StackFLOW: Monocular Human-Object Reconstruction by Stacked Normalizing Flow with Offset

Chaofan Huo[1,2] , **Ye Shi**[1,2] , **Yuexin Ma**[1,2] , **Lan Xu**[1,2] , **Jingyi Yu**[1,2] and **Jingya Wang**[*1,2]

[1]ShanghaiTech University
[2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
{huochf, shiye, mayuexin, xulan1, yujingyi, wangjingya}@shanghaitech.edu.cn,

Abstract

Modeling and capturing the 3D spatial arrangement of the human and the object is the key to perceiving 3D human-object interaction from monocular images. In this work, we propose to use the Human-Object Offset between anchors which are densely sampled from the surface of human mesh and object mesh to represent human-object spatial relation. Compared with previous works which use contact map or implicit distance filed to encode 3D human-object spatial relations, our method is a simple and efficient way to encode the highly detailed spatial correlation between the human and object. Based on this representation, we propose Stacked Normalizing Flow (StackFLOW) to infer the posterior distribution of human-object spatial relations from the image. During the optimization stage, we finetune the human body pose and object 6D pose by maximizing the likelihood of samples based on this posterior distribution and minimizing the 2D-3D corresponding reprojection loss. Extensive experimental results show that our method achieves impressive results on two challenging benchmarks, BEHAVE and InterCap datasets. Our code has been publicly available at https://github.com/huochf/StackFLOW.

1 Introduction

For a decade, the 3D information recovery of the human and the object from the image was studied alone, without considering their interaction. Recent studies suggest that the integration of humans and surrounding objects can produce physically consistent results [Hassan *et al.*, 2019], and that the reconstruction accuracy of both can be improved even more [Chen *et al.*, 2019; Sun *et al.*, 2021; Zhang *et al.*, 2023a]. In the monocular human-object reconstruction task, which aims at reconstructing human mesh and object mesh jointly from a single-view image, the interaction plays an important role in providing constraints for human pose and object position. However, how to utilize the interaction between the human and the object to refine themselves still remains unsolved.
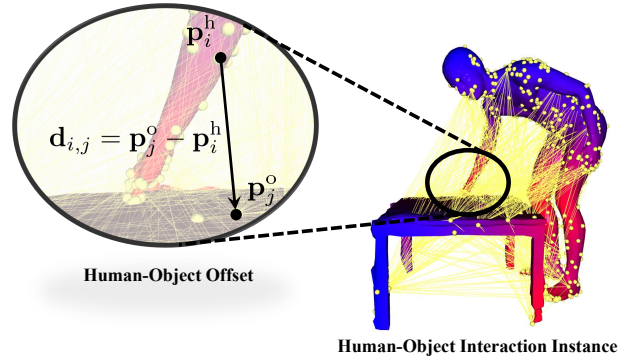
---

[*]Corresponding author.

Figure 1: Human-Object Offset $\mathbf{d}_{i,j}$ describes how far from the human anchor point $\mathbf{p}_i^{\mathrm{h}}$ to object anchor point $\mathbf{p}_j^{\mathrm{o}}$ through the direction of the vector $\mathbf{d}_{i,j}$. They are calculated between two sets of anchors which are densely sampled from the surface of human mesh and object mesh beforehand. The dense offset captures a highly-detailed correlation between human parts and object parts. It is a quantitative representation to encode the 3D spatial relationship between the human and the object given human-object interaction instance.

The visible presentation of interaction between the human and the object in the 3D world is their spatial arrangement, which involves the posture of the human and the 6D pose of the object. Creating an appropriate representation for human-object spatial arrangement is vital for both human-object interaction capture from images and post-optimization for refinement. Contact map is a recently popular fine-grained representation to model the interaction between the human and the object. It has been applied to model human-scene interaction [Huang *et al.*, 2022a] and human-object interaction [Zhang *et al.*, 2020a; Xie *et al.*, 2022]. The contact map defines the contact regions in human mesh and object mesh and is suitable to be applied in the post-optimization step to generate plausible results by drawing closer contact points in human and object mesh. However, it only preserves the local contact information and cannot model the non-contact interaction types. Moreover, it relies on plausible initialization of the human and the object during optimization, and therefore it is not an independent representation to encode human-object spatial arrangement. Another way to encode human-object spatial arrangement is using the implicit distance field, which is a neural function that maps 3D

points to point-to-face distances [Karunratanakul *et al.*, 2020; Xie *et al.*, 2022]. It is suitable to model 3D object shapes but some shortcomings may show up in modeling human-object spatial relationships. First, it is low-efficient since we need to sample many points to approximate the surface of the mesh. Moreover, spatial arrangement is encoded using functionalized representation implicitly rather than vectorized representation which results in applying probabilistic models to model the distribution of human-object spatial arrangement is difficult and indirect. In this paper, we pursue an efficient and unifying representation to encode the 3D spatial relationship between the human and the object.

The relative distance frequently assumes a prominent role in numerous descriptions of 3D spatial relations, serving as a conspicuous and widely employed method of encoding. But in the scenario of human-object interaction, things may become complicated because of the articulated human body. In this work, we present a novel representation to encode human-object spatial relations using human-object offset. In order to involve all human body parts and cover various object shapes, we randomly sample anchor points from the surface of human mesh and object mesh. The offsets are calculated between all human anchors and object anchors for a given human-object pair as depicted in Figure 1. We treat the offset as the numerical description of the spatial relation pattern for a target human-object pair. These offsets are representative since they encode highly detailed correlations between human parts and object parts. We can use them to recover the posture of the human and the position of the object by adjusting the position of human anchors and object anchors. Due to the regular topological structure of human mesh and rigid object mesh, these offsets are very redundant. Then we use PCA to transform these offsets from high-dimension offset space to low-dimension latent space by linear projection. The human-object offset is a generalized representation from the contact map since the contact map only keeps the anchors with zero offsets.

Regressing accurate offsets from the image is hard due to the variety of spatial relations, the ambiguity of monocular capture, the indeterminacy of viewports, and the diversity of object scale. To tackle these problems, we design our method from two aspects. First, we use a probabilistic model to infer the distribution of spatial relationship instead of single point regression following [Kolotouros *et al.*, 2021]. This distribution narrows down the search space of human-object spatial relations during the post-optimization step and more convincing results can be produced. Moreover, we decouple the process of inferring human-object spatial relation into two stacked subprocesses: human pose estimation and pose-conditioned distribution inference. With the guide of human pose, distribution for human-object spatial relations can be learned more stably and efficiently.

Our contributions can be summarized as:

1. A new 3D spatial relation encoding technique is proposed to encode highly informative global correlation between the human and the object. The proposed Human-Object Offset (HO-offset) is densely sampled from the surface of human mesh and object mesh to construct latent spatial relation space.

2. We propose a novel Stacked Normalizing Flow to infer the posterior distribution of human-object spatial relation for an input image. During inference, a new post-optimization process is designed with relative offset loss to constrain the body pose of the human and the 6D pose of the object.

3. Our method outperforms the previous SOTA method with 16% relative accuracy improvement and 88% relative optimization time reduction.

2 Related Works

Monocular 3D human-object reconstruction. Although there are extensive works in 3D human mesh recovery [Kanazawa *et al.*, 2018; Kolotouros *et al.*, 2019; Lin *et al.*, 2021; Liang *et al.*, 2023; Zhang *et al.*, 2023b] and 6D object pose estimation [Kehl *et al.*, 2017; Li *et al.*, 2019; Chen *et al.*, 2022a], reconstructing human and object jointly is yet a newly proposed problem. 3D human-object reconstruction can be divided into various settings, we only focus on reconstructing 3D human-object from a single-view RGB image. Towards reconstructing and understanding human activity in 3D scenes, [Chen *et al.*, 2019] presents the 3D holistic scene understanding task, which combines 3D scene reconstruction and 3D human pose estimation. Physical commonsense about human-object interaction is utilized to improve the performance of these two tasks. Its follow-up work [Weng and Yeung, 2021] extends this direction to holistic human-object mesh reconstruction. They present an end-to-end trainable model that reconstructs both the human body and object meshes from a single RGB image. In the other direction, the scale of observed objects is zoomed in from the global human-scene to local human-object pairs. [Zhang *et al.*, 2020a] tackles the problem of reconstructing human-object spatial arrangement in the wild. They propose an optimization-based framework which incorporates predefined 3D commonsense constraints to reduce the likely 3D spatial layout between the human and the object. [Xie *et al.*, 2022] presents a unified data-fitting model that learns human-object spatial configuration priors from dataset [Bhatnagar *et al.*, 2022] which is collected using multi-view capture systems. More recently, [Xie *et al.*, 2023] tackles the challenge of single-view human-object tracking under heavy occlusion. Our work focuses on the second direction.

Spatial relationship modeling. Modeling and capturing human-object spatial relationships are inescapable topics throughout various human-object interaction tasks. In the 2D human-object interaction detection task, spatial relation is encoded using relative 2D coordinates between the human bounding box and the object bounding box [Gkioxari *et al.*, 2018], 2-channels mask map of human and object [Ulutan *et al.*, 2020], relative locations between human parts and the center point of the object [Wan *et al.*, 2019] and two-direction spatial distributions between human body parts and object parts [Liu and Tan, 2022]. Similarly, in 3D human-object reconstruction tasks it can be encoded using 3D space positions of human center and object center [Chen *et al.*, 2019] or 3D relative spatial positions and orientations between object and person parts [Savva *et al.*, 2016]. But recently, more works

admit that the contact map is a more fine-grained way to describe how humans and objects interact. [Zhang *et al.*, 2020a] uses commonsense knowledge to define which parts in human and object mesh are participated in the interaction. [Xie *et al.*, 2022] utilizes contact loss between human and object to get a more physically plausible and accurate reconstruction. This idea is also applied in human-scene interaction [Huang *et al.*, 2022a]. Another popular way to model spatial relationships is using an implicit relative distance field. [Karunratanakul *et al.*, 2020] proposes the grasping field that is a continuous function mapping any points in 3D space to two point-to-surface signed distances. Towards hand-object reconstruction, they use a variational encoder-decoder network to learn it from data. This similar idea is also applied to human-scene interaction [Zhang *et al.*, 2020b] and human-object interaction [Xie *et al.*, 2022]. Different from these works, we encode spatial relations using offset vectors between anchors densely sampled from the surface of human mesh and object mesh, while previous works just use the coarse relative distance between human parts and the center of the object or just focus on the local regions in contact.

Probabilistic models in 3D reconstruction. Due to the inherent ambiguity of monocular 3D reconstruction, probabilistic models are more appropriate for inferring distribution from partial observation rather than deterministic prediction. [Bui *et al.*, 2020] devises a multi-hypotheses method that continuously models the orientation of camera pose using Bingham distribution and camera position using multivariate Gaussian. [Sengupta *et al.*, 2021] infers multivariate Gaussian distribution of occluded or invisible body from a single image. Except for multivariate Gaussian, normalizing flow is another popular probabilistic model which is proposed in the context of variational inference [Rezende and Mohamed, 2015] and density estimation [Dinh *et al.*, 2017]. In the context of 3D reconstruction, more recent works utilize normalizing flow for human pose estimation [Wandt *et al.*, 2022], human mesh recovery [Kolotouros *et al.*, 2021], two-hand reconstruction [Wang *et al.*, 2022], conditioned human pose generation [Aliakbarian *et al.*, 2022] and human motion synthesis [Henter *et al.*, 2020]. Following these previous works, we deploy normalizing flow to learn the distribution of potential spatial arrangement between the human and the object from monocular images.

3 Method

Given an input image and target object category, we aim at predicting the SMPL parameters including person shape $\boldsymbol{\beta}$, person pose $\boldsymbol{\theta}$ and object 6D pose i.e. rotation matrix $\boldsymbol{R}$ and translation $\mathbf{t}$. Since predicting these parameters in isolation will produce inconsistent results such as a floating object in the air or interpolation between the human and the object, we propose to use directed offset to place constraints on the body pose of the person and the relative position of the object in 3D space. As shown in Figure 2, our method can be divided into three steps: 1) human-object spatial relation encoding, 2) posterior distribution inference, and 3) post-optimization. In the first step, we construct latent spatial relation space to get a vectorized representation for human-object spatial relation,

which is shown in Sec. 3.1. In Sec. 3.2, we present how to infer a coarse distribution for all possible 3D human-object relative arrangements by using normalizing flow. During the optimization stage, we attempt to get a harmonious result that is both aligned well with the image by minimizing the 2D-3D reprojection loss and coherent with posterior knowledge by maximizing the likelihood of potential spatial relation. The details for this optimization process are shown in Sec. 3.3.

3.1 Spatial Relation Encoding with Human-Object Offset

Human-object interaction instance. To study how human interacts with object in 3D space, we consider the human and the object as a whole and treat this human-object pair as the minimal atomic unit, which is named as human-object interaction instance (HOI instance). For a given human-object pair, there is a trivial way to model it using three components: 1) human mesh modeled by a parametric human body model SMPL [Loper *et al.*, 2015] which defines a mapping $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ from pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$ to a body mesh $\boldsymbol{M}_{\mathrm{SMPL}} \in \mathbb{R}^{6890 \times 3}$, 2) a pre-scanned object mesh template $\boldsymbol{M}_{\mathrm{object}}$ for target object category, 3) spatial arrangement which is parameterized by the relative translation $\mathbf{t}$ and rotation $\boldsymbol{R}$ of object mesh with respect to the root joint of SMPL. We assume the SMPL is rooted at the origin with zero translation and identity rotation since we actually do not care about the global orientation and translation of SMPL mesh in the context of human-object spatial relation encoding. In this representation, an HOI instance is parameterized using human shape $\boldsymbol{\beta}$, human pose $\boldsymbol{\theta}$, object relative translation $\mathbf{t}$ and object relative rotation $\boldsymbol{R}$. Since the human and the object are treated separately, the relation between the human and the object in 3D space cannot be captured clearly using only relative translation and rotation between human mesh and object mesh. Based on this observation, we propose to use the dense offsets between anchors in human mesh and object mesh to capture a highly detailed correlation between human parts and object parts.

Human-object offset vector. Since humans can interact with objects in different ways, there will be quite diverse human-object spatial relation patterns. To cover all possible interaction types, we design a simple but general way to encode this. First, we randomly sample m points from the surface of a human mesh to form human anchor set $\mathcal{A}_{\mathrm{SMPL}}$ and n points from the surface of object mesh to form object anchor set $\mathcal{A}_{\mathrm{object}}$. These anchors are sampled only once and we keep them fixed across all human-object interaction instances. Given a human-object interaction instance, the offsets between human anchors and object anchors are calculated by

$$\mathbf{d}_{i,j} = \mathbf{p}_j^{\mathrm{o}} - \mathbf{p}_i^{\mathrm{h}}, \mathbf{p}_j^{\mathrm{o}} \in \mathcal{A}_{\mathrm{object}}, \mathbf{p}_i^{\mathrm{h}} \in \mathcal{A}_{\mathrm{SMPL}}, \qquad (1)$$

where $\mathbf{p}_i^{\mathrm{h}}$ is i-th anchor in SMPL mesh M_{SMPL} and $\mathbf{p}_j^{\mathrm{o}}$ is the j-th anchor in object mesh M_{object}. We connect all anchors in human anchor set with all anchors in object anchor set to get $m \times n$ offsets. These offsets are concatenated together to form human-object offset vector $\mathbf{x} = (\mathbf{d}_{i,j}) \in \mathbb{R}^{3mn}$. This spatial relationship between the human and the object is encoded within the offset vector.
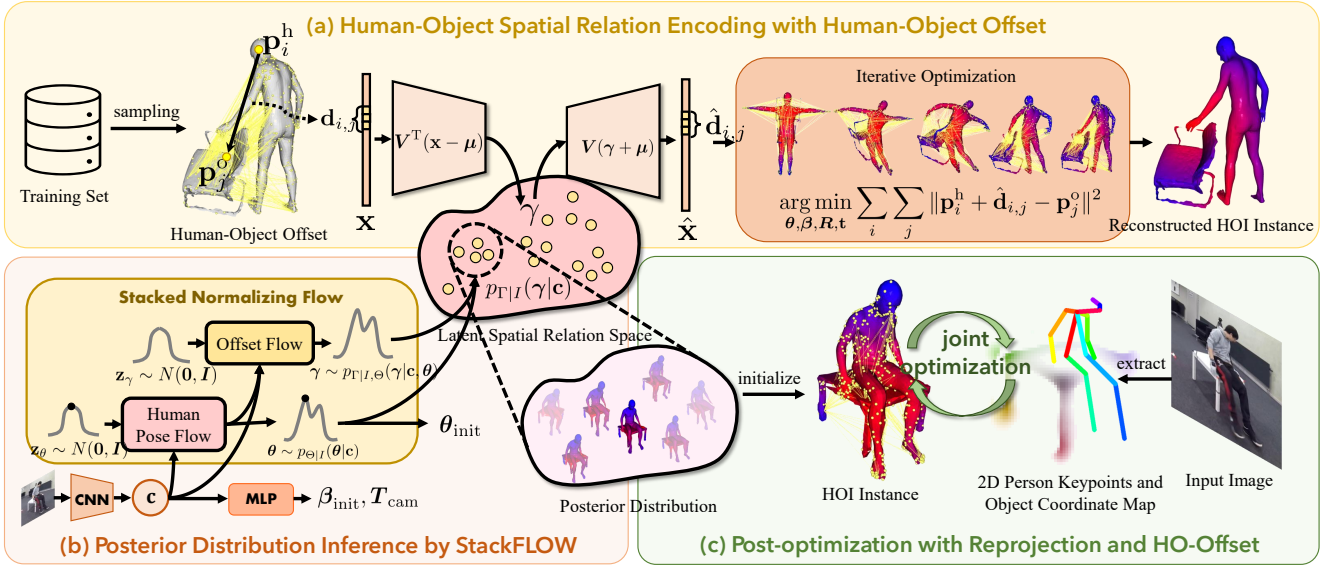
Figure 2: Main framework for our method. (a) We use human-object offset to encode the spatial relation between the human and the object. For a human-object pair, offsets are calculated and flattened into an offset vector $\mathbf{x}$. Based on all offset vectors calculated from training set, the latent spatial relation space is constructed using principle component analysis. To get a vectorized representation for human-object spatial relation, the offset vector is projected into this latent spatial relation space by linear projection. Inversely, given a sample γ from this latent spatial relation space, we can reproject it to recover offset vector $\hat{\mathbf{x}}$. The human-object instance can be reconstructed from $\hat{\mathbf{x}}$ by iterative optimization. (b) With pre-constructed latent spatial relation space, we use stacked normalizing flow to infer the posteriori distribution of human-object spatial relation for an input image. The details are shown in Sec. 3.2. (c) In post-optimization stage, we further finetune the reconstruction results using 2D-3D reprojection loss and offset loss which is illustrated in Sec. 3.3.

Latent spatial relation space construction. To obtain a more compact representation for human-object spatial relation, auto-encoder [Hinton and Salakhutdinov, 2006] has been used. However, principal components analysis (PCA) [Wold *et al.*, 1987] is more adequate in some cases due to its simplicity and efficiency. Based on these considerations, we use PCA to construct latent spatial relation space. We first collect all human-object instances from the training dataset and calculate offsets between anchors using Eq. (1). For each HOI instance, offsets are concatenated together to form an offset vector $\mathbf{x}$. If there are t HOI instances in the training dataset, we will get a matrix $\boldsymbol{X} \in \mathbb{R}^{t \times 3mn}$. PCA is then applied to this matrix to extract the top k component vectors which are mutually orthogonal. These main component vectors form the basis for the latent spatial relation space. Given $\mathbf{x}$, we can project it onto this latent space by linear projection, i.e.

$$\boldsymbol{\gamma} = \boldsymbol{V}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}), \tag{2}$$

where $\boldsymbol{V} \in \mathbb{R}^{3mn \times k}$ is the projection matrix composed by these component vectors, $\boldsymbol{\mu}$ is the mean vector for offset vector $\mathbf{x}$, and γ is a latent vector in this latent spatial relation space. Inversely, we can reproject an arbitrary sample $\boldsymbol{\gamma}$ from latent space $\mathbb{R}^k$ to offset space $\mathbb{R}^{3mn}$ as follows,

$$\hat{\mathbf{x}} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{\mu}. \tag{3}$$

By constructing latent spatial relation space in this way, compactness and continuity can be satisfied because of linear projection. Another benefit is that the latent space can be constructed efficiently using PCA technique and there is no need to train a complex neural network.

Recover HOI instance from HO-offset. An important characteristic of a good representation is that original information can be recovered from it. Human-object offset vector encodes not only spatial arrangement between the human and the object, but also provides constraints on human pose which indicates that we can recover human body pose and object 6D pose from dense human-object offsets by controlling the position of anchors in the surface of the human and the object mesh. Given an arbitrary sample from latent spatial relation space, the offset vectors $\hat{\mathbf{x}}$ can be recovered according to Eq. (3). Variables $\{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{R}, \mathbf{t}\}$ are calculated from this offset vector approximately by minimizing the ℓ_2 norm between target offset $\hat{\mathbf{d}}_{i,j}$ and actual offset from the i-th anchor of human $p_i^{\mathrm{h}}(\boldsymbol{\theta}, \boldsymbol{\beta})$ to the j-th anchor object $p_j^{\mathrm{o}}(\boldsymbol{R}, \mathbf{t})$, i.e.

$$\mathcal{L}_{\text{HO-offset}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R}, \mathbf{t}) = \sum_i \sum_j \|p_i^{\mathrm{h}}(\boldsymbol{\theta}, \boldsymbol{\beta}) + \hat{\mathbf{d}}_{i,j} - p_j^{\mathrm{o}}(\boldsymbol{R}, \mathbf{t})\|^2.$$
$$\tag{4}$$

Note that the positions of human anchor points are controlled by human shape $\boldsymbol{\beta}$ and human pose $\boldsymbol{\theta}$, since they are sampled from SMPL mesh and the position of object anchor points are calculated by:

$$p_j^{\mathrm{o}}(\boldsymbol{R}, \mathbf{t}) = \boldsymbol{R}\hat{\mathbf{p}}_j^{\mathrm{o}} + \mathbf{t}, \tag{5}$$

where $\hat{\mathbf{p}}_j^o$ is the j-th anchor points in object template mesh. Directly optimizing Eq. (4) needs many iterations and may be stuck in local minimum points. The optimization steps can be greatly reduced if we initialize $\{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{R}, \mathbf{t}\}$ properly. We first use the neural network to predict human shape $\boldsymbol{\beta}_{\text{init}}$ and

human pose $\boldsymbol{\theta}_{\text{init}}$ as described in Sec. 3.2 and then substitute β_{init} and $\boldsymbol{\theta}_{\text{init}}$ into (4) to obtain the initial value of $\boldsymbol{R}_{\text{init}}$ and $\mathbf{t}_{\text{init}}$, i.e., solving the following optimization problem:

$$\boldsymbol{R}_{\text{init}}, \mathbf{t}_{\text{init}} = \arg\min_{\boldsymbol{R}, \mathbf{t}} \mathcal{L}_{\text{HO-offset}}(\boldsymbol{\theta}_{\text{init}}, \beta_{\text{init}}, \boldsymbol{R}, \mathbf{t}). \quad (6)$$

Note that Eq. (6) admits a closed-form solution as described in [Choy *et al.*, 2020].

3.2 Posterior Distribution Inference by Stacked Normalizing Flow

Given an image I, we attempt to recover the spatial arrangement of the human-object pair which is encoded using HO-offset. Reconstructing human-object interaction instances from a single-view image is ambiguous due to self-occlusion and mutual-occlusion. Instead of regressing latent human-object spatial arrangement features from the image directly, we follow previous work [Kolotouros *et al.*, 2021] to model it as probabilistic distribution inference. This distribution inference process requires us to predict the conditional probability $p_{\Gamma|I}(\boldsymbol{\gamma}|\mathbf{c})$ using a bijective function f_{offset}, which transforms a random variable $\mathbf{z}_\gamma$ sampled from normal distribution to latent spatial relation feature $\boldsymbol{\gamma}$ with the input image I as condition, i.e.

$$\boldsymbol{\gamma} = f_{\text{offset}}(\mathbf{z}_\gamma|\mathbf{c}), \quad (7)$$

where $\mathbf{c}$ is visual feature extracted from input image I using CNN encoder. However, we find that it is not easy to learn this distribution from images directly in practice. To ease the training process, we decouple it into two stacked conditional probabilities:

$$p_{\Gamma|I}(\boldsymbol{\gamma}|\mathbf{c}) = \int_{\boldsymbol{\theta}} p_{\Gamma|I,\Theta}(\boldsymbol{\gamma}|\mathbf{c}, \boldsymbol{\theta}) p_{\Theta|I}(\boldsymbol{\theta}|\mathbf{c}) \mathrm{d}\boldsymbol{\theta}. \quad (8)$$

We model it using two different flows: (1) human pose flow conditioned on the input image, (2) offset flow conditioned on the human pose and input image, i.e.

$$\boldsymbol{\theta} = f_{\text{SMPL}}(\mathbf{z}_\theta|\mathbf{c}), \mathbf{z}_\theta \sim N(0, I), \quad (9)$$

and

$$\boldsymbol{\gamma} = f_{\text{offset}}(\mathbf{z}_\gamma|\mathbf{c}, \boldsymbol{\theta}), \mathbf{z}_\gamma \sim N(0, I). \quad (10)$$

The structure of these stacked normalizing flows is depicted in Figure 2(b). Given an input image, the CNN is used to extract visual feature $\mathbf{c}$ from the image I. The initial human shape $\boldsymbol{\beta}$ and the translation of camera $\boldsymbol{T}_{\text{cam}}$ are predicted from $\mathbf{c}$. To infer the posterior distribution of $\boldsymbol{\gamma}$ after observing image I, StackFLOW is employed. StackFLOW contains two normalizing flows: human pose flow and offset flow. As formulated in Eq. (9) and Eq. (10), the human pose flow takes visual feature $\mathbf{c}$ as condition to transform a random variable $\mathbf{z}_\theta$ sampled from normal distribution to human pose distribution $p_{\Theta|I}(\boldsymbol{\theta}|\mathbf{c})$. We take $\boldsymbol{\theta}_{\text{init}} = \arg\max_{\boldsymbol{\theta}} p_{\Theta|I}(\boldsymbol{\theta}|\mathbf{c})$ as initial value for human pose. Human pose $\boldsymbol{\theta}$ is combined with visual feature $\mathbf{c}$ as the conditions for offset flow to transform random variable $\mathbf{z}_\gamma$ to distribution $p_{\Gamma|I,\Theta}(\boldsymbol{\gamma}|\mathbf{c}, \boldsymbol{\theta})$. Combining these two distributions, we can get the posterior distribution $p_{\Gamma|I}(\boldsymbol{\gamma}|\mathbf{c})$ according to Eq. (8).

To train these two normalizing flows, we optimize the network by minimizing the negative log-likelihood of ground-truth $\boldsymbol{\theta}_{\text{gt}}$ and $\boldsymbol{\gamma}_{\text{gt}}$, i.e. the loss function is

$$\mathcal{L}_{\text{NLL}} = -\ln p_{\Gamma|I;\Theta}(\boldsymbol{\gamma}_{\text{gt}}|\mathbf{c}, \boldsymbol{\theta}_{\text{gt}}) - \ln p_{\Theta|I}(\boldsymbol{\theta}_{\text{gt}}|\mathbf{c}). \quad (11)$$

In addition to the loss $\mathcal{L}_{\text{SMPL}}$ for supervising SMPL parameters shown in [Kolotouros *et al.*, 2021], we introduce another loss for spatial relation feature $\boldsymbol{\gamma}$:

$$\mathcal{L}_\gamma = \lambda_{exp} \mathbb{E}_{\boldsymbol{\gamma} \sim p_{\Gamma|I}}[\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_{\text{gt}}\|_1] + \|\boldsymbol{\gamma}^\star - \boldsymbol{\gamma}_{\text{gt}}\|_1, \quad (12)$$

where $\boldsymbol{\gamma}^\star = \arg\max_{\boldsymbol{\gamma}} p_{\Gamma|I}(\boldsymbol{\gamma}|\boldsymbol{c})$. The total training loss is

$$\mathcal{L}_{\text{train}} = \lambda_{\text{SMPL}} \mathcal{L}_{\text{SMPL}} + \lambda_{\text{NLL}} \mathcal{L}_{\text{NLL}} + \lambda_\gamma \mathcal{L}_\gamma. \quad (13)$$

3.3 Joint Optimization with Reprojection and Human-Object Offset

During inference, we begin with $\mathbf{z}_\theta = \mathbf{0}, \mathbf{z}_\gamma = \mathbf{0}$ to get initial human pose $\boldsymbol{\theta}_{\text{init}}$ and latent relation feature $\boldsymbol{\gamma}^\star$. We then use Eq. (3) to project latent spatial relation feature $\boldsymbol{\gamma}^\star$ back to offset vector $\mathbf{x}$. The offset $\mathbf{d}_{i,j}$ can be obtained from $\mathbf{x}$ by taking corresponding elements. Finally, we obtain the initial prediction $\{\boldsymbol{\theta}_{\text{init}}, \beta_{\text{init}}, \boldsymbol{R}_{\text{init}}, \mathbf{t}_{\text{init}}\}$ from Eq. (6). This initial prediction is based on the distribution with the most likelihood. To make results aligned well with the input image, we need to finetune results with 2D-3D reprojection loss.

Let $\mathbf{J}^{\text{3D}} \in \mathbb{R}^{K \times 3}$ be the 3D joints of human body and $\hat{\mathbf{J}}^{\text{2D}} \in \mathbb{R}^{K \times 2}$ be 2D locations of corresponding joints which are extracted using OpenPose [Cao *et al.*, 2019], then the 2D-3D reprojection loss for human is defined as

$$\mathcal{L}_{\text{J}} = \sum_{i=1}^{K} \|\Pi(\mathbf{J}_i^{\text{3D}}) - \hat{\mathbf{J}}_i^{\text{2D}}\|_1, \quad (14)$$

where $\Pi : \mathbb{R}^3 \to \mathbb{R}^2$ is the camera projection function. As for object, we use EPro-PnP [Chen *et al.*, 2022b] to get 3D object coordinates $\mathbf{x}^{\text{3D}} \in \mathbb{R}^{N \times 3}$, 2D image coordinates $\mathbf{x}^{\text{2D}} \in \mathbb{R}^{N \times 2}$ and 2D weights $\mathbf{w}^{\text{2D}} \in \mathbb{R}_+^{N \times 2}$, then the 2D-3D reprojection loss for object is defined as

$$\mathcal{L}_{\text{coor}} = \sum_{i=1}^{N} \|\mathbf{w}_i^{\text{2D}} \circ (\Pi(\mathbf{R}\mathbf{x}_i^{\text{3D}} + \mathbf{t}) - \mathbf{x}_i^{\text{2D}})\|_1. \quad (15)$$

We also place constraints on human 3D body pose by maximizing the posteriori probabilities of the human body pose $\mathcal{L}_{\text{posteriori}}^\theta = -\|\mathbf{z}_\theta\|^2$. The loss for 2D-3D reprojection loss is defined as

$$\mathcal{L}_{\text{2D-3D}} = \lambda_{\text{J}} \mathcal{L}_{\text{J}} + \lambda_{\text{coor}} \mathcal{L}_{\text{coor}} + \lambda_{\text{posteriori}}^\theta \mathcal{L}_{\text{posteriori}}^\theta. \quad (16)$$

The 2D-3D reprojection loss aims at aligning results with image content without considering the interaction between the human and the object. To restrict the relative offset between the human and the object, we add the offset loss $\mathcal{L}_{\text{HO-offset}}$ shown in Eq. (4) and posteriori distribution loss $\mathcal{L}_{\text{posteriori}}^\gamma = -\|\mathbf{z}_\gamma\|^2$, which form the loss for human-object spatial relation

$$\mathcal{L}_{\text{offset}} = \lambda_{\text{HO-offset}} \mathcal{L}_{\text{HO-offset}} + \lambda_{\text{posteriori}}^\gamma \mathcal{L}_{\text{posteriori}}^\gamma. \quad (17)$$

Finally, optimization loss is defined as

$$\mathcal{L}_{\text{optim}} = \mathcal{L}_{\text{2D-3D}} + \mathcal{L}_{\text{offset}}. \quad (18)$$
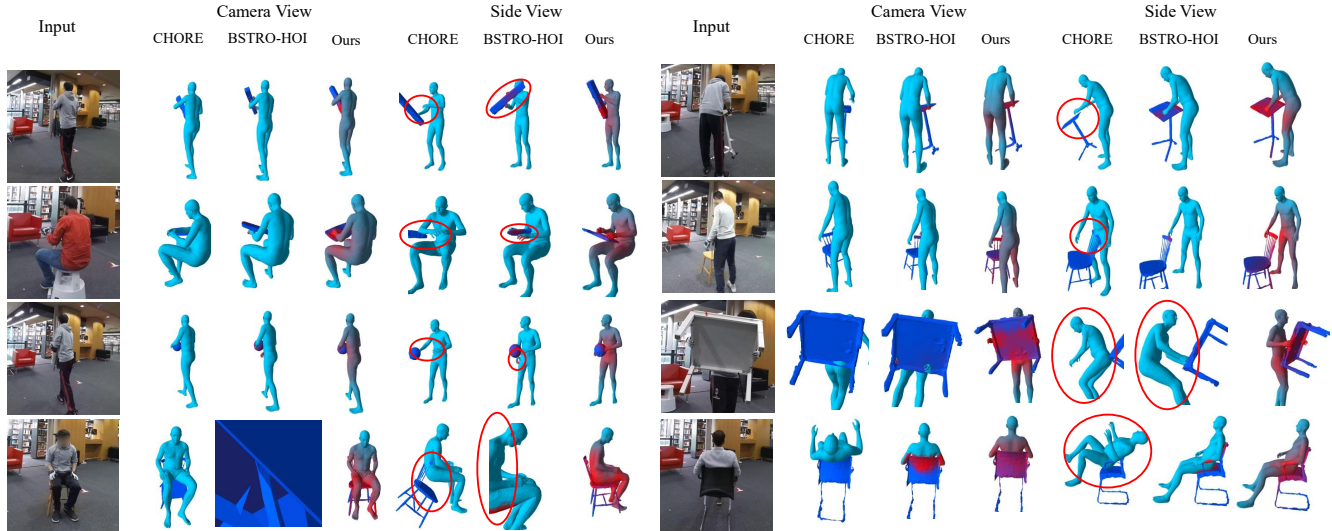
Figure 3: Visualized reconstruction results on BEHAVE dataset. The red regions depict the contact region in BSTRO or the relative distance of our method. The red circles mark the incorrect reconstruction results. These results show that our method performs well in some heavy occlusion cases.

4 Experiments

Dataset. We conduct experiments on two indoor datasets: BEHAVE [Bhatnagar *et al.*, 2022] and InterCap [Huang *et al.*, 2022b]. BEHAVE is a recently released dataset that captures 8 subjects interacting with 20 different objects indoors using a multi-view camera capture system. We follow the official train/test split to train and test our method. Due to the cost of collecting annotations, BEHAVE doesn't provide enough training data which will easily cause the overfitting problem. To prevent this, we render fake images with new viewpoints and new subjects to augment the original training dataset. InterCap is a larger dataset which contains 4M images of 10 subjects interacting with 10 objects. We randomly select 20% sequences for testing and the rest for training which results in 326, 955 images in the training split and 73, 541 images in the testing split.

Free-viewport augmentation. We apply free-viewport data augmentation to generate new images. For each HOI instance sampled from the training dataset, we first use MetaAvatar [Wang *et al.*, 2021] trained on CAPE dataset [Ma *et al.*, 2020; Pons-Moll *et al.*, 2017] to generate clothed human mesh given human pose θ and place it with object mesh template, which is transformed by R and t, in world coordinate. We then render new images by changing the viewport of the camera to simulate all possible occlusions between the human and the object. In the end, we render 12 images with different viewports for each HOI instance in the training dataset. These rendered fake images are used as a supplementary dataset to train our model.

Evaluation metric. Following previous works [Bhatnagar *et al.*, 2022], we use Chamfer distance to evaluate the quality of the reconstructed mesh. For a fair comparison, we assume the object label and bounding box are known before, what we need to predict are the SMPL parameters and the object's 6D

pose. With the reconstructed SMPL mesh and object mesh, we first align them with ground truth meshes using Procrustes analysis, then Chamfer distance is calculated based on the point clouds sampled from reconstructed meshes and ground truth meshes.

4.1 Comparisons with the State-Of-The-Arts

We compare our method with three state-of-the-art methods: PHOSA [Zhang *et al.*, 2020a], CHORE [Xie *et al.*, 2022] and BSTRO [Huang *et al.*, 2022a] on BEHAVE and Inter-Cap dataset. PHOSA is an optimization-based framework that targets at reconstructing human-object spatial arrangement from image in the wild. CHORE is a learning-based method that learns to jointly reconstruct the human and the object from a single RGB image. BSTRO is a powerful model which predicts human-scene contact from a single image. To compare with contact-based models, we adapt it to the task of human-object reconstruction. We name this baseline as BSTRO-HOI. More details about BSTRO-HOI can be found in supplementary Materials.

Method	BEHAVE		InterCap	
	SMPL ↓	Object ↓	SMPL ↓	Object ↓
PHOSA	12.17 ± 11.13	26.62 ± 21.87	6.06 ± 11.13	14.81 ± 11.96
CHORE	5.58 ± 2.11	10.66 ± 7.71	6.86 ± 2.45	15.49 ± 10.13
BSTRO-HOI	4.77 ± 2.46	11.08 ± 13.14	4.80 ± 2.82	9.70 ± 11.05
Ours	4.61 ± 2.04	9.86 ± 9.59	**4.42 ± 1.85**	**8.04 ± 7.37**
Ours†	**4.33 ± 1.83**	**8.87 ± 8.76**	-	-

Table 1: Comparison of mean and standard deviation of Chamfer distance (cm) over all HOI instances on BEHAVE and InterCap datasets. † indicates the model is trained with augmented dataset. **Blod** indicates the best result.

Quantitative evaluation. As shown in Table 1, we compare our method with baseline methods on BEHAVE dataset and InterCap dataset. Our method achieves competitive results compared with state-of-the-art methods. Compared with pure optimization-based method PHOSA, all learning-based methods show incomparable advantages. Compared with other learning-based methods, our method achieves more accurate results, which indicates that human-object offset is a more suitable representation to encode human-object spatial relation.

Qualitative evaluation. We also compare our method against CHORE and BSTRO-HOI qualitatively for heavy occlusion cases in Figure 3. From these cases, we can see that when objects are heavily occluded by human or some human body parts are heavily occluded by object, our method can still draw hints from visible human body parts or objects to guess the potential position of object or potential human body pose by means of HO-offset. As BSTRO-HOI depends on good initilization of human pose and object pose, it fails on the cases in which the object or the human is almost unseen. CHORE also has the same problem. On the contrary, our method is more robust on these heavy occlusion cases.

Method complexity comparison. We compare different methods in terms of space efficiency and time efficiency in Table 2. Our method makes a good balance between space complexity and computation complexity. It is noteworthy that our method outperforms CHORE from 7.90 to 6.60 (with 16% improvement) in terms of reconstruction accuracy with a dramatic reduction from 366.04 to 43.39 (with 88% reduction) during the optimization stage. This dramatic reduction of time consumed in the post-optimization stage benefits from two aspects. First, before post-optimization, we have already got a good initialization which is predicted by StackFLOW, only a few iterations are needed to get the optimal results. The other factor that contributes to dramatic time reduction is the simplicity and efficiency of our optimization loss terms. On the contrary, CHORE relies on multi-stage optimization and complex losses for CHORE field fitting to get accurate reconstruction results.

Method	#Params (M)	GFLOPs	Time (s)	Chamfer Dist.
PHOSA	-	-	14.23	19.40
CHORE	**18.19**	396.39	366.04	7.90
BSTRO-HOI	146.99	40.20	18.90	7.40
Ours (w/o optim.)	77.02	**5.50**	**1.15**	9.34
Ours (w optim.)	77.02	**5.50**	43.39	**6.60**

Table 2: Time and space complexity comparisons on BEHAVE dataset. The second and third columns compare the size and computation of neural network during inference. The fourth column compares the time spent processing each image. The time is tested on a single NVIDIA GeForce RTX 2080 Ti GPU. The last column compares the reconstruction error of different methods. The chamfer distance is averaged between SMPL and object.

4.2 Ablation Study

Effectiveness of offset loss. To demonstrate the effectiveness of offset loss in the stage of post-optimization, we report the results with and without the offset loss in Table 3. Without any optimization, our method can already achieve comparable performance. If we optimize only with reprojection loss, the accuracy of reconstruction becomes worst due to the incorrect of coordinate map predicted by EPro-PnP [Chen *et al.*, 2022b]. Only if we jointly optimize with offset loss and 2D-3D reprojection loss, the best performance can be achieved.

		BEHAVE		InterCap	
$\mathcal{L}_{offset}$	$\mathcal{L}_{\text{2D-3D}}$	SMPL $\downarrow$	Object $\downarrow$	SMPL $\downarrow$	Object $\downarrow$
		4.83 ± 2.06	13.85 ± 11.88	4.96 ± 2.26	11.53 ± 10.56
$\checkmark$		5.68 ± 2.25	13.85 ± 12.17	5.75 ± 2.52	12.25 ± 10.83
	$\checkmark$	4.79 ± 2.44	15.15 ± 18.19	5.71 ± 3.35	17.27 ± 15.30
$\checkmark$	$\checkmark$	$\mathbf{4.33 \pm 1.83}$	$\mathbf{8.87 \pm 8.76}$	$\mathbf{4.42 \pm 1.85}$	$\mathbf{8.04 \pm 7.37}$

Table 3: Effectiveness of different loss during the process of post-optimization. $\mathcal{L}_{offset}$ denotes the loss about offset loss and $\mathcal{L}_{\text{2D-3D}}$ denotes the loss about 2D-3D reprojection in Eq. (18).

Effectiveness of data augmentation. In Table 4, we list the performance of different methods trained with augmented dataset or without augmented dataset. After trained along with our augmented dataset, the performance can be improved across all methods. Whether using generated data or not, our method outperforms other state-of-the-art methods.

Method	Data AUG.	SMPL $\downarrow$	Object $\downarrow$
CHORE	✗	5.58 ± 2.00	10.66 ± 7.71
	$\checkmark$	$\mathbf{5.52 \pm 2.00}$	$\mathbf{10.27 \pm 7.75}$
BSTRO-HOI	✗	4.77 ± 2.46	11.08 ± 13.14
	$\checkmark$	$\mathbf{4.50 \pm 2.28}$	$\mathbf{10.29 \pm 12.09}$
Ours	✗	4.61 ± 2.04	9.86 ± 9.59
	$\checkmark$	$\mathbf{4.33 \pm 1.83}$	$\mathbf{8.87 \pm 8.76}$

Table 4: Ablation studies on BEHAVE dataset for the effectiveness of data augmentation.

5 Conclusion

In this work, we show how to encode and capture highly detailed 3D human-object spatial relations from single-view images using Human-Object Offset. Towards monocular human-object reconstruction, a Stacked Normalizing Flow is proposed to infer the posterior distribution of human-object spatial relation from a single-view image. During the optimization stage, offset loss is proposed to constrain the body pose of humans and the relative 6D pose of objects. Our method outperforms state-of-the-art models on two challenging benchmarks including BEHAVE or InterCap dataset. Especially, our model is good at handling heavy occlusion cases. Even if the objects are heavily occluded by the human, our method can still draw cues from visible human pose to infer the potential pose of the objects.

Acknowledgments

References

[Aliakbarian *et al.*, 2022] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J. Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13253–13262, June 2022.

[Bhatnagar *et al.*, 2022] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15935–15946, June 2022.

[Bui *et al.*, 2020] Mai Bui, Tolga Birdal, Haowen Deng, Shadi Albarqouni, Leonidas Guibas, Slobodan Ilic, and Nassir Navab. 6d camera relocalization in ambiguous scenes via continuous multimodal inference. In *Eur. Conf. Comput. Vis.*, pages 139–157, 2020.

[Cao *et al.*, 2019] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Chen *et al.*, 2019] 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 *Int. Conf. Comput. Vis.*, pages 8648–8657, October 2019.

[Chen *et al.*, 2022a] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2781–2790, June 2022.

[Chen *et al.*, 2022b] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2781–2790, June 2022.

[Choy *et al.*, 2020] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2514–2523, June 2020.

[Dinh *et al.*, 2017] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

[Gkioxari *et al.*, 2018] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8359–8367, June 2018.

[Hassan *et al.*, 2019] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Int. Conf. Comput. Vis.*, pages 2282–2292, October 2019.

[Henter *et al.*, 2020] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics*, 39(4):236:1–236:14, 2020.

[Hinton and Salakhutdinov, 2006] Geoffrey E. Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504 – 507, 2006.

[Huang *et al.*, 2022a] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13274–13285, June 2022.

[Huang *et al.*, 2022b] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *Pattern Recognition*, pages 281–299. Springer International Publishing, 2022.

[Kanazawa *et al.*, 2018] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, June 2018.

[Karunratanakul *et al.*, 2020] 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, 2020.

[Kehl *et al.*, 2017] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Int. Conf. Comput. Vis.*, pages 1521–1529, Oct 2017.

[Kolotouros *et al.*, 2019] 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 *Int. Conf. Comput. Vis.*, pages 2252–2261, October 2019.

[Kolotouros *et al.*, 2021] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Int. Conf. Comput. Vis.*, pages 11605–11614, October 2021.

[Li *et al.*, 2019] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Int. Conf. Comput. Vis.*, pages 7678–7687, October 2019.

[Liang *et al.*, 2023] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. In *AAAI*, February 2023.

[Lin *et al.*, 2021] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1954–1963, June 2021.

[Liu and Tan, 2022] Lu Liu and Robby T. Tan. Human object interaction detection using two-direction spatial enhancement and exclusive object prior. *Pattern Recognition*, 124:108438, 2022.

[Loper *et al.*, 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[Ma *et al.*, 2020] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6469–6478, June 2020.

[Pons-Moll *et al.*, 2017] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017.

[Rezende and Mohamed, 2015] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[Savva *et al.*, 2016] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning Interaction Snapshots from Observations. *ACM Trans. Graph.*, 35(4), 2016.

[Sengupta *et al.*, 2021] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16094–16104, June 2021.

[Sun *et al.*, 2021] 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.

[Ulutan *et al.*, 2020] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13617–13626, June 2020.

[Wan *et al.*, 2019] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Int. Conf. Comput. Vis.*, pages 9469–9478, October 2019.

[Wandt *et al.*, 2022] Bastian Wandt, James J. Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6635–6645, June 2022.

[Wang *et al.*, 2021] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems*, volume 34, pages 2810–2822. Curran Associates, Inc., 2021.

[Wang *et al.*, 2022] Jiayi Wang, Diogo Luvizon, Franziska Mueller, Florian Bernard, Adam Kortylewski, Dan Casas, and Christian Theobalt. HandFlow: Quantifying View-Dependent 3D Ambiguity in Two-Hand Reconstruction with Normalizing Flow. *Vision, Modeling, and Visualization*, 2022.

[Weng and Yeung, 2021] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 334–343, June 2021.

[Wold *et al.*, 1987] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[Xie *et al.*, 2022] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *Eur. Conf. Comput. Vis.*, page 125–145, October 2022.

[Xie *et al.*, 2023] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2023.

[Zhang *et al.*, 2020a] 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 *Eur. Conf. Comput. Vis.*, page 34–51, 2020.

[Zhang *et al.*, 2020b] 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, 2020.

[Zhang *et al.*, 2023a] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2023.

[Zhang *et al.*, 2023b] Juze Zhang, Ye Shi, Lan Xu, Jingyi Yu, and Jingya Wang. Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In *AAAI*, February 2023.