BEHAVE: Dataset and Method for Tracking Human Object Interactions

Bharat Lal Bhatnagar[1,2], Xianghui Xie[2], Ilya A. Petrov[1], Cristian Sminchisescu[3], Christian Theobalt[2], and Gerard Pons-Moll[1,2]

[1]University of Tübingen, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3]Google Research
{i.petrov, gerard.pons-moll}@uni-tuebingen.de, {bbhatnag, xxie, theobalt}@mpi-inf.mpg.de,
sminchisescu@google.com

Abstract

Modelling interactions between humans and objects in natural environments is central to many applications including gaming, virtual and mixed reality, as well as human behavior analysis and human-robot collaboration. This challenging operation scenario requires generalization to vast number of objects, scenes, and human actions. Unfortunately, there exist no such dataset. Moreover, this data needs to be acquired in diverse natural environments, which rules out 4D scanners and marker based capture systems. We present BEHAVE dataset, the first full body human-object interaction dataset with multi-view RGBD frames and corresponding 3D SMPL and object fits along with the annotated contacts between them. We record ∼15k frames at 5 locations with 8 subjects performing a wide range of interactions with 20 common objects. We use this data to learn a model that can jointly track humans and objects in natural environments with an easy-to-use portable multi-camera setup. Our key insight is to predict correspondences from the human and the object to a statistical body model to obtain human-object contacts during interactions. Our approach can record and track not just the humans and objects but also their interactions, modeled as surface contacts, in 3D. Our code and data can be found at: http://virtualhumans.mpi-inf.mpg.de/behave.

1. Introduction

The last decade has seen rapid progress in modelling the appearance of humans ranging from body pose, shape [52, 58, 60, 61, 81], faces [74] and even detailed clothing [5, 7, 11, 57, 65]. With various practical use cases like virtual try-on, personalised avatar creation, and several applications

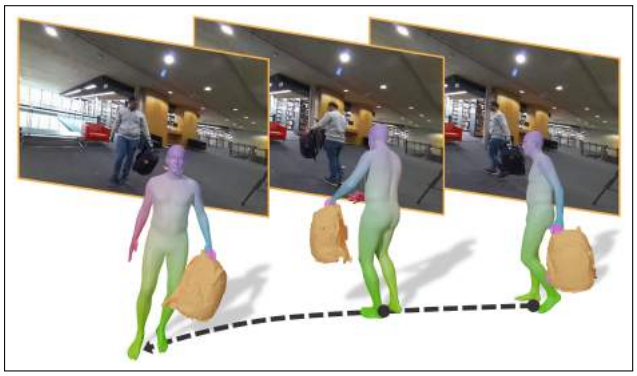

Figure 1. Given a multi-view RGBD sequence, our method tracks the human, the object and their contacts in 3D.

in augmented and mixed reality, or human-robot collaboration, the focus on humans is justified. Beyond modelling appearance, few methods have focused on capturing and synthesizing *human interactions* (human-object/scene interaction). There exists work to capture humans in a static 3D scene [33], even without using external cameras [30], and work to synthesize static poses [34, 49], or full body movement [32, 32, 50, 68] in a 3D scene.

These methods show growing interest in modelling human behavior, highlighting a need to capture real human interactions. Existing methods [32, 68] however are learned from high quality curated data captured using optical marker based motion capture systems or wearable sensors. Unfortunately, such commercial systems are expensive, drastically limit the interactions that can be captured, and often fail when tracking humans and objects under occlusion. In addition, the recording volume is spatially confined and difficult to re-locate, thus limiting the activities, scenes, and objects that can be captured. Wearable sensors [30] are not restricted in volume, but close range interaction can not be accurately captured. Altogether, the

lack of diverse 3D interaction data, and the lack of accurate and flexible capture methods both constitute barriers in modelling human behavior.

With the goal of simplifying the data capture process and hence allowing faster progress in the field, we propose BEHAVE, a method to capture diverse 3D human interactions in natural environments, using a setup comprising of portable, cheap, and easy to use RGBD cameras. Tracking human interactions from sparse consumer grade cameras is however extremely challenging. Depth data is inherently noisy and incomplete. Moreover, the person and object occlude each other frequently during interactions. Furthermore, capturing interactions requires estimating human-object contacts accurately, which is difficult because contacts represent small regions in the image, close to the observable (resolution) limit. This requires innovation that goes significantly beyond the current state of the art trackers. We propose to track the human using a parametric human model (such as SMPL [52]) and track objects using template meshes. Naively fitting the human model and an object 3D template to the point-cloud completely fails due to the aforementioned challenges. Our key idea is to train a neural model which jointly completes the human and object shape, represented with implicit surfaces, while predicting a correspondence field to the human, as well as an object orientation field. These rich outputs allow us to formulate a powerful human-object fitting objective which is robust to missing data, noise and occlusion.

To train and evaluate BEHAVE, we capture the *largest* dataset of human-object interactions in natural environments. The BEHAVE dataset contains 20 3D objects, 8 subjects (5 male, 3 female), 5 different locations and totals around 15.2k frames of recording. We provide ground truth SMPL and 3D object meshes as well as contacts.
Our contributions can be summarized as follows:

- We propose the first approach that can accurately 3D track humans, objects and contacts in natural environments using multi-view RGBD images.
- We collect the *largest* dataset of multi-view RGBD sequences and corresponding human models, object and contact annotations. See Sec. 3 for details regarding its usefulness to the community.
- Since there exists no publicly available code and datasets to accurately track human-object interactions in natural environments, we will release our code and data for further research in this direction.

2. Related Work

In this section, we first briefly review work focused on object and human reconstruction, in isolation from their environmental context. Such methods focus on modelling appearance and do not consider interactions. Next, we cover methods focused on humans in static scenes and finally discuss closer-related work to ours, for modelling dynamic human-object interactions.

2.1. Appearance modelling: Humans and objects without scene context

Human reconstruction and performance capture Perceiving humans from monocular RGB data [12, 29, 31, 41, 43, 44, 58, 59, 64, 87] and under multiple views [37–40, 62] settings has been widely explored. Recent work tends to focus on reconstructing fine details like hand gestures and facial expressions [20, 25, 85, 91], self-contacts [27, 54], interactions between humans [26], and even clothing [6, 11]. These methods benefit from representing human with parametric body models [52, 58, 81], thus motivating our use of recent implicit diffused representations [8, 10] as backbone for our tracker.
Following the success of pixel-aligned implicit function learning [64, 65], recent methods can capture human performance from sparse [38, 80] or even a single RGB camera [47,48]. However, capturing 3D humans from RGB data involves a fundamental ambiguity between depth and scale. Therefore, recent methods use RGBD [56, 69, 73, 76, 84] or volumetric data [9,10,19] for reliable human capture. These insights motivate us to build novel trackers based on multi-view RGBD data.

Object reconstruction Most existing work on reconstructing 3D objects from RGB [21, 46, 53, 75, 78] and RGBD [45, 55, 82] data does so in isolation, without the human involvement or the interaction. While challenging, it is arguably more interesting to reconstruct objects in a dynamic setting under severe occlusions from the human.

2.2. Interaction modelling: Humans and objects with scene context

Humans in static scenes Modelling how humans act in a scene is both important and challenging. Tasks like placement of humans into static scenes [34, 49, 90], motion prediction [15,32] or human pose reconstruction [16,33,77,86, 89] under scene constrains, or learning priors for human-object interactions [66], have been investigated extensively in recent years. These methods are relevant but restricted to modelling humans interacting with *static* objects. We address a more challenging problem of jointly tracking human-object interactions in *dynamic* environments where objects are manipulated.

Dynamic human object interactions Recently, there has been a strong push on modeling hand-object interactions based on 3D [42,72], 2.5D [13,14] and 2D [22,24,28,35,83]

Dataset	RGBD	Hum.	Ob.Cont.	Qual.	Scal.
NTU [51]	✓	Jts.	X	NA	***
PiGr [66]	✓	Jts.	X	NA	**
GRAB [72]	X	✓	✓	***	*
PROX [33]	✓	✓	Stat.	*	**
Ours	✓	✓	✓	**	***

Table 1. We compare the proposed BEHAVE dataset with existing ones containing human-object interactions. Our criteria are based on availability of RGB input, 3D human, 3D contact with the object, quality (more stars, better), and scalability to capture at diverse locations (more stars, better). NTU-RGBD [51] and PiGraphs [66] do not provide full 3D human and object contacts and are hence unsuitable for modelling dynamic 3D interactions. GRAB [72] uses a marker based capture system and hence contains the highest quality data but this also makes it difficult to scale. PROX [33] is easier to scale as it uses a single Kinect based capture setup (although, scene needs to be pre-scanned) but this reduces the overall quality. More importantly it does not contain dynamic interactions. Ours is the first dataset that captures dynamic human-object interactions in diverse environments.

data. Although powerful, these methods are currently restricted to modelling only *hand-object* interactions. In contrast, we are interested in *full body* capture. Methods for dynamic full body human object interaction approach the problem via 2D action recognition [36, 51] or reconstruct 3D object trajectories during interactions [23]. Despite being impressive, such methods either lack full 3D reasoning [36, 51] or are limited to specific objects [23].

More recent work reconstructs and tracks human-object interactions from RGB [71] or RGBD streams [70], but does not consider contact prediction, thus missing a component necessary for accurate interaction estimates.

Very relevant to our work, PHOSA [88] reconstructs humans and objects from a single image. PHOSA uses hand crafted heuristics, instance specific optimization for fitting, and pre-defined contact regions, which limits generalization to diverse human-object interactions. Our method on the other hand learns to predict the necessary information from data, making our models more scale-able. As shown in the experiments, the accuracy of our method is significantly higher to PHOSA.

3. BEHAVE Dataset

We present BEHAVE dataset, the *largest* dataset of human-object interactions in natural environments, with 3D human, object and contact annotation, to date. See Tab. 1 for comparison with other datasets. Our dataset contains multi-view RGBD frames, with accurate pseudo-ground truth SMPL [52], object fits, human and object segmentation masks, and contact annotations.

Recording multi-view RGBD data We setup and calibrate 4 Kinects at 4 corners of our square recording volume where all interactions are performed by 8 subjects (5 male, 3 female). Interactions are captured at 5 disparate indoor locations with 20 commonly used, yet diverse objects: 5 different boxes, 2 chairs, 2 tables, crate, backpack, trashcan, monitor, keyboard, suitcase, basketball, exercise ball, yoga mat, stool and a toolbox. We include common interactions such as lifting, carrying, sitting, pushing and pulling with hands and feet, as well as free interactions. See our supplementary video for sample sequences. In total, our dataset contains 10.7k frames for training and 4.5k frames for testing respectively.

Human segmentation and SMPL fitting We segment the human in our images by running DetectronV2 [79] followed by manual correction with [67] on the segmentation masks. These masks are then used to segment multi-view depth maps and lift human point clouds from 2D to 3D. We use FrankMocap [63] to initialize SMPL's pose from the images and then use instance specific optimization [6] to fit the SMPL model to the segmented human point cloud. For more accurate fitting, we additionally obtain the SMPL shape parameters of each subject from 3D scans using [9]. We report a chamfer error of 1.80cm between the segmented kinect point cloud and our SMPL fits.

Object segmentation and fitting To obtain object segmentation, we pre-scan objects using a 3D scanner [1, 3]. We then use multi-view object keypoints, marked manually by AMT [2] annotators in images, to optimize the 6D pose of the pre-scanned object mesh to the given frame. We obtain the chamfer error of 2.42cm between the segmented Kinect point cloud and object fit. The segmentation masks are then obtained by projecting fitted object meshes to the images.

Contact annotation Based on the pseudo-GT SMPL and object fits as described above, we automatically detect contacts if a point on the human surface (registered SMPL) is closer than 2cm to the object surface. For every object point, we store a binary contact label (whether there is a contact or not) and correspondence to the human (contact location on the surface).

See supplementary for more details on data acquisition.

How will this dataset be useful to the community? We devote significant effort in recording the *largest*, so far, dataset of natural, full body, day-to-day human interactions with common objects in different natural environments. We propose following challenges with BEHAVE dataset:

Figure 2. We present BEHAVE dataset, the *largest* dataset of human-object interactions in natural environments. BEHAVE contains multi-view RGBD sequences and corresponding 3D object and SMPL fits along with 3D contacts.

- **Tracking human-object interactions.** Track humans and objects using multi-view RGBD data. This can further be extended to track with just multi-view RGB, no-depth, and eventually just a single camera.
- **Reconstruction from a single image.** Joint 3D reconstruction of 3D humans and objects from a single RGB image. Currently, there is no dataset that can be used for benchmarking let alone to learn such a model.
- **Pose and shape estimation.** Benchmarking pose and shape estimation methods in challenging natural environments where the person is heavily occluded by the interacting object.

Apart from these tasks, the research community is free to explore other applications of the BEHAVE dataset.

4. Method: Tracking human, object and contacts

We present BEHAVE, a method to jointly track humans, objects and their interactions (represented as surface contacts) based on multi-view RGBD input. We formulate our method as an extended per-frame registration problem: we register the human (using SMPL [52]) and the object (using its pre-scanned object mesh), and predict contacts as correspondences between SMPL and object meshes. See Fig. 3 for the overview of our method.

Our formulation must obey three properties, (i) the SMPL model $M(\cdot)$ should fit the human in the multi-view input, (ii) the object mesh $\mathbf{W}^o$ should fit the input object and, (iii) SMPL model and object should satisfy contacts.

To facilitate joint reasoning of the human, object and contacts directly in 3D, we lift the human $\mathcal{S}^h$, and object $\mathcal{S}^o$ point clouds to 3D using multi-view depth and semantic segmentation. Our joint formulation fits SMPL $M(\cdot)$ and the object $\mathbf{W}^o$ to multi-view RGB-D data at each time step, using explicit contacts. This takes the following form

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{R}^o, \mathbf{t}^o) = E^{\text{SMPL}}(\mathcal{S}^h, M(\boldsymbol{\theta}, \boldsymbol{\beta})) + E^{\text{obj}}(\mathcal{S}^o, \mathbf{W}^o) + E^{\text{contact}}(\mathbb{1}^c \mathbf{W}^o, M(\boldsymbol{\theta}, \boldsymbol{\beta})). \tag{1}$$

The SMPL model is parameterized by pose $\boldsymbol{\theta}$, and shape $\boldsymbol{\beta}$. For notation brevity, we include the global SMPL translation into the pose parameters. We assume the template object $\mathbf{W}$ be rigid and only estimate the rotation $\mathbf{R}^o$, and translation $\mathbf{t}^o$, to fit the object mesh $\mathbf{W}^o = \mathbf{R}^o\mathbf{W} + \mathbf{t}^o$, to the object point cloud.

The indicator matrix, $\mathbb{1}^c$, selects the vertices on the object mesh $\mathbf{W}^o$, that are in contact with the SMPL model. This ensures that contact locations on the object and the human mesh adequately align in 3D.

The term $E^{\text{SMPL}}(\mathcal{S}^h, M(\boldsymbol{\theta}, \boldsymbol{\beta}))$ is designed to accurately fit SMPL to the human point cloud $\mathcal{S}^h$. The term $E^{\text{obj}}(\mathcal{S}^o, \mathbf{W}^o)$ is designed to fit the object mesh to the object point cloud and $E^{\text{contact}}(\mathbb{1}^c\mathbf{W}^o, M(\boldsymbol{\theta}, \boldsymbol{\beta}))$ ensures that contacts between the human and object match (align). We explain each term in detail next.

4.1. Fitting human model to the human point cloud

Fitting SMPL to the human point cloud $\mathcal{S}^h$ requires, (i) that distance between the SMPL model and the human point cloud should be minimized and (ii) the correct SMPL parts fit the corresponding body parts of the point cloud. The latter is important to avoid degenerate cases such as $180°$ flipped fitting, where the left hand is erroneously matched to the right side of the body or vice-versa [9]. With these considerations, we design our SMPL fitting objective as:

$$E^{\text{SMPL}} = \boldsymbol{d}(\mathcal{S}^h, M(\boldsymbol{\theta}, \boldsymbol{\beta})) + E^{\text{corr}} + E^{\text{reg}}, \tag{2}$$

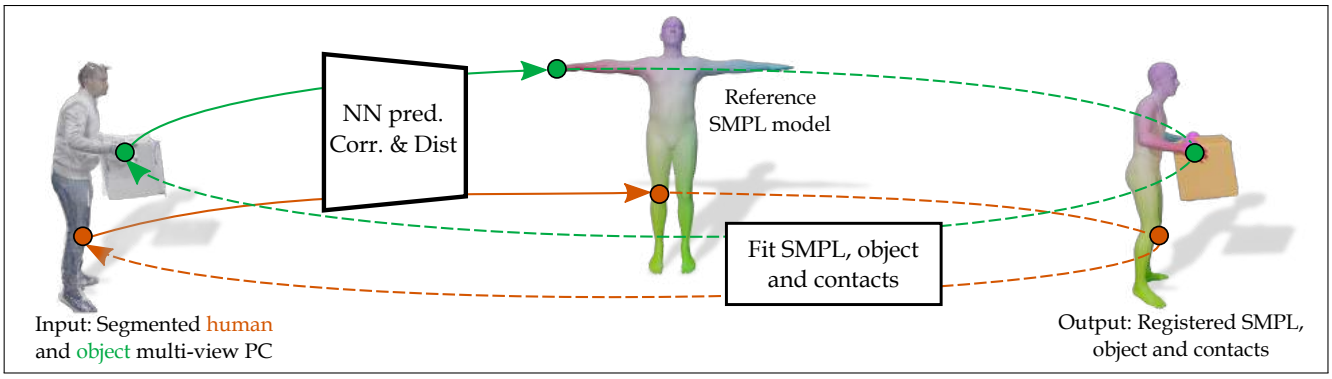

Figure 3. Given a sequence of multi-view images, we track the human and the object using SMPL and a template object mesh. We lift the segmented multi-view RGBD frames to 3D and obtain a human and object point cloud. As shown here, our network predicts correspondences between the human point cloud and the body model, which allows us to fit SMPL. We also predict correspondences from the object to the body model, thus allowing us to model contacts. Our network predictions (see Sec. 4) allow us to register SMPL and the object mesh to a video, making an accurate joint tracking of the human and object possible.

where $\boldsymbol{d}(\mathcal{S}^h, M(\boldsymbol{\theta}, \boldsymbol{\beta}))$ minimizes the point-to-mesh distance between the input human point cloud $\mathcal{S}^h$ and the SMPL model. To avoid sub-optimal local minima during fitting [9, 10], we train a neural network that predicts dense correspondences from the input to the SMPL model. This ensures that correct SMPL parts explain corresponding input regions, using the term E^{corr}.

Specifically, we train an encoder network similar to [17, 18] that takes the segmented and voxelized human $\mathcal{S}^h$ and object $\mathcal{S}^o$ point cloud as inputs, and generates a voxel aligned grid of features $\mathbf{F} = f_\phi^{\text{enc}}(\mathcal{S}^h, \mathcal{S}^o)$. We then sample N 3D query points, $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}, \mathbf{p}_i \in \mathbb{R}^3$ and for each point $\mathbf{p}_i = (x, y, z)$ obtain the corresponding point feature $\mathbf{F}_i = \mathbf{F}(x, y, z)$. We pass this point feature through a decoder network f_ϕ^{udf}, to predict the unsigned distance to object and human surfaces, $u_i^o, u_i^h = f_\phi^{\text{udf}}(\mathbf{F}_i), u_i^o, u_i^h \in \mathbb{R}$, respectively. We use a second decoder network f_ϕ^{corr}, to predict the correspondence of point $\mathbf{p}_i$ to the SMPL model, $\mathbf{c}_i = f_\phi^{\text{corr}}(\mathbf{F}_i), \mathbf{c}_i \in \mathbb{R}^3$.

E^{corr} enforces that the distance between the input point $\mathbf{p}_i$ and the corresponding point $\mathbf{c}_i$ after transforming it with the SMPL model is same as the distance predicted by the network u_i^h. Under a slight abuse of notation we use $M(\mathbf{c}_i, \boldsymbol{\theta}, \boldsymbol{\beta})$ to transform $\mathbf{c}_i$ with the SMPL function.

$$E^{\text{corr}} = \sum_{i=1}^{N} ||\mathbf{p}_i - M(\mathbf{c}_i, \boldsymbol{\theta}, \boldsymbol{\beta})|_2 - u_i^h|. \tag{3}$$

If the correspondences predicted by the network $\mathbf{c}_i$ deviate from the SMPL surface, these cannot be skinned using the SMPL model as its function is only defined on the body surface. To alleviate this issue, we use the LoopReg [10] formulation that allows us to pose and shape off-the-surface correspondences as well.

The final term $E^{\text{reg}} = E^{\text{J2D}} + E^{\boldsymbol{\theta}} + E^{\boldsymbol{\beta}}$, adds regularisation for SMPL joints, $E^{\text{J2D}} = \sum_{k=1}^{K} |\pi_k M^J(\boldsymbol{\theta}, \boldsymbol{\beta}) - \mathbf{J}_{2D}^k|_2$, where π_k is the camera projection matrix of camera k, $M^J(\cdot)$ are the 3D body joints and $\mathbf{J}_{2D}^k$ are the 2D joints detected in the k^{th} Kinect image. $E^{\boldsymbol{\theta}}$ and $E^{\boldsymbol{\beta}}$ are regularisation terms on SMPL pose and shape similar to [12].

4.2. Fitting the object mesh to the object point cloud

In order to fit the object mesh, we must ensure that distance from the input object point cloud to the object mesh is minimized. Minimizing this one-sided distance is necessary but not sufficient. Since severe occlusions are common in our interaction setting, large parts of object might be missing from the object point cloud, making fitting difficult. To alleviate this issue we must also ensure that all the vertices of the object mesh are correctly placed w.r.t. the input, even when the point cloud is incomplete. To do so, we take the object mesh vertices $\mathbf{v}_j^o \in \mathbf{W}^o, j \in \{1, \ldots, L\}$ and obtain the corresponding point feature $\mathbf{F}_j$, same as Sec. 4.1, where L is the number of object mesh vertices. We then obtain the unsigned distances to the object and human surfaces using the point feature $u_j^o, u_j^h = f_\phi^{\text{udf}}(\mathbf{F}_j)$. Since $\mathbf{v}_j^o$ is a vertex on the object mesh, its distance to the object surface u_j^o must be zero for a correct fit. This allows us to accurately fit the object vertices to the point data even when corresponding parts are missing from the object point cloud.

$$E^{\text{obj}} = \boldsymbol{d}(\mathcal{S}^o, \mathbf{W}^o) + \sum_{j=1}^{L} |u_j^o|, \tag{4}$$

where, $\boldsymbol{d}(\mathcal{S}^o, \mathbf{W}^o)$ minimizes the point-to-mesh distance between the object point cloud and the object mesh, and the term $\sum_{j=1}^{L} |u_j^o|$ uses implicit unsigned distance prediction to reason about missing object parts.

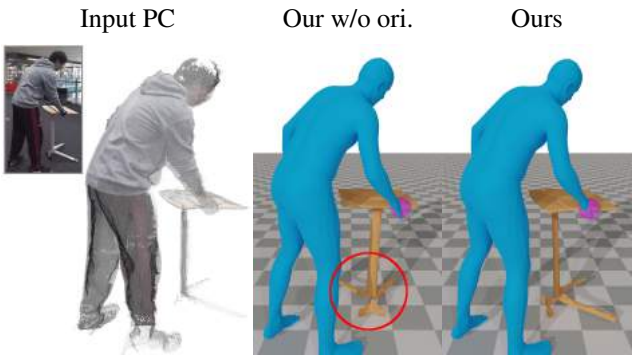

Figure 4. We show that our network predicted orientation is important for accurate object fitting. Without our orientation prediction the fitting gets stuck in a local minima.

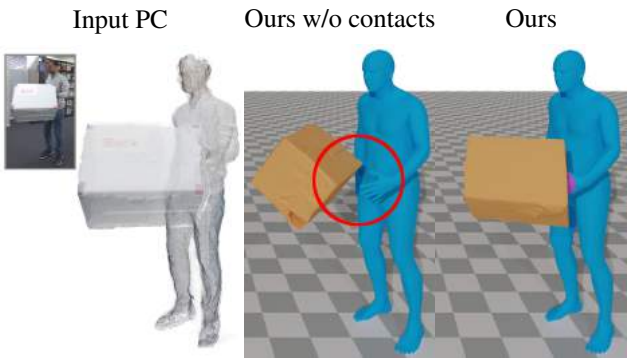

Figure 5. Without our network predicted contacts we observe artefacts like floating objects, leading to unrealistic tracking.

Predicting object orientation. Although the terms in Eq. (4) minimize the bi-directional distance between the object point cloud and the object mesh, they do not guarantee that parts of the object point cloud are explained by the semantically corresponding parts on the object mesh, e.g. in Fig. 4, the legs of the table are not aligned correctly. This issue can be fixed if we obtain the global object orientation during fitting. We represent the orientation of the object with the principal components obtained by running PCA on the object vertices.

We train a neural network f^a_ϕ, that uses the point feature $\mathbf{F}_j$ (same as Sec. 4.1) corresponding to each query point $\mathbf{p}_j$ and predicts the global orientation of the object $\mathbf{a}_j = f^a_\phi(\mathbf{F}_j), \mathbf{a}_j \in \mathbb{R}^9$. We find that orientation prediction is unreliable if the query point is far from the object surface, hence we filter out points whose unsigned distance from the object surface u^o_j, is greater than a threshold $\epsilon = 2cm$. The global orientation of the object is obtained by averaging the orientation predictions from the filtered points, $\mathbf{a}^o = \frac{1}{M}\sum_{j=1}^{M}\mathbf{a}_j$ where M is the number of filtered points. Next, we compute the relative rotation between the current object orientation $\bar{\mathbf{a}}$ and the predicted object orientation $\mathbf{a}^o$, and use this to initialise the object rotation $\mathbf{R}^o = \mathbf{a}^o(\bar{\mathbf{a}}^T\bar{\mathbf{a}})^{-1}\bar{\mathbf{a}}^T$. We further run SVD on $\mathbf{R}^o$ and only keep the rotation matrix.

Initialising $\mathbf{R}^o$ with the network predicted object orientation is crucial to avoid local minima during object fitting, as can be seen in Fig. 4 and Tab. 3.

4.3. Refining human & object models using contacts

Our formulation above gives reasonably good human and object fits but does not ensure that human and object meshes satisfy the contacts predicted by the network. This often leads to floating objects and hovering hands see Fig. 5, as human and object models are not in contact. In this section we explicitly optimize the human and object meshes to fit the contacts predicted by the network. We model contacts as vertices in the registered object mesh $\mathbf{v}^o_j \in \mathbf{W}^o$, that are very close to the input human $u^h_j < \epsilon$ and object $u^o_j < \epsilon$ surface. Similarly to Sec. 4.2, we use f^{udf}_ϕ to obtain the unsigned distances u^o_j, u^h_j and f^{corr}_ϕ to obtain the correspondences $\mathbf{c}_j$ of these points to the SMPL model, respectively. In order to filter query points close to human and object surfaces we compute a binary indicator matrix $\mathbb{1}^c \in \mathbb{R}^N$ such that $\mathbb{1}^c_j = 1$ iff $u^o_j < \epsilon, u^h_j < \epsilon$.

$$E^{\text{contact}} = \sum_{j=1}^{N} \mathbb{1}^c_j |\mathbf{v}^o_j - M(\mathbf{c}_j, \boldsymbol{\theta}, \boldsymbol{\beta})|_2. \tag{5}$$

E^{contact} allows us to jointly optimise the SMPL model and the object parameters $\mathbf{R}^o, \mathbf{t}^o$ to satisfy the contacts predicted by the network.

4.4. Network training

In this section we elaborate on training our networks.

Feature encoding. We use a 3D CNN similar to IF-Net [17] to obtain a voxel aligned multi-scale grid of features $\mathbf{F} = f^{\text{enc}}_\phi(\mathcal{S}^h, \mathcal{S}^o)$.

Unsigned distance prediction. To train the network f^{udf}_ϕ, we sample N query points $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ in 3D. For each query point $\mathbf{p}_j$ we obtain its point feature $\mathbf{F}_j$ (Sec. 4.1) and use this to predict the unsigned distance [18] to human and object surface $u^o_j, u^h_j = f^{\text{udf}}_\phi(\mathbf{F}_j)$.

We jointly train $f^{\text{enc}}_\phi, f^{\text{udf}}_\phi$ with standard L2 loss. The GT for u^o_j, u^h_j is easily available as our dataset contains GT SMPL and object fits allowing us to obtain GT distance of point $\mathbf{p}_j$ from the SMPL and object mesh.

SMPL correspondence prediction. To train f^{corr}_ϕ, we use the point feature $\mathbf{F}_j$ of sampled query point $\mathbf{p}_j$ to predict its correspondence to the SMPL model $\mathbf{c}_j = f^{\text{corr}}_\phi(\mathbf{F}_j)$.

We jointly train $f^{\text{enc}}_\phi, f^{\text{corr}}_\phi$ using a standard L2 loss. Since we have the GT SMPL fit in our dataset we simply find the closest SMPL surface point for the query point $\mathbf{p}_j$ and use this as the GT correspondence.

Method	SMPL v2v (cm)	obj. v2v (cm)
IP-Net [9]	6.61	NA
LoopReg [10]	9.12	NA
Fit to input	16.15	26.09
PHOSA [88]	13.73	34.73
Ours	**4.99**	**21.20**

Table 2. We compare our method to obtain SMPL and object fits with IP-Net, LoopReg and PHOSA. We also show that directly fitting SMPL and object meshes to the input leads to sub-optimal performance. Our method not only obtains better fits, but unlike LoopReg and IP-Net, we can also fit the object.

Object orientation prediction. To train the network f^a_ϕ, we use the point feature $\mathbf{F}_j$ of a sampled query point $\mathbf{p}_j$ to predict the global object orientation $\mathbf{a}_j = f^a_\phi(\mathbf{F}_j)$.
We jointly train $f^{\text{enc}}_\phi, f^a_\phi$ with standard L2 loss. We find that points far away from the object surface are unreliable in predicting the object orientation. Hence we only apply this loss to points that are close to the object, i.e. the GT $u^o_j < \epsilon$. Since we have the GT object fit, we obtain the GT orientation by running PCA on the object mesh vertices and use the 3 principal axes in $\mathbb{R}^9$.

5. Experiments

In this section we compare our approach with existing methods. Our experiments show that we clearly outperform existing baselines. Next, we ablate our design choices and highlight the importance of contact and object orientation prediction in capturing human-object interactions.

5.1. Comparing with PHOSA

We find PHOSA [88], a method to reconstruct humans and objects from a single image, quite relevant to our work. Although PHOSA uses only a single image whereas we use multi-view images, thus giving our method an advantage, it is still the closest competing method. We run Procrustes alignment on PHOSA results to remove depth ambiguity.
It should be noted that PHOSA depends on pre-defined fixed contact regions whereas our approach can freely predict full-body contacts and PHOSA uses hand crafted heuristics to model contacts whereas our approach learns contact modelling from data, making our method more scalable. We compare our method with PHOSA in Fig. 6 and Tab. 2, and clearly outperform it.

5.2. Why not fit human and object models directly to point clouds?

Since there are no existing methods that can jointly track humans, objects and the contacts from a multi-view input, we create an obvious baseline where we fit the SMPL and object meshes directly to the input point cloud. We show (Tab. 2) that direct fitting easily gets stuck in local minima.

Method	SMPL v2v (cm)	obj. v2v (cm)
A) Ours w/o ori.	4.98	24.02
B) Ours w/o cont.	**4.96**	21.28
C) Ours	4.99	**21.20**

Table 3. We analyse the importance of (A) object orientation prediction and (B) contact prediction for our method. It can be seen that object orientation prediction noticeably improves object localisation error. The effect of contact loss is not significant quantitatively but makes noticeable difference qualitatively see Fig. 5.

This is because the point clouds are very noisy and large parts are missing due to heavy occlusion between the person and the object during interactions. Our network, on the other hand, can implicitly reason about missing parts, thus generating more accurate results.

5.3. Why can't existing human registration approaches be extended to our setting?

There are no direct baselines that can jointly track humans, objects, and contacts from multi-view input. There are works [9, 10] that pursue similar ideas of predicting correspondences and fitting SMPL to the human point cloud. In this subsection we explore their suitability in our setting.

Comparison to IPNet [9] IPNet takes as input a human point cloud and predicts an implicit reconstruction of the human and sparse correspondences to the SMPL model, which enables its fitting to the implicit reconstruction.
This approach has three major disadvantages. First, querying occupancies for a 128^3 grid to obtain implicit reconstruction is expensive. Second, it predicts occupancies which requires water-tight surfaces. And third, running traditional Marching Cubes makes occupancy prediction non-differentiable w.r.t. SMPL fitting.
Our formulation in Eqs. (2) and (3) alleviates these problem as we can fit SMPL by only querying $N = 30k$ points instead of $128^3 (\sim 2M)$ points. Since we use unsigned distance prediction, our method can work with non-water tight surfaces. We can also fit SMPL directly to unsigned distance predictions, thus removing the requirement for Marching Cubes. We compare our approach with IP-Net [9] (trained on our dataset) in Tab. 2 and show that we obtain better performance than IP-Net at much lower cost ($30k$ (ours) vs. $\sim 2M$ (IP-Net) query points and no Marching Cubes). This shows that our formulation is superior than IPNet even for human registration. We can additionally handle objects and interactions. Qualitative comparisons are given in the supplementary material.

Comparison with LoopReg [10] LoopReg fits SMPL to the input point cloud by explicitly predicting correspondences. We find the idea interesting and use their diffused

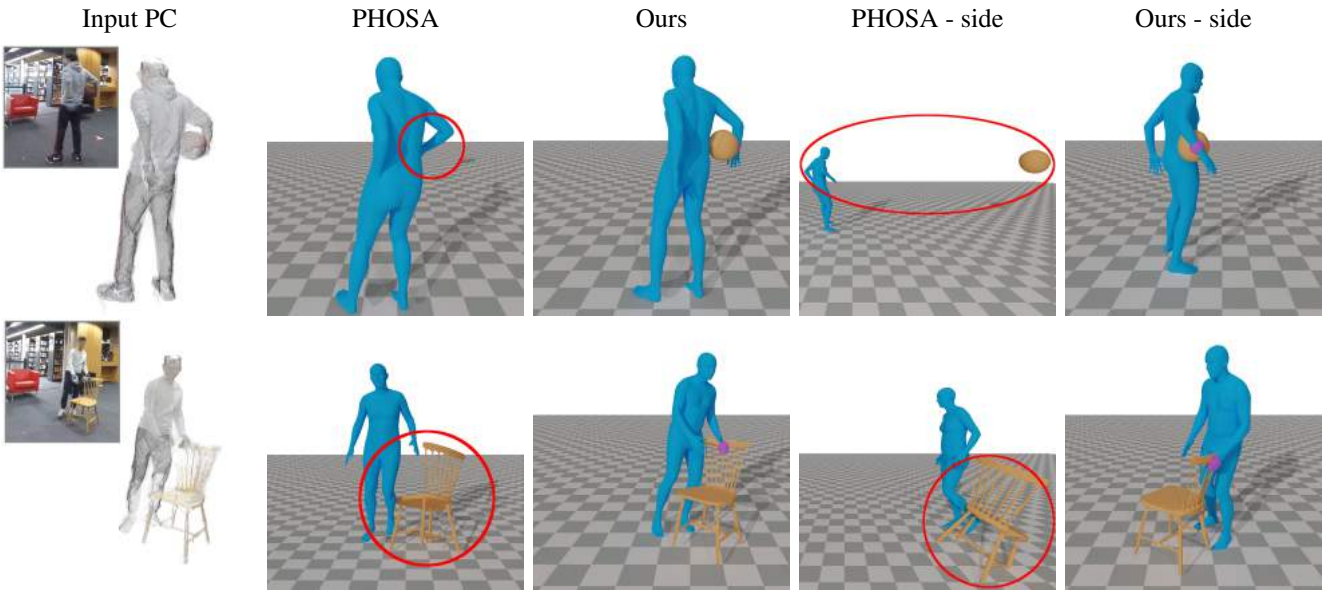

Figure 6. We compare our method to track human, object and contacts with PHOSA [88]. It can clearly be seen that our method can reason about the human-objects contacts and produces more accurate results.

SMPL formulation in our method. LoopReg is, however, not directly applicable in our setting as it assumes a noise free and complete human point cloud. When the point cloud is incomplete due to occlusions, no correspondences are predicted for missing parts. Since LoopReg can only use surface points for fitting, this makes registration inaccurate. BEHAVE handles this case by using distances to the SMPL surface(Eq. (3)) predicted for each of the sampled query points to fit the body model, thus allowing the use of non-surface points for fitting. This is important as the Kinect point cloud is noisy. We outperform LoopReg [10] (trained on our dataset) and show (Tab. 2) that our formulation is robust to missing parts and noisy input.

5.4. Importance of contacts

In this experiment we show that our network predicted contacts are key for physically plausible tracking. Even though quantitative difference is not significant (Tab. 3), it can be seen in Fig. 5 that without contact information, the human and the object do not lock into the correct location. Hence, we notice unnatural results like floating objects. Using our contact prediction alleviates such issues.

We encourage the readers to see our supplementary document for detailed discussion regarding limitations and future work with BEHAVE.

6. Conclusions

We have presented BEHAVE, the first methodology to jointly track humans, objects and explicit contacts in natural environments. By introducing neural networks to predict correspondences to a 3D human body model along with unsigned distance fields defined over human and object surfaces, we are able to accurately model human-object contacts. We further integrate such neural predictions into a proposed joint registration method resulting in the robust 3D tracking of human-object interactions.

Along with our proposed method we also provide BEHAVE, the *largest* dataset of RGBD sequences and annotated humans, objects, and contacts to date. BEHAVE dataset is the *first* benchmark for the part of the research community interested in modelling human-object interactions. We propose real-world challenges like reconstructing humans and object from a single RGB image, tracking human-object interactions from multiple and single-view RGB(D) input, pose estimation etc. Our dataset together with our code is released in order to stimulate future research in this important emerging domain.

Acknowledgements Special thanks to RVH team members [4], and reviewers, their feedback helped improve the manuscript. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and ERC Consolidator Grant 4DRepLy (770784). Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] Agisoft metashape, https://www.agisoft.com/. 3

[2] https://www.mturk.com. 3

[3] https://www.treedys.com/. 3

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

[5] Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conf. on Pattern Recognition*, pages 347–360, 2017. 1

[6] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[7] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conf. on 3D Vision*, sep 2018. 1

[8] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3D human shape and articulated pose. In *IEEE International Conf. on Computer Vision*, 2021. 2

[9] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2, 3, 4, 5, 7

[10] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020. 2, 5, 7, 8

[11] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2

[12] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conf. on Computer Vision*. Springer International Publishing, 2016. 2, 5

[13] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, 04 2019. 2

[14] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, August 2020. 2

[15] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. *ArXiv*, abs/2007.03672, 2020. 2

[16] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[17] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2020. 5, 6

[18] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Neural Information Processing Systems (NeurIPS)*. , December 2020. 5, 6

[19] Julian Chibane and Gerard Pons-Moll. Implicit feature networks for texture completion from partial 3d data. In *European Conference on Computer Vision*, pages 717–725. Springer, 2020. 2

[20] 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*, pages 20–40. Springer, 2020. 2

[21] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 2

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

[23] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. *arXiv preprint arXiv:2108.08844*, 2021. 3

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

[25] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, Dec. 2021. 2

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

[27] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Thirty-Fifth AAAI Conf. on Artificial Intelligence (AAAI'21)*, 2021. 2

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

[29] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[30] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 1

[31] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2

[32] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021. 1, 2

[33] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision*, 2019. 1, 2, 3

[34] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, June 2021. 1, 2

[35] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[36] JF Hu, WS Zheng, J Lai, and J Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2186–2200, 2017. 3

[37] Yinghao Huang, Federica Bogo, Christoph Classner, Angjoo Kanazawa, Peter V Gehler, Ijaz Akhter, and Michael J Black. Towards accurate markerless human shape and pose estimation over time. In *International Conf. on 3D Vision*, 2017. 2

[38] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–354, 2018. 2

[39] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 2

[40] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[41] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018. 2

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

[43] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262. IEEE, June 2020. 2

[44] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[45] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3559–3568, 2018. 2

[46] Jiahui Lei, Srinath Sridhar, Paul Guerrero, Minhyuk Sung, Niloy Mitra, and Leonidas J. Guibas. Pix2surf: Learning parametric 3d surface models of objects from images. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[47] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, SIGGRAPH 2020, New York, NY, USA, 2020. Association for Computing Machinery. 2

[48] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. *arXiv preprint arXiv:2007.13988*, 2020. 2

[49] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[50] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4):40, 2020. 1

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

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. ACM, 2015. 1, 2, 3, 4

[53] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[54] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 2

[55] Norman Muller, Yu-Shiang Wong, Niloy J Mitra, Angela Dai, and Matthias Nießner. Seeing behind objects for 3d multi-object tracking in rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6071–6080, 2021. 2

[56] Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, et al. Volumetric capture of humans with a single rgbd camera via semi-parametric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9709–9718, 2019. 2

[57] Chaitanya Patel, Zhouyincheng Liao, and Gerard Pons-Moll. The virtual tailor: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2020. 1

[58] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[59] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2

[60] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. 1

[61] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 1

[62] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018. 2

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

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

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

[66] 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), 2016. 2, 3

[67] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 3

[68] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209:1–209:14, 2019. 1

[69] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 246–264. Springer, 2020. 2

[70] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *arXiv preprint arXiv:2104.14837*, 2021. 3

[71] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complexhuman-object interactions. *arXiv preprint arXiv:2108.00362*, 2021. 3

[72] 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

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

[74] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *IEEE International Conf. on Computer Vision*, 2017. 1

[75] Dimitrios Tzionas and Juergen Gall. 3d object reconstruction from hand-object interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 729–737, 2015. 2

[76] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *European Conference on Computer Vision*, pages 430–446. Springer, 2020. 2

[77] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. *arXiv preprint arXiv:2012.01591*, 2020. 2

[78] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *Advances In Neural Information Processing Systems*, 2017. 2

[79] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[80] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Neural Information Processing Systems*, 2021. 2

[81] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *IEEE International Conf. on Computer Vision and Pattern Recognition*, 2020. 1, 2

[82] Bo Yang, Hongkai Wen, Sen Wang, Ronald Clark, Andrew Markham, and Niki Trigoni. 3d object reconstruction from a single depth view with adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 679–688, 2017. 2

[83] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 2

[84] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 2

[85] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. *IEEE International Conf. on Computer Vision and Pattern Recognition*, 2020. 2

[86] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2

[87] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12971–12980, October 2021. 2

[88] 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. 3, 7, 8

[89] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11343–11353, 2021. 2

[90] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, Nov. 2020. 2

[91] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4822, 2021. 2