module output is

$$\tilde{\boldsymbol{X}}_T = \text{sa}(\boldsymbol{Q}_T, \boldsymbol{K}_T, \boldsymbol{V}_T). \tag{10}$$

See Fig. 4 (Right) for an illustration.

The Joint Displacement Network consists of interleaving spatial and temporal self-attention modules. The output of the last module is fed to a linear layer to produce the joint displacement vectors $\bar{\boldsymbol{d}}$ in template frame. As the last step, we utilize the pose transformation derived from IK previously to transform $\bar{\boldsymbol{d}}$ back to global frame:

$$\boldsymbol{d}_k = \left(\prod_{i \in A(k)} \boldsymbol{R}_{i,\text{pa}(i)}\right)\bar{\boldsymbol{d}}_k. \tag{11}$$

The training loss for this module is defined by

$$L_{\text{disp}} = \left\|\boldsymbol{j}_{\text{disp}} + \boldsymbol{d} - \boldsymbol{j}_{\text{gt}}\right\|_2^2. \tag{12}$$

3.4. Hand Fitting

With the predicted sequence of hand joints $\boldsymbol{j}$, we need to recover the hand meshes. This is done by minimizing

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \left\|\mathcal{J}\left(H\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right)\right) - \boldsymbol{j}\right\|_2^2 + \mathcal{L}_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\theta}), \tag{13}$$

where $\mathcal{J}$ is the function which takes hand vertices as input and outputs joint coordinates. The second term of (13) regularizes the shape and pose parameters of MANO,

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = w_1 \left\|\boldsymbol{\beta}\right\|^2 + w_2 \sum_{t=1}^{T} \left\|\boldsymbol{\theta}^t\right\|^2 + w_3 \sum_{t=1}^{T-1} \left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right\|^2 + w_4 \sum_{t=2}^{T-1} \sum_{i=1}^{J} \left\|\ddot{\boldsymbol{j}}_i^t\right\|, \tag{14}$$

where we enforce temporal smoothness by regularizing both the first and the second time derivatives of the joints.

3.5. Data Synthesis

Accurately capturing hand motion sequences, especially in presence of interacting objects, is a particularly challenging task. Sophisticated solutions usually involve expensive marker-based MoCap systems. As a result, there are only few dynamic hand-object interaction datasets available for use. Nevertheless, capturing the hand in a static pose while grasping an object is relatively straightforward. We can have a much larger training set if we are able to utilizing the widely available static hand grasping datasets. In the following, we introduce a simple yet efficient way to synthesize hand sequences from static poses.

Given mesh of a static hand grasping an object, we first fit MANO model to the hand to get the target joint rotations $\boldsymbol{P}^T$, global orientation $\mathbf{R}^T$ and translation $\boldsymbol{d}^T$, where T is the desired sequence length. We then generate a source hand as the first frame of the sequence, where the pose is generated by adding a small random Gaussian noise to the mean MANO pose. Similarly, we perturb $\mathbf{R}^T$ with Gaussian noise to get the global orientation of the initial hand. Next, we compute the average distance moved by the hand per frame from GRAB. The initial translation is determined by moving along the negative normal direction of the target hand palm by this distance.

To obtain hand meshes in intermediate time steps, we apply linear interpolation to hand translation and spherical linear interpolation to joint rotations:

$$\boldsymbol{d}^t = (1-t)\boldsymbol{d}^0 + t\boldsymbol{d}^T \tag{15}$$

$$\boldsymbol{P}^t = \text{SLERP}(\boldsymbol{P}^0, \boldsymbol{P}^T, t) \tag{16}$$

$$\boldsymbol{R}^t = \text{SLERP}(\boldsymbol{R}^0, \boldsymbol{R}^T, t). \tag{17}$$

Generating sequences in this way could result in hand-object intersections. Rather than relying on path planning algorithms to prevent collisions, we simply compute the highest intersection volume of a sequence and eliminate sequences with intersection volume surpassing a predefined threshold. See Figure. 5 for a sample sequence.

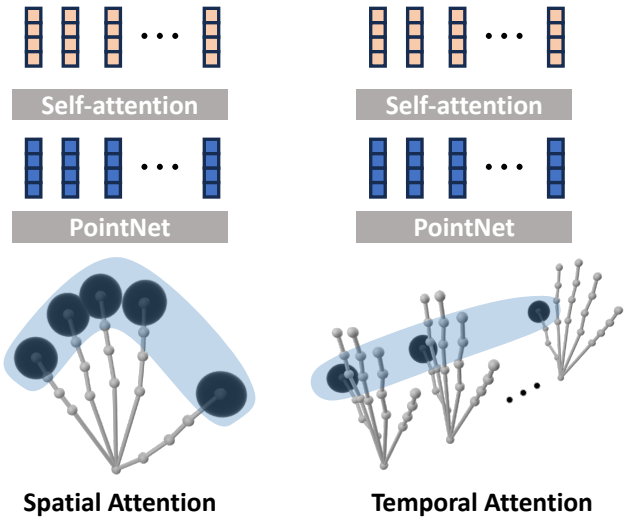

Figure 4. An illustration of spatial and temporal attention networks. We first process the features of each joint by PointNet. For spatial attention, every joint attends to every other joint of the same hand. While for temporal attention, a joint in one frame attends to the same joint in every other frame.

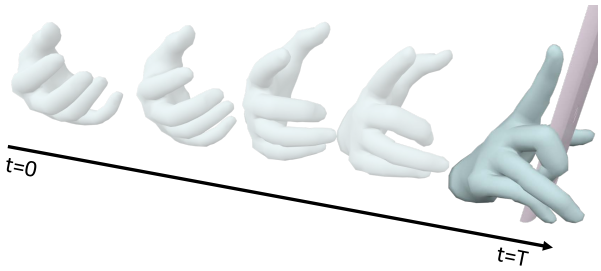

Figure 5. A sample training sequence synthesized by our heuristic rule. At the rightmost side of the time axis is a static grasping pose from ObMan [12]. We synthesize intermediate poses by interpolating joint angles from the mean MANO pose.

3.6. Implementation Details

For Joint Initialization Network, the side length of the cube sensor is 18cm. We sample 2000 points on the partial object mesh as input to PointNet. We uniformly sample 10 frames in the past and in the future within a 1 second time window to compute the trajectory feature. When querying for joint-local object points, we use the sphere sensor with a radius of 2.5cm. A maximum of 300 points are sampled in the neighbourhood of each joint.

4. Experiments

4.1. Datasets

GRAB. We train GEARS on GRAB [30], a large-scale MoCap dataset for whole-body grasping. GRAB contains interaction sequences with 51 objects. Following the official protocol, we select 10 objects for validation and testing, and train with the rest. Due to symmetry of the two hands, we flip left hands to increase the amount of training data. We further augment the training set by transferring grasps to objects of varying sizes, following [48].

InterCap. InterCap is a dataset of whole-body human-scene interaction captured by multiview RGB-D cameras. It features frame-wise pseudo-groundtruth annotations for body, hand and 6D object poses, which are reconstructed by jointly reasoning about human and object contact areas. As we solely focus on hand-object interaction, we consider a subset of objects where the hand is in interaction.

ObMan. ObMan [12] is a static hand grasping dataset. It consists of object models taken from Shapenet [5] and synthetic hand grasps generated by the robotic grasping software GraspIt [23]. Since it only has static hand poses, we cannot directly train on it. Instead, we apply the data synthesis technique and generate 200 sequences for training and testing. Each sequence has a fixed length of 60 frames.

4.2. Metrics

Mean Per-Joint Position Error (MPJPE). We report the average Euclidean distance between predicted and groundtruth 3D hand joints.

Penetration Depth (PD). Penetration depth is the minimum distance required for moving a mesh to make it no longer in intersection with another mesh. We approximate it by finding the maximum vertex-to-object distance for all the penetrating hand vertices.

Intersection Volume (IV). We measure hand-object interpenetration by voxelizing hand and object meshes and reporting the volume of voxels occupied by both. However, interpreting this metric in isolation could be misleading, since it does not account for non-effective grasping artifacts.

Contact IoU (C-IoU). This metric evaluates the Intersection-over-Union between the groundtruth binary hand-object contact map and the contact map of predicted hands. The contact map is defined on the object surface. It takes a value of 1 if a hand vertex is within $\pm 2\text{mm}$ of an object vertex and 0 otherwise.

4.3. Baselines

TOCH [48] is an object-centric model designed for refining noisy hand-object interaction sequences. We tailor it to our task by feeding it with the groundtruth hand trajectory and replacing the noisy hands in the training set with flat hands.

ManipNet [40] relies on both occupancy-based and distance-based sensors to generate dexterous hand motions. Since the original work assumed a different hand model, we adapt it to MANO to compare on a fair ground.

GRIP [32] takes body arm trajectory as input to generate hand poses. It employs a standalone module to denoise the

	MPJPE (mm) ↓	PD (mm) ↓	IV (cm^3) ↓	C-IoU (%) ↑
TOCH	8.18	5.37	2.72	20.1
ManipNet	9.32	5.66	3.21	18.3
GRIP	7.71	4.80	2.51	19.9
GEARS (ours)	**7.24**	**4.36**	**2.24**	**22.7**

Table 1. We quantitatively compare GEARS to other baselines on the GRAB dataset. Each model is trained with the same amount data, including the synthetic sequences generated from ObMan.

	PD (mm) ↓	IV (cm^3) ↓
ManipNet	8.22	6.15
GRIP	7.92	5.68
GEARS (ours)	**7.44**	**5.21**

Table 2. Quantitative comparison on InterCap. We evaluate on a selected subset of objects where hand interaction is involved.

	MPJPE (mm) ↓	C-IoU (%) ↑
$r = 0$	9.34	14.8
$r = 0.02$	7.28	22.1
$r = 0.03$	7.24	21.9
w/o displacement	9.63	13.2
w/o attention	7.85	18.4
w/o synthetic	7.31	20.6
Ours (iterative)	7.37	19.2
Ours (full)	**7.24**	**22.7**

Table 3. Ablaion studies evaluated on GRAB. The variable r refers to the radius of joint-local sensor in millimeters.

arm trajectory and obtain hand trajectory. For fair comparison, we directly provide GRIP with input hand trajectories.

4.4. Quantitative and Qualitative Evaluation

To verify that our GEARS generates realistic interaction sequences, we first evaluate our method on GRAB and compare with the aforementioned baselines. The results are reported in Table. 1. GEARS outperforms other baselines on all four metrics, which clearly demonstrates the advantage of our method. We can observe that although both ManipNet and GRIP rely on distance-based sensors, GRIP achieves better performance both in terms of joint accuracy and inter-penetration score. We hypothesize that it could attribute to the two-stage approach followed by GRIP. Similar to us, GRIP generates a coarse hand first and subsequently refined it. Moreover, TOCH incurs a higher MPJPE but also achieves higher contact IoU than GRIP. This observation shows that a higher joint error doesn't necessarily indicate worse grasping quality. TOCH leverages an object-centric interaction representation, which naturally encourages hand-object contact. See Figure. 6 (top row) for qualitative results on GRAB.

GRAB contains mostly small-to-medium sized household objects. To assess our model's generalization capability to larger objects, we evaluate on the InterCap dataset. We exclude TOCH from this comparison because the object-centric contact map used by TOCH is highly sensitive to object size. Since the groundtruth hand pose annotations of InterCap are not accurate enough, we only report penetration depth and intersection volume, see Table. 2. Compared to GRIP and ManipNet, GEARS incurs less penetration with the objects. Note that all three methods report higher numbers than on GRAB. It can be partially explained by the fact that the input hand trajectory provided by InterCap may exhibit a certain degree of noise. See Figure. 6 (bottom rows) for a qualitative comparison on InterCap.

4.5. Ablation Studies

We ablate different components of GEARS and report the change in performance on GRAB to further justify our proposed method, see Table 3. We first evaluate how sensitive is the model to different sensor radius. Zero radius means that sensor features are neglected by the network. We can observe that as long as the radius is set within a reasonable range, it doesn't have a significant impact on performance.

Moreover, we train three baseline models, for which i) the Joint Displacement Network is removed; ii) spatio-temporal attention is replaced by fully-connected layers; iii) additional synthetic training sequences are not used. It's clear that the Joint Displacement Network plays the most important role in our architecture. This agrees with our intuition that local object geometry features are essential to fine-grained placement of joints.

Lastly, we design an iterative baseline, where at inference time the output of the Joint Displacement Network is fed back to itself as input. We expect that one more round of pose refining would further improve the generation quality. Surprisingly, the iterative refining approach doesn't bring any benefit. We hypothesize that the underlying reason could be distributional shift of test data, since the Joint Displacement Network has only seen the output of Joint Initialization Network during training.

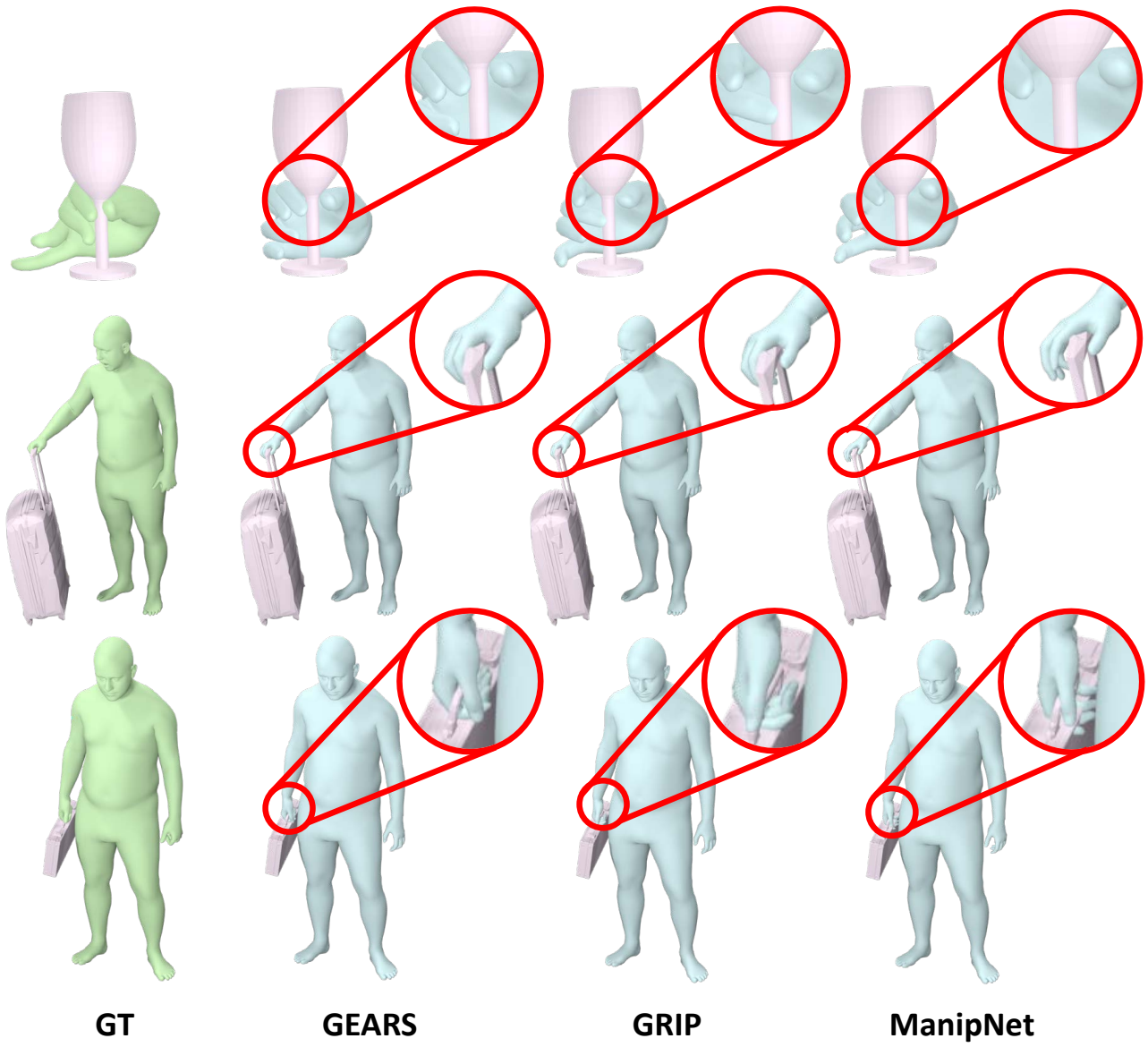

Figure 6. Qualitative results on GRAB (top row) and InterCap (bottom two rows). GEARS makes effective contact with the objects while avoiding hand-object inter-penetration.

5. Conclusion

We present GEARS, a learning-based method for generating hand interaction sequences given hand and object trajectories. The main insight which makes GEARS effective is the novel joint-centered point-based sensor which captures local geometry properties of the target object. Furthermore, we design a spatio-temporal self-attention architecture to process joint-local features and learn the correlation among hand joints during interaction. GEARS is capable of generalizing across objects of varying sizes and categories. We show that GEARS outperforms previous methods in terms of generation quality and generalizability.

Acknowledgements This work is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans). Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The project was made possible by funding from the Carl Zeiss Foundation.

References

[1] Antonio Bicchi. On the closure properties of robotic grasping. *The International Journal of Robotics Research*, 14(4): 319–334, 1995. 3

[2] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2013. 3

[3] 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. 3

[4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 3

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[6] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[7] 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. 3

[8] S El-Khoury, A Sahbani, and P Bidaud. 3d objects grasps synthesis: A survey. In *13th World Congress in Mechanism and Machine Science*, pages 573–583, 2011. 3

[9] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023. 3

[10] 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. 3

[11] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 3

[12] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 3, 6

[13] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. *arXiv preprint arXiv:2104.03304*, 2021. 3

[14] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 3

[15] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *Robotics: science and systems*, 2021. 3

[16] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[17] 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. 3

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

[19] Beatriz León, Stefan Ulbrich, Rosen Diankov, Gustavo Puche, Markus Przybylski, Antonio Morales, Tamim Asfour, Sami Moisio, Jeannette Bohg, James Kuffner, et al. Opengrasp: a toolkit for robot grasping simulation. In *International Conference on Simulation, Modeling, and Programming for Autonomous Robots*, pages 109–120. Springer, 2010. 3

[20] 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. 3

[21] C Karen Liu. Dextrous manipulation from a grasping pose. In *ACM SIGGRAPH 2009 papers*, pages 1–6. 2009. 3

[22] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[23] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 3, 6

[24] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024. 3

[25] Igor Mordatch, Zoran Popović, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 137–144, 2012. 3

[26] Van-Duc Nguyen. Constructing force-closure grasps. *The International Journal of Robotics Research*, 7(3):3–16, 1988. 3

[27] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bod-

ies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3

[28] 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. 3

[29] Karun B Shimoga. Robot grasp synthesis algorithms: A survey. *The International Journal of Robotics Research*, 15(3): 230–266, 1996. 3

[30] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pages 581–600. Springer, 2020. 3, 6

[31] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[32] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using latent consistency and spatial cues. 2024. 2, 3, 6

[33] 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

[34] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 3

[35] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022. 3

[36] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[37] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023. 3

[38] 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. 3

[39] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 3

[40] 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, 3, 6

[41] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *International Conference on 3D Vision (3DV)*, 2024. 3

[42] 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. 3

[43] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022. 3

[44] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. Robust realtime physics-based motion control for human grasping. *ACM Transactions on Graphics (TOG)*, 32 (6):1–12, 2013. 3

[45] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 585–594, 2023. 3

[46] Yu Zheng and Wen-Han Qian. Coping with the grasping uncertainties in force-closure analysis. *The international journal of robotics research*, 24(4):311–327, 2005. 3

[47] Yanzhao Zheng, Yunzhou Shi, Yuhao Cui, Zhongzhou Zhao, Zhiling Luo, and Wei Zhou. Coop: Decoupling and coupling of whole-body grasping pose generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2163–2173, 2023. 3

[48] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3, 6

[49] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15741–15751, 2021. 3