Full-Body Articulated Human-Object Interaction

Nan Jiang[1,2⋆†], Tengyu Liu[2⋆], Zhexuan Cao[3†], Jieming Cui[1,2,†], Zhiyuan Zhang[2,3,†],
Yixin Chen[2], He Wang[4], Yixin Zhu[5✉], Siyuan Huang[2✉]

https://jnnan.github.io/project/chairs/ ✉ yixin.zhu@pku.edu.cn, syhuang@bigai.ai

⋆Equal contributors † Work done during an internship at BIGAI
[1] School of Intelligence Science and Technology, Peking University
[2] Beijing Institute of General Artificial Intelligence (BIGAI)
[3] Department of Automation, Tsinghua University
[4] Center on Frontiers of Computing Studies, Peking University
[5] Institute for Artificial Intelligence, Peking University

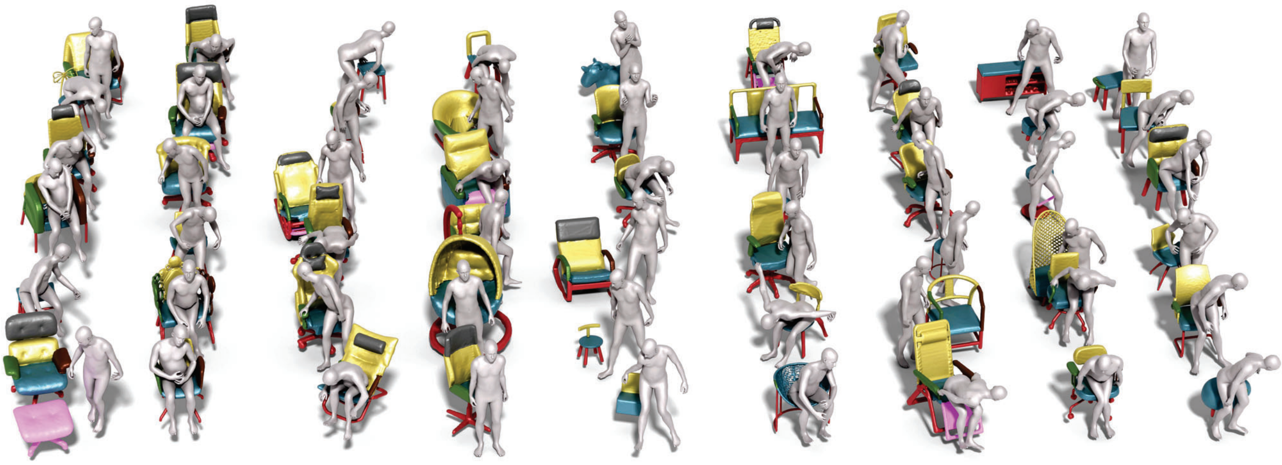

Figure 1: **Examples of the proposed Capturing Human and Articulated-object InteRactionS (`CHAIRS`) dataset.** It contains fine-grained interactions between 46 participants and 81 sittable objects with drastically different kinematic structures, providing multi-view RGB-D sequences and ground-truth 3D mesh of humans and articulated objects for over 17.3 hours of recordings.

Abstract

*Fine-grained capture of 3D Human-Object Interactions (HOIs) enhances human activity comprehension and supports various downstream visual tasks. However, previous models often assume that humans interact with rigid objects using only a few body parts, constraining their applicability. In this paper, we address the intricate challenge of Full-Body Articulated Human-Object Interaction (f-AHOI), where complete human bodies interact with articulated objects having interconnected movable joints. We introduce `CHAIRS`, an extensive motion-captured f-AHOI dataset comprising 17.3 hours of diverse interactions involving 46 participants and 81 articulated as well as rigid sittable objects. The `CHAIRS` provides 3D meshes of both humans and articulated objects throughout the interactive sequences, offering **realistic** and **physically plausible** full-body interactions. We demonstrate the utility of `CHAIRS` through object pose estimation. Leveraging the geometric relationships inherent in HOI, we propose a pioneering model that employs human pose estimation to address articulated object pose and shape estimation within whole-body interactions. Given an image and an estimated human pose, our model reconstructs the object's pose and shape, refining the reconstruction based on a learned interaction prior. Across two evaluation scenarios, our model significantly outperforms baseline methods. Additionally, we showcase the significance of `CHAIRS` in a downstream task involving human pose generation conditioned on interacting with articulated objects. We anticipate that the availability of `CHAIRS` will advance the community's understanding of finer-grained interactions.*

1. Introduction

In the realm of computer vision and robotics, the fundamental comprehension of Human-Object Interaction (HOI) [30, 31, 64, 62] lies at the core of dissecting intricate human activities. This paper embarks on unraveling the complex challenge of Full-Body Articulated Human-Object

Table 1: **Comparisons between `CHAIRS` and other HOI datasets.**

Dataset	# object	# participants	# instructions	# hours	fps	# view	articulated objects	human	annotation type
GRAB [46]	51	10	4	3.8	120	0	No	Whole-body	mocap
D3D-HOI [58]	24	5	/	0.6	3	1	Yes	Whole-body	manual
BEHAVE [2]	20	8	6	4.2	30	4	No	Whole-body	multi-kinect
ARCTIC [11]	10	9	1	1.2	30	8+1	Yes	Two hands	mocap
COUCH [67]	4	6	6	3	60	4	No	Whole-body	mocap
`CHAIRS` (Ours)	81	46	32	17.3	30	4	Yes	Whole-body	mocap

Interaction (f-AHOI). This endeavor mandates tackling two pivotal dimensions: (i) fashioning kinematic-agnostic representations for **articulated** objects and (ii) delving into the intricate spatial-temporal tapestry interweaving objects with human **whole-bodies**. Our primary focus resides in addressing the intricate task of object pose estimation within the realm of f-AHOI, considering that reconstructing 3D human poses from frontal viewpoints is comparably uncomplicated.

The crux of object pose estimation within the context of f-AHOI is punctuated by three principal challenges:

The dearth of comprehensive f-AHOI datasets Existing strides in 3D HOI predominantly either assume interactions with rigid objects or confine themselves to involving specific segments of the human anatomy [2, 46, 65, 29, 67, 11, 58, 14]. Regrettably, these assumptions drastically oversimplify genuine human interactions that span diverse body parts engaging with articulated objects embodying moveable elements such as cabinets and office chairs. A more intricate level of interaction necessitates a richer dataset.

The multifaceted landscape of object kinematic structures Objects constituting the realm of f-AHOI exhibit notable disparities in their kinematic frameworks, even when categorized under the same umbrella. Prevailing methodologies often lean towards uniform structures [58, 32, 11, 14], thereby disregarding the diverse tapestry that constitutes real-world scenarios. The endeavor of accurately reconstructing objects manifesting divergent geometries and structures is plagued with its own set of challenges.

The intricacies of complex interactions Engaging with articulated objects entails grappling with intricate spatial and physical relationships, often entailing occlusions and intricate points of contact. The intricacy of these dynamics thrusts conventional pose estimation mechanisms reliant on point cloud template-matching [65, 49, 19, 39, 28] into the realm of insufficiency. The prominence of contacts further compounds the endeavor of precise reconstruction, as slight inaccuracies can swiftly usher implausible interactions into the picture.

The trajectory of this research endeavors to navigate the above three challenges through the prism of three principal solutions, respectively:

To confront the scarcity of f-AHOI datasets, we introduce `CHAIRS`, a multi-view RGB-D dataset. Illustrated in Fig. 1, `CHAIRS` chronicles a diverse tapestry of interactions, seamlessly intertwining 46 participants with 81 sittable objects (*e.g.*, chairs, sofas, stools, and benches). 28 of these objects are endowed with moveable parts. Each frame encapsulates 3D meshes of both human **whole-bodies** and objects, casting a spotlight on interactions with sittable objects that encompass a diverse spectrum of structures and distinctive movable elements conducive to multifarious human interactions.

To traverse the labyrinth of kinematic diversity, `CHAIRS` meticulously selects representative objects characterized by an eclectic array of structures. Unlike traditional datasets and methodologies tethered to uniform kinematics [55, 51, 32], we champion real-world heterogeneity, encompassing the gamut from rigid stools to swivel chairs boasting up to 7 movable components. Each component is linked to its parent through a nexus of revolute, prismatic, or composite joints.

To unravel the enigma of complex interactions, we proffer an innovative approach to articulated object pose estimation, one that harnesses the subtle interplay of fine-grained interaction relationships to reconstruct the object in question. This approach diverges from the conventional recourse of manually labeling contact maps corresponding to human body parts [65, 16, 2]. Instead, our approach melds the intricacies of these relationships with a reconstruction model and an interaction prior, the latter of which is imbued with the essence of a conditional Variational Auto-Encoder (cVAE). This evolution sidesteps the need for predefined knowledge grounded in laborious annotation. Moreover, the significance of these intricate relationships is showcased through our venture into learning human poses within the ambit of articulated objects. By juxtaposing the generative prowess harnessed from `CHAIRS` with that stemming from a dataset centered on rigid objects [67], we underscore the pivotal role played by the nuanced geometrical relationships encapsulated within `CHAIRS` in the broader canvas of downstream tasks.

Our **contributions** are four-fold: (i) `CHAIRS`, a sprawling multi-view RGB-D repository infused with diverse 3D meshes. (ii) The seamless extension of articulated object pose estimation to the arduous landscape of f-AHOI. (iii) An object pose estimation approach that transcends the strictures of structure. (iv) An overarching interaction prior that captures the subtleties of fine-grained interactions, acting as a catalyst for the journey of pose estimation.

(a) sequences of objects articulating over time (b) frames of diverse interactions

Figure 2: **Samples from the CHAIRS dataset.** CHAIRS encompasses diverse AHOIs captured through precisely calibrated multi-view RGB-D cameras, offering detailed 3D meshes of human participants and articulated objects. The figure showcases (a) RGB frames alongside corresponding ground-truth mesh sequences and (b) an array of varied AHOI instances.

2. Related Work

3D Human-Object Interaction (HOI) The evolution of HOI research spans from 2D image-based interaction detection [4, 40, 13, 30, 31, 64, 62, 22] to 3D interaction reconstruction [43, 16, 5, 53, 58, 63, 66] and generation [17, 50, 57, 21, 52, 67] within 3D scenes. Notably, PiGraph [43] and D3D-HOI [16, 58] capture daily activities and reconstruct interactions, often relying on visual observations. In contrast, MoCap systems [46, 2, 11, 67] offer fine-grained 3D human-object interactions. GRAB [46] and ARCTIC [11] emphasize interactions with small objects, while BEHAVE [2] and COUCH [67] involve interactions with everyday objects. However, these works often focus on rigid objects or hand-object interactions with articulated objects. In contrast, our CHAIRS dataset captures realistic *whole-body* interactions with diverse articulated objects.

Articulated Human-Object Interaction Articulated Human-Object Interactions (AHOIs) build on part-level object representations, modeling intricate spatial-temporal interactions between humans and articulated objects [14]. Noteworthy contributions include D3D-HOI [58], ARCTIC [11], and 3DADN [41]. D3D-HOI [58] captures humans interacting with containers, ARCTIC [11] focuses on motion-captured RGB-D hand-object interactions, and 3DADN [41] annotates movable object parts from internet videos. However, these works often emphasize hand-object interactions, whereas our focus extends to AHOIs encompassing diverse articulated objects and multiple body parts.

Contact-Rich HOI The realm of f-AHOI requires a deeper HOI understanding. While 3D HOI literature has expanded, few works address full-body contacts through reconstruction [16] or generation [69, 52, 15, 67]. However, these works often focus on static scenes with limited interactions. In contrast, our CHAIRS dataset encompasses diverse articulated objects and interactions.

Articulated Object Pose Estimation The estimation of 6-DOF poses for rigid objects has garnered attention [25, 19, 3, 48, 39, 37, 9]. Template-based methods [20, 60, 49, 26] and regression models [1] are common, with recent strides in articulated object pose estimation [8, 33, 28] leveraging these techniques. Regression and implicit function models [36, 47, 59, 24] are also explored. Despite progress, these methods often assume consistent kinematic structures within object categories. In contrast, our CHAIRS dataset features diverse kinematic structures and models capable of handling various parts and kinematics of 3D objects.

3. The CHAIRS Dataset

A significant challenge in modeling AHOIs is the lack of accurate 3D annotations. To address this gap, we introduce CHAIRS, a comprehensive AHOI dataset featuring multi-view RGB-D sequences. CHAIRS offers precise 3D meshes of humans and articulated objects during interactions, captured through a hybrid motion capture (MoCap) system that combines inertial and optical tracking techniques. The data collection process prioritizes realism and physical authenticity, resulting in a dataset that significantly advances interaction understanding. A detailed comparison between CHAIRS and previous HOI datasets is outlined in Tab. 1.

3.1. Data Collection

Overview CHAIRS encompasses a total of 1390 sequences depicting articulated interactions involving humans and sittable objects like chairs, sofas, stools, and benches. Exemplary sequences from CHAIRS and a showcase of the object variety can be seen in Fig. 2. Each object's exploration involves 6 distinct participants, each contributing three interaction sequences, resulting in 18 sequences for each object.

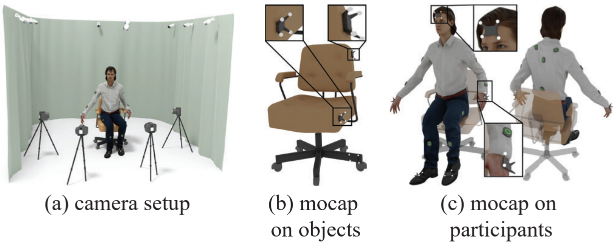

Figure 3: **Setup for data collection in the CHAIRS dataset.** Our data collection setup consisted of (a) four front-facing RGB-D cameras supplemented by a network of motion capture cameras surrounding the recording area, (b) hybrid trackers affixed to various movable parts of objects, and (c) a configuration incorporating five hybrid trackers and seventeen IMUs distributed on the participants.

In every sequence, participants execute 6 diverse actions drawn randomly from a pool of 32 interactions, such as shifting a stool, reclining on a sofa, or rotating a chair; please refer to the Supplementary Material Appx. C.1 for further details. Participant instructions were kept deliberately high-level to ensure authentic and natural performances.

Object diversity The object gallery in CHAIRS boasts an array of objects, each possessing a range of appearances and kinematic structures. Objects were curated by sourcing them online, with a focus on maximizing stylistic diversity. Notably, 28 objects incorporate at least one articulated joint, contributing to rich interaction scenarios. The 3D meshes of these objects were captured using the Scaniverse app on an iPad Pro (11-inch, 2nd generation) and subsequently refined manually to eliminate any imperfections. The 3D meshes were further segmented using the annotation tool [35] into eight functional parts. Participants received context-specific instructions tailored to the object they were interacting with.

Camera and hardware setup As depicted in Fig. 3, all sequences were exclusively captured in a controlled laboratory setup, encompassing a designated area of 5m$\times$4m ensuring complete visibility of all actions for the cameras. Four Kinect Azure DK cameras, strategically positioned to capture front-facing multi-view perspectives, were employed to record the interactions. These cameras were meticulously calibrated and synchronized. To ensure the precision of ground-truth poses for both humans and objects, a commercial inertial-optical hybrid MoCap system was incorporated alongside the Kinect setup; for further specifics, see the subsequent section.

3.2. Motion Capture (MoCap) System

Hybrid MoCap Our MoCap system is composed of a MoCap suit outfitted with 5 hybrid trackers and 17 wearable Inertial Measurement Units (IMUs), alongside a pair of gloves equipped with 12 IMUs each. The setup further includes supplementary hybrid trackers and a collection of 8 high-speed cameras. A hybrid tracker, which encompasses 4 optical markers and an IMU, is capable of accurately measuring its own 6D pose even in conditions of substantial occlusion. The arrangement of our data collection setup is illustrated in Fig. 3. For capturing the pose of a human or an object part, either an IMU or a hybrid tracker can be utilized to record the global orientation or 6D pose, respectively.

Capturing articulated object poses The recording process of articulated object poses in the context of interactions unfolds across three phases. First, we position the object into its canonical pose and affix a hybrid tracker to each of its movable components. Subsequently, we calculate the relative transformation between the object part and the trackers. During the recording process, the real-time ground-truth 6D pose of each object part is computed based on the tracker poses. Finally, we match the rigid parts to the kinematic structure of the object to yield high-fidelity object poses.

Capturing human body poses For human poses and shapes, we adopt the SMPL-X [38] representation. Participants are attired in a MoCap suit incorporating 17 IMUs, don a pair of MoCap gloves, and have 5 hybrid trackers affixed to their heads, hands, and feet. It is noteworthy that while hybrid trackers capture 6D poses, IMUs solely measure global orientations. The optimization of human model shape parameters is undertaken to ensure that the reconstructed SMPL-X mesh aligns with the positions of the hybrid trackers. As a result, the MoCap system delivers real-time estimated human poses and shapes during recording.

3.3. Post-processing

Data alignment Due to the disparate 3D coordinates and temporal clocks of Kinect cameras and the MoCap system, alignment becomes crucial. This alignment is achieved by correlating the 3D coordinates of Kinect sequences with MoCap reconstructions through plane-to-plane correspondences [44], a technique that mitigates the influence of outliers, disturbances, and partial overlaps. For the synchronization of temporal sequences from Kinect and MoCap, time-lagged cross-correlation [45] is applied, a common approach for aligning two sequences with relative time shifts.

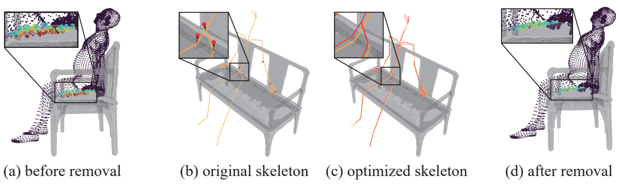

Figure 4: **Illustration of the penetration removal process.** In panels (a) and (d), the small purple points represent human vertices devoid of penetration, while the larger colored points indicate instances of penetration. The red points symbolize the most pronounced penetration, whereas the blue points signify minimal contact. Panels (b) and (c) feature yellow lines that depict the original skeleton configuration, red markers denoting the target joints undergoing optimization, and red lines illustrating the resultant optimized skeleton configuration.

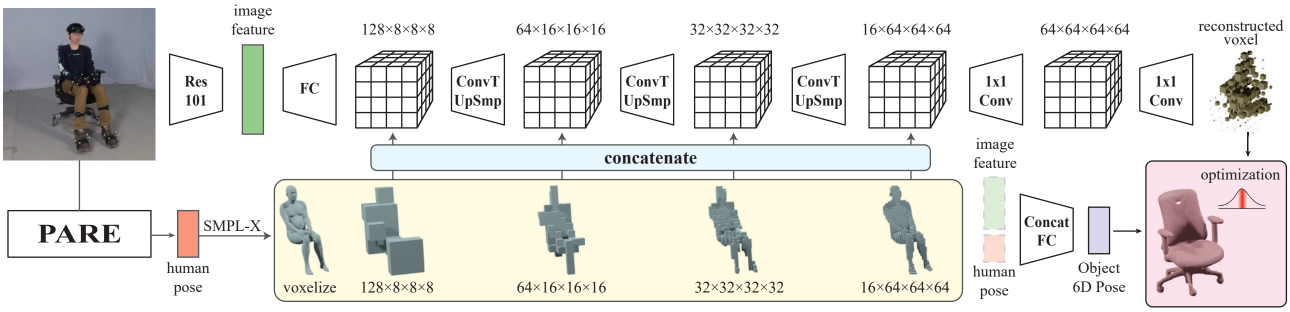

Figure 5: **Model architecture.** The reconstruction model leverages the predicted voxelized representation of the human to facilitate the estimation of pose for the interacting object. We undertake root 6D pose regression for the object utilizing the image feature in conjunction with SMPL-X parameters. The predictions along with an interaction prior are harnessed for the refinement of the final estimated pose.

Penetration removal Owing to the limited sensor count and discrepancies in limb lengths, unrealistic contacts and penetrations persist in captured 3D interactions. To address this issue, we rectify these physical anomalies with a carefully devised optimization algorithm, as depicted in Fig. 4. Given a parameterized human body and an articulated object point cloud, we compute penetration depths between the human and object point clouds. Subsequently, we utilize the transpose of the linear-blend-skinning weights of SMPL-X to aggregate the maximum penetration depth and direction to the human skeleton joints. This information is then employed to calculate a target skeleton that mitigates the penetration. Finally, we employ a gradient-based optimization technique to adjust the human model to the new skeleton while maintaining proximity to the MoCap reconstruction. This process reduced the average penetration depth in `CHAIRS` from 3.5 cm to 2.6 cm, with an average contact value of 0.2 cm.

Ensuring data quality Following data alignment and penetration removal, the Chamfer distances between annotations and observations in `CHAIRS` are measured at 2.8 cm for objects and 1.9 cm for humans. This level of quality is comparable to a recent dataset [2], which reports Chamfer distances of 2.4 cm and 1.8 cm, respectively.

Privacy protection To safeguard identities, we apply face blurring [34] to all participant faces. Furthermore, we informed all participants that they retain the right to have their data removed from `CHAIRS` at any time.

4. Articulated Object Pose Estimation

`CHAIRS` offers extensive potential for various AHOI tasks, including detection, motion generation, physics-based analysis, and even language-guided motion generation with additional annotations. We highlight the value of `CHAIRS` by focusing on the task of articulated object pose estimation. Despite recent advancements in articulated object pose estimation [58, 11, 14, 68] and HOI reconstruction [5, 46, 69, 54], the challenge of articulated object pose estimation in the context of f-AHOI remains largely unaddressed. This specific context demands accurate object pose estimation in scenarios involving heavy occlusion and dense contact.

4.1. Task Definition

Given an observed image I, the parameterized human model $H = (\beta, \theta_b, \theta_h, R_b, T_b)$, and the meshes $X = \{X_i, i = 1, \cdots, N\}$ representing the interacting object with N parts, our task involves estimating the object pose $O = \{(R_i, T_i), i = 0, \cdots, N\}$. Here, $\beta \in \mathbb{R}^{10}$, $\theta_b \in \mathbb{R}^{21\times6}$, $\theta_h \in \mathbb{R}^{30\times6}$, $R_b \in \mathbb{R}^6$, and $T_b \in \mathbb{R}^3$ represent the shape and pose parameters of the SMPL-X [38] model. Specifically, $(R_0 \in \mathbb{R}^6, T_0 \in \mathbb{R}^3)$ corresponds to the root pose of the object, while $\{(R_i \in \mathbb{R}^6, T_i \in \mathbb{R}^3)\}$ denotes the global rotation and translation for each part X_i. The orthogonal 6D representation [70] is used for representing rotations in both human and object poses.

4.2. Model Architecture

Our approach for object pose estimation is rooted in an interaction-aware framework that harnesses the fine-grained geometric relationships present in HOIs, along with learned interaction priors. This approach comprises two key stages. Given an image and estimated SMPL-X [38] parameters, we first estimate object occupancy grids and root poses using a reconstruction model. Subsequently, we fine-tune the reconstructed human-object pair using a learned interaction prior. The overall framework of our model is illustrated in Fig. 5, while Fig. 6 showcases the interaction prior model.

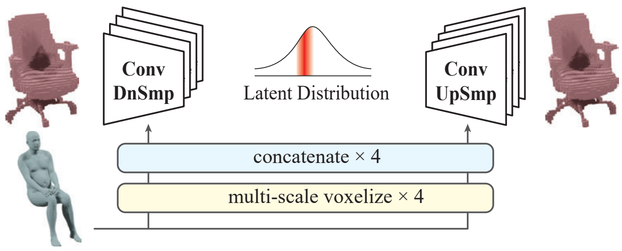

Figure 6: **Illustration of the interaction prior model.** The interaction prior model, realized as a cVAE, generates object voxels based on conditioning information from human voxels. During optimization, we aim to minimize the norm of the latent code.

4.3. Object Reconstruction and Pose Initialization

Given an observation I, we estimate human pose and shape using an off-the-shelf estimator. These estimated human shapes H' are then voxelized into four resolutions with Kaolin [23]. To better exploit the geometric relationship between human-object pairs, we guide the estimation of object shape and pose using human pose. Specifically, we begin by extracting ResNet-101 [18] features from the image and subsequently estimate object voxels based on these image features using a 3D decoder. This decoder comprises three 3DConvT layers, along with upsampling layers at distinct resolutions, and two additional 1x1 3DConv layers. Furthermore, we fuse the convolutional feature grids with human voxels at each resolution, amplifying the influence of human pose. The final 3DConv layer generates the estimated object occupancy grid $\mathcal{V}_O'$. Additionally, we concatenate image features extracted from ResNet-101 with SMPL-X parameters, and employ an additional MLP to regress the root pose (R_0', T_0') of the object. This root pose also serves as the initialization for the optimization process.

To train the reconstruction model, we first initialize the human shape estimator with pre-trained weights from the PARE model [38], followed by fine-tuning using the `CHAIRS`. Subsequently, we fix the weights of the PARE model and proceed to train the reconstruction model, utilizing the object pose estimation loss $\mathcal{L}^{\mathcal{O}}$. This loss is characterized by the L1 loss computed on object voxels.

4.4. Interaction Prior

To capture the nuanced relationship between humans and interacting objects, we introduce an interaction prior model based on a cVAE. This model learns the conditional distribution of object occupancy given the human shape.

In this context, the cVAE prior model conditions on a multi-resolution voxelized human, with the objective of reconstructing a voxelized object. The architecture employs 3DConvNets as both encoder and decoder components. During training, we feed the voxelized object into the encoder to acquire object features at multiple scales. These object features are then combined with the multi-resolution human voxels corresponding to each layer. An MLP estimates the latent Gaussian distribution $\mathcal{N}(\mu, \sigma)$, which is used to parameterize the latent code z through re-parameterization. This latent code is subsequently decoded using the decoder. The feature grids at each decoder layer are concatenated with the corresponding human voxel condition.

Training the prior model occurs on `CHAIRS` and involves four distinct loss components:

$$\mathcal{L}_{\mathrm{P}} = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{pene}} + \mathcal{L}_{\mathrm{contra}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{recon}}$ and $\mathcal{L}_{\mathrm{KL}}$ denote the standard reconstruction and KL divergence losses, respectively. $\mathcal{L}_{\mathrm{pene}}$ constitutes a penetration loss, penalizing voxel grids occupied by both humans and objects. $\mathcal{L}_{\mathrm{contra}}$ serves to maximize the distance of latent variables between original and augmented noisy data. Augmentation of training data involves introducing random noise to a portion of the samples.

4.5. Pose Optimization with Interaction Prior

To reconstruct the intricate human-object relationship and refine the object poses, we employ an optimization stage that builds upon initialized poses, utilizing kinematic insights and the interaction prior. This process involves the object's CAD model, Unified Robot Description Format (URDF), estimated SMPL-X parameters H', and object voxels $\mathcal{V}_O'$ from the reconstruction model. We initiate the object model $\hat{O}$ using estimated root transformations and random part states. We iteratively update $\hat{O}$'s parameters by minimizing the combined objective $\mathcal{J}_{\mathrm{recon}} + \mathcal{J}_{\mathrm{z}}$:

$$\mathcal{J}_{\mathrm{recon}} = \|V(\hat{O}) - \mathcal{V}_O'\|_2, \quad \mathcal{J}_{\mathrm{z}} = \|\mathtt{Enc}(H', \hat{O})\|, \tag{2}$$

where $V(\cdot)$ is the voxelization function. $\mathcal{J}_{\mathrm{recon}}$ measures the voxelized object model's distance from the estimated object voxels. $\mathcal{J}_{\mathrm{z}}$ enforces a small norm for the latent predicted by the cVAE encoder, regulating proximity to the interaction prior. The process of pose optimization with interaction prior is illustrated in Fig. 7.

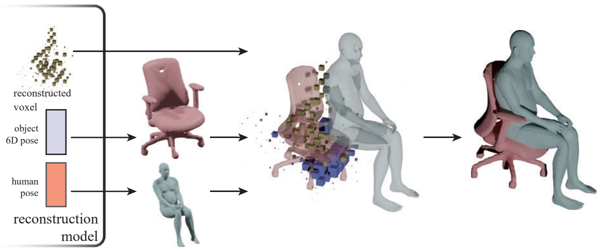

Figure 7: **An illustration of pose estimation with interaction prior.** Starting with the reconstruction output, we optimize the object according to the **reconstructed voxel** and **interaction prior**.

Object pose optimization Optimized parameters include the root 6D pose R, T of the object and joint parameters Φ (if applicable), controlling part rotation and shift under kinematic constraints. For joints (except the root), we consider revolute, prismatic, or combined revolute-prismatic configurations. The latter, such as the joint linking a chair's base and seat, restricts rotation and shift along the same axis.

During optimization, we initiate the root pose R, T using the estimated root 6D pose from the object pose estimation model. All joint parameters Φ are set to zero. Optimization involves minimizing the reconstruction loss $\mathcal{J}_{\mathrm{recon}}$ and interaction prior loss $\mathcal{J}_{\mathrm{z}}$ through gradient descent. Calculating losses necessitates determining object part occupancy post-application of R, T, and Φ for each optimization step. As direct voxelization lacks differentiability regarding object parameters R, T and Φ, we employ trilinear interpolation on the affined $(0, 1)$ voxel grid to resample voxel occupancy. This permits gradient flow for root and joint parameter updates. Post-optimization, parameters yield an updated 3D object model and enhanced representation (*e.g.*, mesh) with reduced geometric error.

Contrastive loss We intend our interaction prior to grasp a comprehensive human-object interaction distribution through a conditional Gaussian model. Reasonable and common spatial relationships between human and object latent codes should cluster near the Gaussian mean, in contrast to unreasonable ones. A contrastive loss aids training of the interaction prior model alongside penetration, reconstruction, and KL-divergence losses. Positive examples $(H, \mathcal{V}_O)$ involve an observed human H and object voxel $\mathcal{V}_O$. Corresponding negative examples $(H, \mathcal{V}'_O)$ are generated by perturbing the object, adding noise to root and articulated poses, and voxelizing $\mathcal{V}'_O$. The contrastive loss $\mathcal{L}_{\text{contra}}$, defined as $\mathcal{L}_{\text{contra}} = max(0, ||\texttt{Enc}(\mathcal{V}_O, H)|| - ||\texttt{Enc}(\mathcal{V}'_O, H)||)$, guides latent codes of perturbed human-object pairs away from the distribution centroid. Here, `Enc` represents the conditional encoder of our proposed cVAE-based prior model.

5. Experiments

Experimental settings We split `CHAIRS` into training, testing, and validation sets; 70% of objects are used for training, 20% for testing, and the remaining for validation. We evaluate our model under two settings: with (*w/ opt*) and without optimization (*w/o opt*). In the *w/ opt.* setting, we report the chamfer distance between objects posed with ground truth and estimated transformation parameters. In the *w/o opt.* setting, we do not have the estimated transformation parameters. Thus, we report the chamfer distance between the ground-truth object mesh and the mesh obtained by running the marching cube algorithm on the reconstructed voxels.

Evaluation metrics We evaluate object pose estimation using mean rotation and translation errors for each object part. Object shape reconstruction is evaluated with chamfer distance and intersection over union (IoU). For reconstructed f-AHOI, we assess penetration depth and contact scores between the human and object. Penetration depth is the maximum depth of the object's surface within the human's body, while contact value is the shortest distance between the human and object. Contact values are clipped to [0,20cm] for distant human-object pairs.

Baseline methods We compare articulated object pose estimation with LASR [59] and ANCSH [28] as baselines, where we use depth maps as input for ANCSH. Both methods are *fine-tuned* on `CHAIRS`. Additionally, we compare our model with D3D-HOI [58], PHOSA [65], and CHORE [56] that jointly estimate human and object poses. We adapted D3D-HOI's optimization objectives to better fit `CHAIRS`'s data distribution.

5.1. Results and Analyses

Quantitative results are presented in Tab. 2. Our model, leveraging geometrical relationships, exhibits substantial improvements in pose estimation and shape reconstruction compared to existing methods. In the *w/o opt.* setting where the object is unknown, our model surpasses the state-of-the-art LASR method by a significant margin. While D3D-HOI and ANCSH excel, they assume known object structures. Remarkably, our model outperforms all baselines when provided with the object structure in the *w/ opt.* setting.

Table 2: **Comparisons against existing methods.** $*$: method requires knowledge of object structure and/or geometry; †: method does not rely on object-related knowledge.

Method	Object				HOI	
	Rot.↓ (°)	Transl.↓ (mm)	CD↓ (mm)	IoU↑ (%)	Pene.↓ (mm)	Cont.↓ (mm)
LASR† [59]	/	/	205.2	/	/	/
Ours (w/o opt.)†	/	/	**160.2**	**11.03**	**4.530**	**2.720**
ANCSH* [28]	/	/	90.36	/	/	/
PHOSA* [65]	29.31	175.2	177.9	7.60	2.046	1.689
D3D-HOI* [58]	27.31	119.2	126.9	16.60	7.472	**1.163**
CHORE* [56]	21.82	87.58	95.40	16.44	**1.050**	1.742
Ours (w/ opt.)*	**19.35**	**66.23**	**72.30**	**21.57**	1.143	1.562

We present qualitative results in Fig. 8, where columns (a)-(h) illustrate the reconstruction outcomes on the test set. In these columns, we display the reconstructed meshes prior to optimization using the marching cubes algorithm. It is evident from the visualizations that our model successfully produces plausible and accurate interaction representations even before the optimization process. Notably, the optimization step enhances the finer interaction details.

5.2. Ablations

We conduct three ablation studies to assess the efficacy of our model's design choices. Quantitative results of these ablation studies are presented in Tab. 3.

Table 3: **Ablation of interaction, prior, and contrastive loss.**

Method	Object				HOI	
	Rot.↓ (°)	Transl.↓ (mm)	CD↓ (mm)	IoU↑ (%)	Pene.↓ (mm)	Cont.↓ (mm)
Full†	/	/	**160.2**	**11.03**	4.530	**2.720**
− prior†	/	/	165.3	10.52	**4.377**	3.295
Full*	19.35	**66.23**	**72.30**	**21.57**	1.143	**1.562**
− prior*	19.97	83.39	87.90	18.81	1.749	2.081
− contr.*	21.52	81.90	87.28	18.93	1.265	2.393
− inter.*	**17.88**	69.53	78.12	19.50	**1.022**	2.320

Prior We conduct an experiment where we remove the interaction prior model and solely optimize object poses by minimizing $\mathcal{L}_{\text{recon}}$. In both $*$ and † settings, we observe a substantial performance drop. This underscores the critical role played by the interaction prior in accurately estimating

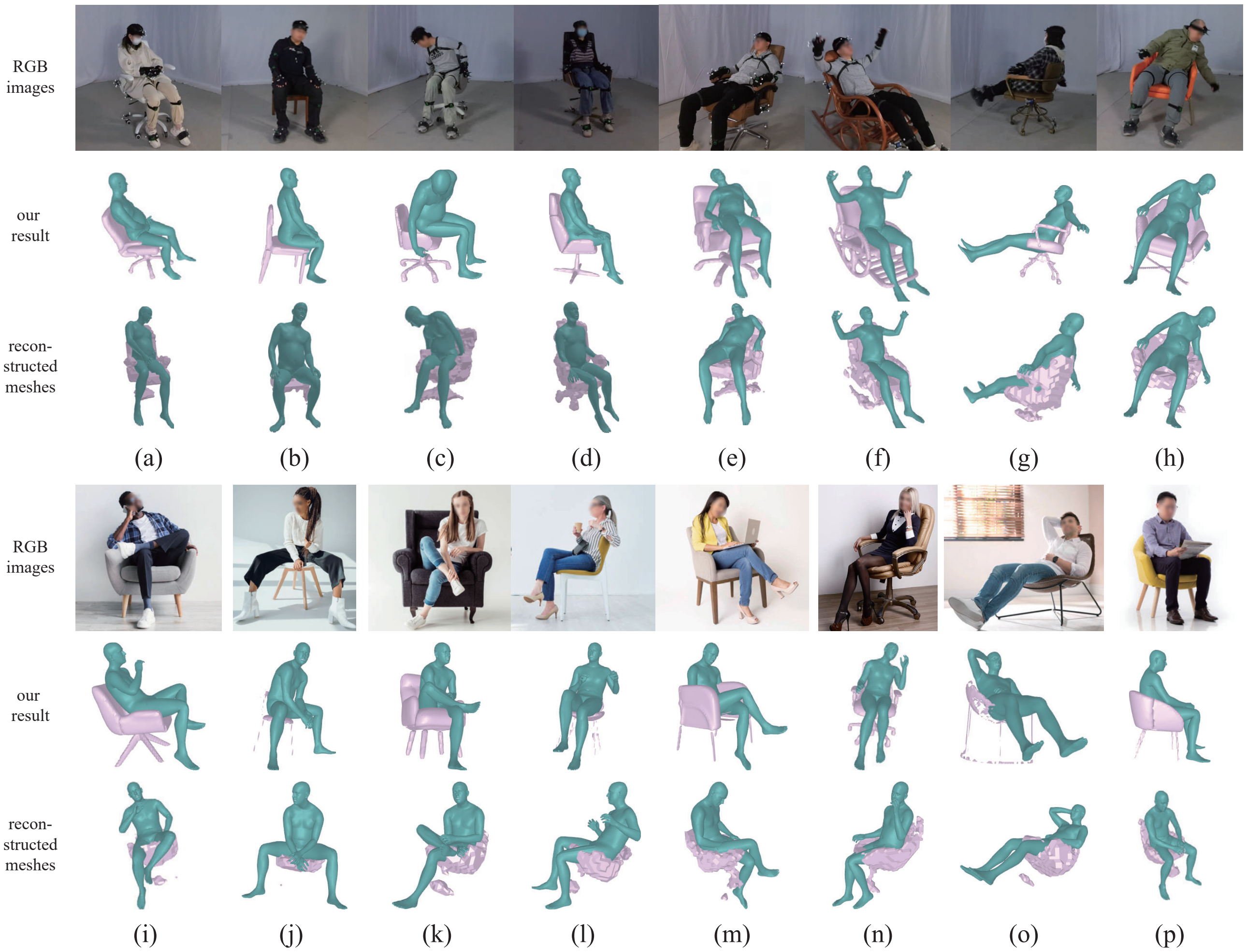

Figure 8: **Qualitative results.** (a)-(h) Test set results. (i)-(p) Wild images results. RGB images, optimized poses, and mesh obtained by running marching cubes on reconstructed voxels are shown. Please refer to Fig. A6 in Supplementary Materials for more qualitative results.

object poses. It is important to note that both settings involve an optimization step, and the primary distinction is that the $*$ model possesses access to the object's geometry and structure during optimization. When the prior model is omitted in the † setting, we observe a decrease in penetration and a larger increase in contact value. This observation suggests that our interaction prior model exerts influence by pulling the object closer to the human when they are not in contact.

Contrast In this ablation, we exclude the contrastive loss $\mathcal{L}_{\text{contra}}$ from the training of the prior model. The results are analogous to those of the −prior experiment. This outcome underscores the crucial role that the contrastive loss plays in facilitating the learning of a robust interaction prior.

Interaction We proceed to remove the concatenation of human voxels in the 3DConv layers of both the reconstruction model and the interaction prior model. This removal eliminates the interaction awareness in our model. We observe a modest degradation across all object reconstruction metrics, underscoring the importance of interaction awareness in our approach. Interestingly, the removal of interaction awareness leads to increased contact values and decreased penetration, resembling the outcomes of the −prior ablation in the *w/o opt.* setting. This suggests that interaction awareness also contributes to bringing the human and object into closer proximity. Lastly, we note an unexpected low rotation error, which we attribute to the presence of rotation symmetries in the dataset.

Additionally, we assess our method's performance under varying qualities of human pose estimation in Tab. 4. The results reveal notable improvement in object pose estimation as human poses become more accurate, thus validating our initial hypothesis. Notably, the pose estimation model [27] effectively predicts most challenging poses accurately, leaving the avenue of leveraging interactions to enhance human pose estimation as a potential future research direction.

In summary, our analysis highlights the substantial contributions of all three model components to object pose and shape reconstruction.

Table 4: **Ablation of human pose estimation quality.** GT denotes using ground-truth human poses to optimize the object poses, No prior denotes not considering human-object interaction prior.

Method	Human		Object	
	MPJPE↓(mm)	PA-MPJPE↓(mm)	CD↓(mm)	IOU↑(%)
No prior	/	/	87.90	18.81
PARE [27]	81.09	47.19	73.79	21.66
PARE(finetune)	74.50	43.99	72.30	21.57
GT	**0**	**0**	**65.50**	**23.16**

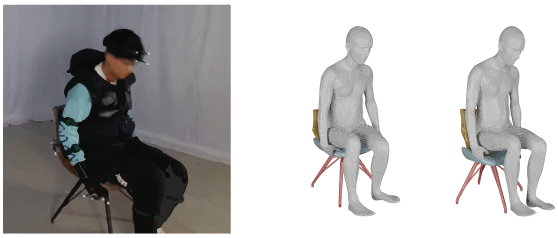

Figure 9: **Failure cases.** In situations involving rotation-symmetrical objects, our model encounters challenges with estimating rotation while maintaining a relatively low visual error.

In-the-wild generalization We also assess the model's generalizability on a limited set of internet images. As depicted in Fig. 8 (i-p), the qualitative results illustrate that our model successfully generalizes its capabilities to images captured outside of controlled laboratory settings.

Failure cases Our model encounters challenges in accurately estimating the orientation of object parts when those parts exhibit geometric similarity under specific rotations. Rotation symmetry is commonly observed in spherical and cylindrical object components, such as the base of a stool or a round seat. An illustrative example of this symmetry is presented in Fig. 9. Notably, existing methods [10, 49] address this challenge through (i) accepting multiple equally-valid ground truths and (ii) employing a min-of-N loss to calculate the smallest distance to any of these ground truths. However, implementing such methods necessitates a meticulous classification of symmetry types for each object.

Furthermore, we observe that our model's performance diminishes in scenarios where no f-AHOI is present, such as instances where a human is situated far away from an object like a chair. Under these circumstances, our model is unable to leverage interactions to enhance object pose estimation.

6. Application: Generating Interacting Humans

We further investigate the intricate relationships within AHOI by exploring the generation of interacting human poses in the presence of articulated objects. To this end, we employ a 3D conditional diffusion model known as SceneDiffuser [22], trained on our `CHAIRS`. To evaluate the quality of the generated poses, we compare them with poses generated using the same model trained on COUCH [67], a recent dataset featuring humans seated on *rigid* chairs. We use the feature extracted from the point cloud of the object via Pointnet++ as a conditioning input and flatten the SMPL-X parameters of the human to form tokens for input to a Transformer. The implementation details closely follow those of the human pose generation task described in Huang *et al*. [22]. Qualitative comparisons of the generated human poses are shown in Fig. 10. Notably, the model trained with `CHAIRS` captures more nuanced and natural geometrical relationships when interacting with articulated objects. For a more in-depth analysis of this downstream application, we direct readers to the Appx. B in Supplementary Materials.

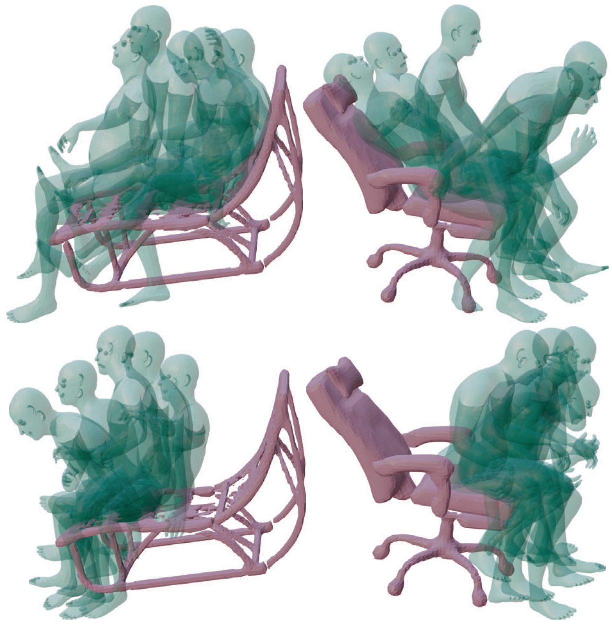

Figure 10: **Generated human poses given articulated objects.** The models are trained on `CHAIRS` **(a, b)** and COUCH [67] **(c, d)**.

7. Conclusion

We advance the study of HOI to encompass fine-grained, articulated interactions with (i) `CHAIRS`, an extensive dataset, (ii) a challenging object reconstruction problem under f-AHOI, and (iii) a strong baseline. Our `CHAIRS` captures diverse, natural AHOIs involving various sittable objects. The object reconstruction problem confronts kinematic assumptions, with our model effectively leveraging intricate interactions to resolve ambiguities.

Limitations One limitation of our work lies in the fact that the parametric human model used in `CHAIRS` does not account for clothing, leading to misalignments between the 3D annotations and the images. Consequently, the usage of pixel-aligned features may be compromised.

Acknowledgment The authors would like to thank four anonymous reviews for constructive feedback and NVIDIA for their generous support of GPUs and hardware. This work is supported in part by the National Key R&D Program of China (2022ZD0114900) and the Beijing Nova Program.

References

[1] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. Learning to generalize kinematic models to novel objects. In *Conference on Robot Learning (CoRL)*, 2019. 3

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 1, 10

[3] Markus Braun, Qing Rao, Yikang Wang, and Fabian Flohr. Pose-rcnn: Joint object detection and pose estimation using 3d object proposals. In *Intelligent Transportation Systems Conference (ITSC)*, 2016. 3

[4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[5] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5

[6] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992. 2

[7] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[8] Karthik Desingh, Shiyang Lu, Anthony Opipari, and Odest Chadwicke Jenkins. Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In *International Conference on Robotics and Automation (ICRA)*, 2019. 3

[9] Thanh-Toan Do, Ming Cai, Trung Pham, and Ian Reid. Deep-6dpose: Recovering 6d object pose from a single rgb image. *arXiv preprint arXiv:1802.10367*, 2018. 3

[10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 9

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Articulated objects in free-form hand interaction. *arXiv preprint arXiv:2204.13662*, 2022. 2, 3, 5

[12] Haw-ren Fang and Yousef Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 16(3):197–221, 2009. 2

[13] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[14] Sanjay Haresh, Xiaohao Sun, Hanxiao Jiang, Angel X Chang, and Manolis Savva. Articulated 3d human-object interactions from rgb videos: An empirical analysis of approaches and challenges. In *International Conference on 3D Vision (3DV)*, 2022. 2, 3, 5

[15] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[16] 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 (ICCV)*, 2019. 2, 3

[17] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[19] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[20] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *International Conference on Computer Vision (ICCV)*, 2011. 3

[21] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 3

[22] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 3, 9, 1

[23] Krishna Murthy Jatavallabhula, Edward Smith, Jean-Francois Lafleche, Clement Fuji Tsang, Artem Rozantsev, Wenzheng Chen, Tommy Xiang, Rev Lebaredian, and Sanja Fidler. Kaolin: A pytorch library for accelerating 3d deep learning research. *arXiv preprint arXiv:1911.05063*, 2019. 6

[24] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[25] Jia Kang, Wenjun Liu, Wenzhe Tu, and Lu Yang. Yolo-6d+: single shot 6d pose estimation using privileged silhouette information. In *IEEE International Conference on Image Processing (ICIP)*, 2020. 3

[26] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[27] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8, 9, 3

[28] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object

pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 7

[29] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[30] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[31] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[32] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[33] Frank Michel, Alexander Krull, Eric Brachmann, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Pose estimation of kinematic chain instances via object coordinate regression. In *British Machine Vision Conference (BMVC)*, 2015. 3

[34] Asmaa Mirkhan. Blurry faces. `https://github.com/asmaamirkhan/BlurryFaces`, 2020. 5

[35] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[36] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[37] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 5, 6, 3

[39] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[40] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[41] Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F Fouhey. Understanding 3d object articulation in internet videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

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

[43] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 3

[44] A. Segal, Dirk Hhnel, and S. Thrun. Generalized-icp. In *Robotics: Science and Systems (RSS)*, 2009. 4, 2

[45] Chenhua Shen. Analysis of detrended time-lagged cross-correlation between two nonstationary time series. *Physics Letters A*, 2015. 4, 3

[46] 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, 3, 5

[47] Wei-Cheng Tseng, Hung-Ju Liao, Lin Yen-Chen, and Min Sun. Cla-nerf: Category-level articulated neural radiance field. In *International Conference on Robotics and Automation (ICRA)*, 2022. 3

[48] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[49] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 9

[50] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[51] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[52] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[53] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[54] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[55] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[56] 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)*, 2022. 7, 1

[57] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, X. Wang, and Trevor Darrell. Hierarchical style-based networks for motion synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[58] 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, 3, 5, 7, 1

[59] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7

[60] Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. 3

[61] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[62] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 1, 3

[63] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[64] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3

[65] 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, 7, 1

[66] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[67] 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)*, 2022. 2, 3, 9, 1, 6

[68] Zeyu Zhang, Lexing Zhang, Zaijin Wang, Ziyuan Jiao, Muzhi Han, Yixin Zhu, Song-Chun Zhu, and Hangxin Liu. Part-level scene reconstruction affords robot interaction. In *International Conference on Intelligent Robots and Systems (IROS)*, 2023. 5

[69] 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 (ECCV)*, 2022. 3, 5

[70] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5