International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-01984-1

InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction from Multi-view RGB-D Images

Yinghao Huang[1] · **Omid Taheri**[1] · **Michael J. Black**[1] · **Dimitrios Tzionas**[2]

Received: 23 April 2023 / Accepted: 1 January 2024

Abstract
Humans constantly interact with objects to accomplish tasks. To understand such interactions, computers need to reconstruct these in 3D from images of whole bodies manipulating objects, e.g., for grasping, moving and using the latter. This involves key challenges, such as occlusion between the body and objects, motion blur, depth ambiguities, and the low image resolution of hands and graspable object parts. To make the problem tractable, the community has followed a divide-and-conquer approach, focusing either only on interacting hands, ignoring the body, or on interacting bodies, ignoring the hands. However, these are only parts of the problem. On the contrary, recent work focuses on the whole problem. The GRAB dataset addresses whole-body interaction with dexterous hands but captures motion via markers and lacks video, while the BEHAVE dataset captures video of body-object interaction but lacks hand detail. We address the limitations of prior work with InterCap, a novel method that reconstructs interacting whole-bodies and objects from multi-view RGB-D data, using the parametric whole-body SMPL-X model and known object meshes. To tackle the above challenges, InterCap uses two key observations: (i) Contact between the body and object can be used to improve the pose estimation of both. (ii) Consumer-level Azure Kinect cameras let us set up a simple and flexible multi-view RGB-D system for reducing occlusions, with spatially calibrated and temporally synchronized cameras. With our InterCap method we capture the InterCap dataset, which contains 10 subjects (5 males and 5 females) interacting with 10 daily objects of various sizes and affordances, including contact with the hands or feet. To this end, we introduce a new data-driven hand motion prior, as well as explore simple ways for automatic contact detection based on 2D and 3D cues. In total, InterCap has 223 RGB-D videos, resulting in 67,357 multi-view frames, each containing 6 RGB-D images, paired with pseudo ground-truth 3D body and object meshes. Our InterCap method and dataset fill an important gap in the literature and support many research directions. Data and code are available at https://intercap.is.tue.mpg.de.

Keywords Computer vision · Computer graphics · 3D virtual human · Human-object interaction · Machine learning · SMPL · SMPL-X

1 Introduction

A long-standing goal of Computer Vision is to understand human actions from videos. Given a video, people effortlessly figure out what objects exist in it, the spatial layout of objects, and the pose of humans. Moreover, they deeply understand the depicted action. What is the subject doing? Why are they doing this? What is their goal? How do they achieve this? To empower computers with the ability to infer such abstract concepts from pixels, we need to capture rich datasets and to devise appropriate algorithms.

Since humans live in a 3D world, their physical actions involve interacting with objects. Think of how often one goes to the kitchen, grabs a cup of water, and drinks from it. This involves contacting the floor with the feet, contacting the cup with the hand, moving the hand and cup together while maintaining contact, and drinking while the lips contact the cup. Thus, to understand human actions, it is necessary to reason in 3D about humans and objects *jointly*.

There is significant prior work on estimating 3D humans without taking into account objects (Bogo et al. 2016) and estimating 3D objects without taking into account humans (Zollhöfer et al. 2018). There is even recent work on inserting

Communicated by Bastian Goldluecke.

✉ Yinghao Huang
yinghao.huang@tuebingen.mpg.de

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany

[2] University of Amsterdam, Amsterdam, The Netherlands

Published online: 06 February 2024

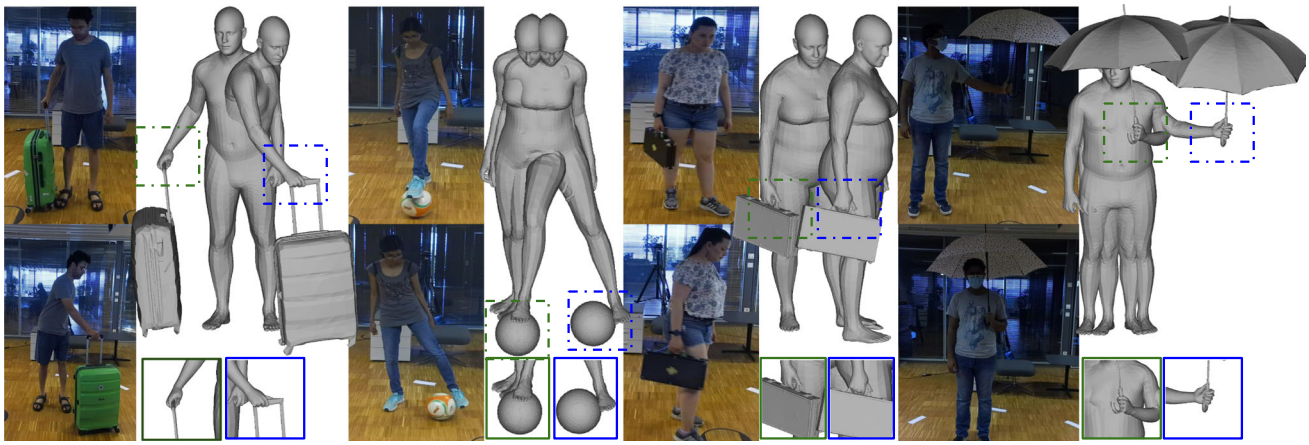

Fig. 1 Humans interact with objects to accomplish tasks. To understand such interactions we need the tools to reconstruct them from whole-body videos in 4D, i.e., as 3D meshes in motion. Existing methods struggle, due to the strong occlusions, motion blur, and low-resolution of hands and object structures in such videos. Moreover, they mostly focus on the main body, ignoring the hands and objects. We develop InterCap, a novel method that reconstructs plausible interacting whole-body and object meshes from multi-view RGB-D videos, using contact constraints to account for strong ambiguities. With this we capture the rich InterCap dataset of 223 RGB-D videos (67,357 multi-view frames, with 6 Azure Kinects) containing 10 subjects (5 fe-/males) interacting with 10 objects of various sizes and affordances; note the hand-object grasps

bodies into 3D scenes such that their interactions appear realistic (Zhang et al. 2020c; Li et al. 2019; Hassan et al. 2021). But there is little work on estimating 3D humans interacting with scenes and moving objects, in which contact is explicitly modeled and exploited. To study this problem, we need a dataset of videos with rich human-object interactions and reliable 3D ground truth.

PROX

(Hassan et al. 2019) takes a step in this direction by estimating the 3D body in a known 3D scene. The scene mesh provides information that helps resolve human pose ambiguities commonly encountered when a single camera is used. However, PROX involves only coarse interactions of bodies, static scenes with no moving objects, and no dexterous fingers. The recent BEHAVE dataset (Bhatnagar et al. 2022) uses multi-view RGB-D data to capture humans interacting with objects, but does not include detailed hand pose or fine hand-object contact. Finally, the GRAB dataset (Taheri et al. 2020) captures the kind of detailed hand-object and whole-body-object interaction that we seek but is captured using marker-based Motion Capture (MoCap) and, hence, lacks images.

We argue that what is needed is a new dataset of RGB videos containing natural human-object interaction in which the whole body is tracked reliably, the hand pose is captured, objects are also tracked, and the hand-object contact is realistic; see Fig. 1. This is challenging, and requires technical innovation to create. To that end, we design a system that uses multiple RGB-D sensors that are spatially calibrated and temporally synchronized. To build this data we fit the SMPL-X body model, which has articulated hands, by extending the PROX (Hassan et al. 2019) method to use multi-view data and grasping hand-pose priors. We also track the 3D objects with which the person interacts. The objects used in this work are representative of items one finds in daily life. We obtain accurate 3D models for each object with a handheld Artec scanner. Altogether we collect 223 sequences (67,357 multi-view frames, each containing 6 RGB-D images), with 10 subjects (5 males, 5 females) interacting with 10 everyday objects.

The problem, however, is that separately estimating the body and objects is not sufficient to ensure accurate 3D body-object contact. Consequently, a key innovation of this work is to estimate these *jointly*, while exploiting information about *contact*. Objects do not move independently, so, when they move, it means the body is in contact. We define likely contact regions on objects and on the body. Then, given frames with known likely contacts, we enforce contact between the body and the object when estimating the body and object poses. The resulting method produces natural body poses, hand poses, and object poses. Uniquely, it provides detailed pseudo ground-truth contact information between the whole body and objects in RGB video.

In summary, our major contributions are as follows: (1) We develop a novel Motion-Capture method utilizing multiple RGB-D cameras. It is relatively lightweight and flexible, yet accurate enough, thus suitable for data capture of daily scenarios. (2) We extend previous work on fitting SMPL-X to images to fit it to multi-view RGB-D data while taking into account body-object contact. (3) We capture a novel dataset that contains whole-body human motions and interaction with objects, as well as multi-view RGB-D imagery. (4) We train a new hand motion prior that improves the smoothness and realism of the reconstructed motion. (5) We explore

automatic human-object interaction detection by developing two baselines; their accuracy is around 80% on our dataset.

This article is an extension of our InterCap conference paper (Huang et al. 2022b); the two latter contributions above are new over the conference paper, while we also provide additional discussion and technical details. Our InterCap data and code are available at https://intercap.is.tue.mpg.de.

2 Related Work

There is a large literature on estimating 3D human pose and shape from images or videos (Bogo et al. 2016; Pavlakos et al. 2019; Choutas et al. 2020; Kanazawa et al. 2018; Kocabas et al. 2020; Varol et al. 2017; Mehta et al. 2017; Omran et al. 2018; Kolotouros et al. 2019; Kanazawa et al. 2019; Rempe et al. 2021; Dwivedi et al. 2024). For an exhaustive discussion, please see the surveys by (Tian et al. 2022; Wang et al. 2021; Sarafianos et al. 2016). Here we focus on the work most closely related to ours, particularly as it concerns, or enables, capturing human-object interaction.

MoCap from Multi-view Videos and IMUs. Markerless MoCap from multi-view videos (Liu et al. 2011; De Aguiar et al. 2008; Huang et al. 2017, 2022a; Joo et al. 2018) is widely studied and commercial solutions exist (e.g., Theia Markerless, DeepMotion, The Captury). Compared with traditional marker-based MoCap, markerless MoCap offers advantages of convenience, applicability in outdoor environments, non-intrusiveness, and greater flexibility. However, traditional MoCap methods, both marker-based and markerless ones, focus on extracting a 3D skeleton. This is useful for several applications, such as biomechanics, gaming or fitness. However, skeletons do not suffice for our goal of reasoning about body-scene contact. To enable that, we need to capture the full body *surface*.

Various 3D human representations have been proposed, with recent work focused on learning a parametric mesh-based model of body shape from large-scale collections of 3D scans (Anguelov et al. 2005; Loper et al. 2015; Romero et al. 2017; Pavlakos et al. 2019; Osman et al. 2020; Xu et al. 2020; Osman et al. 2022). Here we use the SMPL-X model (Pavlakos et al. 2019) because it contains fully-articulated hands, which are critical for reasoning about object manipulation. The body parameters are often estimated by fitting the 3D generative model to various 2D cues, such as joints detected by neural networks (Cao et al. 2019; Wei et al. 2016; Newell et al. 2016) or silhouettes (Rhodin et al. 2016; Xu et al. 2018; Alldieck et al. 2018). Though effective, these monocular video-based methods suffer from depth ambiguity and occlusions. To address this, researchers combine IMUs with videos to obtain better results (von Marcard et al. 2018; Pons-Moll et al. 2010), reaching even real-time performance (Malleson et al. 2017).

Many methods estimate 3D bodies from multi-view images but focus on skeletons and not 3D bodies (He et al. 2020; Iskakov et al. 2019; Qiu et al. 2019; Tu et al. 2020; Dong et al. 2019, 2021a; Zhang et al. 2020b). Recent work addresses 3D body shape estimation from multiple views (Huang et al. 2017; Dong et al. 2021b; Zhang et al. 2021b). Most related to our work are two recent datasets. The RICH dataset (Huang et al. 2022a), fits SMPL-X bodies to multi-view RGB videos taken both indoors and outdoors. The method uses a detailed 3D scan of the scene and models the contact between the body and the world. RICH does not include any object motion; the scenes are completely rigid. In contrast, BEHAVE (Bhatnagar et al. 2022) contains SMPL bodies interacting with 3D objects that move. We go beyond this to integrate novel contact constraints and to capture hand pose, which is critical for human-object interaction. Moreover, BEHAVE focuses on large objects like boxes and chairs, whereas we have a wider range of object sizes, including smaller objects like cups.

Human-Object Interaction. There has been a lot of work on modeling or analyzing human-object interactions (Yao and Fei-Fei 2010; Hamer et al. 2009; Oikonomidis et al. 2011; Rogez et al. 2015; Tzionas et al. 2016; Hampali et al. 2020; Hasson et al. 2019; Karunratanakul et al. 2020; Bhatnagar et al. 2022). A detailed discussion is out of the scope of this work. Here, we focus on modeling and analyzing human-object interaction in 3D space. Most existing work, however, only focuses on estimating hand pose (Hasson et al. 2019; Hampali et al. 2020; Hasson et al. 2020; Romero et al. 2010; Tzionas and Gall 2013), ignoring the strong relationship between body motion, hand motion, and object motion. Recent work considers whole-body motion. For example, the GRAB (Taheri et al. 2020) and ARCTIC (Fan et al. 2023) datasets provide detailed whole-body motion (in a parametric SMPL-X body format) and object motion, for rigid and articulated objects, respectively. Unfortunately, these methods are based on marker-based MoCap and usually do not include videos. Here we focus on tracking the whole-body motion, object motion, and the detailed hand-object contact to provide ground-truth 3D information in RGB video.

Joint Modeling of Humans and Scenes. There is some prior work addressing human-object contact in both static images and video. For example, PHOSA (Zhang et al. 2020a) estimates a 3D body and a 3D object with plausible interaction from a single RGB image. Our focus here, however, is on dynamic scenes. Motivated by the observation that natural human motions always happen inside 3D scenes, researchers have proposed to model human motion jointly with the surrounding environment (Hassan et al. 2019; Savva et al. 2016; Cao et al. 2020; Yi et al. 2022; Taheri et al. 2024, 2022). In PROX (Hassan et al. 2019) the contact between humans and scenes is explicitly used to resolve ambiguities in pose estimation. The approach avoids bodies interpenetrat-

ing scenes while encouraging contact between the scene and nearby body parts. Recently, IPMAN (Tripathi et al. 2023b) extends PROX with a body-stability intuitive-physics term. However, this works only for single frames and interaction with the ground. Finally, HOT (Chen et al. 2023) detects contact automatically as 2D heatmaps in the image, while DECO (Tripathi et al. 2023a) detects 3D body contact given a natural color image.

Prior work also infers the most plausible position and pose of humans given a 3D scene (Zhang et al. 2020c; Li et al. 2019; Hassan et al. 2021). Recently, MOVER (Yi et al. 2022) estimates the 3D scene and the 3D human directly from a static monocular video in which a person interacts with the scene. While the 3D scene is ambiguous and the human motion is ambiguous, by exploiting contact, the method resolves ambiguities, improving the estimates of both the scene and the person. Unfortunately, this assumes a static scene and does not model hand-object manipulation.

Datasets. Traditionally, MoCap is performed using marker-based systems inside lab environments. An approach for this uses MoSh (Loper et al. 2014) to fit a SMPL or SMPL-X body to the markers (Mahmood et al. 2019). An advanced version of this is used for GRAB (Taheri et al. 2020), for capturing interaction and contact with rigid objects, by also fitting object meshes to markers. Such approaches typically lack synchronized RGB videos. Recently, ARCTIC (Fan et al. 2023) extends GRAB's approach not only for interactions with articulated objects, but also for capturing synchronized multi-view RGB videos (including an egocentric camera). Moreover, MoYo (Tripathi et al. 2023b) captures SMPL-X meshes and RGB videos together with synchronized pressure measurements with an instrumented Yoga mat. The HumanEva (Sigal et al. 2010) and Human3.6M (Ionescu et al. 2014) datasets combine multi-view RGB video capture with synchronized ground-truth 3D skeletons from marker-based MoCap. These datasets lack ground-truth 3D body meshes, are captured in a lab setting, and do not contain human-object manipulation. 3DPW (von Marcard et al. 2018) is the first in-the-wild dataset that jointly features natural human appearance in video and accurate 3D pose. However, this dataset does not track objects or label human-object interaction.

PiGraphs (Savva et al. 2016) and PROX (Hassan et al. 2019) provide both 3D scenes and human motions but are relatively inaccurate, because they rely on a single RGB-D camera. This makes these datasets ill-suited as evaluation benchmarks. The recent RICH dataset (Huang et al. 2022a) addresses many of these issues with indoor and outdoor scenes, accurate multi-view capture of SMPL-X body meshes, 3D scene scans, and human-scene contact. However, RICH is not appropriate for our task, as it does not include object manipulation.

An alternative approach is the one used by GTA-IM (Cao et al. 2020) and SAIL-VOS (Hu et al. 2019), which generate synthetic human-scene interaction data using either 3D graphics or 2D videos. These datasets feature high-accuracy ground truth but lack visual realism.

In summary, we believe that a 3D human-object interaction dataset needs to have accurate hand poses to be useful, since hands are how people most often interact with objects. We compare our InterCap dataset with other ones in Table 1.

3 InterCap Method

Our goal is to accurately estimate the human and object motion throughout a video, without using instrumentation like IMUs or optical markers. Our markerless motion-capture method is built on top of the PROX-D method (Hassan et al. 2019), which uses a single RGB-D camera to track the human motion in a known 3D scene. To improve the body tracking accuracy we extend this method to use multiple RGB-D cameras; here we use the latest Azure Kinect cameras. The motivation is that multiple cameras observing the body from different angles give more information about the human and object motion. Moreover, commodity RGB-D cameras are much more flexible to deploy out of controlled lab scenarios than more specialized devices.

The key technical challenge lies in accurately estimating the 3D pose and translation of the objects while a person interacts with them. In this work we focus on 10 variously-sized rigid objects common in daily life, such as cups and chairs. Being rigid does not make the tracking of the objects trivial because of the occlusion by the body and hands. This issue is more severe for small handheld objects like a cup, despite using many cameras. While there is a rich literature on 6 DoF object pose estimation, much of it ignores hand-object interaction. Recent work in this direction is promising but still focuses on scenarios that are significantly simpler than ours, cf. (Sun et al. 2022).

Similar to previous work on hand and object pose estimation (Hampali et al. 2020) from RGB-D videos, in this work we assume that the 3D meshes of the objects are known in advance. To this end, we first gather the 3D models of these objects from the Internet whenever possible and scan the remaining objects ourselves. To fit the known object models to image data, we first preform semantic segmentation, find the corresponding object regions in all camera views, and fit the 3D mesh to the segmented object contours via differentiable rendering. Since heavy occlusion between humans and objects in some views may make the segmentation results unreliable, aggregating segmentation from all views boosts the object tracking performance.

In the steps above, both the subject and object are treated separately and processing is conducted per frame, with no

Table 1 Dataset statistics

Name	Real Data	Mov Obj	Accur Poses	Dext Hands	RGB Seq	D Img	# of Videos	# of Views	# of Img	# of Subj
GTA-IM (Cao et al. 2020)	✗	✗	✓	✗	✓	✓	119	14-67	1 M	50
SAIL-VOS (Hu et al. 2019)	✗	✗	✗	✗	✗	✗	201	1	111K	✗
HumanEva (Sigal et al. 2010)	✓	✗	✓	✗	✓	✗	56	4/7	80K	4
Human3.6M (Ionescu et al. 2014)	✓	✗	✓	✗	✓	✗	165	4	3 M	11
AMASS (Mahmood et al. 2019)	✓	✗	✓	✗	✗	✗	11.2K	✗	✗	344
3DPW (von Marcard et al. 2018)	✓	✗	✓	✗	✓	✗	60	1	51K	5
GRAB (Taheri et al. 2020)	✓	✓	✓	✓	✗	✗	1.33K	✗	✗	10
ARCTIC (Fan et al. 2023)	✓	✓	✓	✓	✓	✗	242	8+1	1.2M	9
MoYo (Tripathi et al. 2023b)	✓	✗	✓	✗	✓	✗	200	8	1.7M	1
PiGraphs (Savva et al. 2016)	✓	✗	✗	✗	✓	✓	63	1	100K	5
PROX (Hassan et al. 2019)	✓	✗	✗	✗	✓	✓	20	1	100K	20
RICH (Huang et al. 2022a)	✓	✗	✓	✗	✓	✗	142	6-8	577K	22
BEHAVE (Bhatnagar et al. 2022)	✓	✓	✓	✗	✓	✓	321	4	15K	8
InterCap (ours)	✓	✓	✓	✓	✓	✓	223	6	400K	10

Comparison of InterCap to existing datasets. We define three categories: (top) synthetic data, (middle) marker-based data, (bottom) markerless data. InterCap achieves a practical balance between accuracy and flexibility of deploying the camera setup. Here "#" stands for "number", "Obj." for "Objects", "Seq." for "Sequences", "Img." for "Images", and "Subj." for "Subjects"

temporal smoothness or contact constraint applied. This inevitably produces jittery motions and heavy penetration between objects and the body. Making matters worse, our human pose estimation exploits OpenPose for 2D keypoint detection, which struggles when the object occludes the body or the hands interact with it. To mitigate this issue and still get reasonable body, hand and object pose in these challenging cases, we manually annotate the frames where the body or the hand is in contact with the object, as well as the body, hand and object vertices that are most likely to be in contact. This manual annotation can be tedious; automatic detection of contact is an open problem (we explore this here with early baselines). We then explicitly encourage the labeled body and hand vertices to be in contact with the labeled object vertices. We find that this straightforward idea works well in practice, yielding reasonable hand and object poses. More details are discussed below.

3.1 Multi-kinect Setup

We use 6 Azure Kinects to track the human and object together, deployed in a "ring" layout in an office; see Fig. 2. Multiple RGB-D cameras provide a good balance between body tracking accuracy and applicability to real scenarios, compared with costly professional MoCap systems like Vicon, or cheap and convenient but not-so-accurate monocular RGB cameras. Moreover, this approach does not require applying any markers, making the images natural. Intrinsic camera parameters are provided by the manufacturer. Extrinsic camera parameters are obtained via camera

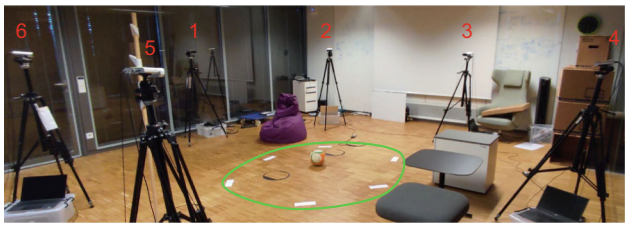

Fig. 2 The setup of our 6 Azure Kinect cameras in an indoor space; the area where the subject moves is highlighted with green color

calibration with Azure Kinect's API (Microsoft 2022). However, these can be a bit noisy, as non-neighbouring cameras in a sparse "ring" layout don't observe the calibration board well at the same time. Thus, we manually refine in MeshLab the extrinsics by comparing the point clouds for neighbouring cameras for several iterations. The hardware synchronization of Azure Kinects is empirically reasonable. Given the calibration information, we choose a camera's coordinate frame as the master frame and transform the point clouds from the other frames into the master one, which is where we fit the SMPL-X and object models.

3.2 Sequential Object-Only Tracking

Object Segmentation. To track an object during interaction, we need reliable visual cues about it to compare with the 3D object model. To this end, we perform semantic segmentation by applying PointRend (Kirillov et al. 2020) to the images. We then extract the object instances that correspond to the

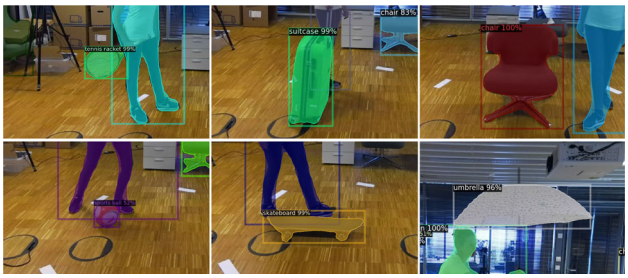

Fig. 3 Object detection and segmentation via PointRend (Kirillov et al. 2020) for all views; images cropped for visualization purposes

categories of our objects; for examples see Fig. 3. We assume that the subject interacts with a single object. Note that, in contrast to previous approaches where the objects occupy a large portion of the image (Hampali et al. 2020; Hassan et al. 2019; Tzionas et al. 2016; Oikonomidis et al. 2011), in our case the entire body is visible, thus, the object takes up a small part of the image and is often occluded by the body and hands; our setting is much more challenging. We observe that PointRend works reasonably well for large objects like chairs, even with heavy occlusion between the object and the human, while for small objects, like a bottle or a cup, it struggles significantly due to occlusion.

In extreme cases, it is possible for the object to not be detected in most of the views. But even when the segmentation is good, the class label for the objects may be wrong. To resolve this, we take two steps: (1) For every frame, we detect all possible object segmentation candidates and their labels. This step takes place offline and only once. (2) During the object tracking phase, for each view, we compare the rendering of the tracked object from the ith frame with all the detected segmentation candidates for the $(i+1)$th frame, and preserve only the candidate with the largest overlap ratio. This render-compare-and-preserve operation takes place iteratively during tracking; we empirically find that this works well in practice.

Object Tracking. Given object masks via semantic segmentation over the whole sequence, we track the object by fitting its model to observations via differentiable rendering (Kato et al. 2018; Loper and Black 2014). This is similar to past work for hand-object tracking (Hampali et al. 2020). We assume that the object is rigid and its mesh is given. The configuration of the rigid object in the tth frame is specified via a 6D rotation and translation vector ξ. For initialization, we manually obtain the configuration of the object for the first frame by matching the object mesh to the measured point clouds; the rest of the frames are processed automatically. Let R_S and R_D be functions that render a synthetic mask and depth image for the tracked 3D object mesh, M. Let also $S = \{S_v\}$ be the observed object masks and $D = \{D_v\}$ be corresponding depth values for the current frame, where v is the camera view. Then, we minimize $E_O(\xi; S, D) =$

$$\sum_{\text{view } v} \lambda_{segm} \|(R_S(\xi, M, v) - S_v) * S_v\|_F^2 + \lambda_{depth} \|(R_D(\xi, M, v) - D_v) * S_v\|_F^2, \tag{1}$$

where the two terms compute how well the rendered object mask and depth image match the detected mask and observed depth over all views; the symbol $*$ is an element-wise multiplication, $\|.\|_F$ is the Frobenius norm, and λ_{segm} and λ_{depth} are steering weights set empirically. For simplicity, we assume that transformations from the master to other camera frames are encoded in the rendering functions R_S, R_D; we do not denote these explicitly here.

3.3 Sequential Human-Only Tracking

We estimate body shape and pose over the whole sequence from multi-view RGB-D videos in a per-frame manner. This is similar in spirit with PROX-D (Hassan et al. 2019), but, in our case, there is no 3D scene constraint and multiple cameras are used. The human pose and shape are optimized independently in each frame. We use the SMPL-X (Pavlakos et al. 2019) model to represent the 3D human body. SMPL-X is a function that returns a water-tight mesh given parameters for shape, β, pose, θ, facial expression, ψ, and translation, γ. We follow the common practice of using a 10-dimensional space for shape, β, and a 32-dimensional latent space in VPoser (Pavlakos et al. 2019) to present body pose, θ.

We minimize the loss defined below. For each frame we essentially extend the major loss terms used in PROX (Hassan et al. 2019) to multiple views:

$$\begin{aligned} E_B(\beta, \theta, \psi, \gamma; K, J_{est}) = E_J + \lambda_D E_D + \\ \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_h} E_{\theta_h} + \lambda_{\theta_f} E_{\theta_f} + \\ \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{P} E_\mathcal{P}, \end{aligned} \tag{2}$$

where $E_\mathcal{E}$, E_β, E_{θ_b}, E_{θ_h}, E_{θ_f} are prior loss terms for facial expressions ($\mathcal{E}$), whole-body shape (β), and for the pose (θ) of the body (b), hand (h), and face (f). And E_α is a prior for extreme elbow and knee bending; for detailed definitions see (Hassan et al. 2019). E_J is a 2D keypoint re-projection loss:

$$E_J(\beta, \theta, \gamma; K, J_{\text{est}}) = \sum_{view\ v} \sum_{joint\ i} k_i^v w_i^v \rho_J\big(\Pi_K^v\big(R_{\theta\gamma}(J(\beta)_i)\big) - J_{est,i}^v\big), \tag{3}$$

where $\theta = \{\theta_b, \theta_h, \theta_f\}$, v and i iterate through views and joints, k_i^v and w_i^v are the per-joint weight and detection confidence, ρ_J is a robust Geman-McClure error function (Geman and McClure 1987), Π_K^v is the projection function with K camera parameters, $R_{\theta\gamma}(J(\beta)_i)$ are the posed 3D joints of SMPL-X, and $J_{est,i}^v$ the detected 2D joints. The term E_D is:

$$E_D(\beta, \theta, \gamma; K) = \sum_{view\ v} \sum_{p \in P^v} \min_{v \in V_b^v} \|v - p\|, \tag{4}$$

where P^v is Azure Kinect's segmented point cloud for the vth view, and V_b^v are SMPL-X vertices that are visible in

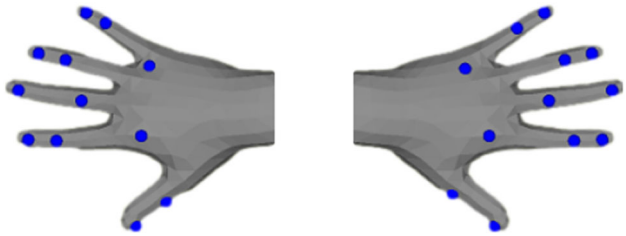

Fig. 4 Virtual marker configuration used in our work to train hand motion priors on the GRAB dataset (Taheri et al. 2020). The blue spheres indicate the vertices chosen as the proxies for the markers

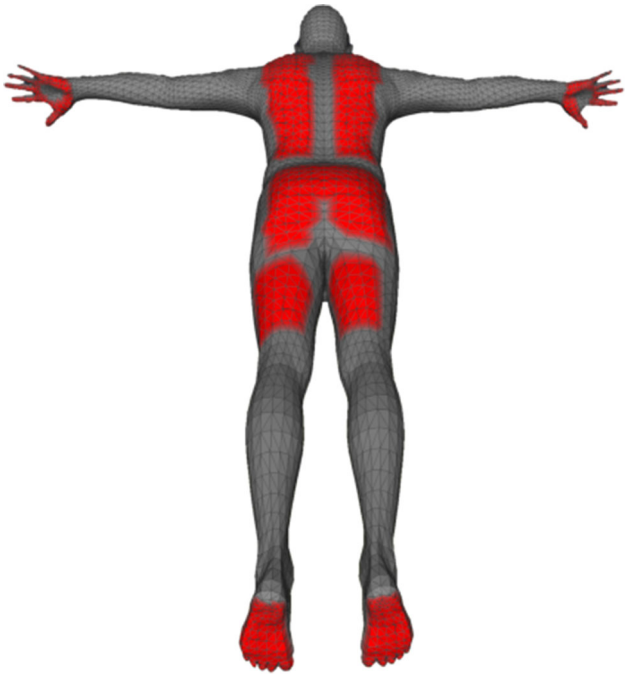

Fig. 5 Annotation of likely body contact areas (red color)

this view. This term measures how far the estimated body mesh is from the combined point clouds, so that we minimize this discrepancy. Note that, unlike PROX, we have multiple point clouds due to the multiple camera views, i.e., our E_D is a multi-view extension of PROX's (Hassan et al. 2019) loss. For each view we dynamically compute the visible body vertices, and "compare" them against the segmented point cloud for that view.

Finally, the term $E_{\mathcal{P}}$ penalizes self-interpenetration of the SMPL-X body mesh; see PROX (Hassan et al. 2019) for a more detailed and formal definition of this:

$$E_{\mathcal{P}}(\theta, \beta, \gamma) = E_{\mathcal{P}_{self}}(\theta, \beta). \tag{5}$$

3.4 Joint Human-Object Tracking Over All Frames

We treat the result of the first rounds of optimization (Sects. 3.2, 3.3) as initialization for refinement via *joint* optimization of the body and the object *over all frames*, subject to *contact* constraints.

For this we fix the body shape parameters, β, as the mean body shape computed over all frames from the first stage, as done in Huang et al. (2017). Then, we jointly optimize the object pose and translation, ξ, body pose, θ, and body translation, γ, over all frames. We add a temporal smoothness loss to reduce jitter for both the human and the object. We also penalize the body-object interpenetration, as done in PROX (Hassan et al. 2019). A key difference is that in PROX the scene is static, while here the object is free to move.

To encourage contact, we annotate the body areas that are most likely to be in contact with the objects and, for each object, we annotate vertices most likely to be contacted. These annotations are shown in Figs. 5 and 6-right, respectively, in red. We also annotate the range (frame IDs) of sub-sequences where the body is in contact with objects, and encourage contact between them explicitly to get reasonable tracking even when there is heavy interaction and occlusion between hands and objects. Note that the latter manual annotation is lightweight, as only the range of frames where contact takes place is recorded.

Formally, we perform global optimization over all T frames, and minimize a loss, E, that is composed of an object (O) fitting loss, E_O, a body (B) fitting loss, E_B, a motion smoothness prior (Zhang et al. 2021a) loss, $E_{\mathcal{S}}$, and a loss penalizing object acceleration, $E_{\mathcal{A}}$. We also use a ground support loss, $E_{\mathcal{G}}$, that encourages the human and the object to be above the ground plane, i.e., to not penetrate it. Moreover, we use a body-object contact loss, $E_{\mathcal{C}}$, that attaches the body to the object for frames with contact. Last, we use two smoothness terms E_L and E_R for the left and right hand, respectively; note that, because hands are smaller than other body parts, the keypoint detections and depth values are noisy. This makes them more prone to jitter. The loss E is defined as:

$$\begin{aligned} E = & \frac{1}{T} \sum_{\text{frame } t} \Big[E_O(\Xi_t; \mathcal{S}_t, \mathcal{D}_t) + E_B(\beta^*, \Theta_t, \Psi_t, \Gamma_t; \mathcal{J}_{est}) \Big] + \\ & \frac{1}{T} \sum_{\text{frame } t} \Big[E_{\mathcal{P}}(\Theta_t, \beta^*, \Gamma_t) + E_{\mathcal{C}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t, \Xi_t, M_o) \Big] + \\ & \frac{\lambda_{\mathcal{G}}}{T} \sum_{\text{frame } t} \Big[E_{\mathcal{G}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t) + E_{\mathcal{G}'}(\Xi_t, M_o) \Big] + \\ & \frac{\lambda_{\mathcal{Q}}}{T} \sum_{\text{frame } t} \Big[Q_t * E_{\mathcal{C}'}(\beta^*, \Theta_t, \Psi_t, \Gamma_t, \Xi_t, M'_o) \Big] + \\ & \lambda_{\mathcal{A}} E_{\mathcal{A}}(\Xi, T, M_o) + \lambda_{\mathcal{S}} E_{\mathcal{S}}(\Theta, \Psi, \Gamma; \beta^*, T) + \\ & \lambda_{\mathcal{S}} \Big[E_L(\Theta, \Psi, \Gamma; \beta^*, T) + E_R(\Theta, \Psi, \Gamma; \beta^*, T) \Big], \end{aligned} \tag{6}$$

where for all frames $t = \{1, \ldots, T\}$ of a sequence, $\Theta = \{\theta_t\}$ and $\Gamma = \{\gamma_t\}$ are the body poses and translations, respectively, $\Psi = \{\psi_t\}$ are the facial expressions, $\Xi = \{\xi_t\}$ is the object rotations and translations, $\mathcal{S} = \{S_t\}$ and $\mathcal{D} = \{D_t\}$ are masks and depth patches, $\mathcal{J}_{est} = \{J_{est,t}\}$ are detected 2D keypoints, M_o is the object mesh, and β^* the mean body

Fig. 6 The objects of our InterCap dataset. Left: Color photos. Right: Annotations (shown in red) for likely contact areas on the objects

shape. The various energy terms are described in detail in the following. The parameters $\lambda_{\mathcal{G}}$, $\lambda_{\mathcal{Q}}$, $\lambda_{\mathcal{S}}$, and $\lambda_{\mathcal{A}}$ are steering weights that are set empirically.

The object fitting term, E_O, comes from Eq. 1 and the body fitting term, E_B, comes from Eq. 2, while, under the hood, both go through all views, ν. The self-penetration term, $E_{\mathcal{P}}$, comes from Eq. 5.

The ground-support terms, $E_{\mathcal{G}}$ and $E_{\mathcal{G}'}$, build on the fact that no human or object vertex, respectively, should be below the ground plane, and penalize any vertex penetrating the ground. We estimate the ground plane surface by fitting a plane to chosen floor points in the observed point clouds. Let $p_{\mathcal{G}}$ be a point on the ground plane and $n_{\mathcal{G}}$ be the corresponding normal; both are defined once and offline. Then, the term $E_{\mathcal{G}}$ for body-ground penetration is defined as:

$$E_{\mathcal{G}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t) = \left\| RL\Big(n_{\mathcal{G}} * \big(p_{\mathcal{G}} - W(\beta^*, \Theta_t, \Psi_t, \Gamma_t)\big)\Big)\right\|^2, \tag{7}$$

where RL is the ReLU function, and $*$ here is the inner product of vectors. The term $E_{\mathcal{G}'}$ for object-ground penetration follows a similar formulation:

$$E_{\mathcal{G}'}(\Xi_t, M_o) = \left\| RL(n_{\mathcal{G}} * (p_{\mathcal{G}} - W'(\Xi_t, M_o)))\right\|^2. \tag{8}$$

where W' denotes the operation of first rigidly deforming the object according to Ξ_t and then concatenating the vertices into a single vector.

The contact term, $E_{\mathcal{C}}$, encourages the annotated likely contact areas of the body (see Fig. 5) to contact the object as in PROX (Hassan et al. 2019):

$$E_{\mathcal{C}}(\beta^*, \Theta_t, \Psi_t, \Gamma_t, \Xi_t, M_o) = CD\Big(H\big(W(\Theta_t, \Psi_t, \Gamma_t, \beta^*)\big), H'\big(W'(\Xi_t, M_o)\big)\Big), \tag{9}$$

where CD refers to the Chamfer Distance function, H is a function that returns only the annotated body-contact vertices of Fig. 5, H' returns for these body-contact vertices the closest points on the object (taking into account the entire object), W' deforms rigidly the object as explained in the previous paragraph, and W similarly (non-rigidly) deforms the SMPL-X mesh and concatenates the vertices into a single vector. Note that this term considers the entire object for establishing "general" contacts, but the hands are likely to contact only certain object parts.

Then, the contact term, $E_{\mathcal{C}'}$, focuses only on the hands and only on a subset of each object according to its affordances. Since grasps are delicate, they need a higher accuracy than "general" contact (described in the above paragraph), thus, we need to accurately specify the frames that contain such contact. Thus, we manually annotate binary vectors $Q = \{Q_t\}, t = \{1, \ldots, T\}$; Q_t is set to 1 if in the tth frame there is contact with a "graspable" object, and set to 0 otherwise.

The vertex acceleration term, $E_{\mathcal{A}}$, is a simple hand-crafted motion prior that encourages smooth motion trajectories for the object:

$$E_{\mathcal{A}}(\Xi; T, M_o) = \frac{1}{T-2}\sum_{t=2}^{T-1}\Big\| W'(\Xi_{t-1}, M_o) + W'(\Xi_{t+1}, M_o) - 2 * W'(\Xi_t, M_o)\Big\|^2 \tag{10}$$

where M_o is the object mesh, and W' deforms the object as described above.

The motion smoothness loss, $E_{\mathcal{S}}$, penalizes abrupt position changes for body vertices. $E_{\mathcal{S}}$ employs the learned motion prior of LEMO (Zhang et al. 2021a) and is defined as:

$$E_{\mathcal{S}}(\Theta, \Psi, \Gamma, A; T, \beta^*) = \frac{\sum_{t=1}^{T-1}\left\| z_{t+1}^{opt} - z_t^{opt}\right\|^2}{Q(T-2)}, \tag{11}$$

where T is the sequence length, and Q is a constant representing the number of virtual body-markers of LEMO; see the paper of (Zhang et al., 2021a) for an explanation (note that they use a different symbol). Moreover, z_t^{opt} is the latent vector for the t-th frame from LEMO's pre-trained motion auto-encoder (F_S):

$$Z^{opt} = F_S(X_{\Delta}^{opt}) = [z_1^{opt}, z_2^{opt}, ..., z_{T-1}^{opt}], \tag{12}$$

where X_{Δ}^{opt} is a (concatenated) vector containing the temporal position change of LEMO's virtual body-markers. For more details, please refer to LEMO (Zhang et al. 2021a).

Hands typically suffer more from jitter compared to the rest of the body, due to noisy keypoint detections and depth values. Therefore, we add two additional smoothness loss

terms for the left hand, E_L, and the right hand, E_R. These are defined similarly to the LEMO-style E_S (see above paragraph), and are trained separately for the left and right hand on the GRAB (Taheri et al. 2020) dataset, which contains accurate and realistic hand grasping motions. We adopt the hand marker configuration proposed in GRAB, as shown in Fig. 4. We use the default network structure and parameter setting as in LEMO (and our E_S term described above). These two terms share the same steering-weight value as E_S in Eq. 6.

3.5 Optimization Details

Similar to the per-frame optimization in the first stages (Sects. 3.2, 3.3), for the second stage (Sect. 3.4) we use the L-BFGS optimization method (Nocedal and Wright 2006) with strong Wolfe line search. The optimization stops when the loss plateaus (relative decrease less than a threshold) or when the maximum number of steps is reached; see values in the code.

For the second stage, the body shape parameters are fixed as the mean of all per-frame shape parameters obtained in the first stage. Moreover, the body pose for each frame is initialized with the corresponding per-frame pose obtained in the first stage.

In our experiments the first stages take 3 to 5 min for a single frame, while trivially supporting parallel per-frame computation. In contrast, the second stage takes around 20 h for 1000 frames.

4 Automatic Interaction Detection

Contact has been used to improve 3D human pose reconstruction (Hassan et al. 2019; Zhang et al. 2021a; Rempe et al. 2021). Data-driven methods are still at their infancy (Chen et al. 2023; Shimada et al. 2022). Instead, typically a distance heuristic is used to "detect" contact (Hassan et al. 2019). However, our setting features several challenging objects with a small size, e.g., a cup, or thin parts, e.g., an umbrella. In these cases, the initially reconstructed hands and objects (Sects. 3.2, 3.3) are not accurate enough for heuristics to work successfully. Thus, we manually label each frame by visually inspecting the multi-view images. Though effective, this practice does not scale. Thus, below we explore ways for automatic interaction "detection" with two baselines.

"2D" baseline. A simple way is to compare the segmentation masks of the human and the object for all views of a frame. More formally, for each view, v, of a certain frame, we detect the binary body mask, S_v^b, and object mask, S_v^o, where a value of 1 denotes that the pixel belongs to the body/object. Then, the number of intersecting mask pixels for the view v is given by:

$$x_v = \| S_v^b \odot S_v^o \|_F^2, \tag{13}$$

where $\odot$ is the Hadamard product and F the Frobenius norm. Then, a binary flag, f_v, indicating whether there is contact for the v-th view can be obtained by comparing x_v to an empirically-set threshold T_{view}:

$$f_v = \begin{cases} 1, & if\ x_v > T_{\text{view}} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

A frame contains contact if the number of its views with contact is bigger than an empirically-set threshold, T_{frame}, namely: $\sum_v f_v > T_{\text{frame}}$. This method depends on the quality of mask segmentation, the image resolution, camera intrinsics and distance of the subject to the camera. When object segmentation fails due to heavy occlusions, the method is not applicable. In our settings, we set $T_{\text{view}} = 10$ and $T_{\text{frame}} = 2$.

"3D" baseline. The first optimization stages (Sects. 3.2, 3.3) produce an initial per-frame 3D reconstruction for both the human and the object. Thus, another criterion for detecting contact, can be whether the 3D meshes of the human and object lie close enough to each other in 3D space. More formally, given the reconstructed 3D body mesh, M_b, and object mesh, M_o, for a certain frame, we consider that these are in contact when the closest Chamfer distance between them is below an empirically-set threshold $T_d = 1mm$, namely: $min(CD(M_b, M_o)) < T_d$.

5 InterCap Dataset

We use the proposed InterCap algorithm (Sect. 3) to capture the InterCap dataset, which uniquely features whole-body interactions with objects in multi-view RGB-D videos.

Data-capture Protocol. We use 10 everyday objects, shown in Fig. 6-left, that vary in size and "afford" different interactions with the body, hands or feet; we focus mainly on hand-object interactions. We recruit 10 subjects (5 males and 5 females) that are between 25 and 40 years old. The subjects are recorded while interacting with 7 or more objects, according to their time availability. Subjects are shown a sample motion for each object and are instructed to interact with objects as naturally as possible. However, they are asked to avoid very fast interactions that cause severe motion blur (Azure Kinect supports only up to 30 FPS), or misalignment between the RGB and depth images for each Kinect (due to technicalities of RGB-D sensors). We capture up to 3 sequences per object depending on object shape and functionality, and by picking an interaction intent from the list below, as in GRAB (Taheri et al. 2020):

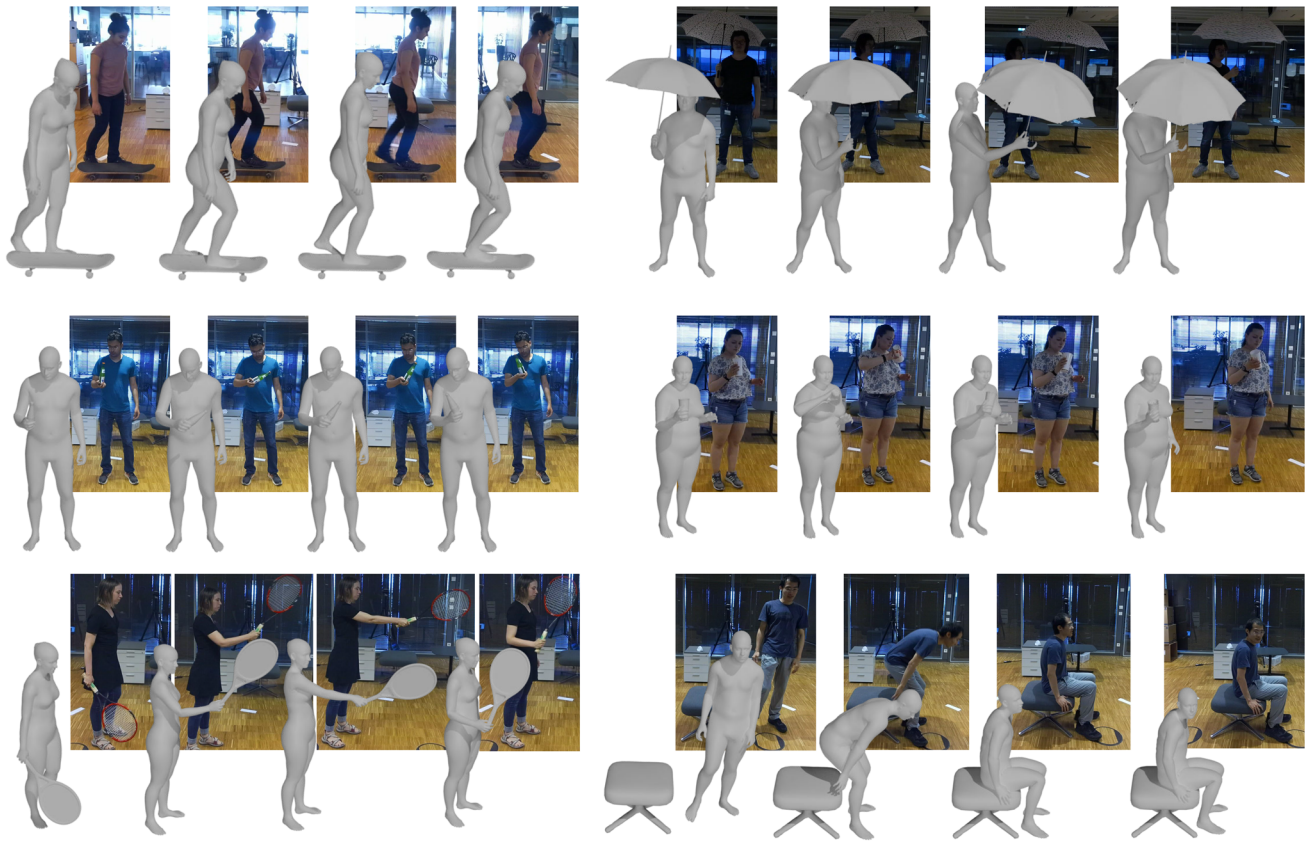

Fig. 7 Samples from our InterCap dataset, drawn from four sequences with different subjects and objects. The estimated 3D object and SMPL-X human meshes have plausible contacts that agree with the input images. Best viewed zoomed in

- "**Pass**": The subject passes the object on to another imaginary person standing on their left/right side; a graspable area needs to be free for the other person to grasp.
- "**Check**": The subject inspects visually the object from several viewpoints by first picking it up and then manipulating it with their hands to see several sides of it.
- "**Use**": The subject uses the object in a natural way that "agrees" with the object's affordances and functionality for everyday tasks.

We also capture each subject performing a freestyle interaction of their choice. All subjects gave informed written consent to publicly share their data for research.

4D Reconstruction. Our InterCap method (Sect. 3) takes as input multi-view RGB-D videos and outputs 4D meshes for the human and object, i.e., 3D meshes over time. Humans are represented as SMPL-X meshes (Pavlakos et al. 2019), while object meshes are acquired with an Artec hand-held scanner. Some dataset frames along with the reconstructed meshes are shown in Figs. 1 and 7; see also the video on our website. Reconstructions look natural, with plausible contact between the human and the object.

Dataset Statistics. InterCap has 223 RGB-D videos with a total of 67,357 multi-view frames (6 RGB-D images each). For a comparison with other datasets, see Table 1.

6 Experiments

Contact Heatmaps. Figure 8-left shows contact heatmaps on each object, across all subjects. We follow the protocol of GRAB (Taheri et al. 2020), which uses a proximity metric on reconstructed human and object meshes. First, we compute per-frame binary contact maps by thresholding (at 4.5mm) the distances from each body vertex to the closest object surface point. Then, we integrate these maps over time (and subjects) to get "heatmaps" encoding contact likelihood. InterCap reconstructs human and object meshes accurately enough so that contact heatmaps agree with object affordances, e.g., the handle of the suitcase, umbrella and tennis racquet are likely to be grasped, the upper skateboard surface is likely to be contacted by the foot, and the upper stool surface by the buttocks.

Figure 8-right shows heatmaps on the body, computed across all subjects and objects. Heatmaps show that most of InterCap's interactions involve mainly the right hand. Contact on the palm looks realistic, and is concentrated on the fingers and MCP joints. The "false" contact on the dorsal side is attributed to our challenging camera setup and interaction scenarios, as well as some reconstruction jitter.

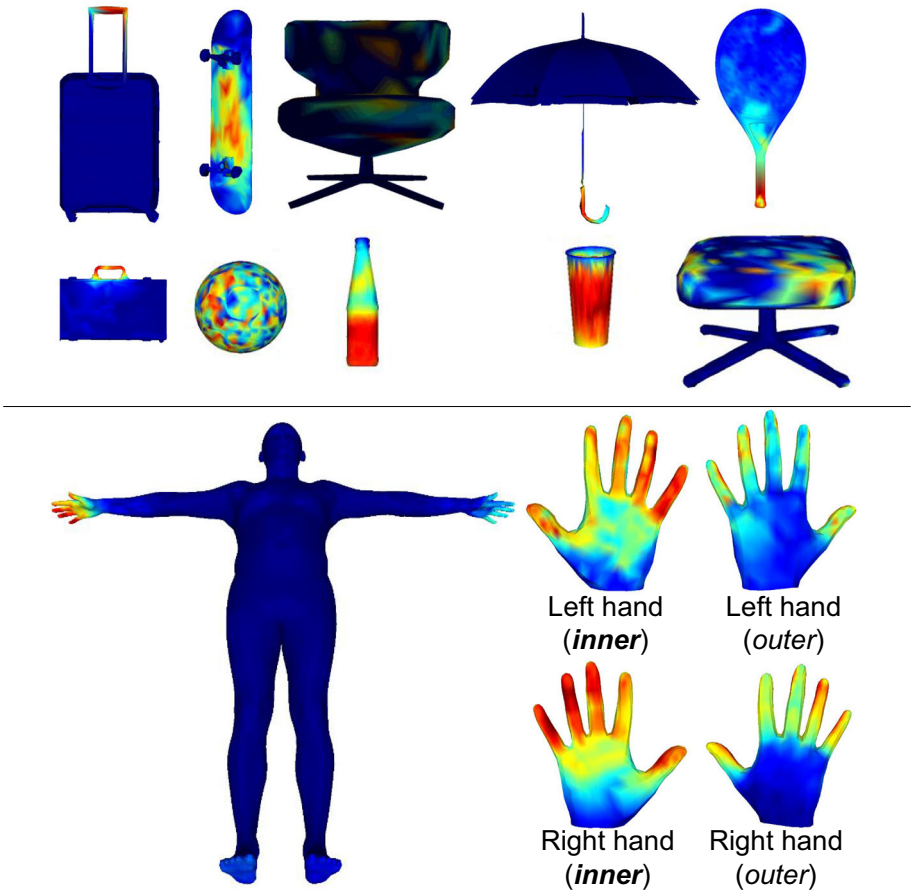

Fig. 8 Contact heatmaps for each object (across all subjects) and the human body (across all objects and subjects). Contact likelihood is color-coded; high likelihood is shown with red, and low with blue. Color-coding is normalized separately for each object, the body, and each hand

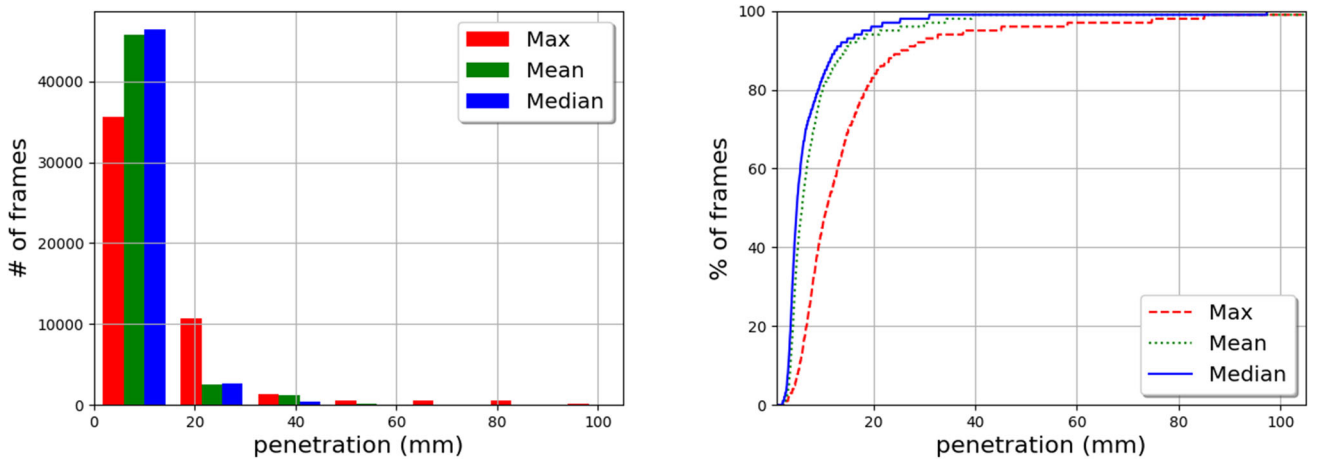

Fig. 9 Statistics of human-object mesh penetration for all InterCap sequences. **Left:** The number of frames (Y-axis) with a certain penetration depth (X-axis). **Right:** The percentage of frames (Y-axis) with a penetration depth below a threshold (X-axis). In the legend, "Max", "Mean" and "Median" refer to three ways of reporting the penetration for each frame, i.e., taking the maximum, mean and median value of the penetration depth of all vertices, respectively

Penetration. We evaluate the penetration between human and object meshes for all sequences of our dataset. We follow the protocol of GRAB et al. (Taheri et al. 2020); we first find the "contact frames" for which there is at least minimal human-object contact, and then report statistics for these. In Fig. 9-left we show the distribution of penetrations, i.e., the number of "contact frames" (Y axis) with a certain mesh penetration depth (X axis). In Fig. 9-right we show the

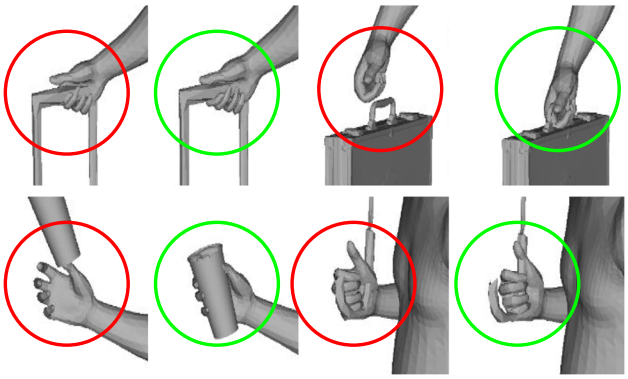

Fig. 10 Ablation of contact term. Each pair of images shows results wo/ (red) and w/ (green) the contact term. Encouraging contact results in more natural hand poses and hand-object grasps

cumulative distribution of penetration, i.e., the percentage of "contact frames" (Y axis) for which mesh penetration is below a threshold (X axis). Roughly 62% of "contact frames" have $\leq$ 7 mm, 80% $\leq$ 9.8 mm, and 98% $\leq$ 35 mm mean penetration. The mean penetration depth over all "contact frames" is 7.2 mm. In theory, being in contact means zero distance between the deformed (compressed) body part and the object. In practice, as SMPL-X does not model deformation due to contact, penetration between body parts and the object is unavoidable. We empirically find that this amount of penetration is normally not noticeable, thus acceptable for most applications that value visual naturalness more than physical correctness.

Fitting Accuracy. For every frame, we compute the distance from each mesh vertex to the closest point-cloud (PCL) point; for each human or object mesh we take into account only the respective PCL area obtained with PointRend (Kirillov et al. 2020) segmentation. The mean vertex-to-PCL distance is 19.05 mm for the body, and 18.14 mm for objects. In comparison, PROX-D (Hassan et al. 2019), our base method, achieves an error of 13.02 mm for the body. This is expected since PROX-D is free to change the body shape to fit each individual frame, while our method estimates a single body shape for the whole sequence. SMPLify-X (Pavlakos et al. 2019) achieves a mean error of 79.54 mm, for VIBE the mean error is 55.59 mm, while ExPose gets an mean error of 71.78 mm. These numbers validate the effectiveness of our method for body tracking. Note that these methods are based on monocular RGB images only, so there is not enough information for them to accurately estimate the global position of the 3D body meshes. Thus we first align the output meshes with the point clouds, then compute the error. Note that the error is bounded from below for two reasons: (1) it is influenced by factory-design imperfections in the synchronization of Azure Kinects, and (2) some vertices reflect body/object areas that are occluded during interaction and their closest PCL point is a wrong correspondence. Despite this, InterCap

Table 2 Evaluation of automatic interaction detection (Sect. 4) on InterCap

Baseline	Sub_1 (%)	Sub_2 (%)	Sub_3 (%)	Sub_4 (%)	Sub_5 (%)	Sub_6 (%)	Sub_7 (%)	Sub_8 (%)	Sub_9(%)	Sub_10 (%)	Mean (%)
"2D"	71	68	71	73	69	59	67	60	75	69	68
	98	94	93	99	97	85	91	100	96	98	95
"3D"	81	77	72	76	70	82	83	82	78	73	77
	100	100	100	100	100	100	100	100	100	100	100

For each cell we report two metrics: (top) the detection accuracy, namely, the percentage of frames correctly classified, and (bottom) the percentage of frames where the method is applicable, by successfully segmenting both the body and the object (sometimes segmentation fails, mostly for the object). The manual contact annotation from InterCap is used as the ground truth to evaluate the proposed two baseline methods: "2D" refers to the first baseline of Sect. 4 that is purely based on 2D visual cues, while "3D" refers to the second baseline that uses the initial 3D body and object mesh reconstructions (from Sects. 3.2 and 3.3)

Table 3 Evaluation of automatic interaction detection (Sect. 4) on the validation set of RICH (Huang et al. 2022a)

Baseline	Sub_1 (%)	Sub_2 (%)	Sub_3 (%)	Sub_4 (%)	Sub_5 (%)	Mean (%)
"2D"	37	62	59	56	90	63
	100	100	100	100	56	94

We report the two metrics of Table 2. Note that there is no "3D" baseline for RICH, as estimating the full 3D scene from 2D images is too challenging

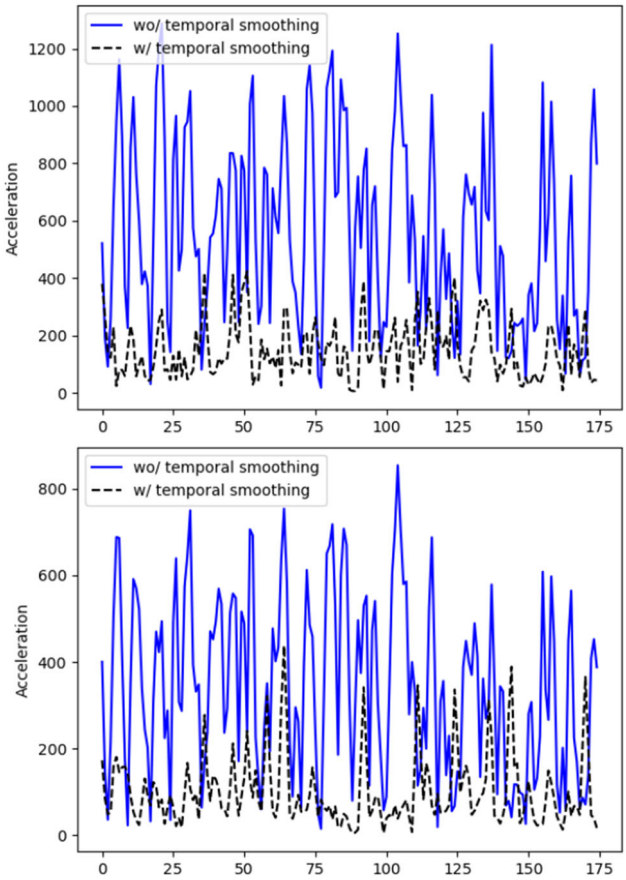

Fig. 11 Ablation of the temporal smoothing term. Acceleration of a random vertex on the back of the main body (upper plot) and the right hand (bottom plot) with (dashed line) and without (solid line) temporal smoothing for a random sequence over the first 175 frames. Dashed lines (w/ temporal smoothing) correspond to lower acceleration, i.e., less motion jitter. The average acceleration value (in m/s^2 for 30 fps sequences) for the upper figure is 564.75 (wo/ smoothing) and 148.69 (w/ smoothing), while for the lower figure it is 343.33 (wo/ smoothing) and 101.20 (w/ smoothing). Thus, smoothing reduces jitter, however, there is still room for improvement

empirically estimates reasonable bodies, hands and objects in interaction, as reflected in the contact heatmaps and penetration metrics above.

The hand smoothness terms E_L and E_R in Eq. 6 help recover more natural and less jittery hand motion at the cost of increased run-time (roughly 15% slower), compared with a simple acceleration penalty loss. One may chose one over the other, depending on the application and its major need (speed or accuracy).

Ablation of the Contact Term. Figure 10 shows results with-/out our term that encourages body-object contact; visualization"zooms" into hand-object grasps. We see that encouraging contact yields more natural hand poses and fewer interpenetrations. This is backed up by the contact heatmaps and penetration metrics discussed above.

Ablation of the Temporal Smoothing Term. Figure 11 shows results with-/out our temporal smoothing term. The plots show the acceleration of a randomly-chosen vertex on the main body (upper plot) and the right hand (bottom plot). For each plot, we show results for 3 different motions, denoted with a different color. The solid lines show results without the temporal smoothing term. The dashed lines of the same color show the same motions with the smoothing term; these are clearly smoother. We empirically find that a learned motion prior in the style of Zhang et al. (2021a), for both the case of the body and the hands, produces more natural motion dynamics than handcrafted ones (Huang et al. 2017).

Discussion on Jitter. Despite the smoothing, some jitter is still inevitable. We attribute this to two factors: (1) OpenPose and PointRend are empirically relatively sensitive to occlusions and illumination (e.g., reflections, shadows, poor lighting); the data terms for fitting 3D models depend on these. (2) Azure Kinects have a reasonable synchronization, yet, there is still a small delay among cameras to avoid depth-camera interference; the point cloud "gathered" across views is a bit "patchy" as information pieces have a small time difference. The jitter is more intense for hands relatively to the body, due to their low image resolution, motion blur, and coarse point clouds. Adding our learned motion priors for the main body and the hands encourages smoother and more natural motion dynamics, however, balancing the data and prior terms in the loss to also preserve contacts is tricky. Despite the aforementioned challenges, InterCap is a good step towards capturing everyday whole-body interactions with commodity hardware. Future work will study advanced motion and grasping priors.

Towards Automatic Interaction Detection. Although we manually annotate the parts of the sequences where the subject interacts with the object, this does not scale. Thus, here we explore the automatic detection of interaction in image sequences with two baselines, as described in Sect. 4. We evaluate the baselines on our InterCap dataset and the RICH dataset (Huang et al. 2022a); the latter features accu-

rate poses and contact between humans and a static scene. We show the results in Tables 2 and 3, where "2D" denotes the first and "3D" the second baseline of Sect. 4. We see that the "3D" baseline outperforms the "2D" one, for both the detection accuracy (percentage of correctly classified frames) and the percentage of frames for which the method is applicable (due to effectively segmenting both the body and object). However, the average accuracy is less than 80%, and the maximal accuracy for all subjects is only slightly greater than 80%. This is not so surprising, given that accurate contact detection is challenging even for human annotators. For the RICH dataset, where no 3D meshes of (segmented) objects are available, only the "2D" baseline is applicable. In this case the average accuracy is around 60%. We conclude that an automatic detection of contact is promising, but more work to this end is necessary in the future.

7 Discussion

Here we focus on whole-body human interaction with everyday rigid objects. We present a novel method, called InterCap, that reconstructs such interactions from multi-view full-body videos, including natural hand poses and contact with objects. With this method, we capture the novel InterCap dataset, with a variety of people interacting with several common objects. The dataset contains reconstructed 3D meshes for the whole body and the object over time (i.e., 4D meshes), as well as plausible contacts between them. In contrast to most previous work, our method uses no special devices like optical markers or IMUs, but only several consumer-level RGB-D cameras. Our setup is lightweight and has the potential to be used in daily scenarios. Our method recovers reasonable hand poses even under strong occlusions from the object.

Extensions over (Huang et al., 2022b): We introduce a new hand smoothness model to reduce the jitter commonly observed for hands; due to the hands' small size, both 2D joint detections and depth observations tend to be noisy. We also explore simple automatic contact detection based on 2D or 3D distances, but conclude that a more involved approach is necessary.

Future work: In future work, we will study reconstructing (Taheri et al. 2020; Fan et al. 2023; Bhatnagar et al. 2022; Lepetit 2020) interactions with smaller objects and dexterous manipulation, as well as synthesizing such interactions (Taheri et al. 2024; Braun et al. 2024; Wu et al. 2022). Finally, we will explore learning-based contact detection from images (Chen et al. 2023; Tripathi et al. 2023a; Brahmbhatt et al. 2020; Narasimhaswamy et al. 2020).

Code and data: See https://intercap.is.tue.mpg.de.

Acknowledgements We thank Chun-Hao P. Huang, Hongwei Yi, Jiaxiang Shang, and Mohamed Hassan for helpful discussions. We thank Yuliang Xiu, Jinlong Yang, Victoria F. Abrevaya, Taylor McConnell, Galina Henz, Marku Höschle, Senya Polikovsky, Matvey Safroshkin and Tsvetelina Alexiadis for data collection and cleaning, and Benjamin Pellkofer for IT and website support. We thank all the participants of our experiments.

Funding Open Access funding enabled and organized by Projekt DEAL.

Financial support. This work was partially supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B.

Conflict of interest. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While MJB was a part-time employee of Amazon during a portion of this project, his research was performed solely at, and funded solely by, the Max Planck Society.

References

Alldieck, T., Magnor, M., Xu, W., et al (2018) Video based reconstruction of 3D people models. In *Computer vision and pattern recognition (CVPR)*, pp. 8387–8397

Anguelov, D., Srinivasan, P., Koller, D., et al. (2005). SCAPE: Shape completion and animation of people. *Transactions on Graphics (TOG), 24*(3), 408–416.

Bhatnagar, B. L., Xie, X., Petrov, I. A., et al. (2022). BEHAVE: Dataset and method for tracking human object interactions. In *Computer vision and pattern recognition (CVPR)*, pp. 15,935–15,946

Bogo, F., Kanazawa, A., Lassner, C., et al. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European conference on computer vision (ECCV)*, pp. 561–578

Brahmbhatt, S., Tang, C., Twigg, C. D., et al. (2020). ContactPose: A dataset of grasps with object contact and hand pose. In *European conference on computer vision (ECCV)*, pp. 361–378

Braun, J., Christen, S. J., Kocabas, M., et al. (2024). Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*

Cao, Z., Gao, H., Mangalam, K., et al. (2020). Long-term human motion prediction with scene context. In *European conference on computer vision (ECCV)*, pp. 387–404

Cao, Z., Hidalgo, G., Simon, T., et al. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 43*(1), 172–186.

Chen, Y., Dwivedi, S. K., Black, M. J., et al. (2023). Detecting human-object contact in images. In *Computer vision and pattern recognition (CVPR)*, pp. 17,100–17,110

Choutas, V., Pavlakos, G., Bolkart, T., et al. (2020). Monocular expressive body regression through body-driven attention. In *European conference on computer vision (ECCV)*, pp. 20–40

De Aguiar, E., Stoll, C., Theobalt, C., et al. (2008). Performance capture from sparse multi-view video. *Transactions on Graphics (TOG), 27*(3), 1–10.

Dong, J., Jiang, W., Huang, Q., et al. (2019). Fast and robust multi-person 3D pose estimation from multiple views. In *Computer vision and pattern recognition (CVPR)*, pp. 7792–7801

Dong, Z., Song, J., Chen, X., et al. (2021b). Shape-aware multi-person pose estimation from multi-view images. In *International conference on computer vision (ICCV)*, pp. 11,158–11,168

Dong, J., Fang, Q., Jiang, W., et al. (2021). Fast and robust multi-person 3D pose estimation and tracking from multiple views. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 14*(8), 1–12.

Dwivedi, S. K., Schmid, C., Yi, H., et al. (2024). POCO: 3D pose and shape estimation using confidence. In *International conference on 3D vision (3DV)*

Fan, Z., Taheri, O., Tzionas, D., et al. (2023). ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer vision and pattern recognition (CVPR)*, pp. 12,943–12,954

Geman, S., & McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th session of the International Statistical Institute, Bulletin of the ISI*

Hamer, H., Schindler, K., Koller-Meier, E., et al. (2009). Tracking a hand manipulating an object. In *International conference on computer vision (ICCV)*, pp. 1475–1482

Hampali, S., Rad, M., Oberweger, M., et al. (2020). HOnnotate: A method for 3D annotation of hand and object poses. In *Computer vision and pattern recognition (CVPR)*, pp. 3193–3203

Hassan, M., Choutas, V., Tzionas, D., et al. (2019). Resolving 3D human pose ambiguities with 3D scene constrains. In *International conference on computer vision (ICCV)*, pp. 2282–2292

Hassan, M., Ghosh, P., Tesch, J., et al. (2021). Populating 3D scenes by learning human-scene interaction. In *Computer vision and pattern recognition (CVPR)*, pp. 14,708–14,718

Hasson, Y., Tekin, B., Bogo, F., et al. (2020). Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Computer vision and pattern recognition (CVPR)*, pp. 568–577

Hasson, Y., Varol, G., Tzionas, D., et al. (2019). Learning joint reconstruction of hands and manipulated objects. In *Computer vision and pattern recognition (CVPR)*, pp. 11,807–11,816

He, Y., Yan, R., Fragkiadaki, K., et al. (2020). Epipolar transformers. In: *Computer vision and pattern recognition (CVPR)*, pp. 7776–7785

Hu, Y. T., Chen, H. S., Hui, K., et al. (2019). SAIL-VOS: Semantic amodal instance level video object segmentation: A synthetic dataset and baselines. In *Computer vision and pattern recognition (CVPR)*, pp. 3105–3115

Huang, Y., Bogo, F., Lassner, C., et al. (2017). Towards accurate marker-less human shape and pose estimation over time. In *International conference on 3D vision (3DV)*, pp. 421–430

Huang, Y., Taheri, O., Black, M. J., et al. (2022b). InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German conference on pattern recognition (GCPR)*, pp. 281–299

Huang, C. H. P., Yi, H., Höschle, M., et al. (2022a). Capturing and inferring dense full-body human-scene contact. In *Computer vision and pattern recognition (CVPR)*, pp. 13,274–13,285

Ionescu, C., Papava, D., Olaru, V., et al. (2014). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 36*(7), 1325–1339.

Iskakov, K., Burkov, E., Lempitsky, V., et al. (2019). Learnable triangulation of human pose. In *International conference on computer vision (ICCV)*, pp. 7717–7726

Joo, H., Simon, T., & Sheikh, Y. (2018). Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer vision and pattern recognition (CVPR)*, pp. 8320–8329

Kanazawa, A., Black, M. J., Jacobs, D. W., et al. (2018). End-to-end recovery of human shape and pose. In *Computer vision and pattern recognition (CVPR)*, pp. 7122–7131

Kanazawa, A., Zhang, J. Y., Felsen, P., et al. (2019). Learning 3d human dynamics from video. In *Computer vision and pattern recognition (CVPR)*, pp. 5614–5623

Karunratanakul, K., Yang, J., Zhang, Y., et al. (2020). Grasping field: Learning implicit representations for human grasps. In *International conference on 3D vision (3DV)*, pp. 333–344

Kato, H., Ushiku, Y., Harada, T. (2018). Neural 3D mesh renderer. In *Computer vision and pattern recognition (CVPR)*, pp. 3907–3916

Kirillov, A., Wu, Y., He, K., et al. (2020). PointRend: Image segmentation as rendering. In *Computer vision and pattern recognition (CVPR)*, pp. 9799–9808

Kocabas, M., Athanasiou, N., Black, M. J. (2020). VIBE: Video inference for human body pose and shape estimation. In *Computer vision and pattern recognition (CVPR)*, pp. 5252–5262

Kolotouros, N., Pavlakos, G., Black, M. J., et al. (2019). Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International conference on computer vision (ICCV)*, pp. 2252–2261

Lepetit, V. (2020). Recent advances in 3D object and hand pose estimation. arXiv:2006.05927

Li, X., Liu, S., Kim, K., et al. (2019). Putting humans in a scene: Learning affordance in 3D indoor environments. In *Computer vision and pattern recognition (CVPR)*, pp. 12,368–12,376

Liu, Y., Stoll, C., Gall, J., et al. (2011). Markerless motion capture of interacting characters using multi-view image segmentation. In *Computer vision and pattern recognition (CVPR)*, pp. 1249–1256

Loper, M. M., Black, M. J. (2014). OpenDR: An approximate differentiable renderer. In *European conference on computer vision (ECCV)*, pp. 154–169

Loper, M., Mahmood, N., & Black, M. J. (2014). MoSh: Motion and shape capture from sparse markers. *Transactions on Graphics (TOG), 33*(6), 1–13.

Loper, M., Mahmood, N., Romero, J., et al. (2015). SMPL: A skinned multi-person linear model. *Transactions on Graphics, 34*(6), 248:1-248:16.

Mahmood, N., Ghorbani, N. F., Troje N, et al. (2019). AMASS: Archive of motion capture as surface shapes. In: *International conference on computer vision (ICCV)*, pp. 5441–5450

Malleson, C., Gilbert, A., Trumble, M., et al. (2017). Real-time full-body motion capture from video and IMUs. In *International conference on 3D vision (3DV)*, pp. 449–457

Mehta, D., Sridhar, S., Sotnychenko, O., et al. (2017). VNect: Real-time 3D human pose estimation with a single RGB camera. *Transactions on Graphics, 36*(4), 44:1-44:14.

Microsoft (2022) Azure Kinect SDK (K4A). https://github.com/microsoft/Azure-Kinect-Sensor-SDK

Narasimhaswamy, S., Nguyen, T., & Hoai, M. (2020). Detecting hands and recognizing physical contact in the wild. In *Conference on neural information processing systems (NeurIPS)*, pp. 7841–7851

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision (ECCV)*, pp. 483–499

Nocedal, J., & Wright, S. J. (2006). Nonlinear equations. *Numerical Optimization* pp. 270–302

Oikonomidis, I., Kyriazis, N., & Argyros, A. A. (2011). Full DOF tracking of a hand interacting with an object by modeling occlusions

and physical constraints. In *International conference on computer vision (ICCV)*, pp. 2088–2095
Omran, M., Lassner, C., Pons-Moll, G., et al. (2018). Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: *International conference on 3D vision (3DV)*, pp. 484–494
Osman, A. A., Bolkart, T., & Black, M. J. (2020). STAR: Sparse trained articulated human body regressor. In *European conference on computer vision (ECCV)*, pp. 598–613
Osman, A. A. A., Bolkart, T., Tzionas, D., et al. (2022). SUPR: A sparse unified part-based human body model. In *European conference on computer vision (ECCV)*, pp. 568–585
Pavlakos, G., Choutas, V., & Ghorbani, N., et al. (2019). Expressive body capture: 3D hands, face, and body from a single image. In *Computer vision and pattern recognition (CVPR)*, pp. 10,975–10,985
Pons-Moll, G., Baak, A., Helten, T., et al. (2010). Multisensor-fusion for 3D full-body human motion capture. In *Computer vision and pattern recognition (CVPR)*, pp. 663–670
Qiu, H., Wang, C., & Wang, J., et al. (2019). Cross view fusion for 3D human pose estimation. In *International conference on computer vision (ICCV)*, pp. 4341–4350
Rempe, D., Birdal, T., Hertzmann, A., et al. (2021). Humor: 3d human motion model for robust pose estimation. In *Computer vision and pattern recognition (CVPR)*, pp. 11,488–11,499
Rhodin, H., Robertini, N., Casas, D., et al. (2016). General automatic human shape and motion capture using volumetric contour cues. In: *European conference on computer vision (ECCV)*, pp. 509–526
Rogez, G., III JSS, & Ramanan, D. (2015). Understanding everyday hands in action from RGB-D images. In *International conference on computer vision (ICCV)*, pp. 3889–3897
Romero, J., Kjellström, H., & Kragic, D. (2010). Hands in action: Real-time 3D reconstruction of hands in interaction with objects. In *International conference on robotics and automation (ICRA)*, pp. 458–463
Romero, J., Tzionas, D., & Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics, 36*(6), 245:1-245:17.
Sarafianos, N., Boteanu, B., Ionescu, B., et al. (2016). 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU), 152*, 1–20.
Savva, M., Chang, A. X., Hanrahan, P., et al. (2016). PiGraphs: Learning interaction snapshots from observations. *Transactions on Graphics, 35*(4), 139:1-139:12.
Shimada, S., Golyanik, V., Li, Z., et al. (2022). HULC: 3D human motion capture with pose manifold sampling and dense contact guidance. In *European conference on computer vision (ECCV)*, pp. 516–533
Sigal, L., Balan, A., & Black, M. J. (2010). HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV), 87*(1–2), 4–27.
Sun, J., Wang, Z., Zhang, S., et al. (2022). OnePose: One-shot object pose estimation without CAD models. In *CVPR*, pp. 6825–6834
Taheri, O., Choutas, V., Black, M. J., et al. (2022). GOAL: Generating 4D whole-body motion for hand-object grasping. In: *Computer vision and pattern recognition (CVPR)*, pp. 13,253–13,263
Taheri, O., Ghorbani, N., Black, M. J., et al. (2020). GRAB: A dataset of whole-body human grasping of objects. In *European Conference on computer vision (ECCV)*, pp. 581–600
Taheri, O., Zhou, Y., Tzionas, D., et al. (2024). GRIP: Generating interaction poses using spatial cues and latent consistency. In *International conference on 3D vision (3DV)*
Tian, Y., Zhang, H., Liu, Y., et al. (2022). Recovering 3d human mesh from monocular images: A survey. arXiv preprint arXiv:2203.01923
Tripathi, S., Chatterjee, A., Passy, J. C., et al. (2023a). DECO: Dense estimation of 3D human-scene contact in the wild. In *International conference on computer vision (ICCV)*, pp. 8001–8013
Tripathi, S., Müller, L., Huang, C. H. P., et al. (2023b). 3D human pose estimation via intuitive physics. In *Computer vision and pattern recognition (CVPR)*, pp. 4713–4725
Tu, H., Wang, C., & Zeng, W. (2020). VoxelPose: Towards multi-camera 3D human pose estimation in wild environment. In *European conference on computer vision (ECCV)*, pp. 197–212
Tzionas, D., Ballan, L., Srikantha, A., et al. (2016). Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV), 118*(2), 172–193.
Tzionas, D., & Gall, J. (2013). A comparison of directional distances for hand pose estimation. In: *German conference on pattern recognition (GCPR)*, pp. 131–141
Varol, G., Laptev, I., & Schmid, C. (2017). Long-term temporal convolutions for action recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 40*(6), 1510–1517.
von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., et al. (2018). Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European conference on computer vision (ECCV)*, pp. 614–631
Wang, J., Tan, S., Zhen, X., et al. (2021). Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding (CVIU), 210*(103), 225.
Wei, S. E., Ramakrishna, V., Kanade, T., et al. (2016). Convolutional pose machines. In *Computer vision and pattern recognition (CVPR)*, pp. 4724–4732
Wu, Y., Wang, J., Zhang, Y., et al. (2022). SAGA: Stochastic whole-body grasping with contact. In *European conference on computer vision (ECCV)*, pp. 257–274
Xu, H., Bazavan, E. G., Zanfir, A., et al. (2020). GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer vision and pattern recognition (CVPR)*, pp. 6183–6192
Xu, W., Chatterjee, A., Zollhöfer, M., et al. (2018). MonoPerfCap: Human performance capture from monocular video. *Transactions on Graphics (TOG), 37*(2), 1–15.
Yao, B., Fei-Fei, L. (2010). Modeling mutual context of object and human pose in human-object interaction activities. In *Computer vision and pattern recognition (CVPR)*, pp. 17–24
Yi, H., Huang, C. H. P., Tzionas, D., et al. (2022). Human-aware object placement for visual environment reconstruction. In *Computer vision and pattern recognition (CVPR)*, pp. 3959–3970
Zhang, Y., An, L., Yu, T., et al. (2020b). 4D association graph for real-time multi-person motion capture using multiple video cameras. In *Computer vision and pattern recognition (CVPR)*, pp. 1321–1330
Zhang, Y., Hassan, M., Neumann, H., et al. (2020c). Generating 3D people in scenes without people. In *Computer vision and pattern recognition (CVPR)*, pp. 6193–6203
Zhang, Y., Li, Z., An, L., et al. (2021b). Light-weight multi-person total capture using sparse multi-view cameras. In *International conference on computer vision (ICCV)*, pp. 5560–5569
Zhang, J. Y., Pepose, S., Joo, H., et al. (2020a). Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European conference on computer vision (ECCV)*, pp. 34–51
Zhang, S., Zhang, Y., Bogo, F., et al. (2021a). Learning motion priors for 4D human body capture in 3D scenes. In *Computer vision and pattern recognition (CVPR)*, pp. 11,323–11,333
Zollhöfer, M., Stotko, P., Görlitz, A., et al. (2018). State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum (CGF), 37*(2), 625–652.