SINGLE-IMAGE COHERENT RECONSTRUCTION OF OBJECTS AND HUMANS

Sarthak Batra
University of Surrey
s.batra@surrey.ac.uk

Simon Hadfield
University of Surrey
s.hadfield@surrey.ac.uk

Partha P. Chakrabarti
Indian Institute of Technology Kharagpur
ppchak@cse.iitkgp.ac.in

Armin Mustafa
University of Surrey
armin.mustafa@surrey.ac.uk

Abstract

Existing methods for reconstructing objects and humans from a monocular image suffer from severe mesh collisions and performance limitations for interacting occluding objects. This paper introduces a method to obtain a globally consistent 3D reconstruction of interacting objects and people from a single image. Our contributions include: 1) an optimization framework, featuring a collision loss, tailored to handle human-object and human-human interactions, ensuring spatially coherent scene reconstruction; and 2) a novel technique to robustly estimate 6 degrees of freedom (DOF) poses, specifically for heavily occluded objects, exploiting image inpainting. Notably, our proposed method operates effectively on images from real-world scenarios, without necessitating scene or object-level 3D supervision. Extensive qualitative and quantitative evaluation against existing methods demonstrates a significant reduction in collisions in the final reconstructions of scenes with multiple interacting humans and objects and a more coherent scene reconstruction.

1. Introduction

Existing methods for human and object reconstructions are either limited to single objects and humans or give limited performance for complex images with multiple people and objects [13, 22, 27, 35, 43, 73]. These methods estimate the 3D poses of humans and objects independently and do not take into account the human-human interactions [88] and even if they do they generally follow a supervised approach [32]. This leads to large collisions between the meshes with incoherent reconstructions. We consider the full scene holistically and exploit information from the human-human and human-object interactions for spatially coherent and more complete 3D reconstruction of in-the-wild images.

PHOSA [88] pioneered the field and proposed the first method that reconstructs humans interacting with objects for in-the-wild images. However, PHOSA does not explicitly model human-human interactions and gives erroneous reconstructions when objects are heavily occluded which leads to reconstructions with incorrect depth ordering and mesh collisions. Multi-human model-free reconstruction from a single image was proposed in [57], however, this method does not deal with interacting humans. Other methods [76, 77] for multi-human reconstructions generate reconstructions with severe mesh collisions because they reconstruct each person independently. To address these challenges, in this paper, we have proposed an optimization-based framework for the spatially coherent reconstruction of scenes with multiple interacting people and heavily occluded objects. The method first reconstructs humans [34] and objects [38] in the image independently. The initial poses of people in the scene are then optimized to resolve any ambiguities that arise from this independent composition using a collision loss, depth ordering loss, and interaction information. To deal with heavily occluded objects, a novel 6 DOF pose estimation is proposed that uses inpainting to refine the segmentation mask of the occluded object for significantly improved pose estimation. Finally, we propose a global objective function that scores 3D object layouts, orientations, collision, and shape exemplars. Gradient-based solvers are used to obtain globally optimized poses for humans and objects. Our contributions are:

- A method for generating a cohesive scene reconstruction from a single image by capturing interactions among humans and between humans and objects within the scene, all without relying on any explicit 3D supervision.
- A collision loss in an optimization framework to robustly estimate 6 DOF poses of multiple people and objects in crowded images.

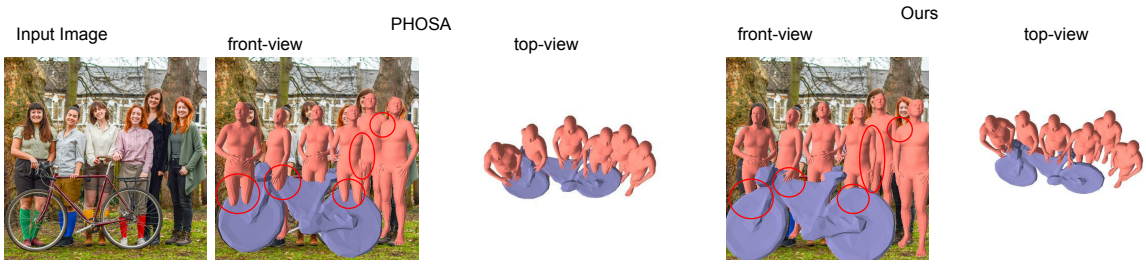

Figure 1. Comparison of the proposed method (right) reconstruction with PHOSA(middle). The proposed method gives a more coherent reconstruction with correct spatial arrangement by reasoning about human-human and human-object interaction

- An inpainting-based method to improve the segmentation mask of heavily occluded objects that greatly boosts the precision of 6 DOF object position estimations.
- Extensive evaluation of the proposed method on complex images with multiple interacting humans and objects from the COCO-2017 dataset [50] against the state-of-the-art demonstrate the effectiveness of our approach.

2. Related Work

3D humans from a single image: Reconstructing 3D human models from images is often achieved through various methodologies. One widely used approach involves fitting parametric models like SMPL[53] to input images [2, 3, 6, 15, 43, 62, 65]. Alternatively, learning-based techniques directly predict model parameters such as pose and shape[4, 25, 32, 35, 42, 59, 61]. [52, 62, 92] use statistical body models and a large number of 3D scans to recover 3D humans from a single image. [6] use 2D poses, [79] uses 2D body joint heatmaps and [44] uses GraphCNN to estimate SMPL model [52]. However, these methods only estimate the 3D of a single person in the scene. Methods like [32, 57, 86, 87] recover the 3D poses and shapes of multiple people focus on resolving ambiguities that arise due to incorrect depth ordering and collisions between people. However, these methods cannot handle large occlusions. Recent advancements in whole-body mesh recovery from images have shifted the focus from solely on regressing body parameters to also accurately estimating hand and face parameters. An example is FrankMocap [68], which employs a modular design. This approach initially runs independent 3D pose regression methods for the face, hands, and body and integrates their outputs through a dedicated module. A one-stage pipeline named OSX [49] has been introduced for 3D whole-body mesh recovery, surpassing existing multi-stage models in accuracy. It introduces a component-aware transformer (CAT) comprising a global body encoder and a local face/hand decoder. KBody[93] represents a methodology for fitting a low-dimensional body model to an image, employing a predict-and-optimize approach. A distinctive feature of KBody is the introduction of virtual joints, enhancing correspondence quality and disentangling the optimization process between pose and shape parameters.

3D objects from a single image: Single-view 3D reconstruction is a complex task, as it necessitates incorporating reliable geometric priors derived from our 3D world. However, these priors often lack in diverse real-world scenarios [30, 31, 63, 64]. Given their robustness and accessibility, learning-based methods have emerged. Deep learning approaches can be categorized based on the employed 3D representations, encompassing voxel-based frameworks [13, 55, 66, 67, 83], point cloud-based methods [10, 18, 26], mesh-based techniques [23, 36, 47, 51], and implicit function-based approaches [12, 71, 72]. The majority of current single-view 3D mesh reconstruction methods employ an encoder-decoder framework. Here, the encoder discerns perceptual features from the input image, while the decoder distorts a template to conform to the desired 3D shape. The pioneering work by [81] introduced deep learning networks to this task. They employed the VGG network [74] as the encoder and utilized a graph convolutional network (GCN) as the decoder. [24] introduced a method wherein a 3D shape is represented as a collection of parametric surface elements, allowing for a flexible representation of shapes with arbitrary topology. [60] addressed topology changes by proposing a topology modification network that adaptively deletes faces. These methods are trained and evaluated on identical object categories.

Recent research has also devised techniques for 3D reconstruction from image collections without explicit 3D supervision. This has been achieved by employing differentiable rendering to supervise the learning process. For instance, [36] proposed a method that reconstructs the underlying shape by learning deformations on top of a category-specific mean shape. [48] developed a differentiable rendering formulation to learn signed distance functions as implicit 3D shape representations, overcoming topological restrictions. [16] learned both deformation and implicit 3D shape representations, facilitating reconstruction in category-specific canonical space. [80] extended category-specific models into cross-category models through distillation. [45] used GNN trained on a synthetic dataset without any humans to deduce an object's layout.

3D human-to-object interaction: Modeling 3D human-object interactions poses significant challenges. Recent

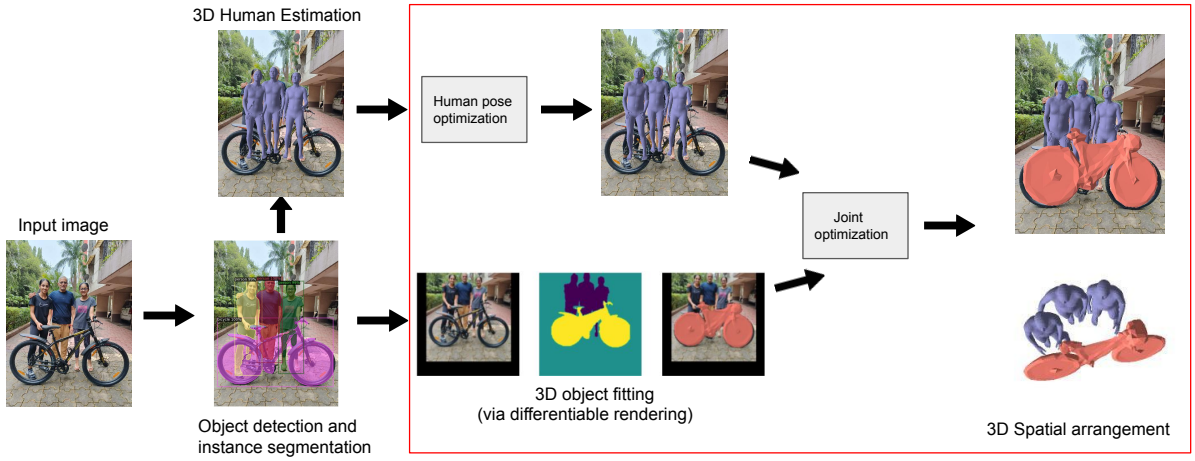

Figure 2. Overview of the proposed method to generate spatially coherent reconstruction from a single image. The steps in red box are novel. The reconstruction before human pose optimization exhibits notable mesh collisions. After human pose optimization, reduced mesh collisions are seen while maintaining relative coherence between humans.

studies have demonstrated remarkable success in capturing hand-object interactions from various perspectives, including 3D [8, 78, 91], 2.5D [7, 9], and images [14, 17, 29, 37]. However, these achievements are limited to hand-object interactions and do not extend to predicting the full body. The complexity increases when considering full-body interactions, with works like PROX [27] successfully reconstructing [27, 82] or synthesizing [28, 89, 90] 3D humans to adhere to 3D scene constraints. Other approaches focus on capturing interactions from multiple views [5, 33, 75] or reconstructing 3D scenes based on human-scene interactions [85]. More recently, efforts have extended to model human-human interactions [19] and self-contacts [20, 56]. [73] used information from the RGBD videos of individuals interacting with interiors to train a model that understands how people interact with their surroundings. Access to 3D scenes gives scene constraints that enhance the perception of 3D human poses [41, 69, 84]. [27] uses an optimization-based method to enhance 3D human posture estimates conditioned on a particular 3D scene obtained from RGBD sensors. Another recent method, [70], creates a 3D scene graph of people and objects for indoor data. [11] represents the optimal configuration of the 3D scene, in the form of a parse graph that encodes the object, human pose, and scene layout from a single image. In our work, we overcome the limitations of existing methods by handling not only on human-object interactions but also capturing human-human interactions and propose a method that deals with major occlusions to significantly improved scene reconstruction.

3. Methodology

The proposed method takes a single RGB image as input and gives a spatially coherent reconstruction of interacting humans and objects in the scene, an overview is shown in Figure 2. We exploit human-human and human-object interactions to spatially arrange all objects in a common 3D coordinate system. First, objects and humans are detected, followed by SMPL-based per-person reconstruction (Sec. 3.1), which gives incorrect spatial reconstructions with collisions between meshes. The human 3D locations/poses are translated into world coordinates and refined through a human-human spatial arrangement optimization using a collision loss (Sec. 3.2). To correctly estimate the 3D object pose (6-DoF translation and orientation) a differentiable renderer is used that fits 3D mesh object models to the predicted 2D segmentation masks [40]. We correct the occluded object mask using image inpainting (in Sec. 3.3) unlike PHOSA [88] which uses an occluded object mask. Lastly, we perform joint optimization that takes into account both human-human and human-object interactions for a globally consistent output. Our framework produces plausible reconstructions that capture realistic human-human and human-object interactions.

3.1. Estimating 3D Humans

Using [34], we estimate the 3D shape and pose parameters of SMPL [52] given a bounding box for a human [54]. The 3D human is parameterized by pose $\theta \in \mathbb{R}^{72}$, shape $\beta \in \mathbb{R}^{10}$, and a weak-perspective camera $\gamma = [\sigma, t_x, t_y] \in \mathbb{R}^3$. To position the humans in the 3D space, γ is converted to the perspective camera projection by assuming a fixed focal length f for all images, and the distance of the person is determined by the reciprocal of the camera scale parameter σ. Thus, the 3D vertices of the SMPL model for the i^{th} human are represented as: $M_i = J(\theta_i, \beta_i) + [t_x, t_y, \frac{f}{\sigma^i}]$, where J is the differentiable SMPL mapping from pose and shape to a human mesh and $t_h^i = [t_x, t_y, \frac{f}{\sigma^i}]$ is the translation of i^{th} human. The person's height and size are regulated by the SMPL shape parameter β. We define scale parameter(s^i)

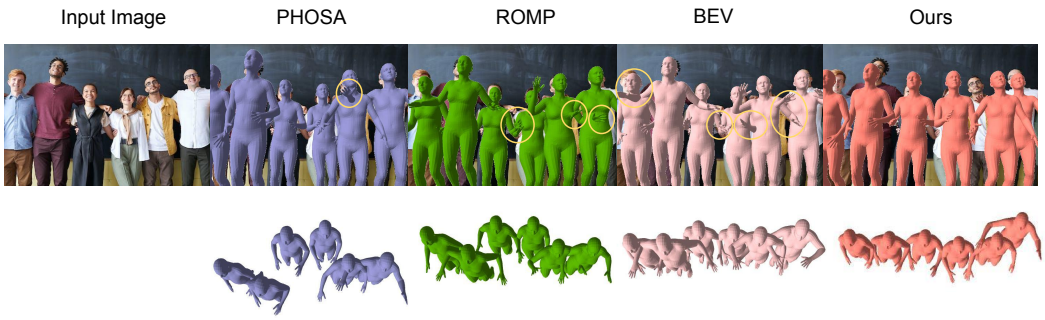

Figure 3. The proposed approach gives spatially coherent reconstructions with a significant reduction in mesh collisions compared to PHOSA [88], ROMP[76], and BEV[77]. Significant collision are shown in highlighted circles.

for each human similar to PHOSA and the final vertices are given by $\bar{M}_i = s^i M_i$.

3.2. Human Pose Optimisation

Independently analyzing human 3D poses results in inconsistent 3D scene configurations. Human-human interactions offer useful information to determine the relative spatial arrangement and not considering this leads to ambiguities like mesh penetration and incorrect depth ordering. We propose an optimization framework that incorporates human-human interactions. We first identify interacting humans in the image and then optimize the pose through an objective function to correctly adjust their spatial arrangements.

Identifying interacting humans - Our hypothesis posits that human interactions are contingent upon physical proximity in world coordinates. Hence we find the interacting humans by identifying the overlap of 3D bounding boxes(More details regarding bounding box overlap criteria can be found in the supplementary material).

Objective function to optimize 3D spatial arrangement Our objective includes collision ($L_{H-collision}$), interaction ($L_{H-interaction}$), and depth ordering loss ($L_{H-depth}$) terms to constraint the pose for interacting humans:

$$L_{HHI-Loss} = \lambda_1 L_{H-collision} + \lambda_2 L_{H-depth} + \lambda_3 L_{H-interaction} \quad (1)$$

We optimize (1) using a gradient-based optimizer [39] w.r.t. translation $\mathbf{t}^i \in \mathbb{R}^3$ and scale parameter s^i and the Rotation $\mathbf{R}^i \in SO3$ for the i^{th} human instance . The human translations are initialized from Sec 3.1. The terms in the objective function are defined below:

Collision Loss ($L_{H-collision}$) - To overcome the problem of mesh collisions, as seen in existing methods in Fig. 3, we introduce a collision loss $L_{H-collision}$ that penalizes interpenetrations in the reconstructed people. Let ϕ be a modified Signed Distance Field (SDF) for the scene that is defined as follows: $\phi(x, y, z) = -min(SDF(x, y, z), 0)$ where ϕ is positive for points inside the human and is proportional to the distance from the surface, and is 0 outside of the human. Typically ϕ is defined on a voxel grid of dimensions $N_p * N_p * N_p$. While it's definitely possible to generate a single voxelized representation for the entire scene, we often find ourselves requiring an extensive fine-grained voxel grid. Depending on the scene's extent, this can pose processing challenges due to memory and computational limitations. To overcome this a separate ϕ_i function is computed for each person by calculating a tight box around the person and voxelizing it instead of the whole scene to reduce computational complexity [32]. The collision penalty of person j for colliding with person i is defined as follows: $P_{ij} = \sum_{v \in M_j} \tilde{\phi}_i(v)$, where $\tilde{\phi}_i(v)$ samples the ϕ_i value for each 3D vertex v in a differentiable way from the 3D grid using trilinear interpolation. If there is a collision between person i and a person j, P_{ij} will be a positive value and decreases as the separation between them increases. If there is no overlap between person i and j, P_{ij} will be zero. Let the translation of person i and person j be T_i and T_j respectively. Then the collision loss between them is defined as:

$$L_{ij} = \begin{cases} \frac{P_{ij}}{\exp(||T_i - T_j + \delta||_2)} & T_i = T_j \\ \frac{P_{ij}}{\exp(||T_i - T_j||_2)} & T_i \neq T_j \end{cases} \quad (2)$$

When the translation values are the same (in case of maximum overlap) we use an extra term δ ($0 < \delta < 1$) to ensure non-zero gradients are not very large to avoid any instabilities during optimization. The final collision loss for a scene with N people is defined as follows:

$$L_{H-collision} = \sum_{j=1}^{N} \Big(\sum_{i=1 i \neq j}^{N} L_{ij} \Big) \quad (3)$$

Interaction Loss ($L_{H-interaction}$) - This is an instance-level to pull the interacting people close together, similar to [88]: $L_{H-interaction} = \sum_{h_i, h_j \in H} \mu(h_i, h_j) ||C(h_i) - C(h_j)||_2$, where $\mu(h_i, h_j)$ identifies whether human h_i and h_j are interacting according to the 3D bounding box overlap criteria. $C(h_i)$ and $C(h_j)$ give the centroid for human i and human j respectively.

Depth-Ordering Loss ($L_{H-depth}$) - This helps to achieve more accurate depth ordering, as in [32]. The loss is defined as: $L_{depth} = \sum_{p \in S} \log(1 + \exp(D_{y(p)}(p) - D_{\bar{y}(p)}(p)))$, where $S = \{p \in I : y(p) > 0, \bar{y}(p) > 0, y(p) \neq \bar{y}(p)\}$ is the pixels in the image I with incorrect depth ordering

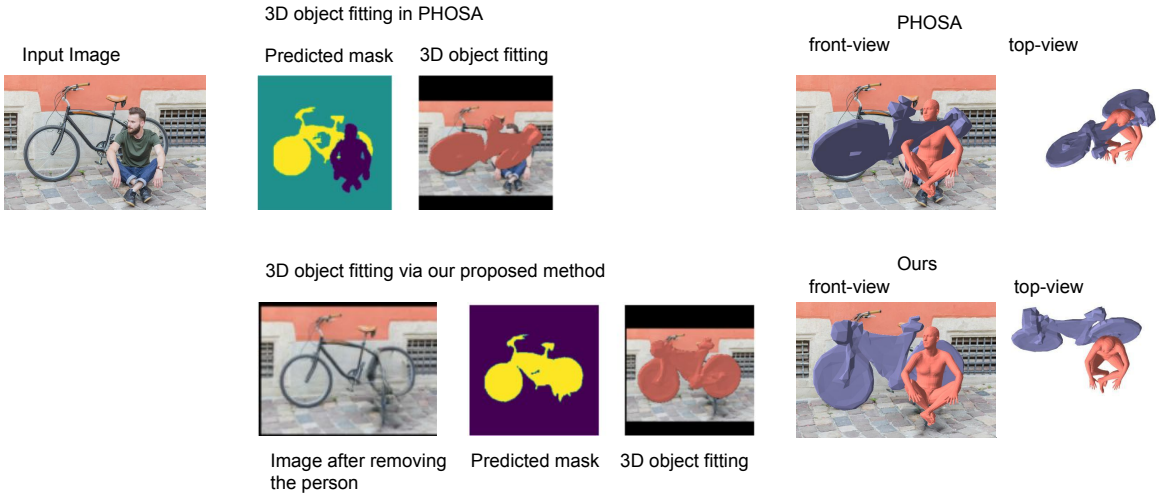

Figure 4. Comparison of the segmentation masks and reconstruction with PHOSA. The segmentation mask of the bicycle is occluded resulting in erroneous reconstruction in PHOSA. The proposed method uses image inpainting to remove the occlusion to generate a better segmentation mask, which leads to a more complete reconstruction.

in the ground truth segmentation, the person index at pixel position p is represented by $y(p)$, and the predicted person index in the rendered 3D meshes is $\bar{y}(p)$ and $y(p) \neq \bar{y}(p)$. $D_{y(p)}(p)$ and $D_{\bar{y}(p)}(p)$ represent the pixel depths.

3.3. 3D Object Pose Estimation

After estimating the shape and pose of humans, the next step is to estimate the same for the objects. To estimate the 3D location $t \in \mathbb{R}^3$ and 3D orientation $R \in SO(3)$ of the objects. For each object category, exemplar mesh models are pre-selected. The mesh models are sourced from [1, 46]. The vertices of j^{th} object are: $V_o^j = s^j(R^j O(c^j, k^j) + t^j)$, where c^j is the object category from MaskRCNN [54], and $O(c^j, k^j)$ determines the $k^j - th$ exemplar mesh for category c^j. The optimization framework chooses the exemplar that minimizes re-projection error to determine k^j automatically and s^j is the scale parameter for j^{th} object.

Our first objective is to estimate the 6 DOF pose of each object independently. It is difficult to estimate 3D object pose in the wild as there are: (1) no parametric 3D models for objects; (2) no images of objects in the wild with 2D/3D pose annotations; and (3) occlusions in cluttered scenes with humans. We address these challenges by proposing an optimization-based approach that uses a differentiable renderer [38] to fit the 3D object to instance masks from [40] in a manner that is robust to minor/major occlusions.

As defined in [88] we calculate a pixel-wise L2 loss over rendered silhouettes S versus predicted masks M but the quality of the predicted mask M is impacted by occlusions as seen in [88], which results in a poorly estimated 6 DOF pose. To address problems due to occlusions, we propose a novel method that improves the masks as shown in Fig. 4.

Given an image I, a total number of objects N, and bounding boxes for rigid B_r and non-rigid B_{nr} objects, along with their masks - M_r for rigid and M_{nr} for non-rigid objects. Each i^{th} object can be occluded by maximum $N-1$ objects. To identify occluding objects we calculate the Intersection over Union(IOU) between all pairs of bounding boxes. Objects with $IOU > 0.3$ (Our selection of this threshold stems from our empirical observations, wherein we found that objects with $IOU > 0.3$ led to noticeable improvements in reconstruction quality. Conversely, when IOU was less than 0.3, the reconstruction results obtained using our method closely resembled those produced by PHOSA [88], more details in supplementary) are occluding objects M for each object. Occluding objects can be removed in numerous ways, for e.g remove only one object at a time. The total possible combinations, in this case, are $\binom{M}{1}$, or you remove a pair of objects at a time and the total possible combinations, in this case, are $\binom{M}{2}$ and so on. The total number of all possible combinations can be described as $\binom{M}{0} + \binom{M}{1} + \binom{M}{2} + ...\binom{M}{M} = 2^M$. To remove j occluding objects where $j \leq M$ we need a single mask $M_{occ-mask}$ that is a combination of the j masks, so $M_{occ-mask} = M_1 + M_2 + ... + M_j$. Now we use the image-inpainting approach proposed by [58] to remove the occluding objects. We pass the current image I and the mask $M_{occ-mask}$ to get a new image without occlusions and use this image to get the new segmentation masks and bounding boxes:

$$I_{new} = EC(I, M_{occ-mask})$$
$$B_r^{new}, B_{nr}^{new}, M_r^{new}, M_{nr}^{new} = OD(I_{new}) \quad (4)$$

where EC is the image inpainting algorithm and OD is the object detection algorithm. Sometimes, the i^{th} object in I may not correspond to the same object in I_{new}. Let's

say the index of the i^{th} object in I_{new} be k. We iterate over the list of new bounding boxes and calculate the IOU of these boxes with $B_r[i]$ and, k corresponds to the index of the bounding box for which IOU is closest to 1. We use the mask $M_r^{new}[k]$ to determine object pose. Estimating a reliable pose also depends heavily on the boundary details. To incorporate this we augment the L2 mask loss with a modified version of the symmetric chamfer loss [21]. Given a no-occlusion indicator η (0 if pixel only corresponds to a mask of a different instance, else 1), the loss is:

$$L_{occ-sil} = \sum(\eta \circ S - M_r^{new}[k])^2 + \sum_{p \in E(\eta \circ S)} \min_{\bar{p} \in E(M)} ||p - \bar{p_2}|| \quad (5)$$

We generate masks $M_{occ-mask}$ for different values of j and a 3D pose corresponding to that mask is chosen that results in a minimum value of $L_{occ-sil}$. The edge map of mask M is computed by E(M). To estimate the 3D object pose, we minimize the occlusion-aware silhouette loss:

$$(R^j, t^j)^* = \underset{R,t}{argmin}(L_{occ-sil}(\Pi_{sil}(V_o^j), M_r^{new}[k])) \quad (6)$$

where Π_{sil} is the silhouette rendering of a 3D mesh model via a perspective camera with a fixed focal length f (Sec 3.1) and M_j is a 2D instance mask for the j^{th} object. Instance masks are computed by PointRend [40].

3.4. Joint Optimization

The joint optimization refines both the human and object poses estimated above, exploiting both human-human and human-object interactions through joint loss functions. Estimating 3D poses of people and objects in isolation from one another leads to inconsistent 3D scene reconstruction. Interactions between people and objects provide crucial clues for correct 3D spatial arrangement, which is done by identifying interacting objects and humans and proposing an objective function for refining human/object poses.

Identifying human-object interaction. Our hypothesis posits that human-object interactions are contingent upon physical proximity in world coordinates. We use 3D bounding box overlap between the human and object to determine whether the object is interacting with a person.

Objective function to optimize 3D spatial arrangements. We define a joint loss function that takes into account both human-human and human-object interactions. It is crucial to include both of them because if you simply optimize with regard to human-object interactions, it may result in erroneous relative positions between interacting people even if it would enhance the relative spatial arrangement between the interacting humans and objects.

$$L_{joint-loss} = L_{HOI-Loss} + L_{HHI-Loss} \quad (7)$$

where $L_{HHI-Loss}$ is same as Eq. 1 and

$$L_{HOI-Loss} = \lambda_1 L_{HO-collision} + \lambda_2 L_{HO-depth} + \lambda_3 L_{HO-interaction} + \lambda_4 L_{occ-sil} \quad (8)$$

Depth-Ordering Loss ($L_{HO-depth}$) is same as Section 3.2. We optimize (8) using a gradient-based optimizer [39] w.r.t. translation $t^i \in \mathbb{R}^3$ and intrinsic scale $s^i \in \mathbb{R}$ for the i^{th} human and, rotation $R^j \in SO(3)$, translation $t^j \in \mathbb{R}^3$ and $s^j \in \mathbb{R}$ for the j^{th} object instance jointly. The object poses are initialized from Sec. 3.3. $L_{occ-sil}$ is the same as (5) except without the chamfer loss which didn't help during joint optimization.

Interaction loss ($L_{HO-interaction}$): This loss handles both coarse and fine interaction between humans and objects as in [88], defined as: $L_{HO-interaction} = L_{coarse-inter} + L_{fine-inter}$.

The coarse interaction loss is: $L_{coarse-inter} = \sum_{h \in H, o \in O} \mu(h,o)||C(h) - C(o)||_2$, where $\mu(h,o)$ identifies whether human h and object o are interacting according to the 3D bounding box overlap criteria. $C(h)$ and $C(o)$ give the centroid for human and the object respectively. To handle human interactions, the fine interaction loss is defined as:

$L_{fine-inter} = \sum_{h \in H, o \in O}(\sum_{P_h, P_o \in P(h,o)} \mu(P_h, P_o)||C(P_h) - C(P_o)||_2)$, where P_h and P_o are the regions of interaction between the humans and the object, respectively. $\mu(P_h, P_o)$ is the overlap of the 3D bounding box between the interacting objects, recomputed at each iteration.

Collision Loss ($L_{HO-collision}$) **-** The formulation of this loss is similar to the collision loss defined in Section 3.2. The difference is that here we take into account the mesh collision between interacting humans and objects in contrast to interacting humans. Let N_h represent the total number of humans and N_o total number of objects, then the Loss function is defined as: $L_{HO-collision} = \sum_{j=1}^{N_o}\left(\sum_{i=1}^{N_h} L_{h_i o_j} + L_{o_j h_i}\right)$, where h_i represents the i^{th} human and o_j represents the j^{th} object.

4. Results and Evaluation

We perform both quantitative and qualitative assessments of the performance of our technique on the COCO-2017 [50] dataset on images that include interactions of humans and objects against PHOSA [88], ROMP[76], and BEV[77].

4.1. Qualitative and Quantitative Analysis

Figures 5 and 6 show a **qualitative** comparison with PHOSA, ROMP and BEV. PHOSA reconstructs both humans and objects; ROMP and BEV only reconstruct humans. As seen our approach yields improved reconstruction quality by effectively mitigating ambiguities arising from mesh collisions and occlusions.

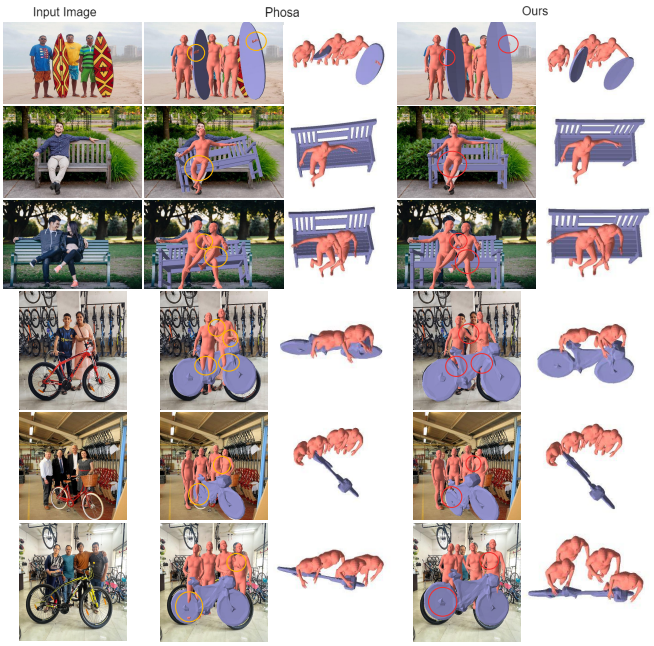

Figure 5. Qualitative comparison on test images from COCO 2017 against PHOSA [88] with human-object interactions. Our method gives better spatial reconstruction while substantially reducing collisions(the golden circles delineate regions characterized by noteworthy mesh collisions, while the red circles delineate areas showcasing enhancements in reconstructions). More qualitative results are shown in 4.3

Methods	E_{H-col}	$E_{H-depth}$	E_{HO-col}	$E_{HO-depth}$
PHOSA	79.42	86.68	78.21	68.84
ROMP	63.51	74.27	-	-
BEV	35.25	56.17	-	-
Ours	**16.46**	**48.37**	**26.65**	**33.77**

Table 1. Quantitative evaluation with PHOSA [88], ROMP[76], and BEV[77]. BEV and ROMP only reconstruct humans. Equations of each evaluation parameter are given in the supplementary.

For **quantitative** evaluation, we employ a forced-choice assessment approach similar to PHOSA[88] on COCO-2017 [50] images since there are no 3D ground truth annotations for people and objects in images in the wild. From the COCO-2017 test set, we randomly selected a sample of images and performed reconstruction on each image. We compare our method with PHOSA, ROMP, and BEV by reconstructing the scenes and comparing the degree of mesh collisions for human-human E_{H-col} and human-object E_{HO-col} and incorrect depth ordering for human-human $E_{H-depth}$ and human-object $E_{HO-depth}$ interactions that results from each method. This is averaged across all images to estimate values in Table 1. Our approach outperforms the state-of-the-art techniques for both multi-human and multi-human-object reconstruction, as well as results in a more coherent and realistic reconstruction with significantly fewer ambiguities.

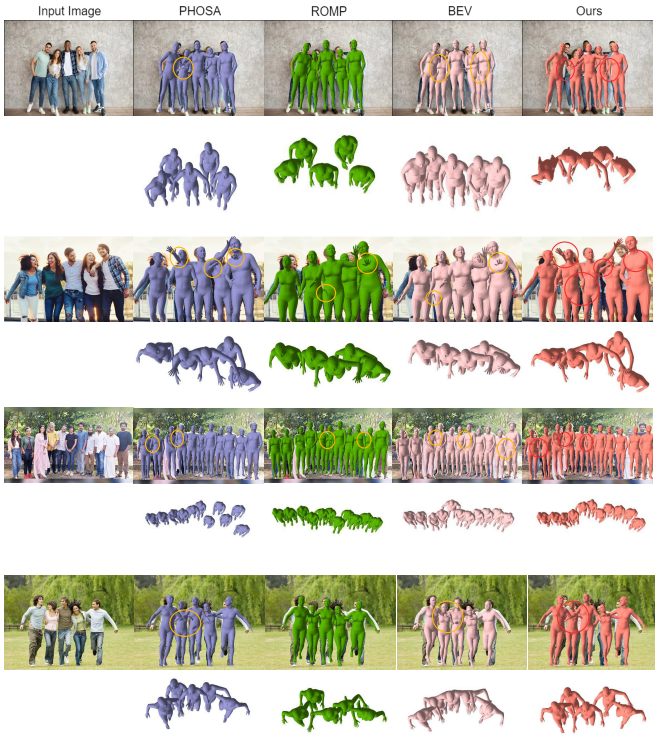

Figure 6. Qualitative results of proposed method on test images from COCO 2017 compared to PHOSA, ROMP, and BEV for human-human interactions. Our method gives more realistic and coherent reconstructions for images with multiple humans.

Ours vs.	PHOSA	ROMP	BEV
	88%	80%	74%

Table 2. User study that gives the average percentage of images for which our method performs better on COCO-2017. 50% implies equal performance.

Ours vs.	No $L_{collision}$	No L_{depth}	No $L_{interaction}$	No $L_{occ-sil}$
	83%	62%	73%	77%

Table 3. In the ablation study we drop loss terms from our proposed method. The higher the percentage, the more the effect of the loss term. No $L_{collision}$ implies the exclusion of both $L_{H-collision}$ and $L_{HO-collision}$. No L_{depth} involves omitting $L_{H-depth}$ and $L_{HO-depth}$. No $L_{interaction}$ means we omitted the $L_{H-interaction}$ and $L_{HO-interaction}$, and lastly No $L_{occ-sil}$ corresponds to dropping the loss term defined in eq. 5

We also perform a subjective study similar to [88], where we show the reconstructions for each image from PHOSA, ROMP, BEV, and our proposed method in a random order to the users and the users mark whether our result looks better than, equal to, or worse than the other methods. We compute the average percentage of images for which our method performs better in Table 2. Overall, the performance of our method is relatively better than the other methods.

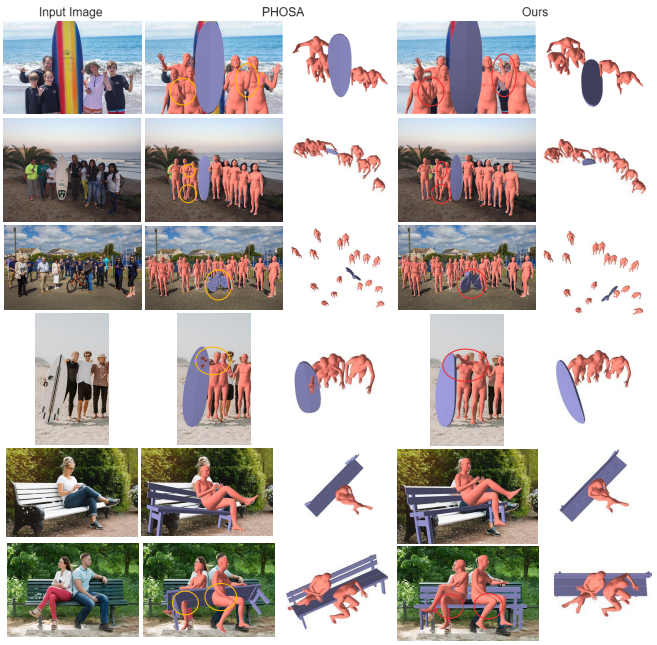

Figure 7. Our method, recovers plausible human-object and human-human spatial arrangements by explicitly reasoning about them. Here we demonstrate reconstruction on images with both humans and objects and compare PHOSA's reconstructions to those produced by our method.

4.2. Ablation Study

An ablative study was conducted to assess the significance of the loss terms in Table 3. The identical forced-choice test similar to PHOSA[88] is conducted for the complete proposed methodology (Equation 7), by omitting loss terms from the proposed method and measuring the performance. Our findings indicate that the exclusion of the collision and occlusion-aware silhouette loss has the most notable effect, with the interaction loss following closely behind. The collision loss prevents mesh intersection and the silhouette loss guarantees that the object poses remain consistent with their respective masks.

4.3. Additional Results

More results are shown Fig. 7 and 8 on challenging Youtube and Google images.

5. Discussion

Current approaches for reconstructing humans/ objects from a single image often produce reconstructions that contain various ambiguities, especially in situations involving multiple interactions between humans and between humans and objects. In this paper, we perform holistic 3D scene perception by exploiting the information from both human-human and human-object interactions in an optimization framework. The optimization makes use of several constraints to provide a full scene that is globally consistent, and reduces collisions, and improves spatial arrangement (Table 1) over other methods. The proposed human optimization framework resolves ambiguities between reconstructed people, and the proposed human-object optimization framework addresses ambiguities between humans and objects. We further introduce a method that significantly improves the pose estimation of heavily occluded objects. We demonstrate via our qualitative and quantitative evaluations that the proposed method outperforms other methods and produces reconstructions with noticeably less ambiguity.

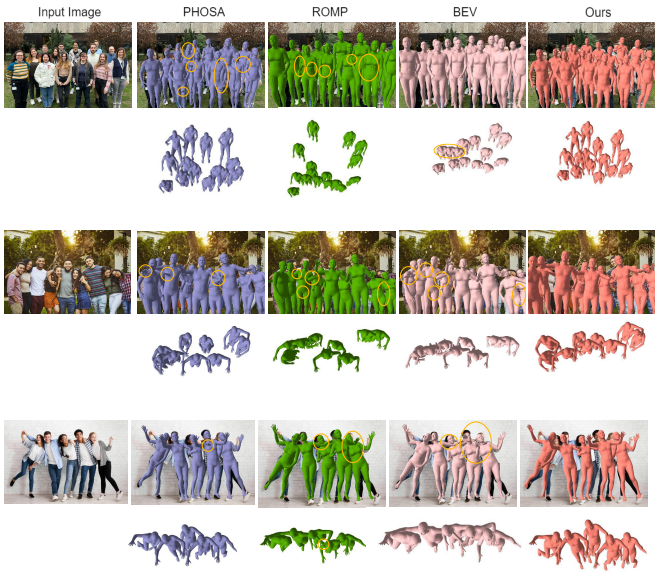

Figure 8. we illustrate the differences in human reconstructions generated by PHOSA, ROMP, BEV, and Our approach when provided with an input image. Our approach produces more plausible reconstructions with a substantial decrease in mesh collisions, all while maintaining relative coherence.

6. Limitations and Future Work

When compared to learning-based techniques, our method demands increased processing time for reconstruction generation and may occasionally yield a slightly inaccurate spatial configuration to mitigate collisions. In terms of future directions, our current implementation only considers coarse interactions among humans. However, in subsequent iterations, we aim to incorporate fine-grained interactions, leveraging this information to refine our estimation of human pose.

References

[1] Free3d. https://free3d.com/. 5

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars

from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3

[6] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[7] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 3

[8] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 3

[9] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 3

[10] Chao Chen, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Unsupervised learning of fine structure generation for 3d point clouds by 2d projections matching. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 12466–12477, 2021. 2

[11] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 3

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5939–5948, 2019. 2

[13] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 1, 2

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 3

[15] Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *European Conference on Computer Vision*, pages 146–165. Springer, 2022. 2

[16] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1536–1546, 2022. 2

[17] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2020. 3

[18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. 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*, pages 605–613, 2017. 2

[19] 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. 3

[20] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1343–1351, 2021. 3

[21] Dariu M Gavrila. Pedestrian detection from a moving vehicle. In *Computer Vision—ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6*, pages 37–49. Springer Berlin Heidelberg, 2000. 6

[22] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 484–499. Springer, 2016. 1

[23] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 2

[24] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[25] 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, pages 10884–10894, 2019. 2

[26] Zhizhong Han, Chao Chen, Yu-Shen Liu, and Matthias Zwicker. Drwr: A differentiable renderer without rendering for unsupervised 3d structure learning from silhouette images. *arXiv preprint arXiv:2007.06127*, 2020. 2

[27] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 1, 3

[28] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 3

[29] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 3

[30] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970. 2

[31] Katsushi Ikeuchi and Berthold KP Horn. Numerical shape from shading and occluding boundaries. *Artificial intelligence*, 17(1-3):141–184, 1981. 2

[32] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2020. 1, 2, 4

[33] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022. 3

[34] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 1, 3

[35] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1, 2

[36] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

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

[38] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 1, 5

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 6

[40] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 3, 5, 6

[41] Hedvig Kjellström, Danica Kragić, and Michael J Black. Tracking people interacting with objects. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 747–754. IEEE, 2010. 3

[42] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[43] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 1, 2

[44] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 2

[45] Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, and Abhinav Gupta. 3d-relnet: Joint object and relational network for 3d prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2212–2221, 2019. 2

[46] 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. 5

[47] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 677–693. Springer, 2020. 2

[48] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. *Advances in Neural Information Processing Systems*, 33:11453–11464, 2020. 2

[49] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 2

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 2, 6, 7

[51] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7708–7717, 2019. 2

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[53] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[54] R Mask. Cnn/he k., gkioxari g., dollar p., girshick r. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 3, 5

[55] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[56] 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. 3

[57] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14474–14483, 2021. 1, 2

[58] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 5

[59] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[60] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 2

[61] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

[62] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[63] Alex Pentland. Shape information from shading: a theory about human perception. In *[1988 Proceedings] Second International Conference on Computer Vision*, pages 404–413. IEEE, 1988. 2

[64] Alex P Pentland. Local shading analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):170–187, 1984. 2

[65] Gerard Pons-Moll and Bodo Rosenhahn. Model-based pose estimation. *Visual Analysis of Humans: Looking at People*, pages 139–170, 2011. 2

[66] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 366–383. Springer, 2020. 2

[67] Stephan R Richter and Stefan Roth. Matryoshka networks: Predicting 3d geometry via nested shape layers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1936–1944, 2018. 2

[68] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 2

[69] Bodo Rosenhahn, Christian Schmaltz, Thomas Brox, Joachim Weickert, Daniel Cremers, and Hans-Peter Seidel. Markerless motion capture of man-machine interaction. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3

[70] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289*, 2020. 3

[71] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[72] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

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

[74] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[75] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4651–4660, 2021. 3

[76] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people.

In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 1, 4, 6, 7

[77] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 1, 4, 6, 7

[78] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 3

[79] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[80] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. *arXiv e-prints*, pages arXiv–2204, 2022. 2

[81] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[82] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021. 3

[83] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[84] Masanobu Yamamoto and Katsutoshi Yagishita. Scene constraints-aided tracking of human body. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, pages 151–156. IEEE, 2000. 3

[85] Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. Human-aware object placement for visual environment reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3959–3970, 2022. 3

[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] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *Advances in Neural Information Processing Systems*, 31, 2018. 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 1, 3, 4, 5, 6, 7, 8

[89] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651. IEEE, 2020. 3

[90] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 3

[91] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3

[92] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. *ACM transactions on graphics (TOG)*, 29(4):1–10, 2010. 2

[93] Nikolaos Zioulis and James F O'Brien. Kbody: Balanced monocular whole-body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3539–3544, 2023. 2