Template Free Reconstruction of Human-object Interaction with Procedural Interaction Generation

Xianghui Xie[1,2,3] Bharat Lal Bhatnagar[3] Jan Eric Lenssen[3] Gerard Pons-Moll[1,2,3]

[1]University of Tübingen, Germany [2]Tübingen AI Center, Germany
[3]Max Planck Institute for Informatics, Saarland Informatic Campus, Germany

https://virtualhumans.mpi-inf.mpg.de/procigen-hdm/

Abstract

Reconstructing human-object interaction in 3D from a single RGB image is a challenging task and existing data driven methods do not generalize beyond the objects present in the carefully curated 3D interaction datasets. Capturing large-scale real data to learn strong interaction and 3D shape priors is very expensive due to the combinatorial nature of human-object interactions. In this paper, we propose ProciGen (Procedural interaction Generation), a method to procedurally generate datasets with both, plausible interaction and diverse object variation. We generate 1M+ human-object interaction pairs in 3D and leverage this large-scale data to train our HDM (Hierarchical Diffusion Model), a novel method to reconstruct interacting human and unseen object instances, without any templates. Our HDM is an image-conditioned diffusion model that learns both realistic interaction and highly accurate human and object shapes. Experiments show that our HDM trained with ProciGen significantly outperforms prior methods that require template meshes, and our dataset allows training methods with strong generalization ability to unseen object instances. Our code and data are released.

1. Introduction

Modelling interactions between humans and their surroundings is important for applications like creating realistic avatars, robotic control and gaming. In this paper, we address the task of jointly reconstructing human and object from a monocular RGB image, without any prior object templates. This is very challenging due to depth-scale ambiguity, occlusions, diverse human pose and object shape variations. Data-driven methods have shown great progress in reconstructing humans [31, 41, 46, 69–71, 85] or objects [55, 112] from monocular inputs thanks to large-scale datasets [1, 9, 12, 19, 37, 63, 94, 108]. However, methods for joint interaction reconstruction are still constrained by the amount of available data. Recent datasets like BEHAVE [7], InterCap [35] capture real interactions with 10 to 20 different objects, which is far away from the number of objects in reality: the chair category from ShapeNet [12] alone has more than 6k different shapes. Training on these real datasets has limited generalization ability to unseen objects (Sec. 4.3). Capturing real interaction data with more objects is prohibitively expensive due to the combinatorial nature: the number of humans times the number of objects leads to a huge number of variations. This motivates us to generate synthetic data which has been shown effective for pre-training reconstruction methods [9, 29, 55, 63, 70].

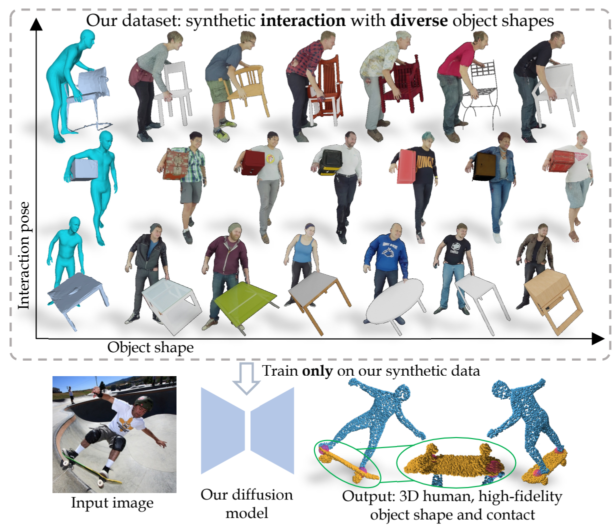

Figure 1. Given a single RGB image, our method trained *only* on our proposed synthetic interaction dataset, can reconstruct the human, object and contacts, without any predefined template meshes.

Synthesizing realistic interaction for different objects is non-trivial due to variations of object topology, geometry details and complex interaction patterns. To address this, we propose **ProciGen**: **Proc**edural **i**nteraction **Gen**eration, a method to generate interaction data with diverse object shapes. We design our method based on the key idea that the way humans interact with objects of the *same* category is similar. And despite the geometry variations, one can still establish semantically meaningful correspondence between different objects. More specifically, we train an autoencoder to obtain correspondences between different objects from the same category, which are then used to transfer contacts from already captured human-object interactions to new object instances. Our method is scalable and allows the *multiplicative combination* of datasets to generate over a million interactions with more than 21k different object instances, which is not possible via real data capture.

Current reconstruction methods [7, 60, 95, 96] are not only bottle-necked by data. Template-based methods [7, 95, 96] cannot generalize to unseen objects as they are *trained only for specific object templates*. Template free methods like PC^2 [60] cannot separate human and object, and have limited shape accuracy. See Tab. 1 for detailed comparison. To alleviate these issues, we propose **HDM**: **H**ierarchical **D**iffusion **M**odel, that predicts accurate shapes and reasons about human-object semantics without using templates. Our key idea is to decompose the combinatorial interaction space into separate human and object sub-spaces while preserving the interaction context. We first use a diffusion model to jointly predict human, object and segmentation labels, and then use two separate diffusion models with cross attention that further refine the separate predictions.

We evaluate our data generation method ProciGen, and model HDM, on BEHAVE [7] and InterCap [35]. Experiments show that HDM with ProciGen significantly outperforms CHORE [95] (which requires object templates) and PC^2 [60]. Our ProciGen dataset also significantly boosts the performance of PC^2 and HDM. Methods trained on our synthetic ProciGen dataset show strong generalization ability to real images even though the objects are unseen.

In summary, our key contributions are:

- We introduce the first procedural interaction generation method for synthesizing large-scale interaction data with diverse objects. With this, we generate 1M+ interaction images with 21k+ objects paired with 3D ground truth.
- We propose a hierarchical diffusion model that can faithfully reconstruct human and object shapes from monocular RGB images without relying on template shapes.
- Our dataset and code are publicly released.

2. Related Work

Interaction Capture. Modelling 3D interactions has been an emerging research field in recent years, with works that model hand-object interaction from RGB [17, 22, 29, 42, 102], RGBD [10, 11, 26] or 3D [42, 80, 83, 114, 116] input, or predict contacts from RGB images [14, 34, 86] and works that model human-scene interaction from single image [8, 28, 47, 74, 101] or video [24, 105]. A recent line of works model full body interacting with dynamic large objects [25, 27, 39, 40, 44, 53, 64, 78, 91, 100, 113]. BEHAVE [7] and follow up works [35, 109] capture interaction datasets, which allow training and benchmarking methods [97] for reconstructing 3D human-object from single RGB images [90, 95, 110] or videos [96]. Despite impressive results, they require predefined mesh templates, which limits applicability to new objects. Our method is template-free and generalizes well to unseen objects.

Method	No-template	Shape acc.	General.	Semantic
CHORE	X	✓*	X	✓
PC^2	✓	X	X	X
PC^2 + Our ProciGen	✓	X	✓	X
Ours	✓	✓	✓	✓

Table 1. **Comparison of different reconstruction methods.** CHORE [95] reconstructs high shape fidelity with known template meshes but does not generalize to new object instances. PC^2 [60] is template-free but its shape predictions lack fidelity and generalization ability is constrained by existing datasets. Training PC^2 with our ProciGen dataset allows better generalization but it cannot reason contacts. Our proposed data generation together with our hierarchical diffusion model can predict accurate shapes, generalize to unseen objects and reason about interaction semantics.

Synthetic Datasets are powerful resources to deep networks. For humans, synthetic rendering of 3D scans [1, 4, 63, 84, 103] are used extensively to train human reconstruction methods [5, 6, 18, 70, 71, 81, 98, 99]. Recent work BEDLAM [9] showed that training purely on synthetic datasets [9, 63] allows strong generalization. Orthogonal to these, large scale 3D object CAD model datasets [12, 94] are also used to pretrain backbone models[38, 51, 62, 107]. Other works [20, 23, 67, 93] consider generating diverse scenes. While being useful for humans, objects or scenes respectively, they do not consider interactions. Our proposed approach can generate millions of interactions with diverse object shapes, allowing for training interaction reconstruction models with great generalization ability.

Diffusion-based Reconstruction. Diffusion models [32, 76] have been shown powerful for 3D reconstruction of human [36, 48] and objects [55, 60, 72]. These works distil pretrained 2D diffusion model [36, 48, 59, 65, 72, 104, 118] or fine-tune diffusion model [54, 55, 73] for 3D reconstruction from images. Recent works also propose image-conditioned point diffusion models for reconstruction [60, 87]. Despite remarkable results, they only model the distribution of single shapes, while our method can learn the complex interaction space with high shape fidelity.

3. Method

We first introduce our method to generate large amounts of interaction data with diverse object shapes in Sec. 3.1. This data allows us to train our novel diffusion model with strong

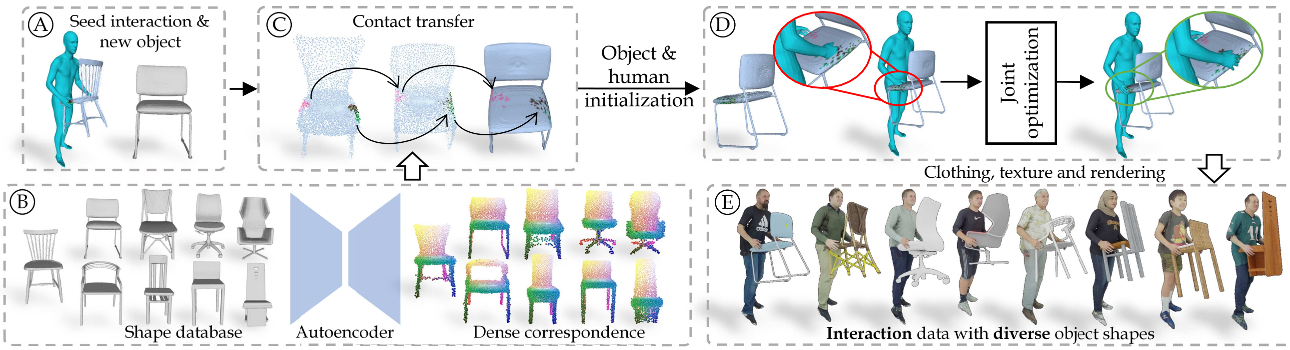

Figure 2. **Our procedural interaction generation method.** Given a seed interaction and a new object from the same category (A), we use a network to compute dense correspondences (B, Sec. 3.1.1), which allows us to transfer contacts and initialize the new object (C, Sec. 3.1.2). We further optimize the human and object poses to avoid interpenetration while satisfying the transferred contacts (D, Sec. 3.1.3). We then add clothing and textures to render images, leading to a large interaction dataset with diverse object shapes (E, Sec. 3.1.4).

generalization ability, which is explained in Sec. 3.2.

3.1. ProciGen: Procedural Interaction Generation

Given a small seed dataset of captured human-object interactions and datasets of various object models, we aim to generate a large-scale interaction dataset with diverse object shapes. Via *multiplicative scaling*, it would allow generating enormous data which is not possible by capturing real data. This is however non-trivial as object geometry varies strongly even within one category. Therefore, we propose a procedural method based on the key observation that humans interact similarly with objects of the same category. By transferring contacts from captured interactions to new object instances, we procedurally scale up the shape variations of real interaction datasets. The task involves solving four different sub-problems, as outlined in Fig. 2:

1. **Establishing dense semantic correspondences** between all objects within one category (Sec. 3.1.1).
2. **Transferring contacts** from real to synthetic objects, using the obtained correspondences (Sec. 3.1.2).
3. **Jointly optimizing human and object** to the newly obtained contacts under a set of constraints (Sec. 3.1.3).
4. **Rendering** novel intersection pairs with textures to make them available as training data (Sec. 3.1.4).

3.1.1 Dense Semantic Correspondence

Given two meshes $\mathcal{M}$ and $\mathcal{M}'$ of two different objects of the same category, the problem of finding dense correspondence amounts to finding a bijective map $\psi : \mathcal{M} \to \mathcal{M}'$, which maps points from one mesh to their semantic counterparts on the other. In cases of arbitrary meshes with changing topology, this problem is heavily ill-posed [3, 21, 79]. Thus, we turn to an approximate solution on discrete surface samples that leverages the regularization and output ordering of MLPs[50] and works well on a wide range of input topologies in practice.

Let $\{\mathcal{M}_i\}_{i=1}^M$ be a dataset of meshes from the same object category and $\mathbf{P}_i \in \mathbb{R}^{N\times 3}$ a point cloud sampled from the surface of $\mathcal{M}_i$. We train an autoencoder $f : \mathbb{R}^{N\times 3} \mapsto \mathbb{R}^{N\times 3}$ on $\{\mathbf{P}_i\}_{i=1}^M$ to minimize the Chamfer distance between predicted and input point clouds. The network f consists of a PointNet [66] encoder and a three-layer MLP decoder that takes unordered points as input and output ordered points. We found that the MLP decoder learns to reconstruct the objects as a mixture of low-rank point basis vectors, thus it automatically provides dense correspondence across objects through the order in the output, as also found in [79, 92, 115]. Effective training of this network requires all shapes to be roughly aligned in a canonical space. When shapes are not aligned, we use ART [115] which uses an additional network to predict an aligning rotation.

To ensure the reconstruction quality, we overfit one network per object category. We show some example reconstructions and correspondences for chairs in Fig. 2B.

3.1.2 Contact Transfer

Given dense correspondences between a set of point clouds, we use them to transfer contact maps from one object to the other. Let $(\mathbf{H} \in \mathbb{R}^{M\times 3}, \mathbf{P} \in \mathbb{R}^{N\times 3})$ be a pair of human and object point clouds from an existing interaction dataset. And let $\mathbf{T} \in SE(3)$ be the non-rigid transformation that brings the object point cloud into canonical space where shapes are roughly aligned. Then, we can find our *contact set* as a set of point pairs from human and object that lie within a distance σ to each other:

$$\mathcal{C} = \{(i,j) \mid ||\mathbf{H}_i - \mathbf{T}^{-1} f(\mathbf{TP})_j)||_2^2 < \sigma\}. \quad (1)$$

We first bring $\mathbf{P}$ into canonical pose, then apply f to obtain a coherent point cloud, which is brought back into interaction pose by $\mathbf{T}^{-1}$. Since our autoencoder f produces coherent point clouds, the obtained contact set can be directly transferred to all other objects $\mathbf{P}'$ within the category, allowing us to pair the human with all other objects, one example transfer is shown in Fig. 2C. Once we transfer the contact points to the new object, we can find the corresponding contact facets in the meshes that have the smallest distances.

3.1.3 Contact-based Joint Optimization

The newly obtained contact sets define how and where a human should interact with the new object. We can also transform the object from canonical to interaction pose with our dense correspondence. However, this naive placement does not guarantee the plausibility of the interaction due to object geometry changes (see Fig. 2D). Hence, we propose a joint optimization to refine the human and object pose such that: **a)** contact points are close to each other, **b)** contact face normals match, and **c)** interpenetration is avoided.

We use the SMPL-H [57, 68] body model $H(\boldsymbol{\theta}, \boldsymbol{\beta})$ to parameterize the human as a function of pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ parameters. The object pose is given as non-rigid transformation $\mathbf{T} \in SE(3)$, and we denote the new object point cloud to which we have transferred contacts as $\mathbf{P}'$. We find the refined human-object poses jointly, by minimizing:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}) = \lambda_c L_c + \lambda_n L_n + \lambda_{\text{colli}} + \lambda_{\text{init}} L_{\text{init}}, \quad (2)$$

where the individual loss terms are given as:

- **Contact:** $L_c = \sum_{(i,j)\in\mathcal{C}} ||\mathbf{H}_i - \mathbf{P}'_j||_2^2$, minimizing the distance between contact points.
- **Normal:** $L_n = \sum_{(i,j)\in\mathcal{C}} ||1 + \mathbf{n}_i^T \mathbf{n}_j||_2^2$, ensuring that normals $\mathbf{n}_i, \mathbf{n}_j$ of contacting faces point in opposite directions.
- **Interpenetration:** L_{colli} penalizing interpenetration based on the bounding volume hierarchy [88].
- **Initialization:** L_{init} is the L2 distance between new and original human pose, regularizing the deformation.

The pose $\boldsymbol{\theta}$ is initialized from the original human pose and $\boldsymbol{\beta}$ is randomly sampled from a set of registered scans [4]. The object pose $\mathbf{T}$ is initialized by Procrustes alignment between the two coherent point clouds $\mathbf{P}'$ and $\mathbf{P}$. After joint optimization we obtain realistic interactions, see Fig. 2D.

3.1.4 Dataset Rendering

Our contact transfer and joint optimization provide us the skeleton of interaction with new objects. To render them as images, we take the optimized SMPL-H parameters from Sec. 3.1.3 and randomly sample the clothing deformation and texture from SMPL+D registrations in MGN [4]. For objects, we use the original texture paired with the mesh. We render the scenes in Blender [16], which is detailed in supplementary. See example renderings in Fig. 2E.

Method Scalability. We emphasize that the proposed procedural generation is a scalable solution that can generate large-scale datasets with only a small amount of effort for data capture: with 2k different interactions (e.g. BEHAVE [7] chair interaction), 6k different objects (e.g. Shapenet chairs [12]) and 100 human scans (e.g. MGN [4]), one can have maximum *1.2 billion* different variations in total, which is not possible with real data capture. The data scale allows for training powerful models that reach performance not obtainable by training on real data only. An example of such a method is detailed in the next section.

3.2. HDM: Hierarchical Diffusion Model

Modelling the joint shape space of humans interacting with objects is difficult since the product of human and object shape variations is huge. One solution is to use two separate networks that reconstruct human and object respectively. However, such a method ignores the interaction cues that have been show important for coherent reconstruction [7, 95, 96, 105]. This motivates us to design a hierarchical solution where we first jointly estimate both human and object(Sec. 3.2.2), and then use separate networks that focus on refining individual shape details (Sec. 3.2.3). An overview of our method can be found in Figure 3.

3.2.1 Preliminaries

Task Overview. Given an input RGB image $\mathbf{I}$ of a person interacting with an object, we aim to jointly reconstruct 3D human and object point clouds $\mathbf{P}^h, \mathbf{P}^o$. Same as prior works [95, 96, 110], we assume known 2D human and object masks, which we consider a weak assumption, given recent advances in 2D segmentation [43, 45, 75].

Due to the ambiguity from monocular input, we adopt a probabilistic approach for 3D reconstruction, which has been proven effective in learning multiple modes given same input [60, 111, 118]. Specifically, we use a diffusion model [32] to learn the distribution of 3D human object interactions conditioned on a single image.

Diffusion models [32, 76] are general-purpose generative models that consist of iterative forward and reverse processes. Formally, given a data point $\mathbf{x}_0$ sampled from a data distribution p_{data}, the forward process iteratively adds Gaussian noise $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ to the sample $\mathbf{x}_0$. The distribution at step t can be computed as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \epsilon\sqrt{1 - \bar{\alpha}_t} \quad (3)$$

where $\bar{\alpha}_t$ controls the noise level at step t and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ [32]. The reverse process starts from Gaussian noise at step T and gradually denoises it back to the original data distribution p_{data} at step 0. At each reverse step, we use a neural network p_θ to approximate the distribution: $p_\theta \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The network is trained with the variational lower bound to maximize the log-likelihood of all

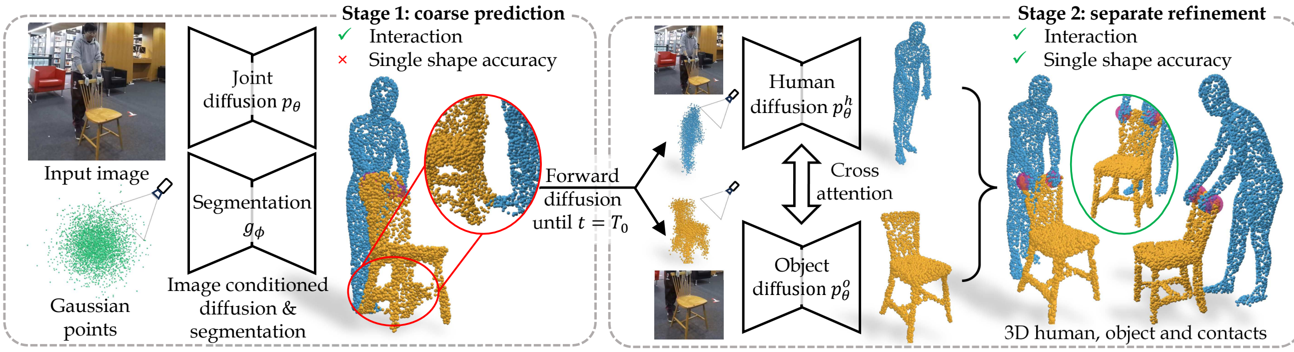

Figure 3. **Our hierarchical diffusion model.** Given an RGB image of a human interacting with an object, we first jointly reconstruct the human and object as one point cloud with segmentation labels (Stage 1, Sec. 3.2.2). This prediction reasons interaction but lacks accurate shapes. We then use two diffusion models for human or object separately with cross attention to refine the initial noisy prediction while preserving the interaction context(Stage 2, Sec. 3.2.3). Our hierarchical design faithfully predicts interaction and shapes.

data points, which is parametrized to minimize the L2 distance between the true noise ϵ and network prediction[32]:

$$\mathcal{L} = E_{t\sim[1,T]}E_{\epsilon_t\sim\mathcal{N}(0,\mathbf{I})}[||\epsilon_t - p_\theta(\mathbf{x}_t, t)||_2^2] \tag{4}$$

3.2.2 Joint Human-object Diffusion

In this first stage, we simultaneously predict both human and object hence the output is one point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$. We adopt PC^2 [60] that diffuses point cloud conditioned on single images. Formally, we use a point voxel CNN [56, 117] $p_\theta : \mathbb{R}^{N\times D} \mapsto \mathbb{R}^{N\times 3}$ as the point diffusion model. Here D is the feature dimension. To obtain per-point input features, we first use a pre-trained encoder [30] to extract feature grid $\mathbf{F} \in \mathbb{R}^{F\times H'\times W'}$ from input image $\mathbf{I}$, here F and H', W' are feature and spatial dimensions respectively. Points $\mathbf{p} \in \mathbf{P}$ are then projected to 2D image plane with $\pi(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}^2$ to extract pixel-aligned feature $\mathbf{F}_{\pi(\mathbf{p})}$. We further concatenate it with point location and diffusion timestamp encodings t_{enc} as the input to the diffusion model: $\mathbf{F}_\mathbf{p} = (\mathbf{F}_{\pi(\mathbf{p})}, \mathbf{p}, t_{\text{enc}})$. To allow generative prediction for points that are occluded, the image features $\mathbf{F}_{\pi(\mathbf{p})}$ are set to zeros when points are occluded [60].

3.2.3 Hierarchical Diffusion for Interaction

Naively using one network to reconstruct interaction leads to noisy point predictions (see Fig. 5), as the combinatorial shape space of human-object interaction is too complex to model. Thus, we propose a second stage to refine human and object shapes separately, by having two additional diffusion models while also preserving the interaction context. In the following, we discuss special aspects of our second stage, namely **1)** how the point cloud is segmented into human and object, **2)** how separate networks are designed to model interaction, **3)** how these models are combined.

Point cloud segmentation. To reason the contacts during interaction and obtain accurate shapes for human and object separately, the combined point cloud needs to be segmented into the points for human and object. To this end, we use an additional network $g_\phi : \mathbb{R}^{N\times D} \mapsto \{0,1\}^N$ that takes point features $\mathbf{F}_\mathbf{p}$ as input and predicts a binary label to indicate whether this is a human or object point. With this prediction, we can segment the point cloud $\mathbf{P}$ predicted by p_θ into human and object points $\mathbf{P}^h, \mathbf{P}^o$.

Preserving interaction context. In our second stage, we use two additional diffusion models p_θ^h, p_θ^o to predict human and object. The networks follow the same design as the joint network p_θ using PVCNN [56, 117]. To encourage the networks to explore interaction cues, we add cross-attention layers between the encoder and decoder layers of human and object branches. Given downsampled points $\mathbf{P}_l \in \mathbb{R}^{N_l\times 3}$ with features $\mathbf{F}_l \in \mathbb{R}^{N_l\times D_l}$ after network layer l, we propagate information from human branch to object branch by computing feature:

$$\mathbf{F}_l^{h\mapsto o} = \text{Attn}(\text{enc}(\mathbf{P}_l^o), \text{enc}(\mathbf{P}_l^h), \mathbf{F}_{\mathbf{P}_l^h}), \tag{5}$$

where $\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is learnable cross attention[89], enc$(\cdot)$ is positional encoding from NeRF [61], and $\mathbf{F}_{\mathbf{P}_l^h} = (\text{enc}(\mathbf{P}_l^h), \mathbf{C})$ is the concatenation of positional encoding and onehot encoding $\mathbf{C}$ indicating these points belong to human. The attention feature $\mathbf{F}_l^{h\mapsto o}$ is then concatenated to the object feature $\mathbf{F}_l^o$ as input to the next layer. We propagate information from object to human branch similarly.

Model Combination. With the separate networks p_θ^h, p_θ^o, one can run the full reverse diffusion process from $t = T$ to $t = 0$ and then combine the denoised points. However, this does not leverage the predicted interaction context from the joint reconstruction stage and is slow. We hence start the reverse diffusion steps from an intermediate step $t = T_0$ instead of T. Specifically, after denoising

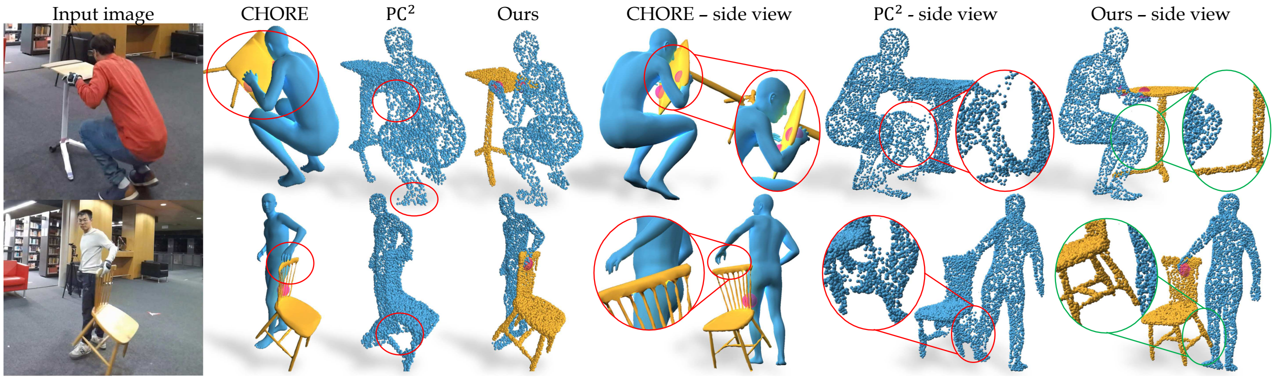

Figure 4. **Comparing reconstruction results on BEHAVE[7] dataset.** CHORE[95] relies on object mesh templates and the prediction is inaccurate for challenging poses. PC^2[60] does not rely on templates but its predicted point clouds are noisy (red circles) and it cannot predict contacts. Ours can reason about human object interaction, and predicts high-fidelity human and object shapes without templates.

and segmentation with the joint model, we apply the forward diffusion process to $\mathbf{P}^h$ and $\mathbf{P}^o$ until step $t = T_0$ using Eq. (3). Then, the individual diffusion models take the noised points as input and gradually denoise them until step $t = 0$. The forward process destroys local noisy predictions but keeps the global structure of human-object interaction. We set $T_0 = \frac{T}{2}$, see supp. for analysis of this value. Our hierarchical design is important to obtain sharp predictions, see Tab. 5 and Fig. 5.

Recall from Eq. (3) that the forward diffusion ends up with a normal distribution. Hence the input and output points of all diffusion models are centered at the origin and scaled to unit sphere, which requires normalization parameters to project them back to image. We estimate it for the first diffusion model p_θ when GT is not available and compute them for separate diffusion models p_θ^h, p_θ^o from the segmented points. We show in Sec. 4.4 that it is better than directly predicting from input image. More details in Supp.

Implementation. We train our diffusion models $p_\theta, p_\theta^h, p_\theta^o$ using the standard loss (Eq. (4)) and segmentation model g_ϕ using L2 distance between predicted and ground truth binary labels. See Supp. for more implementation details.

4. Experiments

In this section, we first describe our data generation and then evaluate the proposed ProciGen data and HDM for reconstruction. Please refer to supp. for implementation details.

Data generation. We leverage the BEHAVE [7], InterCap [35], ShapeNet [12], Objaverse [19], ABO [15] and MGN [4] dataset to generate our synthetic data ProciGen. BEHAVE and InterCap capture multi-view images of humans interacting with 20 and 10 different objects respectively. ShapeNet [12], Objaverse [19] and ABO [15] provide 3D object models as meshes paired with textures. The objects from ShapeNet and ABO are aligned in canonical space while objects from Objaverse are not aligned. MGN [4] consists of 100 human scans paired with SMPL-D registration that allows reposing scans while preserving clothing deformation.

Following the same split from [96], we randomly sample from 380k interactions in BEHAVE and InterCap training set, 21k different shapes in ShapeNet, ABO and Objaverse, and 100 different human shapes and textures in MGN. In total, we generate $\sim$1.1million training images. Please see supplementary for more data distribution details.

Evaluation metric. We evaluate the reconstruction performance using the F-score based on Chamfer distance between point clouds, which is more suitable for measuring the shape accuracy [82]. We compute F-score with a threshold of 0.01m [60] and report the error for human, object and combined point clouds separately, as typically done in interaction reconstruction methods [95, 96].

4.1. Reconstruction on BEHAVE and InterCap

We compare our method with CHORE [95] and PC^2[60] on BEHAVE[7] and InterCap [35] test set in Tab. 2 and Fig. 4. We train CHORE and PC^2 on the training set of BEHAVE and InterCap. Our HDM is trained on our synthetic ProciGen with or without BEHAVE and InterCap training set. We also report per-category accuracy in supplementary.

CHORE is designed for interaction reconstruction and requires *known* object templates. PC^2 is a general shape reconstruction method without templates but it does not separate human and object hence cannot reason the semantics of interaction. Our method trained only on our synthetic ProciGen dataset performs on par with CHORE which already knows the template and PC^2 which already sees the object shapes. After training our HDM on both our ProciGen and real data, our method significantly outperforms baselines.

4.2. Contribution of our ProciGen and HDM

We propose ProciGen for interaction data generation and HDM for interaction reconstruction. To decouple the con-

	Method	Human↑	Object↑	Comb.↑
BEHAVE	CHORE†	0.3454	0.4258	0.3966
	PC2†‡	X	X	0.4231
	Ours‡	**0.3925**	**0.5049**	**0.4604**
	Ours synth. only‡	0.3477	0.4351	0.4110
InterCap	CHORE†	0.4064	0.5135	0.4687
	PC2†‡	X	X	0.5057
	Ours‡	**0.4399**	**0.6072**	**0.5344**
	Ours synth. only‡	0.3851	0.4928	0.4530

Table 2. **Reconstruction results** (F-sc.@0.01m) on BEHAVE [7] and InterCap [35]. † denotes methods with template meshes while ‡ denotes template-free methods. CHORE [95] requires known object templates and is prone to noisy pose predictions. PC^2 [60] does not require templates but cannot predict semantics of human-object and the prediction is inaccurate. Our method separates human and object, does not require any templates and outperforms PC^2 and CHORE. Training only on our synthetic ProciGen data performs on par with CHORE even it has never seen the objects.

Method	Human	Object	Combined
a. PC^2	X	X	0.4231
b. Our HDM	0.3605	0.4575	0.4214
c. PC^2 + our ProciGen	X	X	0.4486
d. Our HDM + ProciGen	**0.3925**	**0.5049**	**0.4604**

Table 3. **Decoupling the contribution** of our ProciGen dataset and reconstruction method. Our ProciGen dataset significantly boosts performance of both PC^2 (c) and our method (d) compared to training on BEHAVE only (a-b). Both our ProciGen and HDM model are important to achieve the best result.

tribution of our data and method, we compare our method against PC^2 [60] trained on BEHAVE [7] only (Tab. 3 a-b) and BEHAVE + our ProciGen dataset (Tab. 3 c-d). The methods are evaluated on the BEHAVE test set. It can be seen that both our proposed data and model are important to obtain the most accurate reconstruction. We also report the model performance vs. data amount in supplementary.

4.3. Generalization Performance

Our ProciGen dataset allows training shape reconstruction methods to generalize to unseen object instances. To evaluate this, we train CHORE[95], PC^2[60] and our HDM model on BEHAVE[7] and our proposed dataset respectively. We then evaluate them on unseen objects of the same categories from InterCap [35] in Table 4. CHORE requires a template to predict 6D pose, which makes it difficult to train on our synthetic dataset with more than 21k different shapes. We hence only train CHORE on BEHAVE dataset.

Methods trained on BEHAVE have limited generalization to InterCap (Tab. 4a-c). An alternative to our ProciGen is to randomly scale and shift the objects from BEHAVE and render new images, which only slightly improves generalization (Tab. 4d). In contrast, our ProciGen significantly boosts the generalization performance (Tab. 4e-f).

Method	Human ↑	Object↑	Combine↑
a. CHORE	0.2263	0.1924	0.2176
b. PC^2	X	X	0.2327
c. Our HDM	0.2389	0.1592	0.2127
d. Our HDM+ augm.	0.3076	0.2089	0.2680
e. PC^2 + Our ProciGen	X	X	0.3843
f. Our HDM+ ProciGen	**0.3502**	**0.4233**	**0.3976**

Table 4. **Generalization performance** of methods trained on BEHAVE [7] (a-c), BEHAVE + random augmentation (d) and our ProciGen (e-f), evaluated on unseen objects from InterCap (F-score@0.01m). CHORE predicts template-specific 6D poses hence does not work on unseen objects from InterCap. PC^2 (b) and our method (c) do not require templates but are constrained by the limited shape variations from BEHAVE. Adding random shape augmentation on BEHAVE objects (d) slightly improves generalization but is still suboptimal. With our proposed ProciGen dataset, both PC^2 and our method can generalize to InterCap and our method achieves better accuracy.

Some qualitative results are shown in Fig. 5. Our method reconstructs human and object with high shape fidelity. We also show the generalization results to COCO dataset [49] in Fig. 6. Our method trained *only* on our ProciGen data generalizes well to in-the-wild images with large object shape variations. See Supp. for more generalization examples.

4.4. Ablating the Hierarchical Diffusion Model

Our HDM predicts interaction semantics and better shapes. In Tab. 5, we ablate alternative designs to our method on the 824 chair images from BEHAVE test set [7] due to resource limit. All methods are trained on our ProciGen dataset.

The human-object segmentation allows us to compute the contacts and manipulate human and object separately. An alternative is projecting the predicted points to 2D image and segment points based on the masks. Due to occlusion and complex interaction, this segmentation is inaccurate, as reflected in the large human and object errors in Tab. 5a. The model that predicts human, object and segmentation with a single model (Tab. 5 b) also does not work as it is difficult to learn high-fidelity interaction shapes.

Another alternative to our first joint diffusion model is to use a network that predicts translation and scale directly from input image and then use them to combine predictions from two separate models. However, such a global prediction does not model interaction with local fine-level details hence the performance is subpar(Tab. 5b). Our cross attention module also improves the performance (Tab. 5d).

5. Conclusion

In this paper, we proposed a procedural generation method to synthesize interaction datasets with diverse human and object shapes. This method allows us to generate 1M+ images paired with clean 3D ground truth and train large image-conditioned diffusion models for reconstruction,

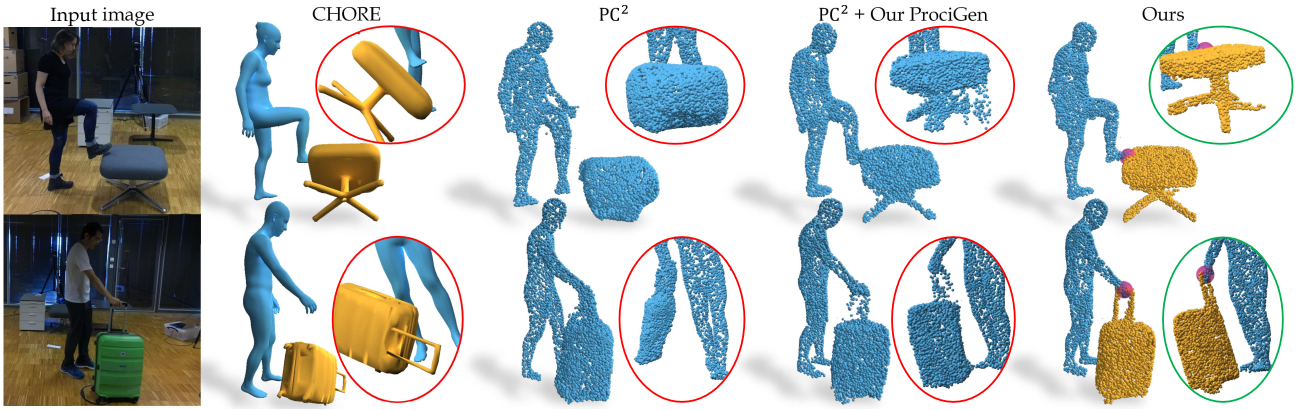

Figure 5. **Generalization results to InterCap [35] dataset.** Note that all object instances are unseen during training time. CHORE [95] predicts template specific object pose hence cannot generalize to new object instances. PC^2[60] does not rely on template but its generalization ability is constrained by limited shape variations from BEHAVE [7]. Training PC^2 on our ProciGen improves its generalization but the predicted point clouds are still noisy. Our method is able to generalize and predicts human and object with high shape fidelity.

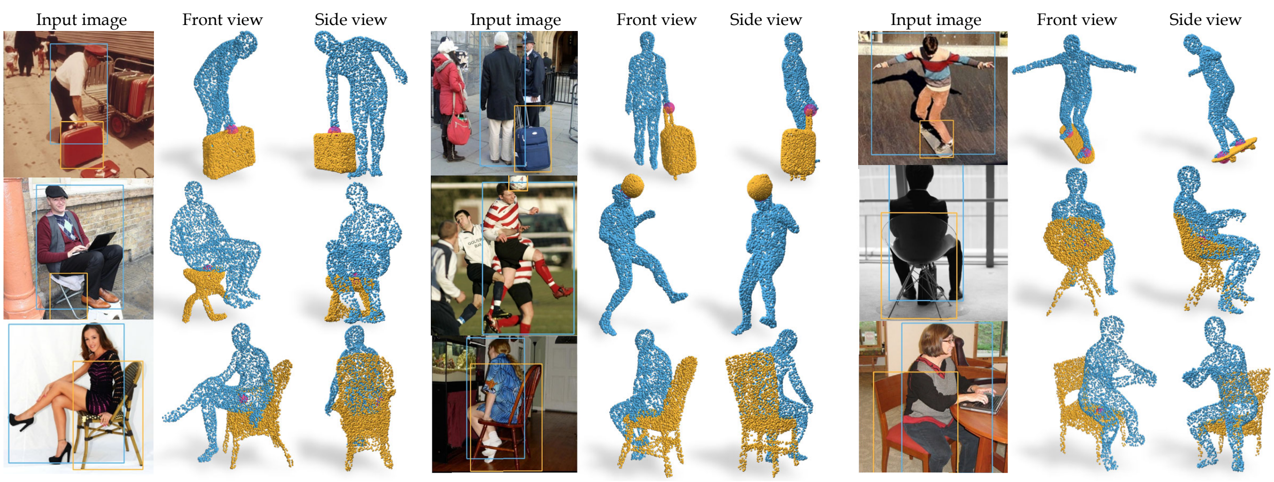

Figure 6. **Testing our method on COCO [49] dataset.** Human and object to be reconstructed are highlighted with blue and yellow box respectively. Our method generalizes to diverse objects from in the wild images *without any shape templates*.

Method (with our ProciGen)	Hum.↑	Obj.↑	Comb.↑
a. PC^2 + projected segm.	0.2961	0.3436	0.3776
b. Single model + segm.	0.3349	0.3638	0.3743
c. Direct pred. + sep. models	0.2809	0.3487	0.3380
d. Ours w/o cross attention	0.3387	0.3806	0.3807
e. Our full model	**0.3433**	**0.3916**	**0.3875**

Table 5. Ablating alternative methods to our HDM (F-score@0.01m). Projecting PC^2 predictions to 2D masks to obtain segmentation (a) is inaccurate and single stage diffusion model (b) cannot learn high-fidelity shapes for both human and object. Combining predictions from separate human and object models using direct translation prediction from images (c) also does not work as it cannot learn fine-grained interactions. Our hierarchical design together with our cross attention module achieves the best result.

without relying on any shape templates. To learn accurate shape space for human and object, we introduce a hierarchical diffusion model that learns both the joint interaction and high fidelity human and object shape subspaces.

We train our method with the proposed synthetic dataset and evaluate it on BEHAVE and InterCap datasets. Results show that our method significantly outperforms CHORE which requires template meshes and PC^2 which does not reason interaction semantics. Ablation studies also show that our synthetic dataset is important to boost the performance and generalization ability of both PC^2 and our model. Our method generalizes well to real images from COCO that have diverse object geometries, which is a promising step toward real in-the-wild reconstruction. Our code and data are released to promote future research.

Acknowledgements. We thank RVH group members [2], especially Yuxuan Xue, for their helpful discussions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), and German Federal Ministry of Education and Research

(BMBF): Tübingen AI Center, FKZ: 01IS18039A, and Amazon-MPI science hub. Gerard Pons-Moll is a Professor at the University of Tübingen endowed by the Carl Zeiss Foundation, at the Department of Computer Science and a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] https://renderpeople.com/. 1, 2

[2] http://virtualhumans.mpi-inf.mpg.de/people.html. 8

[3] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds, 2018. 3

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2, 4, 6

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[7] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 6, 7, 8, 5

[8] Sandika Biswas, Kejie Li, Biplab Banerjee, Subhasis Chaudhuri, and Hamid Rezatofighi. Physically plausible 3d human-scene reconstruction from monocular rgb image using an adversarial learning approach. *IEEE Robotics and Automation Letters*, 8(10):6227–6234, 2023. 2

[9] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 1, 2

[10] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[11] 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 *The European Conference on Computer Vision (ECCV)*, 2020. 2

[12] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1, 2, 4, 6

[13] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 3, 6

[14] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[15] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 6, 2

[16] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 4

[17] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Gregory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[18] Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 1, 6, 2

[20] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award. 2

[21] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *IEEE Computer Vision and Pattern Recognition*, 2021. 3

[22] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *CVPR*, 2020. 2

[23] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 2

[24] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 2

[25] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human–object interaction and scene changes from human motion. In *International Conference on 3D Vision (3DV)*, 2024. 2

[26] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2

[27] Sookwan Han and Hanbyul Joo. Chorus : Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15835–15846, 2023. 2

[28] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision*, 2019. 2

[29] 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. 1, 2

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, New Orleans, LA, USA, 2022. IEEE. 5, 1

[31] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2, 4, 5, 1

[33] JF Hu, WS Zheng, J Lai, and J Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2186–2200, 2017. 5, 8

[34] 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 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 2

[35] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 1, 2, 6, 7, 8, 5

[36] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2

[37] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 1

[38] Li Jiang, Zetong Yang, Shaoshuai Shi, Vladislav Golyanik, Dengxin Dai, and Bernt Schiele. Self-supervised pretraining with masked shape prediction for 3d scene understanding. In *CVPR*, 2023. 2

[39] 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. 2

[40] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream, 2023. 2

[41] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[42] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *8th International Conference on 3D Vision*, pages 333–344. IEEE, 2020. 2

[43] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv:2306.01567*, 2023. 4

[44] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[45] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4

[46] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 1

[47] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)*, 2022. 2

[48] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024. 2

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 7, 8, 5, 9, 10, 11

[50] Feng Liu and Xiaoming Liu. Learning implicit functions for topology-varying dense 3d shape correspondence, 2020. 3

[51] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[52] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5, 8

[53] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Eric Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. *arXiv preprint arXiv:2304.12281*, 2023. 2

[54] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund T, Zexiang Xu, and Hao Su. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[55] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 1, 2

[56] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 5

[57] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics*. ACM, 2015. 4

[58] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 5

[59] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 2

[60] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *CVPR*, 2023. 2, 4, 5, 6, 7, 8, 1, 3

[61] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5

[62] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer, 2022. 2

[63] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[64] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[65] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2

[66] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, abs/1612.00593, 2017. 3, 1

[67] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023. 2

[68] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 4

[69] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 1

[70] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 1, 2

[71] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[72] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 2

[73] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2

[74] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision (ECCV)*, pages 516–533, 2022. 2

[75] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 4

[76] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019. 2, 4

[77] David Stutz and Andreas Geiger. Learning 3d shape completion under weak supervision. *CoRR*, abs/1805.07290, 2018. 3

[78] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[79] Ramana Sundararaman, Riccardo Marin, Emanuele Rodola, and Maks Ovsjanikov. Reduced representation of deformation fields for effective non-rigid shape matching. *Advances in Neural Information Processing Systems*, 35, 2022. 3

[80] 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 (ECCV)*, 2020. 2

[81] Yu Tao, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Dai Quionhai, Hao Li, G. Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2

[82] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? 2019. 6

[83] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[84] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[85] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 1

[86] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8001–8013, 2023. 2

[87] Michal J. Tyszkiewicz, P. Fua, and Eduard Trulls. Gecco: Geometrically-conditioned point diffusion models. *ICCV*, abs/2303.05916, 2023. 2

[88] 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 (IJCV)*, 2016. 4

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5

[90] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *International Conference on 3D Vision (3DV)*, 2022. 2

[91] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *CVPR*, 2023. 2

[92] Christopher Wewer, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. Simnp: Learning self-similarity priors between neural points. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[93] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing, 2018. 2

[94] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 1, 2

[95] 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 (ECCV)*. Springer, 2022. 2, 4, 6, 7, 8

[96] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 6

[97] Xianghui Xie, Xi Wang, Nikos Athanasiou, Bharat Lal Bhatnagar, Chun-Hao P. Huang, Kaichun Mo, Hao Chen, Xia Jia, Zerui Zhang, Liangxian Cui, Xiao Lin, Bingqiao Qian, Jie Xiao, Wenfei Yang, Hyeongjin Nam, Daniel Sungho Jung, Kihoon Kim, Kyoung Mu Lee, Otmar Hilliges, and Gerard Pons-Moll. Rhobin challenge: Reconstruction of human object interaction, 2024. 2

[98] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 2

[99] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[100] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 2

[101] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 12977–12988, 2023. 2

[102] 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 *ICCV*, 2021. 2

[103] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 2

[104] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling pre-trained vision foundation models. *arXiv preprint arXiv:2303.12786*, 2023. 2

[105] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3959–3970, 2022. 2, 4

[106] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[107] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[108] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[109] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 2

[110] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 4

[111] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, 2022. 4

[112] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, William T Freeman, and Jiajun Wu. Learning to Reconstruct Shapes From Unseen Classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1

[113] 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 (ECCV)*. Springer, 2022. 2

[114] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[115] Keyang Zhou, Bharat Lal Bhatnagar, Bernt Schiele, and Gerard Pons-Moll. Adjoint rigid transform network: Task-conditioned alignment of 3d shapes. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022. 3, 1

[116] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[117] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5826–5835, 2021. 5, 1

[118] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2, 4

Template Free Reconstruction of Human-object Interaction with Procedural Interaction Generation

Supplementary Material

In this supplementary, we first discuss in more detail about our implementation for the ProciGen and HDM in Appendix A. We also present the statistics of our generated ProciGen dataset. We then show more results and experimental analysis of our method in Appendix B. We conclude with a discussion of limitations and future works.

A. Implementation Details

We describe in more details of the implementation of our ProciGen and HDM. Our code for both data generation and reconstruction will be made publicly available.

A.1. ProciGen Data Generation

Correspondence estimation. We use the implementation from ART [115] for our autoencoder, which uses PointNet [66] as encoder and 3-layer MLPs as decoder. We sample 8000 points from the mesh surface and train the network with bidirectional Chamfer distance. To ensure reconstruction quality, we overfit one network per category. Each model is trained for 5000 epochs. We report an average reconstruction error of around 7mm for our autoencoders, which indicates highly accurate reconstructions.

Contact transfer and optimization. We use a threshold of $\sigma = 2$cm to find points that are in contact. The loss weights for our contact based loss optimization are: $\lambda_c = 400, \lambda_n = 6.25, \lambda_{\text{colli}} = 9, \lambda_{\text{init}} = 6.25 \cdot 10^4$.

Rendering. We use blender to render our synthesized human-object interactions. We choose one set of 4 camera configurations from BEHAVE [7] and another set of 6 camera configurations from InterCap [35]. For each synthesized interaction, we additionally add small random global rotation and translation to have variations of camera viewpoints. We render the interactions with an empty background since our network also takes images with background masked out as input. We add lights at fixed locations with random light intensities. Our blender scene and rendering code will also be made publicly available.

A.2. HDM: Hierarachical Diffusion Model

We use the modified Point Voxel CNN from [117] as the network for our joint diffusion p_θ, segmentation g_ϕ, and separate diffusion models p^h_θ, p^o_θ. The input images are cropped and resized to 224×224. The joint diffusion model diffuses in total 16384 points while the separate models diffuse 8196 points each. We use the MAE [30] as the image feature encoder. We additionally stack the human and object masks as well as distance transform as additional image features, same as PC^2 [60]. We train our diffusion models for a total of 500000 steps with batch size 20. We use a linear scheduler without warm-up for the forward diffusion process, in which beta increases from $1 \cdot 10^{-5}$ to $8 \cdot 10^{-3}$. For the network optimization, we use AdamW optimizer with linear learning rate decay starting from $3 \cdot 10^{-4}$ and decreasing to 0 during the course of training. The diffusion models are trained with the standard diffusion training scheme [32]. To train the segmentation model, we add small Gaussian noise to the GT point clouds and project them to obtain image features. The loss is then computed between the prediction and recomputed GT labels on the points with noise. To speed up training, we train stage 1 (g_ϕ, p_θ) and stage 2 (p^h_θ, p^o_θ) models separately. For each stage, it takes around 4 days to train on a machine with 4 A40 GPUs.

Camera estimation. Recall from Sec. 3.2.3 that a camera translation is required to project the normalized point clouds back to the image. This needs to be estimated from input when GT camera pose is not available, especially for generalization to diverse datasets. The camera translation consists of three unknowns, which requires at least two point pairs of 3D location and 2D-pixel coordinates. We empirically choose the Gaussian point center and one edge of the point cloud. The idea is to have the initial Gaussian point clouds cover the 2D human object interaction region and the 3D center is projected to 2D crop center.

Formally, let $\mathbf{p}_c = (c_x, c_y)$ be the center coordinate of the 2D interaction region, w be the width of the 2D interaction square crop, $\mathbf{p}_e = (\sigma, 0, z)$ be a 3D point near the edge of the Gaussian sphere with unknown depth z. Given camera projection matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t}_c$, we define the following equations:

$$\mathbf{K}\mathbf{t}_c = \mathbf{p}_c; \;\; \mathbf{K}(\mathbf{p}_e + \mathbf{t}_c) = \mathbf{p}_e^{\text{2D}} \tag{6}$$

The first equation projects origin to $\mathbf{p}_c$ and the second equation projects $\mathbf{p}_e$ to the middle right of the 2D crop $\mathbf{p}_e^{\text{2D}} = (c_x + w/2, c_y)$. This is a linear system of four equations with four unknowns (camera translation and depth z), leading to a unique solution for the translation $\mathbf{t}_c$. We empirically set σ to different values for different categories based on the estimation error on the BEHAVE training set. Furthermore, we compute $\mathbf{p}_c$ as the centroid of all 2D points inside the human and object masks. From Fig. 6, Fig. 17, Fig. 16, Fig. 18, Fig. 19 and , Fig. 20, it can be seen that our method can reconstruct human and object well on different datasets using our estimated translation.

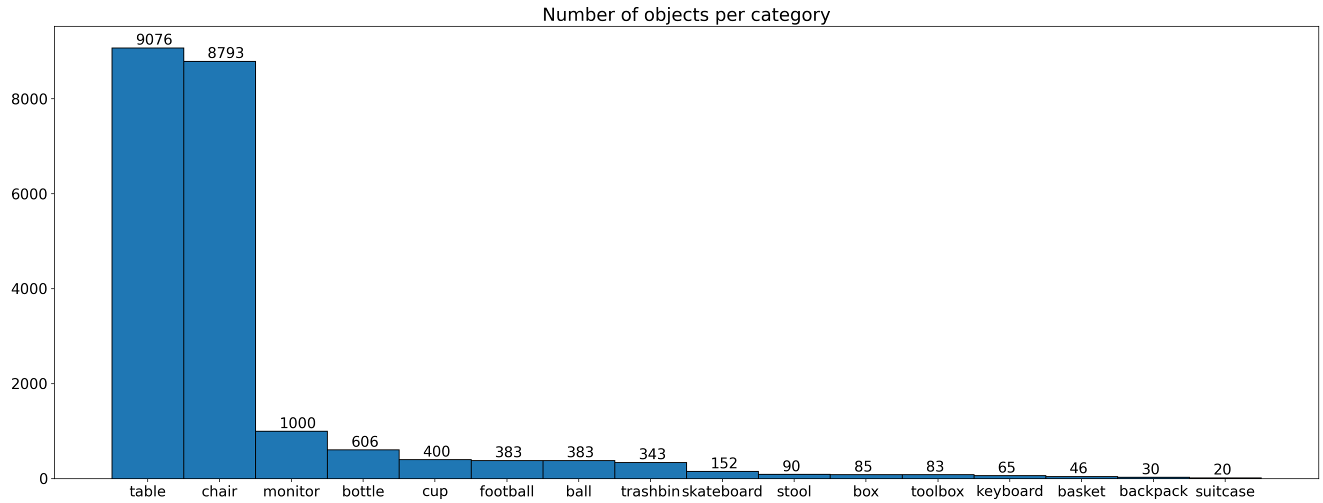

Figure 7. The number of objects per category we used to generate our ProciGen dataset. It can be seen that the shape variations are dominated by tables and chairs, which are also the categories with the most complex shapes.

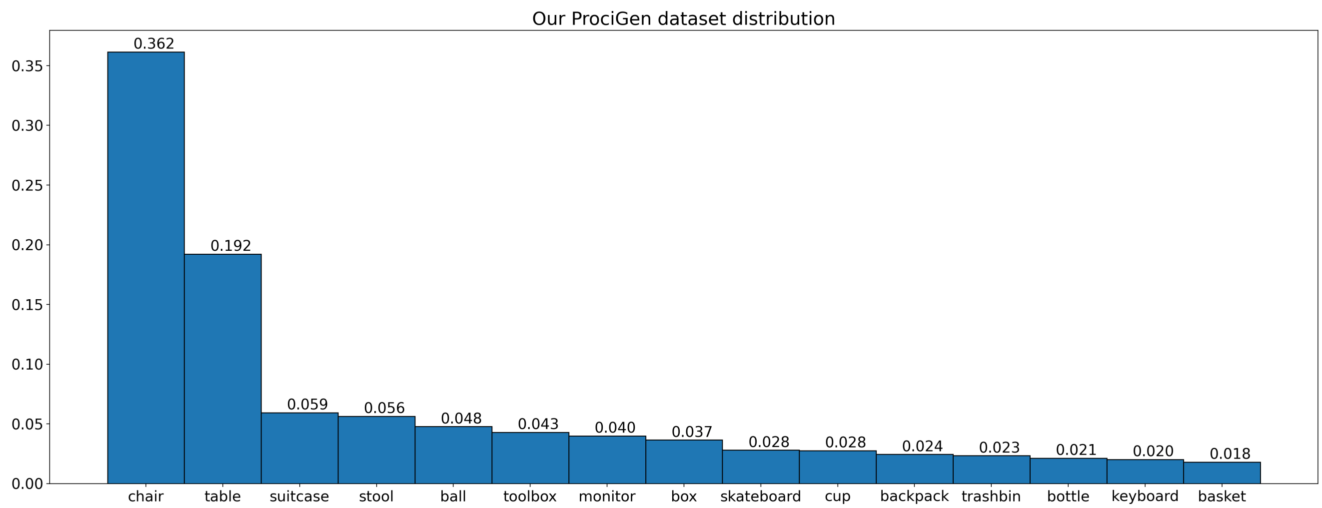

Figure 8. Distribution of interactions per object category. Our dataset features interaction data with very diverse object shapes, which is not possible via real data capture.

A.3. Dataset statistics

We generate our ProciGen dataset based on interactions from BEHAVE [7] and InterCap [35], human scans from MGN [4], object shapes from ShapeNet [12], Objaverse [19] and ABO [15]. When generating our data, we mainly consider the variation of object shapes and interaction poses while the object sizes remain the same. We also try to avoid large imbalances among objects. Therefore, chairs and tables are two dominant categories as they have the most geometry and interaction pose variations (Fig. 7, Fig. 8). Other categories have similar amounts of synthetic data as they have similar amounts of interaction poses. The difference comes from failures in joint optimization due to irregular mesh.

In total we rendered 1.1M interaction images with 21555 different object shapes. The distribution for object shapes and interactions per category are shown in Fig. 7 and Fig. 8. Our dataset has very diverse object shapes, especially for chairs and tables whose geometry also varies a lot in reality. Our procedural generation method is a scalable solution and it allows for generating large-scale interaction datasets with great amount of variations which is not obtainable via capturing real data.

Method	CHORE ↑			PC2 ↑	Ours ↑		
	Hum.	Obj.	Comb.	Comb.	Hum.	Obj.	Comb.
Chair	0.373	0.491	0.443	0.407	**0.384**	**0.521**	**0.463**
Ball	0.330	0.388	0.374	0.424	**0.395**	**0.517**	**0.471**
Backpack	**0.399**	**0.509**	**0.469**	0.436	0.397	0.457	0.444
Table	0.304	0.455	0.389	0.470	**0.379**	**0.642**	**0.517**
Basket	0.301	0.266	0.292	**0.381**	**0.412**	**0.297**	0.364
Box	0.352	0.347	0.362	0.409	**0.414**	**0.401**	**0.424**
Keyboard	0.335	0.412	0.383	0.450	**0.353**	**0.606**	**0.493**
Monitor	0.358	**0.412**	**0.395**	0.377	**0.368**	0.348	0.370
Suitcase	0.400	0.477	0.443	0.404	**0.431**	**0.484**	**0.462**
Stool	0.351	0.479	0.424	0.443	**0.394**	**0.543**	**0.479**
Toolbox	0.281	0.330	0.306	0.398	**0.373**	**0.400**	**0.403**
Trashbin	0.376	0.402	0.398	0.387	**0.407**	**0.414**	**0.422**
BEHAVE all	0.345	0.426	0.397	0.423	**0.392**	**0.498**	**0.457**
Chair	**0.389**	0.468	0.433	0.470	0.384	**0.604**	**0.500**
Cup	0.412	0.538	0.510	0.566	**0.487**	**0.601**	**0.578**
Skateboard	**0.520**	0.684	0.612	0.578	0.491	**0.739**	**0.624**
Bottle	0.426	0.501	0.495	0.592	**0.549**	**0.582**	**0.593**
InterCap all	0.406	0.513	0.469	0.506	**0.440**	**0.607**	**0.534**

Table 6. Per-category F-score@0.01m comparison. Note that PC2 cannot separate human-object hence we only report the combined error, and that CHORE requires template meshes. Our method outperforms baselines for almost all categories.

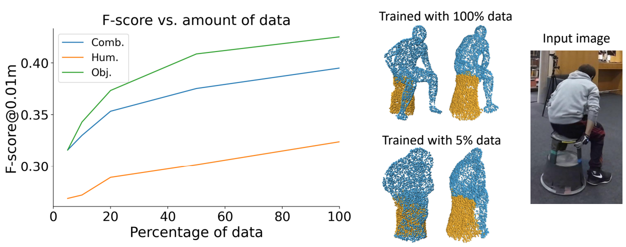

Figure 9. Reconstruction performance vs. amount of data. It can be seen that more data leads to better results.

B. Additional Experiments and Results

B.1. Per-category reconstruction accuracy

We report the accuracy of each category in Tab. 6. Our method consistently outperforms baselines in almost all categories. While the improvements in numbers look small, the visual difference is quite significant, as shown in the paper Fig. 4, Fig. 6.

B.2. Performance vs. amount of data

We show in Table 3 that our data contributes a lot to improve the reconstruction accuracy. To further understand the data contribution, we train our model for the same epochs with different amounts of synthetic data and test on BEHAVE images without fine-tuning. The performance vs. data plot is shown in Fig. 9. More data consistently leads to better performance both quantitatively and qualitatively.

B.3. Analysis of T_0 for our HDM

In our second stage, we first add noise to the clean predictions from stage one until step $t = T_0$, and then run the reverse diffusion process from $t = T_0$ to $t = 0$. We evaluate the performance of our method under different values of T_0 in Figure 10. There is a trade-off for the number of forward steps T_0: with a larger T_0, less interaction information and

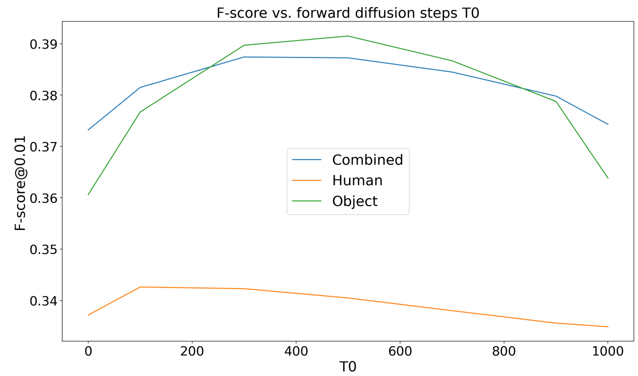

Figure 10. The performance of our method using different intermediate step T_0 for the input to our second stage diffusion. Methods are evaluated using F-score@0.01m. At $T_0 = 500$, we obtain a good balance between human and object performance.

noisy details are preserved and the network predicts sharper detail but less faithful to initial prediction and interaction constraints. It can be seen that $T_0 = 500$ is a good balance between shape fidelity and interaction coherence.

B.4. Shape fidelity

Our method predicts dense and clean point clouds which are ready for accurate surface extraction. We show in Fig. 11 that high-quality meshes can be extracted from our predicted point clouds. More specifically, we use screened Poisson surface reconstruction for the human points using normals estimated by MeshLab. For the object, we first use Delaunay triangulation to obtain triangle mesh. We then run fusion-based waterproofing [77] to obtain a watertight mesh. We also apply Delaunay triangulation and waterproofing to PC^2 [60] predictions and results are shown in Fig. 11. It can be seen that PC^2 predictions have missing structure and noisy point clouds, leading to low-quality meshes. In contract, we can extract high-quality meshes directly from our point cloud reconstructions, without any post processing.

B.5. Interaction semantics

Our method predicts the segmentation of human and object, allowing separate manipulation which is important for downstream applications. To demonstrate this, we use Text2txt [13] to generate textures for the meshes extracted from PC^2 and our predicted point clouds. Other methods such as Paint-it [106] are also applicable here. We show the reconstruction and generated textures in Fig. 13. It can be seen that PC^2 predictions are noisy and it does not reason human and object separately. This leads to low-quality mesh and generating coherent texture for a combined mesh of human and object is difficult. On the contrary, our method separate human and object while also predicting high quality individual shapes. This allows generating high quality texture and changing textures for human and object

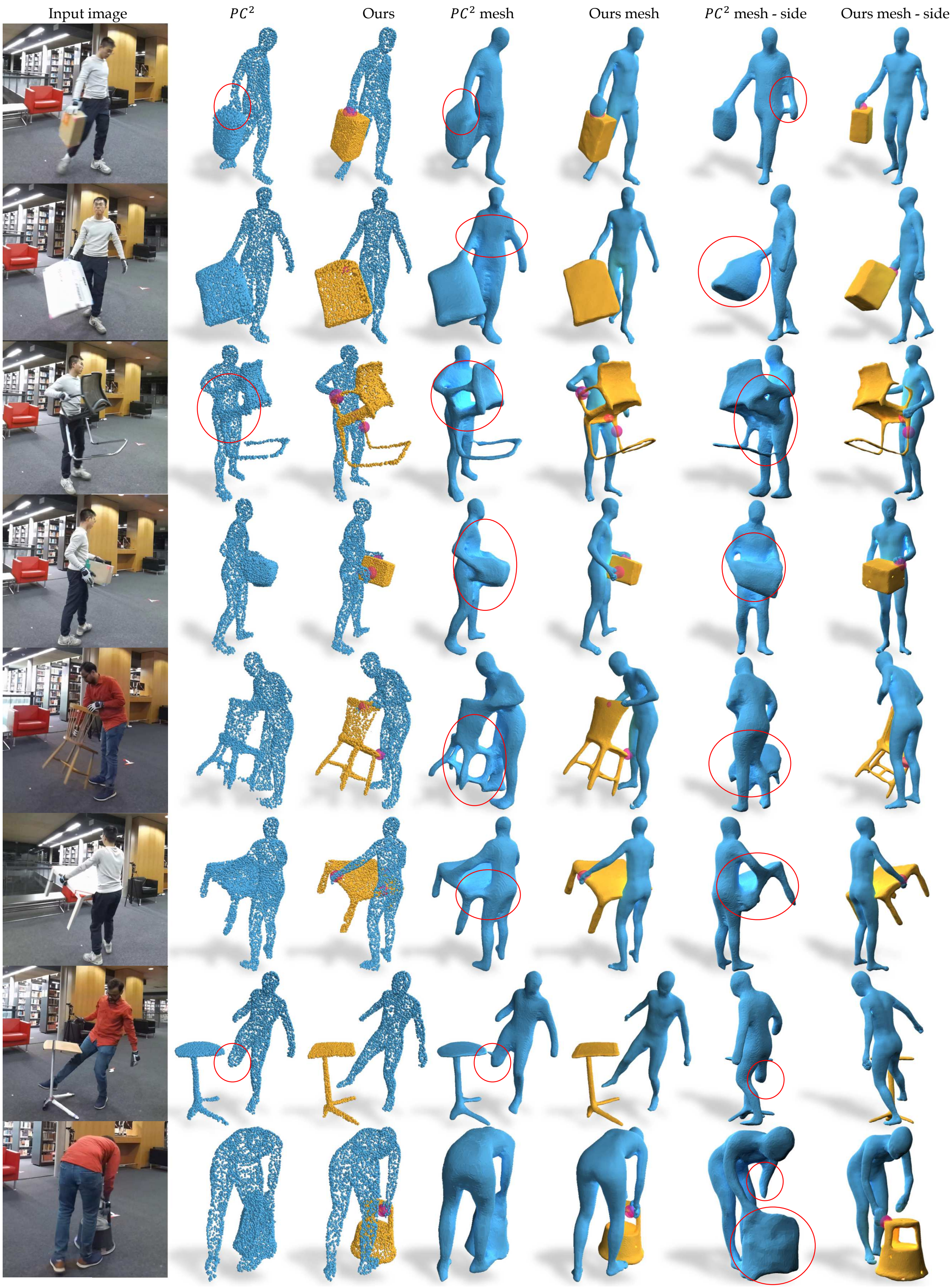

Figure 11. Comparing the shape fidelity of our method with PC^2 on the BEHAVE [7] dataset. PC^2 does not separate human and object and its prediction is noisy, leading to inaccurate meshes. Our method predicts clean point clouds with human object segmentations, allowing us to extract high-quality mesh surfaces.

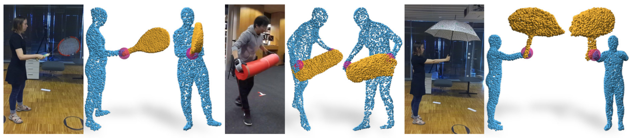

Figure 12. Out of distribution generalization. Our method can reconstruct some categories that are *unseen* in training data.

differently.

B.6. More generalization results

We show more generalization comparison on the InterCap [35] dataset in Fig. 15. Note that all objects from InterCap are unseen during training time. It can be seen that PC^2 trained on BEHAVE [7] only cannot generalize to objects from InterCap. Training PC^2 with our ProciGen dataset allows better generalization ability but its shape prediction is still noisy. Furthermore, PC^2 cannot segment human and object, which is important to reason the interaction semantics and manipulate them separately. Our method generalizes well to InterCap and reconstructs high quality shapes with interaction semantics.

Our method trained *only* on our synthetic ProciGen dataset generalizes well to other datasets. We show results on NTU-RGBD [52], SYSU [33] and challenging in the wild COCO [49] images in figure Fig. 16, Fig. 17 and Fig. 18, Fig. 19, Fig. 20 respectively. Note that our method is trained only on our synthetic ProciGen dataset and not fine-tuned on any images from these datasets. It can be seen that our method generalizes to different datasets with diverse object shapes, without requiring any template meshes.

For quantitative evaluation, we focus on 15 object categories that are seen from our synthetic data (Tab. 6). We test our method on three additional categories from BEHAVE and InterCap that are unseen and have GT data. The F-scores (human/object/combined) are: 0.465/0.453/0.479 (tennis racket), 0.333/0.332/0.361 (yoga mat), 0.375/0.305/0.360 (umbrella), 0.353/0.443/0.420 (all seen categories). We also show example reconstructions in Fig. 12. Our method can reconstruct *unseen* categories.

C. Limitations and Future Work

We present a scalable solution to synthesize large amount of interaction dataset which allows training methods with strong generalization ability. We also propose a model for obtaining high quality human, object shapes and also interaction semantics, without any template shapes. We demonstrate the generalization ability of our method on diverse datasets. Our template-free reconstruction method is a promising first step towards real in-the-wild reconstruction.

Nevertheless, there are still some limitations to the current approach. First, our ProciGen data generation method always starts with a seed interaction pose sampled from an existing interaction dataset. This limits the diversity in terms of interaction poses. Future works can explore generative models such as Object-Popup [64] to further diversify the interaction pose. It is also highly desirable to combine the large human pose variations from AMASS [58], which can further improve the robustness of reconstruction methods to challenging poses.

Secondly, our method struggles to predict accurate human shapes when large chunk of the human body is occluded, see Fig. 14. This is because our method is purely template-free and only use the network to learn the human and object shape priors. Future works can try to further explore human shape or pose constraints to regularize network training and predictions. In addition, our hierarchical diffusion model are designed for human object interaction, which is applicable for general bilateral interaction cases like human-human, hand-hand, and hand-object interactions. However, it cannot handle multi-person or multi-object interactions. We leave these for future works.

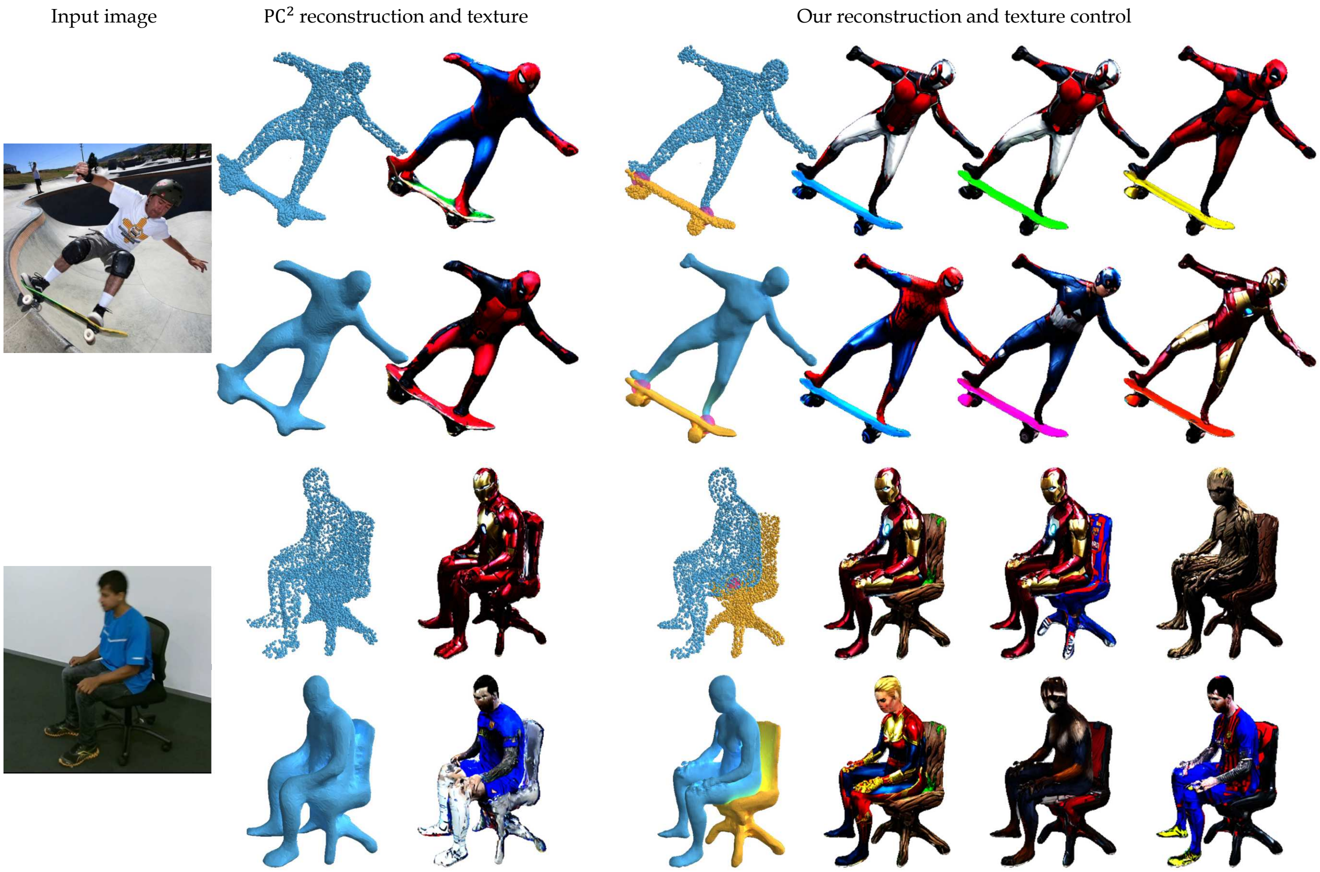

Figure 13. Comparing textures generated for meshes extracted from PC^2 [60] and our predicted point clouds. Textures are obtained using Text2txt [13]. PC^2 predicts human and object as one joint point cloud with noisy points, which leads to inaccurate mesh surfaces and it is difficult to generate textures for this combined mesh. It also does not allow changing human and object textures separately. Our method predicts high quality point clouds with segmentation. This enables us to extract high-fidelity mesh, which is important for generating high-quality texture and manipulating human and object differently.

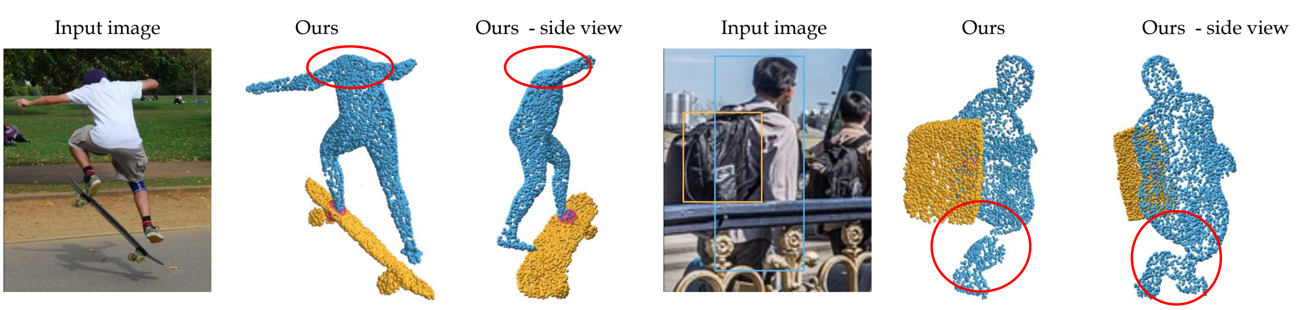

Figure 14. Example failure cases of our method. Our method can fail when large parts of human body are invisible, leading to incoherent human shape reconstructions. Future works can explore human body shape priors to regularize the network predictions.

Figure 15. Comparing generalization performance on InterCap [35]. All objects are unseen during training time. PC^2 trained only on BEHAVE [7] has limited generalization ability. Training PC^2 with our ProciGen improves generalization but it still cannot reason human and object separately and the predicted points are noisy. Our method trained only on our ProciGen generalizes well to InterCap objects even they are completely unseen.

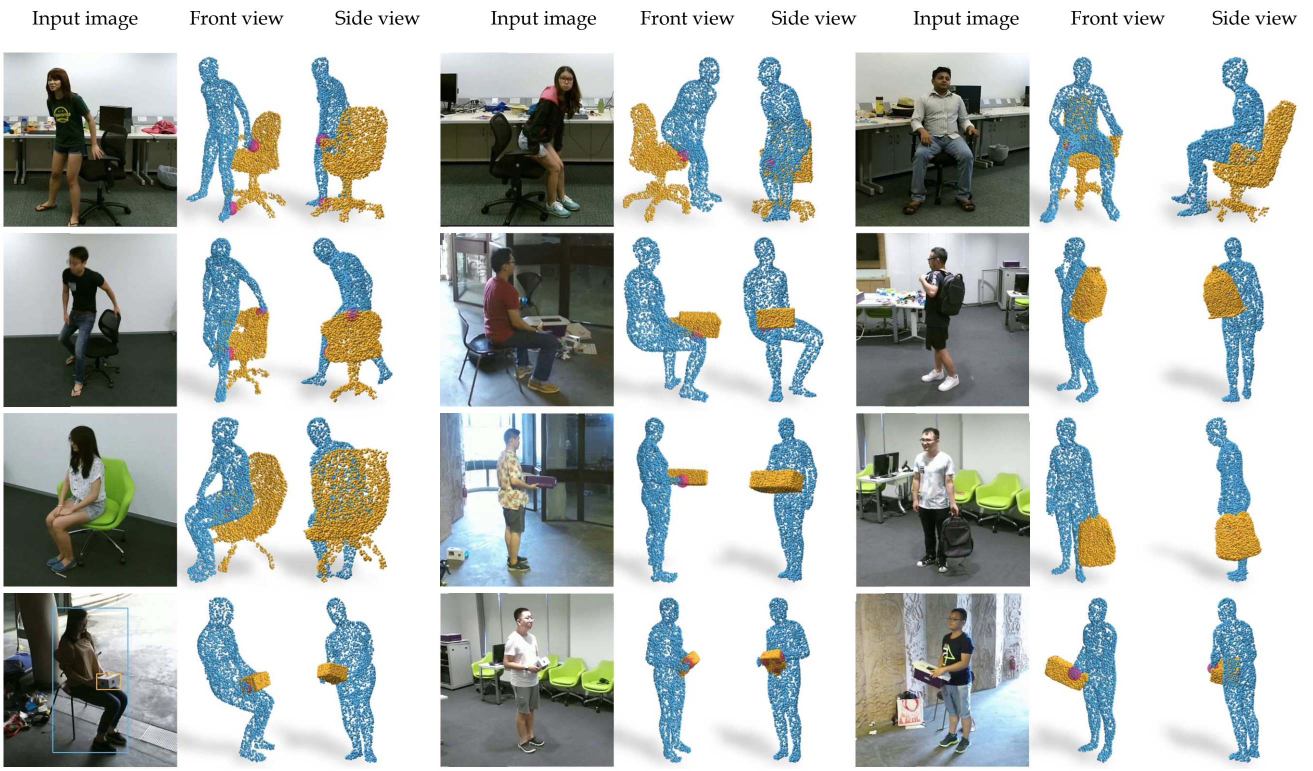

Figure 16. Generalization results on NTU-RGBD [52] dataset. Our method can reconstruct different objects faithfully under various camera viewpoints and lighting conditions, without relying on any template shapes.

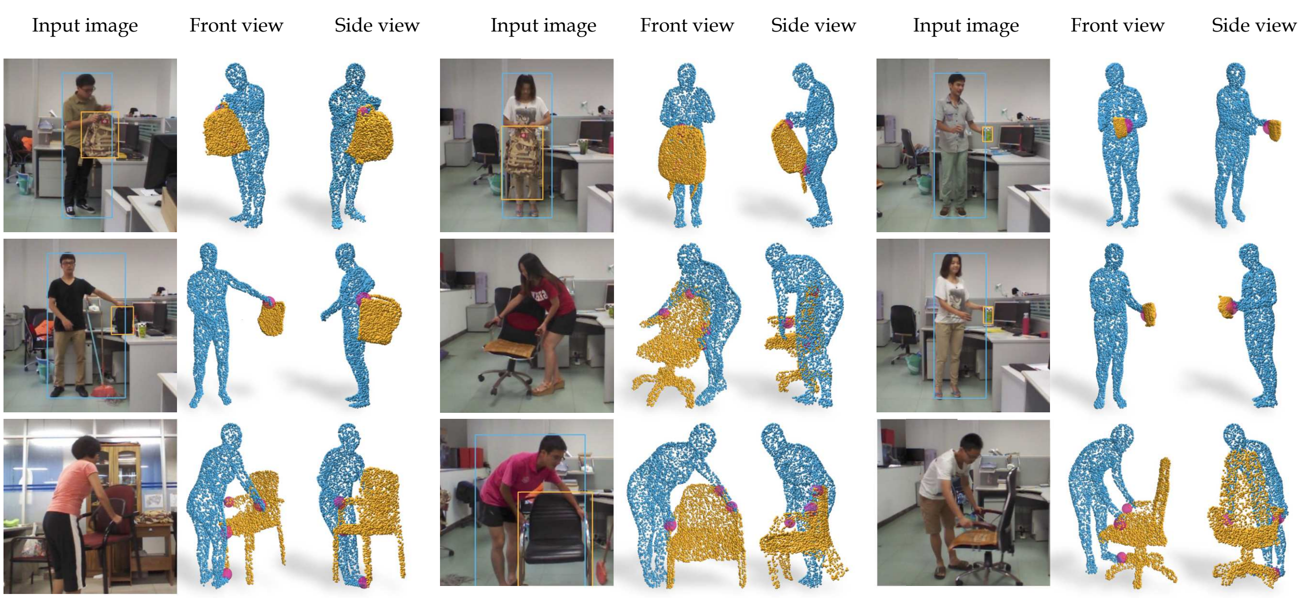

Figure 17. Generalization results on SYSU action [33] dataset. Our method can reconstruct different real-life human and objects during challenging interactions and occlusions.

Figure 18. Generalization results to COCO [49] dataset. Our method can reconstruct high-quality human and object from in the wild images which has very diverse shape variations, without using any template shapes.

Figure 19. Generalization results to COCO [49] dataset. Our method reconstructs diverse object shapes in the wild.

Figure 20. Generalization results to COCO [49] dataset. Our method can reconstruct challenging human and object pose as well as shapes without using any template shapes.