MoCapDeform: Monocular 3D Human Motion Capture in Deformable Scenes

Zhi Li[1,2] Soshi Shimada[1,2] Bernt Schiele[2] Christian Theobalt[2] Vladislav Golyanik[2]

[1]Saarland University, SIC [2]MPI for Informatics, SIC

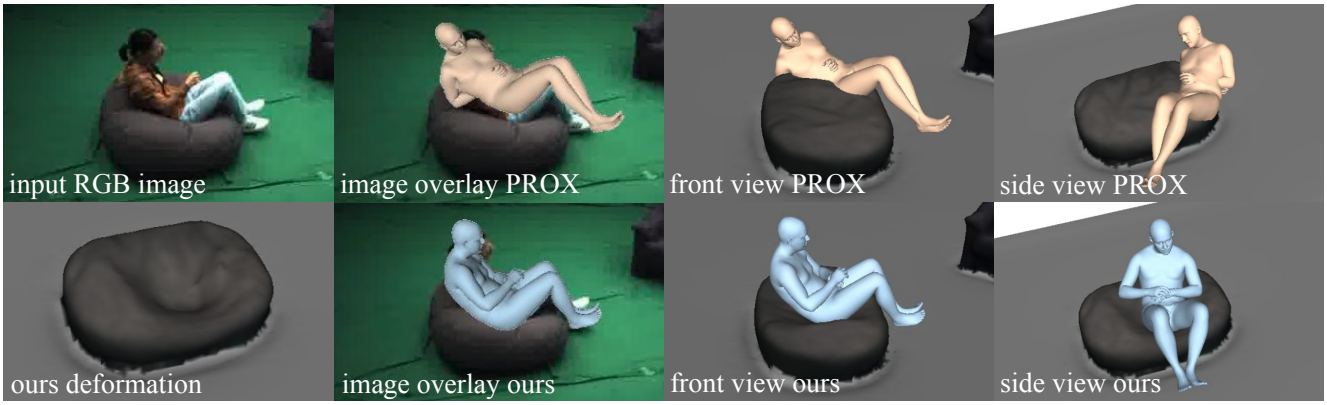


Figure 1: Existing monocular 3D human motion capture methods such as PROX [19] ignore abundant scene deformation when penalising human-scene collisions, resulting in erroneous global poses (top). **Our MoCapDeform algorithm is the first that models non-rigid scene deformations and finds the accurate global 3D poses of the subject by human-deformable scene interaction constraints**, achieving increased accuracy with significantly fewer penetrations (bottom).

Abstract

3D human motion capture from monocular RGB images respecting interactions of a subject with complex and possibly deformable environments is a very challenging, ill-posed and under-explored problem. Existing methods address it only weakly and do not model possible surface deformations often occurring when humans interact with scene surfaces. In contrast, this paper proposes MoCapDeform, i.e., a new framework for monocular 3D human motion capture that is the first to explicitly model non-rigid deformations of a 3D scene for improved 3D human pose estimation and deformable environment reconstruction. MoCapDeform accepts a monocular RGB video and a 3D scene mesh aligned in the camera space. It first localises a subject in the input monocular video along with dense contact labels using a new raycasting based strategy. Next, our human-environment interaction constraints are leveraged to jointly optimise global 3D human poses and non-rigid surface deformations. MoCapDeform achieves superior accuracy than competing methods on several datasets, including our newly recorded one with deforming background scenes.

1. Introduction

3D human motion capture from monocular images is an active research area [27, 40, 55, 30, 44, 22, 6, 49, 7, 56, 1]. Relying solely on monocular RGB inputs is challenging and severely ill-posed, as no explicit 3D cues are provided. As observed in daily life and noticed in the literature [13, 19], the 3D world constrains human actions when they move and interact with it. Thus, environmental constraints and scene priors can provide additional cues for global 3D human motion capture. Leveraging them is a promising route to reduce the ambiguities and, at the same time, infer deformable scene geometry by observing human-scene interactions.

Existing works consider foot-floor interactions to recover physically plausible human motions [49, 43, 48, 42]. Others utilise explicit 3D environment models as constraints [19, 63, 59, 64, 14]. These techniques either do not fully address the scale ambiguity [49, 43, 48, 42, 19, 63, 59] or require RGB-D rather than RGB inputs [64]. Several other methods rely on body-mounted sensors (inertial measurement units) [14, 61] to localise humans globally. Next, virtually all approaches consider the background being static and ignore potential scene changes caused by human-scene interactions. When a subject sits on a couch (lays in a bed), the latter deforms significantly due to its non-rigidity

and forces exerted by the subject. Unfortunately, while existing works [19, 63, 59, 64, 14] use human-environment contact and inter-penetration constraints to avoid collisions, they simply disregard such scene deformations, which results in substantial 3D reconstruction errors.

We argue that 3D scene deformations cannot be ignored if we would like to raise the accuracy of the reconstructed 3D human poses to the next level. Hence, this paper proposes *MoCapDeform*, a new framework for 3D monocular human motion capture with a 3D scene prior (a mesh); see Figs. 1 and 2 for an overview. In contrast to previous works, our method accurately localises the global human position in the scene from RGB inputs. It does so by a raycast contact finding policy to tackle the scale ambiguity problem, and a scene deformation modelling technique to address the limitation of low 3D reconstruction accuracy caused by the negligence of scene deformations.

Our method comprises three stages. In the **first stage**, we initialise the 3D human poses parameterised by the SMPL-X model [38], which can be done by any off-the-shelf 3D human pose estimator. This yields initial root-relative 3D poses that are reasonably accurate and sufficiently satisfy the image observations. Next, contact probability maps are estimated from the initial 3D poses, indicating which vertices on the human mesh are in contact with the 3D environment. The contact points are then re-projected to the image domain and passed through a raycasting operation to find the corresponding contact locations on the scene mesh. In the **second stage**, we register the estimated contact points on the human mesh to the raycasting results, leading to coarse global 3D human poses. **Finally**, these poses are further refined by jointly optimising for pose updates and deformations of the scene with which the body is in contact. In summary, our **contributions** are as follows:

- MoCapDeform—the first framework for joint markerless 3D human motion capture from monocular RGB images and capture of non-rigid 3D scene deformations. Such joint reasoning increases the accuracy of 3D human pose estimation on various benchmarks compared to existing methods (Sec. 4).
- A new raycasting based optimisation algorithm for finding dense contacts between humans and the environment (Sec. 3.3).
- A joint scene deformation and human pose refinement optimisation to recover both accurate human poses and scene deformations (Sec. 3.4).
- A new dataset for the experimental evaluation with human-scene interactions and observable scene deformations (Sec. 4.1).

We compare MoCapDeform to several previous state-of-the-art methods that assume monocular RGB images or videos and pre-scanned scene meshes input [20, 19, 38]. Our approach regresses significantly more accurate global 3D human poses on the PROX [19] and our new datasets, producing reasonable scene deformations (Sec. 4). The new dataset and source code are available at https://github.com/Malefikus/MoCapDeform.

2. Related Work

Kinematic Monocular 3D Human Pose Estimation. Most works on monocular 3D human pose estimation reconstruct human joint positions in local coordinates. Some methods first estimate 2D poses in the 2D image space and then lift them into 3D [5, 27, 55, 35, 10, 6]. Several other approaches learn feature representations for 2D poses (without explicit 2D joint outputs) and perform lifting [30, 39, 18]. Further approaches regress 3D joints directly from RGB images via neural networks [54, 28, 44]. While straightforward, this direct joint position representation has some issues, such as being hard to use in graphics applications, temporal jitter and implausible human skeletons due to varying bone lengths. These limitations can be addressed by estimating parameters of pre-defined body models such as joint angles for kinematic skeletons [68, 30, 29], pose and shape parameters of parametric body models [2, 22, 40, 38, 25, 69], or template-based human performance capture methods [17, 60, 15, 16]. While most works estimate root-relative 3D human poses, several ones attempt to regress 3D human poses with absolute depths in the camera space [34, 46, 29]. Without depth priors, however, it is difficult to obtain accurate and artefact-free human poses in the global reference frame.

3D Human Pose Estimation with Scene Constraints. While significant progress in 3D human pose estimation was made over the last decades [11, 32, 41, 45], utilising scene constraints remains insufficiently explored [19, 43, 64, 42, 7, 47]. Several works consider the ground plane only, and by enforcing volume occupancy exclusions or detecting foot-floor contacts, they impose physical plausibility on the reconstructed motions [62, 46, 43, 49, 42]. Some works consider holistic representations and model complex scenes by placing arranged object meshes (recovered from categorised templates) into the desired coordinate frame [33, 63, 59]. This way, the knowledge about the spatial arrangement can be utilised to coarsely constrain the global position of the human. Another line of work employs pre-scanned meshes of the whole environment [19, 64, 57, 3]. Hassan *et al.* [19] attempt to register empirically assumed contact vertices on the human body to the nearest scene vertices for human-scene collision detection. This approach does not fully resolve the depth ambiguity since the exact contact locations in the scene are still unknown. Zhang *et al.* [64] use RGB-D videos as inputs and focus on the motion plausibility on top of scene constraints.

All in all, the assumption of an available mesh enables accurately locating the human in the global reference frame,

and provides the opportunity to model scene deformations for more accurate scene reconstruction and global human pose recovery. Considering the form of inputs (monocular RGB images + pre-scanned scene meshes) and outputs (global human poses), the PROX approach [19] is most related to our method. Another recent paper proposes HULC, *i.e.,* a framework for 3D human motion capture in complex environments [47]. They estimate dense human-scene contacts with a neural network trained on a new large-scale dataset, and not a learning-free raytracing like we do. In contrast to previous and concurrent works [19, 64, 47], MoCapDeform models scene deformations, which helps to accurately locate humans in the global 3D space.

3. The Proposed Approach

We describe our optimisation framework with three stages; see Fig. 2. The inputs to our framework are RGB images of a static camera and a pre-scanned mesh of the scene at the beginning of capture aligned to the camera coordinate frame; the outputs are posed 3D human bodies and deformed scene meshes. The **first stage** (Sec. 3.2) performs *initial pose estimation*, where we estimate the initial pose from an RGB image, which suffers from scale ambiguity but faithfully overlays onto the images, *i.e.,* the root-relative pose is coarsely accurate, on top of which the contact probability map can be estimated. The **second stage** (Sec. 3.3) is a *global pose optimisation*, in which we utilise the estimated contact points on the human mesh, then cast camera rays through the human contact points to the scene mesh to find the contact points on the environment, and then optimise for the global poses respecting these contacts. Finally, in **stage three** (Sec. 3.4), we perform *joint scene deformation and pose refinement* to obtain accurate global 3D human poses and realistic scene deformations.

3.1. Assumptions and Notations

Our method assumes that a 3D mesh of the scene and its registration in camera space are given. The 3D mesh is represented by $M_s = (\mathbf{V}_s, \mathbf{F}_s)$, with vertices $\mathbf{V}_s \in \mathbb{R}^{N_v \times 3}$ (N_v as the number of vertices) and triangular faces $\mathbf{F}_s \in \mathbb{N}^{N_f \times 3}$ (N_f as the number of faces) containing the vertex indices of each triangle. The static 3D scene mesh can be reconstructed with standard commercial solutions, either leveraging Structure Sensor [53] and the Skanect [51] software [19], or multi-view reconstruction and differentiable rendering techniques from Agisoft Metashape [31]. The reconstructed meshes contain surface normals that correctly indicate the "outside" and the "inside" of the scene. Based on the topology of the pre-defined scene mesh, our method outputs a deformed per-frame scene mesh $M'_s = (\mathbf{V'}_s, \mathbf{F}_s)$ (we omit frame indices in this notation for conciseness).

Following the representation of [19], we adopt a parametric 3D body model SMPL-X [38]. The model is formulated as a differentiable function $M_b(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\gamma})$ parameterised by shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, body, hand and jaw pose $\boldsymbol{\theta} \in \mathbb{R}^{52\times 3}$ of axis-angle representation with three degrees of freedom (DoF) for each joint, holistic facial expression parameters $\boldsymbol{\psi} \in \mathbb{R}^{10}$, and global translation $\boldsymbol{\gamma} \in \mathbb{R}^3$ defining the global position of the pelvis joint. The model output is a 3D human mesh $M_b = (\mathbf{V_b}, \mathbf{F_b})$, with vertices $\mathbf{V_b} \in \mathbb{R}^{10475\times 3}$ and the corresponding triangular faces $\mathbf{F_b} \in \mathbb{N}^{20908\times 3}$ expressing the mesh connectivity. From the mesh vertices, we can regress an underlying rigged skeleton $J(\boldsymbol{\beta})$ with 55 joints defined by linear blend skinning. Following the notation of [2], we denote the posed joints as $R_{\boldsymbol{\theta\gamma}}(J(\boldsymbol{\beta})_i)$ for each joint i, where $R_{\boldsymbol{\theta\gamma}}$ denotes the kinematic tree defined by the pose parameter $\boldsymbol{\theta}$ and translation vector $\boldsymbol{\gamma}$. With this representation, we obtain a globally posed and shaped human body, *i.e.,* both the skinned mesh and the underlying articulation.

3.2. Stage1: Initial Human Pose Estimation

Stage 1 initialises the global human poses from monocular RGB images. This can be done by any of the off-the-shelf 3D human pose estimators [2, 30, 22, 29, 19, 25], which take the RGB images as input, follow closely the image cue and produce the required root-relative poses. As MoCapDeform is not restricted to sequential inputs, we employ a single-frame method for the initialisation. For a fair comparison with the SOTA approaches [19, 20], which are both based on the optimisation-based SMPLify-X [38], we initialise with SMPLify-X. Stage 1 minimises the following objective function:

$$E_1(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = E_J + \lambda_\theta E_\theta + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta. \quad (1)$$

E_J is the RGB data term, *i.e.,* is a re-projection loss seeking to minimise the robust weighted distance between 2D joints—estimated from the RGB image using OpenPose [58, 4, 50]—and the 2D projection of the corresponding posed 3D joints of SMPL-X. We assume a perspective camera model. The re-projections are weighted by the detection confidence scores, and a robust Geman-McClure error function [12] is applied on top to down-weight noisy detections. E_θ is the trained VAE-based body pose prior of VPoser [38], which enforces natural human poses learned from a large 3D human pose corpus [26]. $E_\alpha = \sum_{i\in(\text{elbows,knees})} \exp(\boldsymbol{\theta_i})$ is a prior penalising extreme bending for elbows and knees. E_β is an ℓ_2-regulariser for human shape to penalise deviation from the neutral state. For a more detailed explanation, please refer to [38, 19].

3.3. Stage 2: Global Pose Optimisation

The goal of Stage 2 is to regress an accurate global 3D position of the human body estimated in Stage 1. Since monocular 3D human pose estimation is ill-posed without further priors, it is unlikely to obtain accurate results.

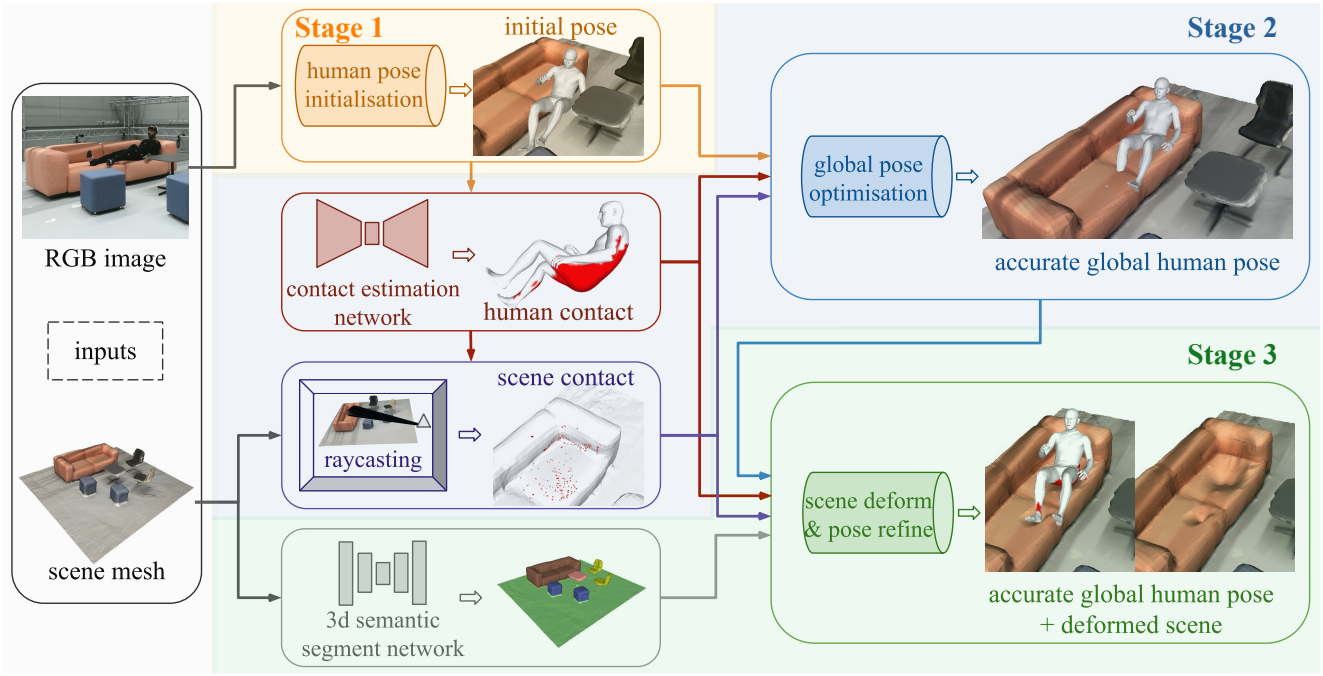


Figure 2: **Overview of MoCapDeform.** We first initialise the human pose and use it to find the contact points on the human mesh. Then, we apply raycasting to find the contact points on the scene mesh surface, which are then used to recover improved global human poses. Finally, we perform joint scene deformation and human pose refinement and obtain accurate global human pose and realistic scene deformations.

Hence, we use the given scene mesh as a constraint to tackle depth ambiguity. We utilise the human-scene contact information by firstly estimating human body contacts on the initialised human body. We then perform raycasting of the re-projected human body contacts into the 3D scene to find the corresponding scene contacts and, finally, register the human body contact points to the scene contacts.

3.3.1 Raycast Contact Estimation

Our raycasting policy has three steps illustrated in Fig. 3: 1) Body-centric contact estimation, 2) Contact re-projection with masking, and 3) Raycast and scene contact estimation.

To find contacts on a scene through our raycast policy, we first need to find contacts on the human body. We thus employ POSA [20], *i.e.,* a conditional variational autoencoder (cVAE) that generates probability maps for different canonical human poses. The learned cVAE decoder takes as input human mesh points (in a root-relative pose and a canonical reference frame) and a latent vector $\boldsymbol{z}\sim\mathcal{N}(0, I)$ that conditions the sampling and can be directly applied to the initial human poses from Stage 1 since the root-relative poses are accurate. Specifically, we canonicalise the initialised human meshes from Stage 1 following the same formulation as in POSA and choose a zero vector $\boldsymbol{z}$ to generate contact probability samples that lie in the peak in the

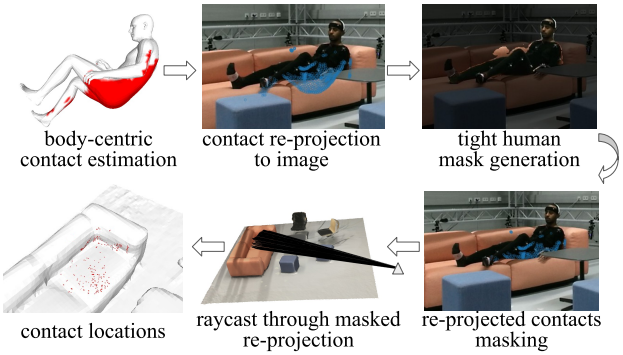


Figure 3: Overview of our raycast contact policy.

learned latent space. Then, by thresholding the generated probability map by 0.5 [20], we obtain the human mesh vertices that are in contact with the environment.

The generated human contacts are then re-projected to the image space. Then, we cast rays from the camera through these re-projections to find intersections with the 3D scene mesh. However, as the scene geometry is complex, there are usually multiple hits along each ray, and the challenge is to find out which hit is at the contact. Intuitively, as illustrated in Fig. 4, when the re-projected contacts fall on the body parts that are not occluded in the image

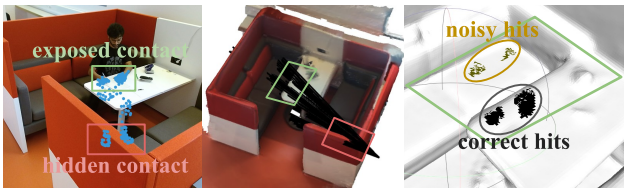


Figure 4: Detection of non-occluded areas and noisy contact label filtering based on the analysis of the ray-mesh hits.

(the green rectangles in Fig. 4), the front-most hits will be at the correct contacts. Otherwise, the rays firstly hit occluders (red rectangles in Fig. 4). To tell apart the seen and occluded body parts, we apply PointRend [24], which finds tight human masks in the images. The masks are then used to segment out the contact re-projections on the occluded body parts. Furthermore, to eliminate the noisy raycast when the contact estimation is slightly off, we apply the DBSCAN clustering method [9] for denoising, as shown in Fig. 4. We empirically set the scanning radius ϵ=0.5 and MinPts=50 of the DBSCAN, to help eliminate the clusters (hit by the ray) that are far away from the main cluster or contain a small number of samples. After these steps, the resulting points are considered the corresponding scene contacts.

3.3.2 Our Objective Function (Stage 2)

With the help of the raycast results, we perform Stage 2 optimisation and register the estimated human contacts to the raycasted scene contacts. This leads to refinement of the initial estimates from Stage 1. We minimise the following objective function:

$$E_2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s) = E_J + \lambda_{\text{obs}} E_{\text{obs}} + \lambda_{\text{un}} E_{\text{un}} + \lambda_{t\theta} E_{t\theta} + \lambda_{t\gamma} E_{t\gamma}, \tag{2}$$

where E_J is as is in (1). E_{obs} is the "seen" contact term, which minimises the distance between estimated contact points on the human mesh and the raycasted contact points on the scene mesh. E_{un} denotes the "unseen" contact term, which registers the rest of the estimated human contacts that do not have raycast hits to the corresponding nearest scene vertices by minimising their Chamfer distance. In E_{obs} and E_{un}, the distances are calculated by a Geman-McClure error function [12] to downweight noisy detections. Finally, we apply temporal smoothness terms $E_{t\theta}$ and $E_{t\gamma}$. $E_{t\theta}$ and $E_{t\gamma}$ are the ℓ_2-distances between the poses $\boldsymbol{\theta}$ and global translations $\boldsymbol{\gamma}$ between frames t and $t-1$, respectively.

3.4. Joint Scene Deformation and Pose Refinement

In Stage 3, the coarse global poses obtained from Stage 2 are further refined, taking into account scene deformations. This stage jointly optimises for plausible 3D scene deformation in interaction regions and more accurate human poses, also resulting in fewer inter-penetrations between the two.

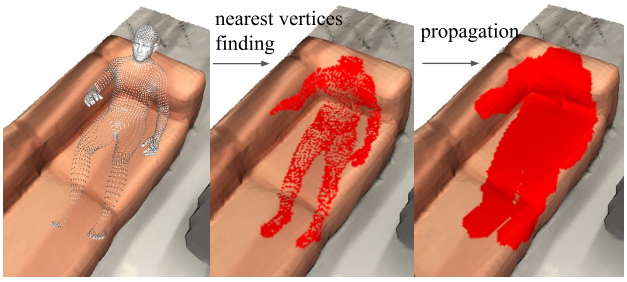


Figure 5: Determination of movable scene points.

3.4.1 Scene Deformation Modelling

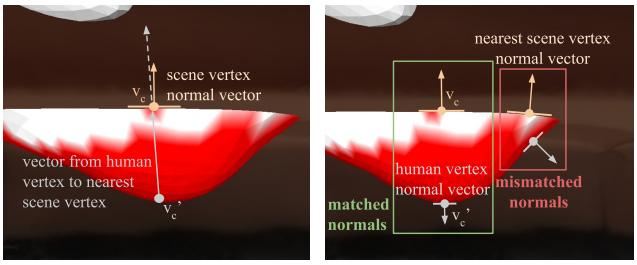
To model deformations of non-rigid objects such as couches or beds, we use an as-rigid-as-possible (ARAP) regulariser [52]. It allows deforming a mesh with guidance by predefined sparse control vertices $c \in \Omega$ (Ω is a subset of contact point indices). The latter are first moved to the target positions, and the neighbouring faces are encouraged to stay rigid as much as possible. The deformed state of the entire surface is then found by optimising

$$E_{\text{ARAP}}(M_s', \mathbf{R}_c) = \sum_{c \in \Omega} \sum_{n \in \mathcal{N}(c)} w_{cn} \left\| (\mathbf{v'}_c - \mathbf{v'}_n) - \mathbf{R}_c(\mathbf{v}_c - \mathbf{v}_n) \right\|, \tag{3}$$

where $\mathbf{R}_c$ is the unknown rotation matrices; $\mathbf{v}_c$ and $\mathbf{v'}_c$ are the control vertex positions before and after the optimisation; $\mathcal{N}_c$ is the set of per-vertex neighbours $\mathbf{v_c}$; $\mathbf{v}_n$ and $\mathbf{v'}_n$ are the neighbouring vertices of $\mathbf{v}_c$ before and after optimisation, and w_{cn} are cotangent weights. To integrate ARAP regulariser in our framework, we define on the scene mesh: 1) A set of control vertices $\mathbf{v_c}$; 2) The corresponding target positions $\mathbf{v'}_c$ for the control vertices; 3) A set of neighbouring vertices $\mathbf{v_n}$, which are allowed to move.

In practice, with the help of the current state of the human body, we first partition the whole scene mesh into movable and static areas and then choose the control points from the movable area and define the target positions accordingly. In the beginning, we need to know which parts of the scene are deformable and which are not. For that purpose, we adopt a 3D scene mesh segmentation network VMNet [21] trained on a large-scale indoor dataset ScanNet [8]. VMNet estimates semantic labels of the furniture. In this paper, we regard "sofa" and "bed" as non-rigid and all other object classes as rigid. After masking out rigid parts, we further define "movable" areas inside the deformable areas. As illustrated in Fig. 5, for every vertex on the human mesh, we find its nearest scene vertex and then propagate the obtained points to their k-th order neighbours. These points are regarded as "movable" during ARAP deformation. Here, we empirically set k=3 according to the scale of the meshes.

Next, we define the control vertices $\mathbf{v}_c$ on the scene mesh and their corresponding target positions $\mathbf{v'}_c$. Assuming that the human induces the scene deformations, $\mathbf{v}_c$ and $\mathbf{v'}_c$ can



(a) collision checking (b) normal matching

Figure 6: Collision checking and normal matching. The viewpoint is "inside" the couch, looking at the colliding hip.

be defined according to the vertices $\mathbf{v}_{c'}$ of the human mesh (and their positions) that satisfy the following conditions:

- $\mathbf{v}_c$ is the nearest scene vertex to the human vertex $\mathbf{v}_{c'}$.
- $\mathbf{v}_{c'}$ are marked as "in contact" by the contact estimation step.
- $\mathbf{v}_{c'}$ collide with the scene mesh. Note that in this step, collision detection cannot be achieved by the pre-computed SDF since the scene is deforming. Hence, we check the collision status with the help of scene surface normals, as shown in Fig. 6a: When the vector from $\mathbf{v}_{c'}$ to $\mathbf{v}_c$ (grey) goes in the same direction as the normal vector of $\mathbf{v}_c$ (orange), $\mathbf{v}_{c'}$ will be classified as colliding with the scene.
- The normals of $\mathbf{v}_{c'}$ should be of opposite directions to the normals of $\mathbf{v}_c$, as is shown in Fig. 6b. This is because the scene usually deforms along the direction of the forces applied by the human.

At last, the nearest scene vertices $\mathbf{v}_c$ to the above human vertices $\mathbf{v}_{c'}$ are chosen as control points, and the control target positions $\mathbf{v'}_c$ are defined by the positions of $\mathbf{v}_{c'}$.

3.4.2 Optimisation (Stage 3)

The final stage is a joint and alternating optimisation for scene deformation and human poses. In every iteration k, we firstly pick the control points of ARAP according to the current human pose using the techniques described in Sec. 3.4.1 and then update the scene mesh using (3). Next, we update the human pose based on the updated scene mesh by minimising the following energy function:

$$E_3(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s^k) = E_2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s^k) + \lambda_{\text{pen}} E_{\text{pen}}. \quad (4)$$

where $E_2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s^k)$ is defined in (2) and E_{pen} is a penetration term that does not use the pre-computed SDF, since the scene mesh is being constantly updated. Instead, it utilises the normal checking technique presented in Fig. 6a to detect collisions and then registers the colliding vertices on the human mesh to their nearest scene mesh vertices by minimising the Geman-McClure error function [12].

3.5. Implementation

In Stage 1, closely following [38, 19], we optimise (1) using a PyTorch [37] implementation of the limited-memory BFGS optimiser [36] with line search satisfying strong Wolfe conditions. (2) in Stage 2 and (4) in Stage 3 are optimised by PyTorch implementations of the Adam optimiser [23]. For the scene deformations, *i.e.,* (3) in Stage 3, we adopt the ARAP implementation from Open3D [67]. The off-the-shelf components we adopt (*i.e.,* PointRend [24], VMNet [21] and POSA human contact estimation [19]) are easily deployable and sufficiently accurate for our task. The λ and w_{cn} weights in (1)-(4) are empirically found and fixed in all experiments; see the source code.

4. Experiments

To evaluate our MoCapDeform framework, we conduct extensive experiments on two datasets (Secs. 4.1-4.2) and show qualitative results (Sec. 4.3).

4.1. Datasets

PROX dataset [19]. The PROX dataset includes a large qualitative and a small quantitative sets. The qualitative set contains monocular videos of 20 human subjects interacting with 12 indoor scenes along with the 3D scene scans: altogether, $100k$ RGB-D frames recorded at 30 fps (without ground-truth 3D human poses). The quantitative set contains 180 static RGB-D frames, with one human subject wearing markers interacting with a mimicked living room containing daily furniture. The pseudo-ground-truth SMPL-X parameters for the quantitative set are fitted by the marker-based MoSh++ [26] method.

MoCapDeform (MCD) dataset. To evaluate all outputs of MoCapDeform, including the deformations, we record a new dataset of people interacting with furniture, *i.e.,* a non-rigid sofa, a deformable stool, and especially a beanbag, which retains its deformed shape after the interaction and allows obtaining ground-truth deformations. We reconstruct accurate human meshes and scene geometry with the multi-view camera setting and a markerless differentiable-rendering-based technique [31]. The human meshes can then be used to fit the SMPL-X model parameters and serve as ground truth. The dataset contains four video sequences at 30 fps of four subjects interacting with the furniture ($16k$ sequential RGB images in total). We utilise the dataset for both quantitative and qualitative experiments.

4.2. Quantitative Evaluation

We evaluate the estimated 3D human poses by computing several quantitative metrics indicating the global and local 3D reconstruction accuracy and the degree of penetrations. For global 3D reconstruction accuracy, we adopt the standard **PJE** and **V2V** metrics and report them in

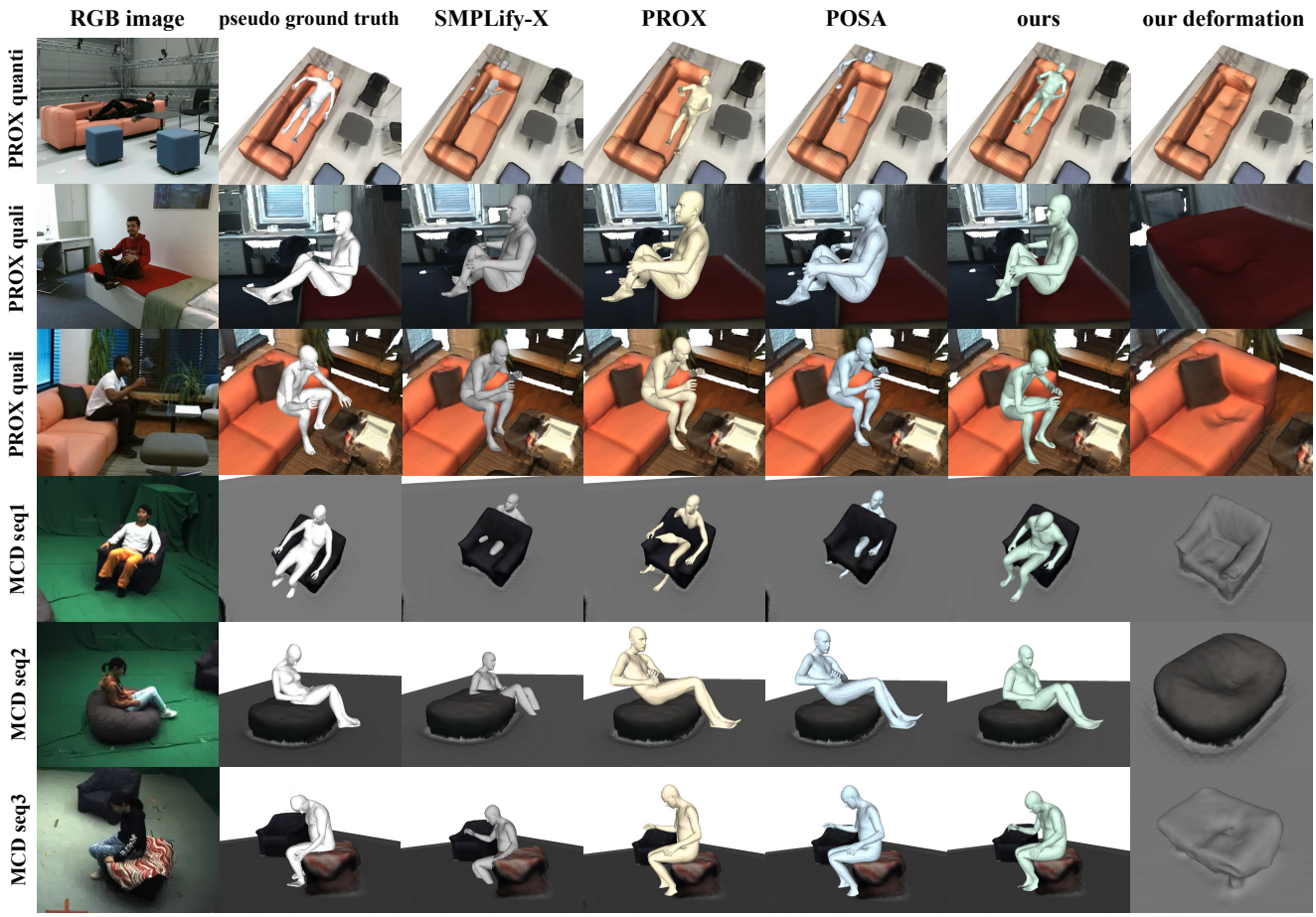


Figure 7: Qualitative results and comparisons on different datasets. Our MoCapDeform achieves more accurate global localisations than the state-of-the-arts, leads to less penetration, and prevents the human bodies from floating when there are large scene deformations. Moreover, it outputs plausible scene deformations not reconstructed by the previous methods.

	PJE	V2V	p.PJE	p.V2V	non-col
SMPLify-X [38]	214.64	219.20	64.04	61.88	92.42%
PROX [19]	167.16	171.35	63.54	63.06	95.63%
POSA [20]	157.11	159.52	63.70	63.23	95.89%
MoCapDeform (s1+s2)	***144.15***	***145.23***	***62.86***	***61.19***	***95.90%***
MoCapDeform (full)	**139.78**	**140.60**	**62.29**	**60.67**	**97.60%**

Table 1: Results on the PROX dataset using RGB inputs. We show our results of Stages 1 and 2 ("s1+s2") and full method and compare them with several state-of-the-arts. Best is indicated in **bold**, and second best in ***bold italic***.

	PJE	V2V	p.PJE	p.V2V	non-col
SMPLify-X [38]	441.86	451.87	**89.73**	**101.53**	97.14%
PROX [19]	375.01	403.22	97.09	107.57	97.99%
POSA [20]	365.91	398.15	97.26	108.67	98.41%
MoCapDeform (s1+s2)	***266.18***	***283.46***	***91.18***	***101.71***	***98.57%***
MoCapDeform (full)	**264.68**	**282.01**	91.91	102.43	**99.04%**

Table 2: Results on MoCapDeform dataset using RGB inputs. We compare outputs of Stages 1 and 2 ("s1+s2") and our full method to several state-of-the-art approaches. Best is indicated in **bold**, and second best in ***bold italic***.

mm. PJE stands for the mean per joint position error, and V2V indicates the mean vertex-to-vertex error. For local 3D reconstruction accuracy, we employ **p.PJE** and **p.V2V** (in mm), which are the PJE and V2V metrics after Procrustes alignment. Furthermore, to evaluate the human-scene penetrations—following the work in this domain [19, 65, 66, 64]—we report the **non-collision score**, which is the percentage of human body mesh vertices that do not penetrate the scene mesh. Note that for MoCapDeform, the non-collision score is calculated over the deformed meshes since they are also an output of our method.

The results on the PROX dataset and our new dataset are summarised in Tables 1 and 2. We report the results of Stages 1+2 and all stages of our framework (full model).

	PJE	V2V	p.PJE	p.V2V	non-col
SMPLify-D [19]	70.63	72.19	44.58	44.33	93.65%
PROX-D [19]	63.03	65.64	39.89	39.74	93.86%
POSA-D [20]	62.44	66.16	39.73	40.11	93.97%
MoCapDeform (s1+s3)	**59.32**	**62.37**	**39.57**	**39.12**	**97.04%**

Table 3: Results on the PROX quantitative dataset using RGB-D inputs. Best is indicated in **bold**.

For our new MoCapDeform dataset, we down-sample the framerate to 5 fps. As can be observed in the tables, both the global pose optimisation and the scene deformation stages contribute to more accurate pose estimation and outperform the previous approaches. Our method achieves significant improvement in terms of the global poses. With the help of the estimated scene deformations, the final output meshes of MoCapDeform have significantly fewer penetrations.

To further evaluate the effectiveness of the scene deformation stage, we conduct experiments on top of the RGB-D inputs from the PROX quantitative dataset; see Table 3. Specifically, we replace the pose initialisation stage with the PROX-D method [19], in which a depth term is used as a constraint during optimisation, supposedly resolving the depth ambiguity. Then we skip the second stage and directly apply our joint scene deformation and pose refinement stage over the PROX-D initialisation, with a depth data term (the same one as used in [19]) added to (4). The numbers in Table 3 show that our scene deformation stage can further improve the accuracy of 3D human pose estimation in all metrics and results in significantly fewer inter-penetrations after the deformation (*cf.* non-collision score).

4.3. Qualitative Results

We show qualitative results on PROX and our MoCapDeform datasets in Fig. 7. SMPLify-X [38] inaccurately localises the human and causes severe penetrations, as it ignores scene information. Both PROX [19] and POSA [20] leverage scene constraints by penalising the human-scene collisions with pre-computed SDF values of the static scenes. Since they do not model scene deformations, the collision penalising terms tend to lift the human above the scene surfaces even in the presence of large scene deformations. This leads to the subject floating (Fig. 7, PROX: rows 1, 2 and 5; POSA: rows 2 and 5), causing inaccurate global positions along the depth channel or severe body-scene penetrations (Fig. 7, PROX: rows 3 and 4; POSA: rows 1, 3 and 4), as the image cues and anti-collision terms cannot be satisfied simultaneously. In contrast, with the help of our raycast contact algorithm and scene deformation modelling, MoCapDeform finds more accurate global human positions without the floating issue, with significantly fewer inter-penetrations, and outputs plausible scene deformations. See Fig. 8 for comparisons between the ground-truth states of the beanbag (reconstructed after the person stands up) and the deformation output from our method.

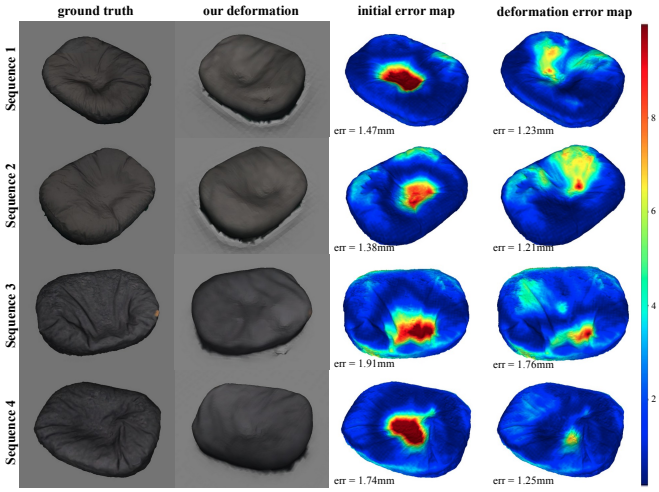


Figure 8: Comparison of ground-truth meshes and our deformations. The error maps show colour-coded per-vertex distances between the ground-truth meshes and the initial shapes or final states estimated by MoCapDeform.

5. Discussion and Conclusion

One limitation of our method is the dependency on reasonable 3D pose initialisation; recovery from starkly erroneous initial poses is unlikely. Moreover, severe scene occlusions blocking all human-scene contacts violate the assumptions of the raycast module, hindering the global pose optimisation. One future direction could be accounting for elastic properties of different objects and the integration of more fine-grained deformation models enabled by segmenting a single object into rigid and non-rigid parts. Next, modelling interactions between subjects wearing loose clothes and a non-rigid environment is an intriguing direction.

Closing Remarks. We present *MoCapDeform*, the first framework for markerless global 3D human motion capture from monocular RGB images with the awareness of non-rigid scene deformations. Benefiting from our new raycast-based contact localisation and joint scene deformation and pose optimisation steps, we find accurate global human poses and, at the same time, reasonable scene deformations. We show significantly improved global 3D human poses compared to several competing approaches. Due to these encouraging results, we expect in future to see more research on human motion capture systems that are aware of scene changes, including non-rigid deformations.

Acknowledgements. This work was supported by the ERC Consolidator Grant 4DRepLy (770784).

References

[1] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. UnrealEgo: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578, 2016. 2, 3

[3] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, pages 387–404, 2020. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 3

[5] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7035–7043, 2017. 2

[6] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *International Conference on 3D Vision (3DV)*, pages 405–414, 2019. 1, 2

[7] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining (KDD)*, 96(34):226–231, 1996. 5

[10] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018. 2

[11] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)*, 73(1):82–98, 1999. 2

[12] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. In *46th Session of the International Statistical Institute (ISI)*, volume 4, pages 5–21, 1987. 3, 5, 6

[13] James J Gibson. *The perception of the visual world.* Houghton Mifflin, 1950. 1

[14] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4318–4329, 2021. 1, 2

[15] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[16] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. A deeper look into deepcap. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 2

[17] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14:1–14:17, 2019. 2

[18] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10905–10914, 2019. 2

[19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 1, 2, 3, 6, 7, 8

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, 2021. 2, 3, 4, 7, 8

[21] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 5, 6

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 2, 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[24] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9799–9808, 2020. 5, 6

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5253–5263, 2020. 2, 3

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 3, 6

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose

estimation. In *International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017. 1, 2

[28] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 2

[29] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Transactions on Graphics (TOG)*, 39(4), 2020. 2, 3

[30] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 1, 2, 3

[31] Metashape. http://www.agisoft.com, 2021. 3, 6

[32] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):90–126, 2006. 2

[33] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019. 2

[34] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *International Conference on Computer Vision (ICCV)*, pages 10133–10142, 2019. 2

[35] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2823–2832, 2017. 2

[36] Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, pages 270–302, 2006. 6

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Neural Information Processing Systems Workshop (NeurIPSW)*, 2017. 6

[38] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 6, 7, 8

[39] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7025–7034, 2017. 2

[40] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 1, 2

[41] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):4–18, 2007. 2

[42] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[43] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision (ECCV)*, pages 71–87, 2020. 1, 2

[44] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 750–767, 2018. 1, 2

[45] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 152:1–20, 2016. 2

[46] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 2

[47] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. HULC: 3d human motion capture with pose manifold sampling and dense body-environment contact awareness. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[48] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (TOG)*, 40(4), 2021. 1

[49] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 1, 2

[50] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1153, 2017. 3

[51] Skanect: 3d scanning. https://skanect.occipital.com. 3

[52] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. 5

[53] Structure sensor: 3d scanning, augmented reality and more. https://structure.io/structure-sensor. 3

[54] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016. 2

[55] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2500–2509, 2017. 1, 2

[56] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[57] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 2

[58] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 3

[59] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 334–343, 2021. 1, 2

[60] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[61] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[62] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 2

[63] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, pages 34–51. Springer, 2020. 1, 2

[64] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 7

[65] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *International Conference on 3D Vision (3DV)*, pages 642–651, 2020. 7

[66] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6194–6204, 2020. 7

[67] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 6

[68] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *European Conference on Computer Vision (ECCV)*, pages 186–201. Springer, 2016. 2

[69] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2