CHORE: Contact, Human and Object REconstruction from a single RGB image

Xianghui Xie[2], Bharat Lal Bhatnagar[1,2], and Gerard Pons-Moll[1,2]

[1] University of Tübingen, Germany
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{xxie, bbhatnag}@mpi-inf.mpg.de, gerard.pons-moll@uni-tuebingen.de

Abstract. Most prior works in perceiving 3D humans from images reason human in isolation without their surroundings. However, humans are constantly interacting with the surrounding objects, thus calling for models that can reason about not only the human but also the object and their interaction. The problem is extremely challenging due to heavy occlusions between humans and objects, diverse interaction types and depth ambiguity. In this paper, we introduce CHORE, a novel method that learns to jointly reconstruct the human and the object from a single RGB image. CHORE takes inspiration from recent advances in implicit surface learning and classical model-based fitting. We compute a neural reconstruction of human and object represented implicitly with two unsigned distance fields, a correspondence field to a parametric body and an object pose field. This allows us to robustly fit a parametric body model and a 3D object template, while reasoning about interactions. Furthermore, prior pixel-aligned implicit learning methods use synthetic data and make assumptions that are not met in the real data. We propose a elegant depth-aware scaling that allows more efficient shape learning on real data. Experiments show that our joint reconstruction learned with the proposed strategy significantly outperforms the SOTA. Our code and models are available at https://virtualhumans.mpi-inf.mpg.de/chore

1 Introduction

In order to deploy robots and intelligent systems in the real world, they must be able to perceive and understand humans interacting with the real world from visual input. While there exists a vast literature in perceiving humans in 3D from single images, the large majority of works perceive humans in isolation [6, 37, 40, 54, 63, 75]. The joint 3D perception of humans and objects has received much less attention and is the main focus of this work. Joint reconstruction of humans and objects is extremely challenging. The object and human occlude each other making inference hard; it is difficult to predict their relative size and spatial arrangement in the 3D world, and the visual evidence for contacts and interactions consists of only a very small region in the image.

Recent work [79] addresses these challenges by reconstructing 3D objects and humans separately, and imposing hand-crafted rules like manually defined object

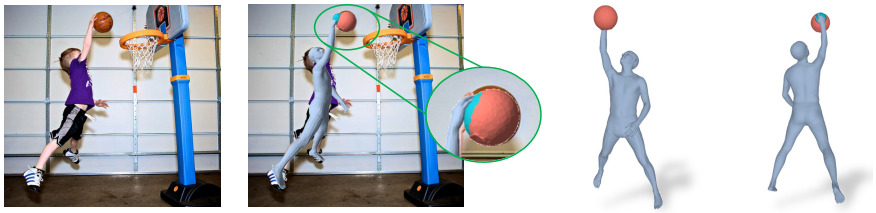

Fig. 1: From a single image, our method jointly reconstructs the 3D human coherently with the object by reasoning about their interactions. Instead of predicting human and object independently and imposing coherency post-hoc [79], we reconstruct both jointly including contacts, leading to significant improvements over the state-of-the-art.

contact regions. However, heuristics cannot scale as it is difficult to specify all possible human-object contacts beforehand. Consider the chair in Fig. 3: one can sit, lean, move it from the handle, back or legs, which constitutes different pairs of human-object contacts that are not easy to manually enumerate. Furthermore, human and object provide strong cues about their relative spatial arrangement for each other during interaction. Reasoning human and object separately does not exploit these cues and can produce errors that cannot be corrected afterwards, see Fig. 4. This motivates a method that can jointly reconstruct humans and objects and learn the interactions from data, without hand-crafted rules.

In this work, we introduce CHORE, a novel method to jointly recover a 3D parametric body model and the 3D object, from a single image. CHORE is built on two high-level ideas. First, instead of reconstructing human and object separately, we argue for jointly reasoning about them in the first place. Secondly, we learn spatial configuration priors of interacting human and object from data without using heuristics. Instead of regressing human parameters directly from image, we first obtain pixel-aligned neural fields predictions (CHORE fields), which consist of two unsigned distance fields to human and object surfaces, a correspondence field to the SMPL [49] body model and an object pose field. This allows us to formulate a more robust optimization energy to fit the SMPL model and the object template to the image. To train our network on real data, we propose a training strategy that enables us to effectively learn pixel-aligned implicit functions with perspective cameras, which leads to more accurate prediction and joint fitting. In summary, our key contributions are:

- We propose CHORE, the first end-to-end learning approach that can reconstruct human, object and contacts from a single RGB image. Our CHORE neural fields predictions allow us to accurately register a controllable body model and 3D object to the image.
- Different from prior works that use weak perspective cameras and learn from synthetic data, we use full perspective camera model, which is crucial to train on real data. To this end, we propose a new training strategy that allows effective pixel-aligned implicit learning with perspective cameras.

– Through our effective training and joint reconstruction, our model achieves a remarkable performance improvement of over 50% compared to prior art [79]. Our code and models are publicly available to foster future research.

2 Related Work

We propose the first learning-based approach to jointly reconstruct 3D human, object and contacts from a single image. In this regard, we first review recent advances in separate human or object reconstructions. We then discuss approaches that can jointly reason about human-object interaction.

Human reconstruction from images. One common approach to reconstruct 3D humans from images is to fit a parametric model such as SMPL [49] to the image [4,5,11,19,44,55,57]. Learning-based approaches have also been used to directly regress model parameters such as pose and shape [3, 26, 38, 40, 43, 54, 56] as well as clothing [3, 9, 27–29, 36, 69]. More recently, implicit function based reconstruction are applied to capture fine-grained details including clothing and hair geometry [34, 62, 63, 75, 80]. These methods however cannot reconstruct 3D objects, let alone human-object interaction.

Object reconstruction from images. Given a single image, 3D objects can be reconstructed as voxels [18, 35, 72], point clouds [22, 45] and meshes [41, 58, 70]. Similar to human reconstruction, implicit functions have shown huge success for object reconstruction as well [50, 53, 76]. The research in this direction has been significantly aided by large scale datasets [15, 66, 74]. We urge the readers to see the recent review [25] of image-based 3D reconstruction of objects. These approaches have shown great promise in rigid object reconstruction but articulated humans are significantly more challenging and human-object interaction even more so. Unlike these methods, our approach can jointly reconstruct humans, objects and their interactions.

Human object interactions. Modelling 3D human-object interactions is extremely challenging. Recent works have shown impressive performance in modelling hand-object interactions from 3D [13, 67, 86], 2.5D [12, 14] and images [20, 21, 32, 42, 77]. Despite being quite impressive, these works have a shortcoming in that they are restricted to only hand-object interactions and cannot predict the full body. Modelling full-body interactions is even more challenging, with works like PROX [30] being able to reconstruct [30, 71] or synthesize [31, 64, 81, 82] 3D humans to satisfy the 3D scene constraints. There are also works capturing interaction from multi-view [10, 39, 65] or reconstructing 3D scene based on human-scene interactions [78]. Recently, this direction has also been extended to model human-human interaction [24], self-contacts [23, 52]. Broadly speaking, these works can model human-object interactions to satisfy scene constraints but cannot jointly reconstruct human-object contacts from single images. Few works directly reason about 3D contacts from images [16, 33, 85] and video [46, 51, 59] but do not reconstruct 3D human and object from a single RGB image.

Closest to our work are [71] and PHOSA [79] . Weng *et al.* [71] predict 3D human and scene layout separately and use scene constraints to optimize

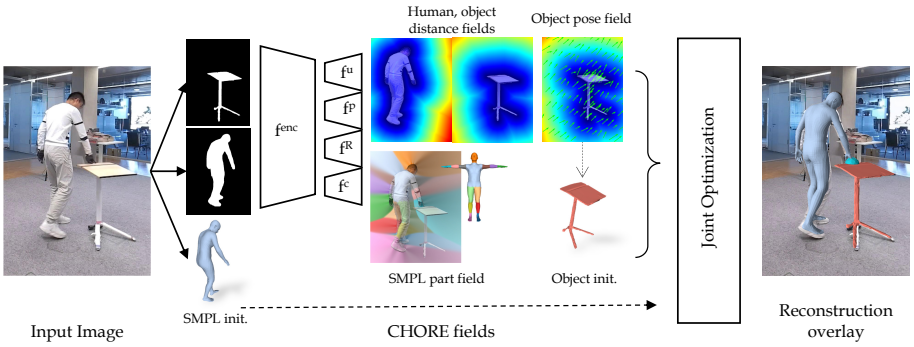

Fig. 2: We present an approach to jointly reconstruct human, object and contacts from a single RGB image. Given an image, human and object masks, we predict CHORE fields: unsigned distance fields to human and object surfaces (f^u), SMPL part (f^p) and object pose (f^R, f^c) fields. We then leverage the neural fields to robustly fit the SMPL model and the object mesh to the image while reasoning about human-object contacts.

human reconstruction. They use predefined contact weights on SMPL-X [55] vertices, which is prone to errors. PHOSA first fits SMPL and object meshes separately and then uses hand crafted heuristics like predefined contact pairs to reason about human-object interaction. These heuristics are neither scale-able nor very accurate. Our method does not rely on heuristics and can learn joint reconstruction and contact priors directly from data. Experiments show that our learning based method easily outperforms heuristic-based PHOSA and [71].

3 Method

We describe here the components of CHORE, a method for 3D reconstruction of the human, object and their contacts from a single RGB image. This work focuses on a single person in close interaction with one dynamic movable object. Our approach is inspired by the classical image based model fitting methods [5, 11], which fit a body model parameterized by shape and pose $\boldsymbol{\theta}, \boldsymbol{\beta}$ by minimizing an energy:

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}) = E_{data} + E_{J2D} + E_{reg}, \tag{1}$$

where E_{data} is a 2D image based loss like a silhouette loss, E_{J2D} is re-projection loss for body joints, and E_{reg} is a loss on body pose and shape priors learned from data [49, 55, 68]. Estimating the human and the object jointly however is significantly more challenging. The relative configuration of human and object needs to be accurately estimated despite occlusions, lack of depth, and hard to detect contacts. Hence, a plain extension of a hand-crafted 2D objective to incorporate the object is far too prone to local minima. Our idea is to jointly predict, based on a single image, 3D neural fields (CHORE fields) for the human, object and their relative pose and contacts, to formulate a learned robust 3D data term. We then fit a parametric body model and a template object mesh to the predicted CHORE neural fields, see overview Fig. 2.

In this section, we first explain our CHORE neural fields in Sec. 3.1, and then describe our novel human-object fitting formulation using learned CHORE neural fields in Sec. 3.2. Unlike prior works on pixel-aligned neural implicit reconstruction [62] which learns from synthetic data and assume a weak perspective camera, we learn from real data captured with full perspective cameras. This poses new challenges which we address with a simple and effective deformation of the captured real 3D surfaces to account for perspective effects (Sec. 3.3).

3.1 CHORE neural fields from single images

As representation, we use the SMPL body model $H(\boldsymbol{\theta}, \boldsymbol{\beta})$ [60] which parameterizes the 3D human as a function of pose $\boldsymbol{\theta}$ (including global translation) and shape $\boldsymbol{\beta}$, and a template mesh for the object. Our idea is to leverage a stronger 3D data term coming from learned CHORE neural fields to fit SMPL and the object templates (one template per class) to the images. In contrast to a plain joint surface reconstruction of human and object, the resulting SMPL and object mesh can be further controlled and edited. To obtain a good fit to the image, we would ideally i) minimize the distance between the optimized meshes and the corresponding ground truth human and object surfaces, and ii) enforce that human-object contacts of the meshes are consistent with input. Obviously, ground truth human and object surfaces are unknown at test time. Hence, we approximate them with CHORE fields. Specifically, a learned neural model predicts 1) the unsigned distance fields of the object and human surfaces, and 2) a part correspondence field to the SMPL model, for more robust SMPL fitting and contact modeling, and 3) an object pose field, to initialize the object pose.

This constitutes a complex multi-task learning problem which we address with a single neural network. We first use a CNN to extract an image-aligned feature grid. For a query point $\mathbf{p} \in \mathbb{R}^3$ we project it to image plane and extract pixel-aligned features, from which we predict the distance, part-correspondence and object pose fields. We explain the details of our network below.

Input encoding f^{enc}. We stack RGB image with human and object masks in separate channels as input and use a leaned feature extractor f^{enc} [62]to extract image feature grid $\mathbf{F}$. Given a query point $\mathbf{p} \in \mathbb{R}^3$, and the image feature $\mathbf{F}$, we extract a pixel-aligned point feature $\mathbf{F}[\pi_{3D \mapsto 2D}(\mathbf{p})]$ by projecting the 3D point to the image. We concatenate the 3D point coordinate $\mathbf{p}$ to the pixel-aligned feature, allowing the network to distinguish points at different depth along the same ray, hence $\mathbf{F}_\mathbf{p} = (\mathbf{F}[\pi_{3D \mapsto 2D}(\mathbf{p})], \mathbf{p})$. The feature extractor f^{enc} is trained with losses defined on neural fields predictions which we explain later.

Human and object distance fields f^u. Based on the point feature $\mathbf{F}_\mathbf{p}$, we use a point decoder f^u to predict the unsigned distances $u_h(\mathbf{p}) \in \mathbb{R}^+$ and $u_o(\mathbf{p}) \in \mathbb{R}^+$ from the point to the human and object surface respectively. This allows to model non-watertight meshes as well. From an unsigned distance field the 3D surface $\mathcal{S}$ can be reconstructed by projecting points to it [17]:

$$\mathcal{S}_i = \{\mathbf{p} - f_i^u(\mathbf{F}_\mathbf{p}) \nabla_\mathbf{p} f_i^u(\mathbf{F}_\mathbf{p}) \,|\, \mathbf{p} \in \mathbb{R}^3\} \tag{2}$$

where f_i^u can be human (f_h^u) or object (f_o^u) UDF predictor. In practice, points $\mathbf{p}$ are initialized from a fixed 3D volume and projection is done iteratively [17].

To train neural distance function f_i^u, we sample a set of points $\mathcal{P}_i$ near the ground truth surfaces and compute the ground truth distance $\text{UDF}_{\text{gt}}(\mathbf{p})$. The decoder f_i^u is then trained to minimize the L_1 distance between clamped prediction and ground truth distance:

$$L_{u_i} = \sum_{\mathbf{p} \in \mathcal{P}_i} |\min(f_i^u(\mathbf{F_p}), \delta) - \min(\text{UDF}_{\text{gt}}(\mathbf{p}), \delta)| \tag{3}$$

where δ is a small clamping value to focus on the volume nearby the surface.

Part field f^p and contacts. While the unsigned distance predictions can reconstruct the human and object surface, they do not tell which human or object points are in contact, and fitting template SMPL meshes using only distance fields often fails [7]. Hence we use another decoder to predict a part correspondence field to the SMPL model with two purposes: i) to formulate a robust SMPL fitting objective [7,8] and ii) to derive which body parts are in contact with the object points. Mathematically, given the point feature $\mathbf{F_p}$ for a point $\mathbf{p} \in \mathbb{R}^3$, we predict its corresponding SMPL part $\mathbf{l_p} = f^p(\mathbf{F_p}) \in \{1, 2, ...K\}$. $K = 14$ is the total number of SMPL parts. The part prediction decoder is trained with standard categorical cross entropy, L_p.

Object pose fields f^R and f^c. Accurate object fitting to the CHORE UDF requires good initialization for the object pose. We learn a predictor $f^R : \mathbf{F_p} \mapsto \mathbb{R}^{3\times3}$ that takes the point feature $\mathbf{F_p}$ as input and predicts the rotation matrix $\mathbf{R}$. Predicting the object translation in camera coordinates directly is more tricky due to the well known depth-scale ambiguity. Hence our predictions are relative to the human. We train a decoder $f^c : \mathbf{F_p} \mapsto \mathbb{R}^5$ which predicts the x, y coordinate of SMPL center (at fixed depth) in camera coordinates, and the object center relative to the SMPL center based on the point feature $\mathbf{F_p}$. Both f^R and f^c are trained with mean squared loss L_R and L_c.

Our feature encoder f^{enc} and neural predictors f^u, f^p, f^R, f^c are trained jointly with the objective: $L = \lambda_u(L_{u_h} + L_{u_o}) + \lambda_p L_p + \lambda_R L_R + \lambda_c L_c$. See supplementary for more details on our network architecture and training.

3.2 3D human and object fitting with CHORE fields

With CHORE fields, we reformulate Eq. (1) to include a stronger data term E_{data} in 3D, itself composed of 3 different terms for human, object, and contacts which we describe next.

3D human term. This term consists of the sum of distances from SMPL vertices to the CHORE neural UDF for the human f_h^u, and a part based term leveraging the part fields f^p discussed in Sec. 3.1:

$$E^h_{\text{data}}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{\mathbf{p} \in H(\boldsymbol{\theta}, \boldsymbol{\beta})} (\lambda_h \min(f_h^u(\mathbf{F_p}), \delta) + \lambda_{p\prime} L_p(l_{\mathbf{p}}, f^p(\mathbf{F_p}))), \tag{4}$$

Here, $L_p(\cdot, \cdot)$ is the standard categorical cross entropy loss function and $l_{\mathbf{p}}$ denotes our predefined part label on SMPL vertex $\mathbf{p}$. $\lambda_h, \lambda_{p\prime}$ are the loss weights.

The part-based term $L_p(\cdot,\cdot)$ is effective to avoid, for example, matching a person's arm to the torso or a hand to the head. To favour convergence, we initialize the human pose estimated from the image using [61].

3D object term. We assume a known object mesh template with vertices $\mathbf{O} \in \mathbb{R}^{3\times N}$ and optimize the object rotation $\mathbf{R}_o \in SO(3)$, translation $\mathbf{t}_o \in \mathbb{R}^3$ and scale $s_o \in \mathbb{R}$ leveraging the neural fields predictions. Given template vertices and object pose, we compute the posed object as $\mathbf{O'} = s_o(\mathbf{R}_o\mathbf{O} + \mathbf{t}_o)$. Intuitively, the distance between object vertices $\mathbf{O'}$ and underlying object surface should be minimized. Furthermore, the projected 2D silhouette should match the observed object mask $\mathbf{M}_o$, yielding the following objective:

$$E^o_{\text{data}}(\mathbf{R}_o, \mathbf{t}_o, s_o) = \sum_{\mathbf{p}\in\mathbf{O'}} (\lambda_o \min(f^u_o(\mathbf{F_p}), \delta) + \lambda_{\text{occ}} L_{\text{occ-sil}}(\mathbf{O'}, \mathbf{M}_o) + \lambda_{\text{reg}} L_{\text{reg}}(\mathbf{O'})), \tag{5}$$

The first term computes the distance between fitted object and corresponding surface represented implicitly with the network f^u_o. The second term $L_{\text{occ-sil}}$ is an occlusion-aware silhouette loss [79]. We additionally use L_{reg} computed as the distance between predicted object center and center of $\mathbf{O'}$. This regularization prevents the object from being pushed too far away from predicted object center.

The object rotation is initialized by first averaging the rotation matrices of the object pose field near the object surface (points with $f^u_o(\mathbf{F_p}) < 4mm$): $\mathbf{R}_o = \frac{1}{M}\sum_{k=1}^{M} \mathbf{R}_k$. We further run SVD on $\mathbf{R}_o$ to project the matrix to $SO(3)$. Object translation $\mathbf{t}_o$ is similarly initialized by averaging the translation part of the pose field nearby the surface. Our experiments show that pose initialization is important for accurate object fitting, see Sec. 4.6.

Joint fitting with contacts. While Eq. (4) and Eq. (5) allow fitting human and object meshes coherently, fits will not necessarily satisfy fine-grained contact constraints. Our key idea here is to jointly reconstruct and fit humans and objects while reasoning about contacts. We minimize a joint objective as follows:

$$E_{\text{data}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{R_o}, \mathbf{t_o}, s_o) = E^h_{\text{data}} + E^o_{\text{data}} + \lambda_c E^c_{\text{data}}, \tag{6}$$

where E^h_{data} and E^o_{data} are defined in Eq. (4) and Eq. (5) respectively. The contact term E^c_{data} consists of the chamfer distance $d(\cdot,\cdot)$ between the sets of human H^c_j and object points O^c_j predicted to be in contact:

$$E^c_{\text{data}}(\mathbf{R}_o, \mathbf{t}_o, s_o) = \sum_{j=1}^{K} d(H^c_j(\boldsymbol{\theta}, \boldsymbol{\beta}), O^c_j). \tag{7}$$

Points on the $j{-}th$ body part of SMPL ($H_j(\boldsymbol{\theta}, \boldsymbol{\beta})$) are considered to be in contact when their UDF $f^u_o(\mathbf{F_p})$ to the object is smaller than a threshold: $H^c_j(\boldsymbol{\theta}, \boldsymbol{\beta}) = \{\mathbf{p} \,|\, \mathbf{p} \in H_j(\boldsymbol{\theta}, \boldsymbol{\beta}) \text{ and } f^u_o(\mathbf{F_p}) \leq \epsilon\}$. Analogously, object points are considered to be in contact with body part j when their UDF to the body is small and they are labelled as part j: $O^c_j = \{\mathbf{p} \,|\, \mathbf{p} \in \mathbf{O'} \text{ and } f^u_h(\mathbf{F_p}) \leq \epsilon \text{ and } f^p(\mathbf{F_p}) = j\}$. The contact term E^c_{data} encourages the object to be close with the corresponding SMPL parts in contact. The loss is zero when no contacts detected. This is important to have a physically plausible and consistent reconstruction, see Sec. 4.5.

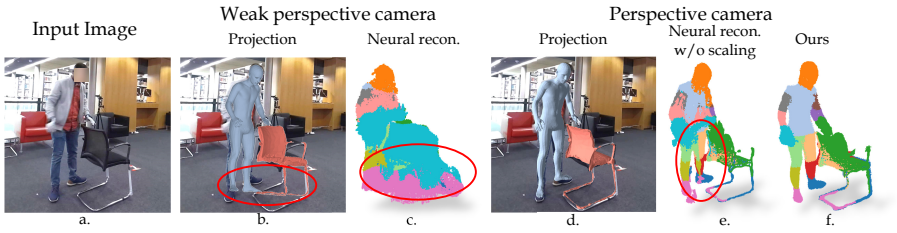

Fig. 3: Learning neural fields on real data. We show projections of GT meshes overlapped with input image using different camera models in b and d. The neural reconstructions (colored based on predicted part labels f^p) are shown in c, e and f. Weak perspective projection is not accurately aligned (b) with image hence network trained with this predicts noisy output (c). Perspective camera projection is accurately aligned (d) but directly applying it, without accounting for depth-scale ambiguity, makes learning hard. Training with this is still not optimal (e). Our proposed method reduces depth ambiguity while maintaining accurate alignment, leading to better predictions (f).

Substituting our data term into Eq. (1), we can now write our complete optimization energy as: $E(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{R_o}, \mathbf{t_o}, s_o) = E^o_{data} + E^h_{data} + \lambda_c E^c_{data} + \lambda_J E_{J2D} + \lambda_r E_{\text{reg}}$. See Supp. for details about different loss weights λ's.

3.3 Learning pixel-aligned neural fields on real data

Our training data consists of real images paired with reference SMPL and object meshes. Here, we want to follow the pixel-aligned training paradigm, which has been shown effective for 3D shape reconstruction. Since both depth and scale determine the size of objects in images this makes learning from single images ambiguous. Hence, existing works place the objects at a fixed distance to the camera, and render them using a weak perspective [62, 63, 84] or perspective camera model [72,73,76,83]. This effectively allows the network to reason about only scale, instead of both scale and depth at the same time.

This strategy to generate training pairs is only possible when learning from *synthetic data*, which allows to move objects and re-render the images. Instead, we learn from *real* images paired with meshes. Re-centering meshes to a fixed depth breaks the pixel alignment with the real images. An alternative is to adopt a weak perspective camera model (assume all meshes are centered at a constant depth), but the alignment is still not accurate as images are recorded with a full perspective camera, leading to inaccurate learning, see Fig. 3 b) and c). Alternatively, attempting to learn a model directly from the original pixel-aligned data with the intrinsic depth-scale ambiguity leads to poor results, as we show in Fig. 3 e). Hence, our idea is to adopt a full perspective camera model, and transform the data such that it is centered at a fixed depth, while preserving pixel alignment with the real images. We demonstrate that this can be effectively achieved by applying depth-dependent scale factor.

Our first observation is that scaling the mesh vertices $\mathbf{V}$ by s does not change its projection to the image. *Proof:* Let $\mathbf{v} = (\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z) \in \mathbf{V}$ be a vertex of the original mesh, and $\mathbf{v}' = (s\mathbf{v}_x, s\mathbf{v}_y, s\mathbf{v}_z) \in s\mathbf{V}$ a vertex of the scaled mesh. For a

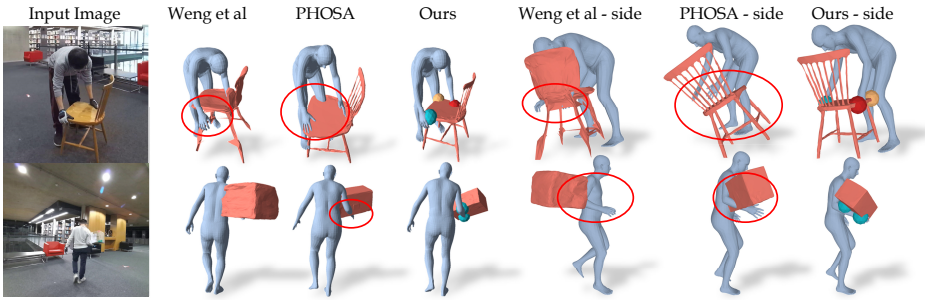

Fig. 4: Comparison with Weng *et al.* [71] and PHOSA [79] on BEHAVE. The contact spheres are placed automatically at the contact centers. [71] uses predefined contact weights and fails to reconstruct spatial arrangement properly. PHOSA's post-hoc also fails to correct errors from separate human-object reconstructions. Our network is trained jointly to reconstruct the human and the object and learns a strong prior on the relative arrangement, allowing for more accurate reconstruction.

camera with perspective focal length f_x, f_y and camera center c_x, c_y, the vertex $\mathbf{v}$ is projected to the image as $\pi(\mathbf{v}) = (f_x \frac{\mathbf{v}_x}{\mathbf{v}_z} + c_x, f_y \frac{\mathbf{v}_y}{\mathbf{v}_z} + c_y)$. The scaled vertices will be projected to the exact same pixels, $\pi(\mathbf{v}') = (f_x \frac{s\mathbf{v}_x}{s\mathbf{v}_z} + c_x, f_y \frac{s\mathbf{v}_y}{s\mathbf{v}_z} + c_y) = \pi(\mathbf{v})$ because the scale cancels out. This result might appear un-intuitive at first. What happens is that the object size gets scaled, but its center is also scaled, hence making the object bigger/smaller and pushing it further/closer to the camera at the same time. This effects cancel out to produce the same projection.

Given this result, all that remains to be done, is to find a scale factor which centers objects at a *fixed* depth z_0 to the camera. The desired scale factor that satisfies this condition is $s = \frac{z_0}{\mu_z}$, where $\mu_z = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_z^i$ is the mean of the un-scaled mesh vertices depth. *Proof:* The proof follows immediately from the linearity property of the empirical mean. The depth of the transformed mesh center is:

$$\mu_z' = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_z^{i\prime} = \frac{1}{n}\sum_{i}^{n} \mathbf{v}_z^i \frac{z_0}{\mu_z} = \mu_z \frac{z_0}{\mu_z} = z_0 \tag{8}$$

This shows that the depth μ_z' of the scaled mesh center is fixed at z_0 and its image projection remains unchanged, as desired.

We scale all meshes in our training set as described above to center them at $z_0 = 2.2m$ and recompute GT labels. To maintain pixel alignment we use the camera intrinsic from the dataset. At test time, the images have various resolution and *unknown* intrinsic with person at large range of depth (see Fig. 6). We hence crop and resize the human-object patch such that the person in the resized patch appears as if they are at z_0 under the camera intrinsic we use during training. We leverage the human mesh reconstructed from [61] as a reference to determine the patch resizing factor. Please see supplementary for details.

4 Experiments

Datasets. We conducted experiments on the BEHAVE [10], COCO [47] and NTU-RGBD [48] dataset. BEHAVE captures 8 subjects interacting with 20 different objects in natural environments using RGBD cameras. Each image is paired with (pseudo) ground truth SMPL and object fits. CHORE is trained on 10.7k frames at 4 locations from BEHAVE and tested on 4.5k frames from one unseen location. All 20 categories are seen in both training and test set but at test time the object instances maybe different. As PHOSA [79] relies on 2D object mask for pose estimation, we exclude images where the object is occluded more than 70% for fair comparison, which leads to 4.1k test frames. To compare with Weng *et al.* [71], we select 10 common object categories between BEHAVE and their method (in total 2.6k images) and compare our method with them on this smaller set, denoted as BEHAVE^{--}.

To verify the generalization ability, we also test our method on COCO [47] where each image contains at least one instance of the object categories in BEHAVE and the object should not be occluded more than 50%. Images where there is no visible contact between human and object are also excluded. There are four overlapping categories between BEHAVE and COCO (backpack, suitcase, chair and sports ball) and in total 913 images are tested.

We additionally test on the NTU-RGBD dataset, which is quite interesting since it features human-object interactions. We select images from the NTU-RGBD dataset following the same criteria for COCO and in total 1.5k images of chair or backpack interactions are tested.

Evaluation and baselines. To the best of our knowledge, PHOSA [79] and Weng *et al.* [71] are our closest baselines. To compare with PHOSA, we annotated object templates with contact part labels and intrinsic scales according to PHOSA's convention. [71] focuses on reconstructing the human in static scenes, hence their objects are mainly from indoor environments. Therefore, we only compare with it on the BEHAVE^{--} set and the indoor NTU-RGBD dataset.

For BEHAVE, we compute the Chamfer distance separately on SMPL and object meshes after Procrustes alignment using the combined mesh. Since [71] uses a different object template, we perform Procrustes alignment using only SMPL vertices for BEHAVE^{--}. As there are no 3D ground truth available in COCO and NTU-RGBD for thorough quantitative comparison, we compare our method with baselines through anonymous user study surveys.

4.1 Comparison on BEHAVE dataset

We compare our method against PHOSA [79] and [71] in Tab. 1. Note that the Procrustes alignment is performed on combined mesh for BEHAVE and on SMPL mesh only for BEHAVE^{--} for a fair comparison with [71]. Tab. 1 clearly shows that our learning based approach outperforms the heuristic optimization based PHOSA and [71].

We also compare our method qualitatively with PHOSA and [71] in Fig. 4. It can be seen that PHOSA fails to reconstruct the object at the correct position

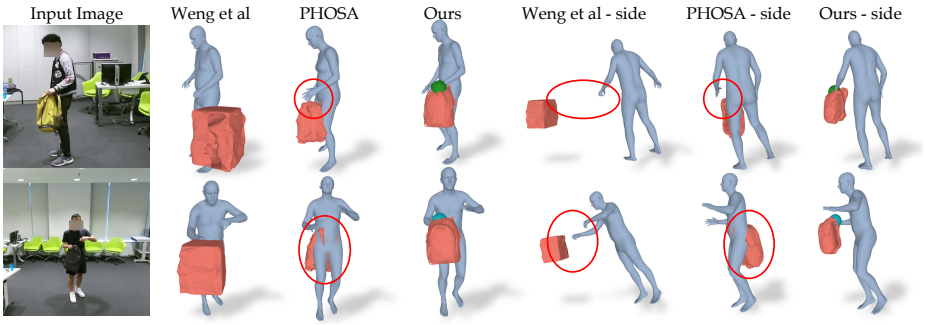

Fig. 5: Comparing our methods with [79] and [71] on NTU-RGBD dataset [48]. Our method trained on BEHAVE generalizes to NTU-RGBD and outperforms baselines.

Dataset	Methods	SMPL ↓	Object ↓
BEHAVE	PHOSA [79]	12.17 ± 11.13	26.62 ± 21.87
	Ours	**5.58 ± 2.11**	**10.66 ± 7.71**
BEHAVE^{--}	PHOSA [79]	6.09 ± 1.98	182.67 ± 319.64
	Weng et al [71]	7.86 ± 3.48	51.86 ± 96.38
	Ours	**5.22 ± 2.03**	**13.15 ± 10.60**

Table 1: The mean and standard deviation of the Chamfer distance (cm) for SMPL and object reconstruction on [10]. Our approach that jointly reasons about the human, object and the contacts outperforms [71] and heuristics based PHOSA [79].

relative to the human. [71] also fails quite often as it uses predefined contact weights on SMPL vertices instead of predicting contacts from inputs. On the other hand, CHORE works well because it learns a 3D prior over human-object configurations based on the input image. See supplementary for more examples.

4.2 Generalisation beyond BEHAVE

To verify our model trained on the BEHAVE dataset can generalize to other datasets, we further test it on the NTU-RGBD [48] as well as COCO dataset [47] and compare with [71, 79].

Comparison with PHOSA [79]. We test our method and PHOSA on the selected COCO images discussed above. We then *randomly* select 50 images and render ours and PHOSA reconstruction results side by side. The results are collected in an anonymous user study survey and we ask 50 people on Amazon Mechanical Turk (AMT [1]) to select which reconstruction is better. Notably, people think our reconstruction is better than PHOSA on 72% of the images. A similar user study for images from the NTU-RGBD dataset is released to AMT and on average 84% of people prefer our reconstruction. Sample images are shown in Fig. 6 and Fig. 5. See supplementary for details on the user study and more qualitative examples.

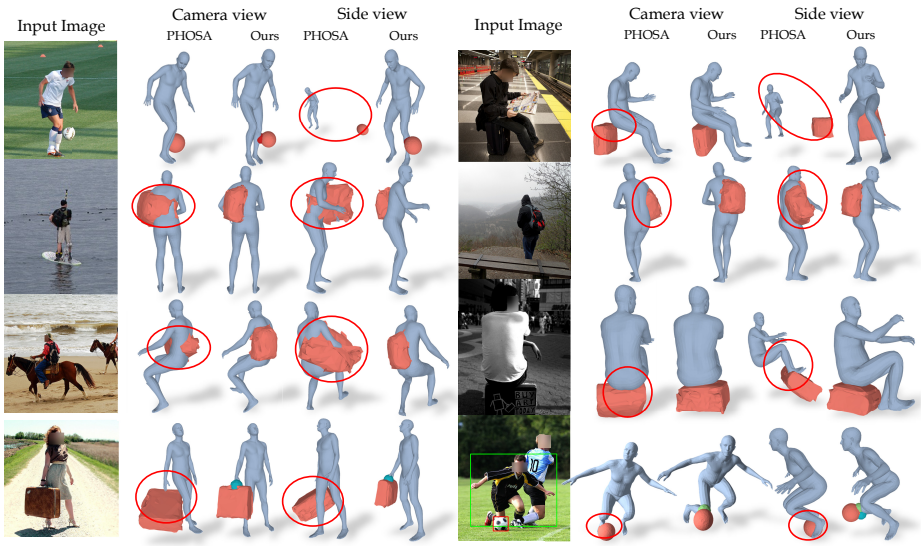

Fig. 6: Comparison with PHOSA [79] on COCO dataset [47]. For images with more than one person, the reconstructed human and object are highlighted by green and red boxes. Our model trained on BEHAVE outperforms PHOSA on 'in the wild' images.

Comparison with Weng *et al.* [71]. One limitation of [71] is that they use indoor scene layout assumptions, which makes it unsuitable to compare on 'in the wild' COCO images. Instead, we only test it on the NTU-RGBD dataset. Similar to the comparison with PHOSA, we release a user study on AMT and ask 50 users to select which reconstruction is better. Results show that our method is preferred by 94% people. Our method clearly outperforms [71], see Fig. 5 and supplementary for more examples.

4.3 Comparison with a human reconstruction method

Our method takes SMPL pose predicted from FrankMocap [61] as initialization and fine tunes it together with object using network predictions. We compare the human reconstruction performance with FrankMocap and PHOSA [79] in Tab. 2. It can be seen that our joint reasoning of human and object does not deteriorate the human reconstruction. Conversely, it slightly improves the human reconstruction as it takes cues from the interacting object into account.

4.4 Pixel-aligned learning on real data

CHORE is trained with the depth-aware scaling strategy described in Sec. 3.3 for effective pixel aligned implicit surface learning on real data. To evaluate its importance, we compare with two baseline models trained with weak perspective and direct perspective (no depth-aware scaling on GT data) camera. Qualitative comparison of the neural reconstruction is shown in Fig. 3. We further fit SMPL and object meshes using these baseline models (with object pose prediction) on

Metric (cm)	PHOSA	FrankMocap	Ours
v2v ↓	6.90 ± 2.68	6.91 ± 2.75	**5.46** ± **3.47**
Chamfer ↓	6.11 ± 1.98	6.12 ± 2.00	**5.27** ± **1.98**

Table 2: Mean vertex to vertex (v2v) and Chamfer distance for human reconstruction results after Proscrustes alignment on SMPL mesh. Our method that jointly reasons about human and object performs on par with human only FrankMocap [61].

BEHAVE test set and compute the errors for reconstructed meshes, see Tab. 3b) and c). It can be clearly seen that our proposed training strategy allows effective learning of implicit functions and significantly improves the reconstruction.

4.5 Contacts modeling

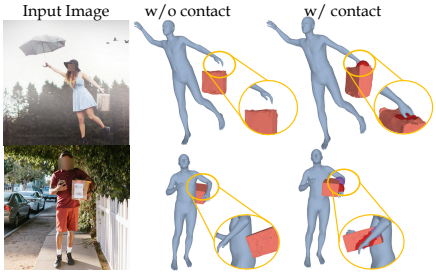

Fig. 7: Accurate human and object reconstruction requires modeling contacts explicitly. Without contact information, the human and the objects are not accurately aligned and the resulting reconstruction is not physically plausible.

Methods	SMPL ↓	Object ↓
a. neural recon.	14.06	15.34
b. weak persp.	9.14	25.24
c. direct persp.	7.63	15.58
d. w/o f^R	6.58	17.54
e. w/o f^c	6.26	14.25
f. w/o contact	**5.51**	10.86
g. full model	5.58	**10.66**

Table 3: Ablation studies. We can see that neither weak perspective (b) nor direct perspective camera (c) can work well on real data. Our proposed training strategy, together with rotation and translation prediction achieves the best results.

We have argued that contacts between human and the object are crucial for accurately modelling interaction. Quantitatively, the contact term E^c_{data} only reduces the object error slightly (Tab. 3f, g) as contacts usually constitute a small region of the full body. Nevertheless, it makes significant difference qualitatively, as can be seen in Fig. 7. Without the contact information, the object cannot be fit correctly to the contact location. Our contact prediction helps snap the object to the correct contact location, leading to a more physically plausible and accurate reconstruction.

4.6 Other ablation studies

We further ablate our design choices in Tab. 3. Note that all numbers in Tab. 3 are Chamfer distances (cm) and computed on meshes after joint fitting except a) which is computed on the neural reconstructed point clouds. We predict object rotation and translation to initialize and regularize object fitting, which are important for accurate reconstruction. In Tab. 3, it can be seen that fitting results without rotation d) or without translation initialization e) are worse

compared to our full model g). Without initialization the object gets stuck in local minima. See supplementary for qualitative examples.

We also compute the Chamfer distance of the neural reconstructed point clouds (Eq. (2)) using alignments computed from fitted meshes. It can be seen that this result is much worse than the results after joint fitting (Tab. 3g). This is because network predictions can be noisy in some regions while our joint fitting incorporates priors of human and object making the model more robust.

5 Limitations and Future Work

We introduce the first learning based method to reconstruct 3D humans and objects from single RGB images. Although the results are promising, there are some limitations. For instance, we assume known object template for the objects shown in the images. Future works can remove this assumption by selecting template from 3D CAD database or reconstructing shape directly from image. Non-rigid deformation, multiple objects, multi-person interactions are exciting avenues for future research. Please also refer to supp. for failure case discussion.

6 Conclusion

In this work we presented CHORE, the first method to *learn* a *joint* reconstruction of human, object and contacts from single images. CHORE combines powerful neural implicit predictions for human, object and its pose with robust model-based fitting. Unlike prior works on pixel-aligned implicit surface learning, we learn from real data instead of synthetic scans, which requires a different training strategy. We proposed a simple and elegant depth-aware scaling that reduces the depth-scale ambiguity while preserving pixel alignment.

Experiments on three different datasets demonstrate the superiority of CHORE compared to the SOTA. Quantitative experiments on BEHAVE dataset evidence that joint reasoning leads to a remarkable improvement of over 50% in terms of Chamfer distance compared to PHOSA. User studies on the NTU-RGBD and the challenging COCO dataset (no ground truth) show that our method is preferred over PHOSA by 84% and 72% users respectively. We also conducted extensive ablations, which reveal the effectiveness of the different components of CHORE, and the proposed depth-aware scaling for pixel-aligned learning on real data. Our code and models are released to promote further research in this emerging field of single image 3D human-object interaction capture.

Acknowledgements. We would like to thank RVH group members [2] for their helpful discussions. Special thanks to Beiyang Li for supplementary preparation. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), and German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

1. https://www.mturk.com
2. http://virtualhumans.mpi-inf.mpg.de/people.html
3. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
4. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: International Conf. on 3D Vision (sep 2018)
5. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: IEEE Conf. on Computer Vision and Pattern Recognition (2018)
6. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
7. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: European Conference on Computer Vision (ECCV). Springer (August 2020)
8. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: 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)
9. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
10. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
11. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: 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)
12. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019), https://contactdb.cc.gatech.edu
13. Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: Contactgrasp: Functional multi-finger grasp synthesis from contact. In: IROS (04 2019)
14. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: The European Conference on Computer Vision (ECCV) (August 2020)
15. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
16. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: 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)
17. Chibane, J., Mir, A., Pons-Moll, G.: Neural unsigned distance fields for implicit function learning. In: Neural Information Processing Systems (NeurIPS). (December 2020)

18. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. Springer International Publishing (2016)
19. Corona, E., Pons-Moll, G., Alenya, G., Moreno-Noguer, F.: Learned vertex descent: A new direction for 3d human model fitting. In: European Conference on Computer Vision (ECCV). Springer (October 2022)
20. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: 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)
21. Ehsani, K., Tulsiani, S., Gupta, S., Farhadi, A., Gupta, A.: Use the force, luke! learning to predict physical forces by simulating effects. In: CVPR (2020)
22. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
23. Fieraru, M., Zanfir, M., Oneata, E., Popa, A., Olaru, V., Sminchisescu, C.: Learning complex 3d human self-contact. CoRR **abs/2012.10366** (2020), https://arxiv.org/abs/2012.10366
24. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
25. Fu, K., Peng, J., He, Q., Zhang, H.: Single image 3d object reconstruction based on deep learning: A review. Multimedia Tools and Applications **80**(1), 463–498 (2021). https://doi.org/10.1007/s11042-020-09722-8, https://doi.org/10.1007/s11042-020-09722-8
26. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
27. Guo, C., Chen, X., Song, J., Hilliges, O.: Human performance capture from monocular video in the wild. In: 2021 International Conference on 3D Vision (3DV). pp. 889–898. IEEE (2021)
28. Habermann, M., Xu, W., , Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020)
29. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Trans. Graph. **38**(2), 14:1–14:17 (Mar 2019). https://doi.org/10.1145/3311970, http://doi.acm.org/10.1145/3311970
30. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: International Conference on Computer Vision (2019)
31. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
32. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)

33. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 13274–13285 (Jun 2022)
34. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. ArXiv **abs/2004.04572** (2020)
35. Häne, C., Tulsiani, S., Malik, J.: Hierarchical surface prediction for 3d object reconstruction. In: 2017 International Conference on 3D Vision (3DV). pp. 412–420 (2017). https://doi.org/10.1109/3DV.2017.00054
36. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: European Conference on Computer Vision. Springer (2020)
37. Jiang, J., Streli, P., Fender, A., Qiu, H., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European conference on computer vision (2022)
38. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020)
39. Jiang, Y., Jiang, S., Sun, G., Su, Z., Guo, K., Wu, M., Yu, J., Xu, L.: Neuralfusion: Neural volumetric rendering under human-object interactions. arXiv preprint arXiv:2202.12825 (2022)
40. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society (2018)
41. Kar, A., Tulsiani, S., Malik, J.: Category-specific object reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
42. Karunratanakul, K., Yang, J., Zhang, Y., Black, M., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 8th International Conference on 3D Vision. pp. 333–344. IEEE (Nov 2020). https://doi.org/10.1109/3DV50981.2020.00043
43. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: Video inference for human body pose and shape estimation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 5252–5262. IEEE (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00530
44. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
45. Lei, J., Sridhar, S., Guerrero, P., Sung, M., Mitra, N., Guibas, L.J.: Pix2surf: Learning parametric 3d surface models of objects from images. In: Proceedings of European Conference on Computer Vision (ECCV) (August 2020), https://geometry.stanford.edu/projects/pix2surf
46. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3d motion and forces of person-object interactions from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
48. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding.

IEEE Transactions on Pattern Analysis and Machine Intelligence (2019). https://doi.org/10.1109/TPAMI.2019.2916873
49. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: ACM Transactions on Graphics. ACM (2015)
50. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
51. Monszpart, A., Guerrero, P., Ceylan, D., Yumer, E., J. Mitra, N.: iMapper: Interaction-guided scene mapping from monocular videos. ACM SIGGRAPH (july 2019)
52. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
53. MÃ¼ller, N., Wong, Y.S., Mitra, N.J., Dai, A., Nie{ss}ner, M.: Seeing behind objects for 3d multi-object tracking in rgb-d sequences. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2021)
54. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: International Conf. on 3D Vision (2018)
55. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
56. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: IEEE Conf. on Computer Vision and Pattern Recognition (2018)
57. Pons-Moll, G., Rosenhahn, B.: Model-Based Pose Estimation, chap. 9, pp. 139–170. Springer (2011)
58. Pontes, J.K., Kong, C., Sridharan, S., Lucey, S., Eriksson, A., Fookes, C.: Image2mesh: A learning framework for single image 3d reconstruction. In: ACCV. Springer International Publishing (2019)
59. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
60. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
61. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: IEEE International Conference on Computer Vision Workshops (2021)
62. Saito, S., , Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
63. Saito, S., Simon, T., Saragih, J., Joo, H.: 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)
64. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: Learning Interaction Snapshots from Observations. ACM Transactions on Graphics (TOG) **35**(4) (2016)

65. Sun, G., Chen, X., Chen, Y., Pang, A., Lin, P., Jiang, Y., Xu, L., Wang, J., Yu, J.: Neural free-viewpoint performance rendering under complex human-object interactions. In: Proceedings of the 29th ACM International Conference on Multimedia (2021)
66. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
67. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020), https://grab.is.tue.mpg.de
68. Tiwari, G., Antic, D., Lenssen, J.E., Sarafianos, N., Tung, T., Pons-Moll, G.: Pose-ndf: Modeling human pose manifolds with neural distance fields. In: European Conference on Computer Vision (ECCV). Springer (October 2022)
69. Tiwari, G., Bhatnagar, B.L., Tung, T., Pons-Moll, G.: Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In: European Conference on Computer Vision (ECCV). Springer (August 2020)
70. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
71. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. arXiv preprint arXiv:2012.01591 (2020)
72. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In: Advances In Neural Information Processing Systems (2017)
73. Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W.T., Tenenbaum, J.B.: Learning 3D Shape Priors for Shape Completion and Reconstruction. In: European Conference on Computer Vision (ECCV) (2018)
74. Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S.: Objectnet3d: A large scale database for 3d object recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. Springer International Publishing (2016)
75. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13296–13306 (Jun 2022)
76. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/39059724f73a9969845dfe4146c5660e-Paper.pdf
77. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: ICCV (2021)
78. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-aware object placement for visual environment reconstruction. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3959–3970 (Jun 2022)
79. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV) (2020)

80. Zhang, S., Liu, J., Liu, Y., Ling, N.: Dimnet: Dense implicit function network for 3d human body reconstruction. Computers & Graphics **98**, 1–10 (2021). https://doi.org/https://doi.org/10.1016/j.cag.2021.04.035
81. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: PLACE: Proximity learning of articulation and contact in 3D environments. In: International Conference on 3D Vision (3DV) (Nov 2020)
82. Zhang, X., Bhatnagar, B.L., Guzov, V., Starke, S., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. arXiv preprint arXiv:2205.00541 (May 2022)
83. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J.B., Freeman, W.T., Wu, J.: Learning to Reconstruct Shapes From Unseen Classes. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
84. Zhao, F., Wang, W., Liao, S., Shao, L.: Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12674–12683 (2021)
85. Zhe, C., Hang, G., Karttikeya, M., Qizhi, C., Minh, V., Jitendra, M.: Long-term human motion prediction with scene context. In: ECCV (2020)
86. Zhou, K., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Toch: Spatio-temporal object correspondence to hand for motion refinement. In: European Conference on Computer Vision (ECCV). Springer (October 2022)