HULC: 3D HUman Motion Capture with Pose Manifold SampLing and Dense Contact Guidance

Soshi Shimada[1] Vladislav Golyanik[1] Zhi Li[1] Patrick Pérez[2]
Weipeng Xu[1] Christian Theobalt[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] Valeo.ai

Abstract. Marker-less monocular 3D human motion capture (MoCap) with scene interactions is a challenging research topic relevant for extended reality, robotics and virtual avatar generation. Due to the inherent depth ambiguity of monocular settings, 3D motions captured with existing methods often contain severe artefacts such as incorrect body-scene inter-penetrations, jitter and body floating. To tackle these issues, we propose HULC, a new approach for 3D human MoCap which is aware of the scene geometry. HULC estimates 3D poses and dense body-environment surface contacts for improved 3D localisations, as well as the absolute scale of the subject. Furthermore, we introduce a 3D pose trajectory optimisation based on a novel pose manifold sampling that resolves erroneous body-environment inter-penetrations. Although the proposed method requires less structured inputs compared to existing scene-aware monocular MoCap algorithms, it produces more physically-plausible poses: HULC significantly and consistently outperforms the existing approaches in various experiments and on different metrics. Project page: `https://vcai.mpi-inf.mpg.de/projects/HULC/`.

Keywords: 3D Human MoCap, dense contact estimations, sampling.

1 Introduction

3D human motion capture (MoCap) from a single colour camera received a lot of attention over the past years [30, 29, 15, 16, 19, 37, 6, 50, 28, 40, 5, 27, 51, 31, 36, 33, 35, 57, 63, 9, 1, 56, 22, 45, 49, 20, 21, 23]. Its applications range from mixed and augmented reality, to movie production and game development, to immersive virtual communication and telepresence. MoCap techniques that not only focus on humans *in a vacuum* but also account for the scene environment—this encompasses awareness of the physics or constraints due to the underlying scene geometry—are coming increasingly into focus [46, 47, 39, 11, 61, 60, 38, 58].

Taking into account interactions between the human and the environment in MoCap poses many challenges, as not only articulations and global translation of the subject must be accurate, but also contacts between the human and the scene need to be plausible. A misestimation of only a few parameters, such as a 3D translation, can lead to reconstruction artefacts that contradict physical reality (*e.g.,* body-environment penetrations or body floating).

On the other hand, known human-scene contacts can serve as reliable boundary conditions for improved 3D pose estimation and localisation. While several algorithms merely consider human interactions with a ground plane [46, 47, 39, 38, 60], a few other methods also account for the contacts and interactions with the more general 3D environment [11, 61]. However, due to the depth ambiguity of the monocular setting, their estimated subject's root translations can be inaccurate, which can create implausible body-environment collisions. Next, they employ a body-environment collision penalty as a soft constraint. Therefore, the convergence of the optimisation to a bad local minima can also cause unnatural body-environment collisions. This paper addresses the limitations of the current works and proposes a new 3D **HU**man MoCap framework with pose manifold samp**L**ing and guidance by body-scene **C**ontacts, abbreviated as HULC. It improves over other monocular 3D human MoCap methods that consider constraints from 3D scene priors [11, 61]. Unlike existing works, HULC estimates contacts not only on the human body surface but also on the environment surface for the improved global 3D translation estimations. Next, HULC introduces a pose manifold sampling-based optimisation to obtain plausible 3D poses while handling the severe body-environment collisions in a *hard manner*. Our approach regresses more accurate 3D motions respecting scene constraints while requiring less-structured inputs (*i.e.,* an RGB image sequence and a point cloud of the static background scene) compared to the related monocular scene-aware methods [11, 61] that require a complete mesh and images. HULC returns physically-plausible motions, an absolute scale of the subject and dense contact labels both on a human template surface model and the environment.

Approach	Inputs	Outputs				
		body pose	$\boldsymbol{\tau}$	absolute scale	body contacts	env. contacts
PROX [11]	RGB + scene mesh	✓	✓	✗	✗	✗
PROX-D [11]	RGBD + scene mesh	✓	✓	✗	✗	✗
LEMO [61]	RGB(D) + scene mesh	✓	✓	✗	✓*	✗
HULC (ours)	RGB + scene point cloud	✓	✓	✓	✓	✓

Table 1: Overview of inputs and outputs of different methods. "$\boldsymbol{\tau}$" and "env. contacts" denote global translation and environment contacts, respectively. "$*$" stands for sparse marker contact labels.

HULC features several innovations which in interplay enable its functionality, *i.e.,* 1) a new learned implicit function-based dense contact label estimator for humans and the general 3D scene environment, and 2) a new pose optimiser for scene-aware pose estimation based on a pose manifold sampling policy. The first component allows us to jointly estimate the absolute subject's scale and its highly accurate root 3D translations.The second component prevents severe body-scene collisions and acts as a hard constraint, in contrast to widely-used soft collision losses [11, 26]. To train the dense contact estimation networks, we also annotate contact labels on a large scale synthetic daily motion dataset: GTA-IM [2]. To summarise, our primary technical contributions are as follows:

– A new 3D MoCap framework with simultaneous 3D human pose localisation and body scale estimation guided by estimated contacts. It is the first method

that regresses the dense body and environment contact labels from an RGB sequence and a point cloud of the scene using an implicit function (Sec. 3.1).
- A new pose optimisation approach with a novel pose manifold sampling yielding better results by imposing hard constraints on incorrect body-environment interactions (Sec. 3.2).
- Large-scale body contact annotations on the GTA-IM dataset [2] that provides synthetic 3D human motions in a variety of scenes (Fig. 1 and Sec. 4).

We report quantitative results, including an ablative study, which show that HULC outperforms existing methods in 3D accuracy and on physical plausibility metrics (Sec. 5). See our video for qualitative comparisons.

2 Related Works

Most monocular MoCap approaches estimate 3D poses alone or along with the body shape from an input image or video [9, 15, 16, 19, 6, 50, 28, 40, 5, 13, 27, 51, 31, 36, 33, 35, 57, 63, 9, 1, 56, 22, 45, 49, 20, 23, 62]. Some methods also estimate 3D translation of the subject in addition to the 3D poses [30, 29, 21, 37]. Fieraru *et al.* [8] propose a multi-person 3D reconstruction method considering human-human interactions. Another algorithm class incorporates an explicit physics model into MoCap and avoids environmental collisions [47, 39, 46, 59]. These methods consider interactions with only a flat ground plane or a stick-like object [25], unlike our HULC, that can work with arbitrary scene geometry.

Awareness of human-scene contacts is helpful for the estimation and synthesis [53, 10] of plausible 3D human motions. Some existing works regress sparse joint contacts on a kinematic skeleton [25, 46, 47, 39, 38, 64] or sparse markers [61]. A few approaches forecast contacts on a dense human mesh surface [12, 32]. Hassan *et al.* [12] place a human in a 3D scene considering the semantic information and dense human body contact labels. Müller *et al.* [32] propose a dataset with discrete annotations for self-contacts on the human body. Consequently, they apply a self-contact loss for more plausible final 3D poses. Unlike the existing works, our algorithm estimates vertex-wise dense contact labels on the human body surface from an RGB input only. Along with that, it also regresses dense contact labels on the environment given the scene point cloud along with the RGB sequence. The simultaneous estimation of the body and scene contacts allows HULC to disambiguate the depth and scale of the subject, although only a single camera view and a single scene point cloud are used as inputs.

Monocular MoCap with scene interactions. Among the scene-aware MoCap approaches [46, 47, 39, 11, 61, 38, 60], there are a few ones that consider human-environment interactions given a highly detailed scene geometry [11, 61, 24]. PROX (PROX-D)[11] estimates 3D motions given RGB (RGB-D) image, along with an input geometry provided as a signed distance field (SDF). Given an RGB(D) measurement and a mesh of the environment, LEMO [61] also produces geometry-aware global 3D human motions with an improved motion quality characterised by smoother transitions and robustness to occlusions thanks

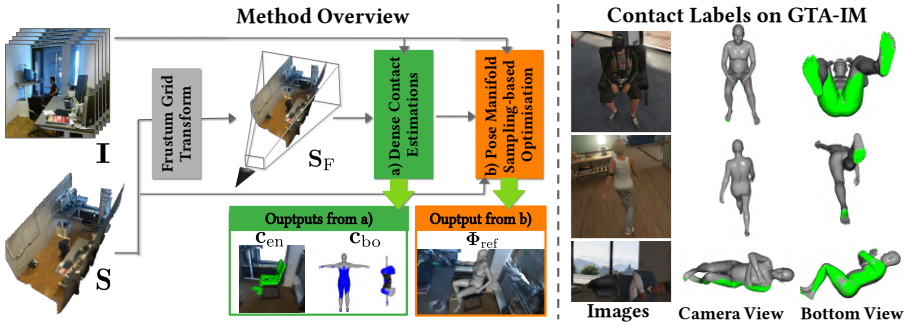

Fig. 1: (*Left*) Given image sequence **I**, scene point cloud **S** and its associated frustum voxel grid $\mathbf{S}_{\mathrm{F}}$, HULC first predicts for each frame dense contact labels on the body $\mathbf{c}_{\mathrm{bo}}$, and on the environment $\mathbf{c}_{\mathrm{en}}$. It then refines initial, physically-inaccurate and scale-ambiguous global 3D poses $\mathbf{\Phi}_0$ into the final ones $\mathbf{\Phi}_{\mathrm{ref}}$ in (b). Also see Fig. 2 for the details of stage (a) and (b). (*Right*) Example visualisations of our contact annotations (shown in green) on GTA-IM dataset [2].

to the learned motion priors. These two algorithms require an RGB or RGB-D sequence with SDF (a 3D scan of the scene) or occlusion masks. In contrast, our HULC requires only an RGB image sequence and a point cloud of the scene; it returns dense contact labels on 1) the human body and 2) the environment, 3) global 3D human motion with translations and 4) absolute scale of the human body. See Table 1 for an overview of the characteristics. Compared to PROX and LEMO, HULC shows significantly-mitigated body-environment collisions.

Sampling-based human pose tracking. Several sampling-based human pose tracking algorithms were proposed. Some of them utilise particle-swarm optimisation [14, 41, 42]. Charles *et al.* [4] employ Parzen windows for 2D joints tracking. Similar to our HULC, Sharma *et al.* [44] generate 3D pose samples by a conditional variational autoencoder (VAE) [48] conditioned on 2D poses. In contrast, we utilise the learned pose manifold of VAE for sampling, which helps to avoid local minima and prevent body-scene collisions. Also, unlike [44], we sample around a latent vector obtained from the VAE's encoder to obtain poses that are plausible and similar to the input 3D pose.

3 Method

Given monocular video frames and a point cloud of the scene registered to the coordinate frame of the camera, our goal is to infer physically-plausible global 3D human poses along with dense contact labels on both body and environment surfaces. Our approach consists of two stages (Fig. 1):

– **Dense Body-environment contacts estimation**: Dense contact labels are predicted on body and scene surfaces using a learning-based approach with a pixel-aligned implicit representation inspired by [43] (Sec. 3.1);

– **Sampling-based optimisation on the pose manifold**: We combine sampling in a learned latent pose space with gradient descent to obtain the absolute scale of the subject and its global 3D pose, under hard guidance by predicted contacts. This approach significantly improves the accuracy of the estimated root translation and articulations, and mitigates incorrect environment penetrations. (Sec. 3.2).

Modelling and Notations. Our method takes as input a sequence $\mathbf{I} = \{\mathbf{I}_1, ..., \mathbf{I}_T\}$ of T successive video frames from a static camera with known intrinsics ($T = 5$ in our experiments). We detect a squared bounding box around the subject and resize the cropped image region to 225×225 pixels. The background scene's geometry that corresponds to the detected bounding box is represented by a single static point cloud $\mathbf{S} \in \mathbb{R}^{M \times 3}$ composed of M points aligned in the camera reference frame in an absolute scale. To model the 3D pose and human body surface, we employ the parametric model SMPL-X [34] (its gender-neutral version). This model defines the 3D body mesh as a differentiable function $\mathcal{M}(\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\beta})$ of global root translation $\boldsymbol{\tau} \in \mathbb{R}^3$, global root orientation $\boldsymbol{\phi} \in \mathbb{R}^3$, root-relative pose $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ of K joints and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ capturing body's identity. For efficiency, we downsample the original SMPL-X body mesh with over 10k vertices to $\mathbf{V} \in \mathbb{R}^{N \times 3}$, where $N = 655$. In the following, we denote $\mathbf{V} = \mathcal{M}(\boldsymbol{\Phi}, \boldsymbol{\beta})$, where $\boldsymbol{\Phi} = (\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\theta})$ denotes the kinematic state of the human skeleton, from which the global positions $\mathbf{X} \in \mathbb{R}^{K \times 3}$ of the $K = 21$ joints can be derived.

3.1 Contact Estimation in the Scene

We now describe our learning-based approach for contact labels estimation on the human body and environment surfaces; see Fig. 1-a) for an overview of this stage. The approach takes $\mathbf{I}$ and $\mathbf{S}$ as inputs. It comprises three fully-convolutional feature extractors, N_1, N_2 and N_3, and two fully-connected layer-based contact prediction networks, Ω_{bo} and Ω_{en}, for body and environment, respectively.

Network N_1 extracts from $\mathbf{I}$ a stack of visual features $\mathbf{f}_{\mathbf{I}} \in \mathbb{R}^{32 \times 32 \times 256}$. The latent space features of N_1 are also fed to Ω_{bo} to predict the vector $\mathbf{c}_{\text{bo}} \in [0, 1]^N$ of per-vertex contact probabilities on the *body* surface.

We also aim at estimating the corresponding contacts on the *environment* surface using an implicit function. To train a model that generalises well, we need to address two challenges: (i) No correspondence information between the scene points and the image pixels are given; (ii) Each scene contains a variable number of points. Accordingly, we convert the scene point cloud $\mathbf{S}$ into a frustum voxel grid $\mathbf{S}_{\text{F}} \in \mathbb{R}^{32 \times 32 \times 256}$ (the third dimension corresponds to the discretised depth of the 3D space over 256 bins, please refer to our supplement for the details). This new representation is independent of the original point-cloud size and is aligned with the camera's view direction. The latter will allow us to leverage a pixel-aligned implicit function inspired by PIFu [43], which helps the networks figure out the correspondences between pixel and geometry information. More specifically, $\mathbf{S}_{\text{F}}$ is fed into N_2, which returns scene features $\mathbf{f}_{\text{S}} \in \mathbb{R}^{32 \times 32 \times 256}$. The third encoder, N_3, ingests $\mathbf{f}_{\text{I}}$ and $\mathbf{f}_{\text{S}}$ concatenated along their third dimension

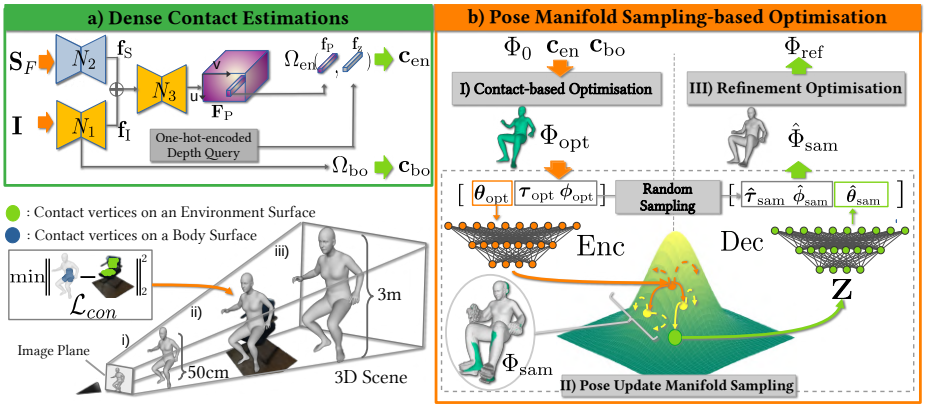

Fig. 2: **Overview of a) dense contact estimation and b) pose manifold sampling-based optimisation.** In b-II), we first generate samples around the mapping from $\boldsymbol{\theta}_{\text{opt}}$ (orange arrows), and elite samples are then selected among them (yellow points). After resampling around the elite samples (yellow arrows), the best sample is selected (green point). The generated sample poses $\boldsymbol{\Phi}_{\text{sam}}$ (in gray color at the bottom left in b-II)) from the sampled latent vectors are plausible and similar to $\boldsymbol{\Phi}_{\text{opt}}$. ***(bottom left of the Figure)*** Different body scale and depth combinations can be re-projected to the same image coordinates (i, ii and iii),*i.e.*, **scale-depth ambiguity**. To simultaneously estimate the accurate body scale and depth of the subject (ii), we combine the body-environment contact surface distance loss $\mathcal{L}_{\text{con}}$ with the 2D reprojection loss.

and returns pixel-aligned features $\mathbf{F}_{\text{P}} \in \mathbb{R}^{32\times32\times64}$. Based on $\mathbf{F}_{\text{P}}$, Ω_{en} predicts the contact labels on the environment surface as follows. Given a 3D position in the scene, we extract the corresponding visual feature $\mathbf{f}_{\text{P}} \in \mathbb{R}^{64}$ at the (u, v)-position in the image space from $\mathbf{F}_{\text{P}}$ (via spacial bilinear interpolation), and query arbitrary depth with a one-hot vector $\mathbf{f}_{\text{z}} \in \mathbb{R}^{256}$. We next estimate the contact labels c_{en} as follows:

$$c_{\text{en}} = \Omega_{\text{en}}(\mathbf{f}_{\text{P}}, \mathbf{f}_{\text{z}}). \tag{1}$$

Given contact ground truths $\hat{\mathbf{c}}_{\text{bo}} \in \{0,1\}^N$ and $\hat{\mathbf{c}}_{\text{en}} \in \{0,1\}^M$ on the body and the environment, the five networks are trained with the following loss:

$$\mathcal{L}_{\text{labels}} = \|\mathbf{c}_{\text{en}} - \hat{\mathbf{c}}_{\text{en}}\|_2^2 + \lambda\ \text{BCE}(\mathbf{c}_{\text{bo}}, \hat{\mathbf{c}}_{\text{bo}}), \tag{2}$$

where BCE denotes the binary cross-entropy and $\lambda = 0.3$. We use BCE for the body because the ground-truth contacts on its surface are binary; the ℓ_2 loss is used for the environment, as sparse ground-truth contact labels are smoothed with a Gaussian kernel to obtain continuous signals. For further discussions of (2), please refer to our supplement. At test time, we only provide the 3D vertex positions of the environment to $\Omega_{\text{en}}(\cdot)$—to find the contact area on the scene

point cloud—rather than all possible 3D sampling points as queries. This significantly accelerates the search of environmental contact labels while reducing the number of false-positive contact classifications. For more details of the network architecture, further discussions of the design choice and data pre-processing, please refer to our supplement.

3.2 Pose Manifold Sampling-based Optimisation

In the second stage of the approach, we aim at recovering an accurate global 3D trajectory of the subject as observed in the video sequence, see Fig. 2-(b) for the overview. An initial estimate $\mathbf{\Phi}_0$ is extracted for each input image using SMPLify-X [34]. Its root translation $\boldsymbol{\tau}$ being subject to scale ambiguity, we propose to estimate it more accurately, along with the actual scale h of the person with respect to the original body model's height, under the guidance of the predicted body-environment contacts (**Contact-based Optimisation**).We then update the body trajectory and articulations in the scene, while mitigating the body-environment collisions with a new sampling-based optimisation on the pose manifold (**Sampling-based Trajectory Optimisation**). A subsequent refinement step yields the final global physically-plausible 3D motions.

I) Contact-based Optimisation Scale ambiguity is inherent to a monocular MoCap setting: Human bodies with different scale and depth combinations in 3D can be reprojected on the same positions in the image frame; see Fig. 2 and supplementary video for the schematic visualisation. Most existing algorithms that estimate global 3D translations of a subject either assume its known body scale [47, 7, 46] or use a statistical average body scale [30]. In the latter case, the estimated $\boldsymbol{\tau}$ is often inaccurate and causes physically implausible body-environment penetrations. In contrast to the prior art, we simultaneously estimate $\boldsymbol{\tau}$ and h by making use of the body-environment dense contact labels from the previous stage (Sec. 3.1).

For the given frame at time $t \in [\![1, T]\!]$, we select the surface regions with $\mathbf{c}_{\text{en}} > 0.5$ and $\mathbf{c}_{\text{bo}} > 0.5$ as effective contacts and leverage them in our optimisation. Let us denote the corresponding index subsets of body vertices and scene points by $\mathcal{C}_{\text{bo}} \subset [\![1, N]\!]$ and $\mathcal{C}_{\text{en}} \subset [\![1, M]\!]$. The objective function for contact-based optimisation is defined as:

$$\mathcal{L}_{\text{opt}}(\boldsymbol{\tau}, h) = \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{3}$$

where the reprojection $\mathcal{L}_{\text{2D}}$, the temporal smoothness $\mathcal{L}_{\text{smooth}}$ and the contact $\mathcal{L}_{\text{con}}$ losses weighted by empirically-set multipliers λ_{2D}, λ_{smooth} and λ_{con}, read:

$$\mathcal{L}_{\text{2D}} = \frac{1}{K}\sum_{k=1}^{K} w_k \|\Pi(\mathbf{X}_k) - \mathbf{p}_k\|_2^2, \tag{4}$$

$$\mathcal{L}_{\text{smooth}} = \|\boldsymbol{\tau} - \boldsymbol{\tau}_{\text{prev}}\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{con}} = \sum_{n \in \mathcal{C}_{\text{bo}}} \min_{m \in \mathcal{C}_{\text{en}}} \|\mathbf{V}_n - \mathbf{P}_m\|_2^2, \tag{6}$$

where $\mathbf{p}_k$ and w_k are the 2D detection in the image of the k-th body joint and its associated confidence, respectively, obtained by OpenPose [3]; Π is the perspective projection operator; $\boldsymbol{\tau}_{\text{prev}}$ is the root translation estimated in the previous frame; $\mathbf{X}_k$, $\mathbf{V}_n$ and $\mathbf{P}_m$ are, respectively, the k-th 3D joint, the n-th body vertex ($n \in \mathcal{C}_{\text{bo}}$) and the m-th scene point ($m \in \mathcal{C}_{\text{en}}$). Note that the relative rotation and pose are taken from $\boldsymbol{\Phi}_0$. The body joints and vertices are obtained from $\mathcal{M}$ using $\boldsymbol{\tau}$ and scaled with h. For $\mathcal{L}_{\text{con}}$, we use a directed Hausdorff measure [18] as a distance between the body and environment contact surfaces. The combination of $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{2D}}$ is key to disambiguate $\boldsymbol{\tau}$ and h (thus, resolving the monocular scale ambiguity). As a result of optimising (3) in frame t, we obtain $\boldsymbol{\Phi}^t_{\text{opt}}$, *i.e.,* the global 3D human motion with absolute body scale. We solve jointly on T frames and optimise for a single h for them.

II-a) Sampling-based Trajectory Optimisation Although the poses $\boldsymbol{\Phi}^t_{\text{opt}}$, $t = 1 \cdots T$, estimated in the previous step yield much more accurate $\boldsymbol{\tau}$ and h compared to existing monocular RGB-based methods, incorrect body-environment penetrations are still observable. This is because the gradient-based optimisation often gets stuck in bad local minima (see the supplementary video for a toy example illustrating this issue). To overcome this problem, we introduce an additional sampling-based optimisation that imposes hard penetration constraints, thus significantly mitigating physically-implausible collisions. The overview of this algorithm is as follows: (i) For each frame t, we first draw candidate poses around $\boldsymbol{\Phi}^t_{\text{opt}}$ with a sampling function $\mathcal{G}$; (ii) The quality of these samples is ranked by a function $\mathcal{E}$ that allows selecting the most promising ("elite") ones; samples with severe collisions are discarded; (iii) Using $\mathcal{G}$ and $\mathcal{E}$ again, we generate and select new samples around the elite ones. The details of these steps, $\mathcal{E}$ and $\mathcal{G}$, are elaborated next (dropping time index t for simplicity).

II-b) Generating Pose Samples. We aim to generate N_{sam} sample states $\boldsymbol{\Phi}_{\text{sam}}$ around the previously-estimated $\boldsymbol{\Phi}_{\text{opt}} = (\boldsymbol{\tau}_{\text{opt}}, \boldsymbol{\phi}_{\text{opt}}, \boldsymbol{\theta}_{\text{opt}})$. To generate samples $(\boldsymbol{\tau}_{\text{sam}}, \boldsymbol{\phi}_{\text{sam}})$ for the global translation and orientation, with 3DoF each, we simply use a uniform distribution around $(\boldsymbol{\tau}_{\text{opt}}, \boldsymbol{\phi}_{\text{opt}})$; see our supplement for the details. However, naïvely generating the relative pose $\boldsymbol{\theta}_{\text{sam}}$ in the same way around $\boldsymbol{\theta}_{\text{opt}}$ is highly inefficient because (i) the body pose is high-dimensional and (ii) the randomly-sampled poses are not necessarily plausible. These reasons lead to an infeasible amount of generated samples required to find a plausible collision-free pose; which is intractable on standard graphics hardware. To tackle these issues, we resort to the pose manifold learned by VPoser [34], which is a VAE [17] trained on AMASS [26], *i.e.,* a dataset with many highly accurate MoCap sequences. Sampling is conducted in this VAE's latent space rather than in the kinematics pose space. Specifically, we first map $\boldsymbol{\theta}_{\text{opt}}$ into a latent pose vector with the VAE's encoder $\text{Enc}(\cdot)$. Next, we sample latent vectors using a Gaussian distribution centered at this vector, with standard deviation $\boldsymbol{\sigma}$ (see Fig. 2-b). Each latent sample is then mapped through VAE's decoder $\text{Dec}(\cdot)$ into a pose that is combined with the original one on a per-joint basis. The complete sampling process reads:

$$\mathbf{Z} \sim \mathcal{N}(\,\text{Enc}(\boldsymbol{\theta}_{\text{opt}}), \boldsymbol{\sigma}), \quad \boldsymbol{\theta}_{\text{sam}} = \mathbf{w} \circ \boldsymbol{\theta}_{\text{opt}} + (1 - \mathbf{w}) \circ \text{Dec}(\mathbf{Z}), \tag{7}$$

where $\circ$ denotes Hadamard matrix product and $\mathbf{w} \in \mathbb{R}^{3K}$ is composed of the detection confidence values w_k, $k = 1 \cdots K$, obtained from OpenPose, each appearing three times (for each DoF of the joint). This confidence-based strategy allows weighting higher the joint angles obtained by sampling, if the image-based detections are less confident (*e.g.,* under occlusions). Conversely, significant modifications are not required for the joints with high confidence values.

Since the manifold learned by VAE is smooth, the poses derived from the latent vectors sampled around $\text{Enc}(\boldsymbol{\theta}_{\text{opt}})$ should be close to $\boldsymbol{\theta}_{\text{opt}}$. Therefore, we empirically set $\boldsymbol{\sigma}$ to a small value (0.1). Compared to the naïve random sampling in the joint angle space, whose generated poses are not necessarily plausible, this pose sampling on the learned manifold significantly narrows down the solution space. Hence, a lot fewer samples are required to escape local minima. At the bottom left of Fig. 2-b contains examples (gray color) of $\boldsymbol{\Phi}_{\text{sam}}$ ($N_{\text{sam}} = 10$) overlayed onto $\boldsymbol{\Phi}_{\text{opt}}$ (green). In the following, we refer to this sample generation process as function $\mathcal{G}(\cdot)$.

II-c) Sample Selection. The quality of the N_{sam} generated samples $\boldsymbol{\Phi}_{\text{sam}}$ is evaluated using the following cost function:

$$\mathcal{L}_{\text{sam}} = \mathcal{L}_{\text{opt}} + \lambda_{\text{sli}}\mathcal{L}_{\text{sli}} + \lambda_{\text{data}}\mathcal{L}_{\text{data}}, \tag{8}$$

$$\mathcal{L}_{\text{sli}} = \|\mathbf{V}_{\text{c}} - \mathbf{V}_{\text{c,pre}}\|_2^2, \tag{9}$$

$$\mathcal{L}_{\text{data}} = \|\boldsymbol{\Phi}_{\text{sam}} - \boldsymbol{\Phi}_{\text{opt}}\|_2^2, \tag{10}$$

where $\mathcal{L}_{\text{sli}}$ and $\mathcal{L}_{\text{data}}$ are contact sliding loss and data loss, respectively, and $\mathcal{L}_{\text{opt}}$ is the same as in (3) with the modification that the temporal consistency (5) applies to the whole $\boldsymbol{\Phi}_{\text{sam}}$; $\mathbf{V}_{\text{c}}$ and $\mathbf{V}_{\text{c,pre}}$ are the body contact vertices (with vertex indices in $\mathcal{C}_{\text{bo}}$) and their previous positions, respectively.

Among N_{sam} samples ordered according to their increasing $\mathcal{L}_{\text{sam}}$ values, the selection function $\mathcal{E}_U(\cdot)$ first discards those causing stronger penetrations (in the sense that the amount of scene points inside a human body is above a threshold γ) and returns U first samples from the remaining ones. If no samples pass the collision test, we regenerate the new set of N_{sam} samples. This selection mechanism introduces the collision handling in a hard manner. After applying $\mathcal{E}_U(\cdot)$, with $U < N_{\text{sam}}$, U elite samples are retained. Then, $\lfloor N_{\text{sam}}/U \rfloor$ new samples are regenerated around every elite sample using $\mathcal{G}$. Among those, the one with minimum $\mathcal{L}_{\text{sam}}$ value is retained as the final estimate. The sequence of obtained poses is temporally smoothed by Gaussian filtering to further remove jittering, which yields the global 3D motion $(\hat{\boldsymbol{\Phi}}_{\text{sam}}^{t})_{t=1}^{T}$ with significantly mitigated collisions.

III) Final Refinement. From the previous step, we obtained the sequence $\hat{\boldsymbol{\Phi}}_{\text{sam}} = (\hat{\boldsymbol{\tau}}_{\text{sam}}, \hat{\phi}_{\text{sam}}, \hat{\boldsymbol{\theta}}_{\text{sam}})$ of kinematic states whose severe body-environment collisions are prevented as hard constraints. Starting from these states as initialisation, we perform a final gradient-based refinement using cost function $\mathcal{L}_{\text{sam}}$ with $\hat{\boldsymbol{\Phi}}_{\text{sam}}$ replacing $\boldsymbol{\Phi}_{\text{opt}}$. The final sequence is denoted $(\boldsymbol{\Phi}_{\text{ref}}^{t})_{t=1}^{T}$.

4 Datasets with Contact Annotations

As there are no publicly-available large-scale datasets with images and corresponding human-scene contact annotations, we annotate several existing datasets. **GTA-IM** [2] dataset contains various daily 3D motions. First, we fit SMPL-X model onto the 3D joint trajectories in GTA-IM. For each frame, we select contact vertices on the human mesh if: i) The Euclidean distance between the human body vertices on and the scene vertices are smaller than a certain threshold; ii) The velocity of the vertex is lower than a certain threshold. In total, we obtain the body surface contact annotations on $320k$ frames, which will be released for research purposes, see Fig. 1 for the examples of the annotated contact labels. **PROX dataset** [11] contains scanned scene meshes, scene SDFs, RGB-D sequences, 3D human poses and shapes generated by fitting SMPL-X model onto the RGB-D sequences (considering collisions). We consider the body vertices, whose SDF values are lower than 5 cm, as contacts. We annotate the environment contacts by finding the vertices that are the nearest to the body contacts. **GPA dataset** [54, 55] contains multi-view image sequences of people interacting with various rigid 3D geometries, accurately reconstructed 3D scenes and 3D human motions obtained from VICON system [52] with 28 calibrated cameras. We fit SMPL-X on GPA to obtain the 3D shapes and compute the scene's SDFs to run other methods [11, 61, 12].

We extract from **GPA** 14 test sequences with 5 different subjects. We also split **PROX** [11] into training and test sequences. The training sequences of **PROX** and **GTA-IM** [2] are used to train the contact estimation networks. For further details of dataset and training, please refer to our supplement.

5 Evaluations

We compare our HULC with the most related scene-aware 3D MoCap algorithms, *i.e.,* PROX[11], PROX-D[11], POSA[12] and LEMO [61]. We also test SMPLify-X [34] which does not use scene constraints. The root translation of SMPLify-X is obtained from its estimated camera poses as done in [11]. To run LEMO [61] on the RGB sequence, we use SMPLify-X[34] to initialise it; we call this combination "LEMO (RGB)". We use the selected test sequences of GPA [54, 55] and PROX [11] dataset for the quantitative and qualitative comparisons. To avoid redundancy, we downsample all the predictions to 10 fps except for the temporal consistency measurement (e_{smooth} in Table 4). Since the 3D poses in PROX dataset are prone to inaccuracies due to their human model fitting onto the RGB-D sequence, we use it only for reporting the body-scene penetrations (Table 4) and for qualitative comparisons.

5.1 Quantitative Results

We report 3D joint and vertex errors (Table 2), global translation and body scale estimation errors (Table 3), body-environment penetration and smoothness

Table 2: Comparisons of 3D error on GPA dataset [54, 55]. "†" denotes that the occlusion masks for LEMO(RGB) were computed from GT 3D human mesh.

	No Procrustes			Procrustes		
	MPJPE [mm]↓	PCK [%]↑	PVE [mm]↓	MPJPE [mm]↓	PCK [%]↑	PVE [mm]↓
Ours	**217.9**	**35.3**	**214.7**	**81.5**	**89.3**	**72.6**
Ours (w/o S)	221.3	34.5	217.2	82.6	**89.3**	73.1
Ours (w/o R)	240.8	31.9	237.3	83.1	86.6	73.6
Ours (w/o SR)	251.1	31.5	245.2	83.9	86.6	74.1
SMPLify-X [34]	550.0	10.0	549.1	84.7	85.9	74.1
PROX [11]	549.7	10.1	548.7	84.6	86.0	73.9
POSA [12]	552.2	10.1	550.9	85.5	85.6	74.5
LEMO (RGB) [61]	570.1	8.75	570.5	83.0	86.4	73.7
LEMO (RGB) [61]†	570.0	8.77	570.4	83.0	86.4	73.6

Table 3: Ablations and comparisons for global translations and absolute body length on GPA dataset.

	global translation error [m] ↓	absolute bone length error [m]↓
Ours (+1m)	**0.242**	0.104
Ours (+3m)	0.244	**0.097**
Ours (+10m)	0.244	0.109
Baseline (+1m)	0.751	0.498
Baseline (+3m)	1.033	0.560
Baseline (+10m)	2.861	1.918
SMPLify-X [34]	0.527	0.156
PROX [11]	0.528	0.160
POSA [12]	0.545	0.136

Table 4: Comparisons of physical plausibility measures on GPA dataset [54, 55] and PROX dataset [11].

		GPA Dataset		PROX Dataset
		non penet. [%]↑	e_{smooth} ↓	non penet. [%]↑
RGB	Ours	**99.4**	20.2	97.0
	Ours (w/o S)	97.6	28.1	93.8
	Ours (w/o R)	**99.4**	24.7	**97.1**
	Ours (w/o SR)	97.6	47.1	93.8
	SMPLify-X [34]	97.7	43.3	88.9
	PROX [11]	97.7	43.2	89.8
	LEMO (RGB)[61]	97.8	**19.9**	-
	POSA [12]	98.0	47.0	93.0
RGB-D	PROX-D [11]	-	-	94.2
	LEMO [61]	-	-	96.4

errors (Table 4) and ablations on the sampling-based optimisation component, *i.e.*, a) Manifold sampling vs. random sampling and b) Different number of sampling iterations in Fig. 3. "Ours (w/o S)" represents our method without the sampling optimisation component, *i.e.,* only the contact-based optimisation and refinement are applied (see Fig. 2-(b) and Sec. 3.2). "Ours (w/o R)" represents our method without the final refinement. "Ours (w/o SR)" denotes ours without the sampling and refinement. For a further ablation study and evaluation of contact label estimation networks, please see our supplement.

3D Joint and Vertex Errors. Table 2 compares the accuracy of 3D joint and vertex positions with and without Procrustes alignment. LEMO also requires human body occlusion masks on each frame. We compute them using the scene geometry and SMPLify-X [34] results. We also show another variant "LEMO (RGB)†" whose occlusion masks are computed using the ground-truth global 3D human mesh instead of SMPLify-X. Here, we report the standard 3D metrics, *i.e.,* mean per joint position error (MPJPE), percentage of correct keypoints (PCK) (@150mm) and mean per vertex error (PVE). Lower MPJPE and PVE

represent more accurate 3D reconstructions, higher PCK indicates more accurate 3D joint positions.

On all these metrics, HULC outperforms other methods both with and without Procrustes. Notably, thanks to substantially more accurate global translations obtained from the contact-based optimisation (Sec. 3.2), HULC significantly reduces the MPJPE and PVE with a big margin, *i.e.,* $\approx 60\%$ error deduction in MPJPE and PVE w/o Procrustes compared to the second-best method. The ablative studies on Table 2 also indicate that both the sampling and refinement optimisations contribute to accurate 3D poses. Note that the sampling optimisation alone ("Ours (w/o R)") does not significantly reduce the error compared to "Ours (w/o SR)". This is because the sampling component prioritises removal of environment penetrations by introducing hard collision handling, which is the most important feature of this component. Therefore, the sampling component significantly contributes to reducing the environment collision as can be seen in Table 4 (discussed in the later paragraph). Applying the refinement after escaping from severe penetrations by the sampling optimisation further increases the 3D accuracy ("Ours" in Table 2) while significantly mitigating physically implausible body-environment penetrations (Table 4).

Global Translation and Body Scale Estimation. Table 3 reports global translation and body scale estimation errors for the ablation study of the contact-based optimisation (Sec. 3.2). More specifically, we evaluate the output Φ_{opt} obtained from the contact-based optimisation denoted "ours". We also show the optimisation result without using the contact loss term (6) ("Baseline"). The numbers next to the method names represent the initialisation offset from the ground-truth 3D translation position (*e.g.,* "+10m" indicates that the initial root position of the human body was placed at 10 meters away along the depth direction from the ground-truth root position when solving the optimisations).

Without the contact loss term—since global translation and body scale are jointly estimated in the optimisation—the baseline method suffers from *up-to-scale* issue (see Fig. 2). Hence, its results are significantly worse due to worse initialisations. In contrast, our contact-based optimisation disambiguates the scale and depth by localising the contact positions on the environment, which confirms HULC to be highly robust to bad initialisations. Compared to the RGB-based methods PROX, POSA and SMPLify-X, our contact-based optimisation result has $\approx 40\%$ smaller error in the absolute bone length, and $\approx 57\%$ smaller error in global translation, which also contributes to the reduced body-environment collisions as demonstrated in Table 4 (discussed in the next paragraph).

Plausibility Measurements. We also report the plausibility of the reconstructed 3D motions in Table 4. *Non penet.* measures the average ratio of non-penetrating body vertices into the environment over all frames. A higher value denotes fewer body-environment collisions in the sequence; e_{smooth} measures the temporal smoothness error proposed in [47]. Lower e_{smooth} indicates more temporally smooth 3D motions. On both GPA and PROX datasets, our full framework mitigates the collisions thanks to the manifold sampling-based optimisations (ours vs. ours (w/o S)). It also does so when compared to other related

works as well. Notably, HULC shows the least amount of collisions even compared with RGBD-based methods on the PROX dataset. Finally, the proposed method also shows the significantly low e_{smooth} (on par with LEMO(RGB)) in this experiment.

More Ablations on Sampling-based Optimisation. In addition to the ablation studies reported in Tables 2, 3 and 4, we further assess the performance of the pose update manifold sampling step (Fig. 2-(b)-(II)) on GPA dataset [54, 55], reporting the 3D error (MPJPE [mm]) measured in world frame. Note that we report MPJPE without the final refinement step to assess the importance of the manifold sampling approach. In Fig. 3-(a), we show the influence of the number N_{sam} of samples on the performance of our manifold sampling strategy vs. a naïve random sampling with a uniform distribution in a kinematic skeleton frame. For the details of the naïve random sampling strategy, please refer to our supplement. In Fig. 3-(a), since the generated samples of the learned manifold return plausible pose samples, our pose manifold sampling strategy requires significantly fewer samples compared to the random sampling (∼15× more samples are required for the random sampling to reach 243 [mm] error in MPJPE). This result strongly supports the importance of the learned manifold sampling. No more than 2000 samples can be generated due to the hardware memory capacity. In Fig. 3-(b), we report the influence of the number of generation-selection steps using functions $\mathcal{G}$ and $\mathcal{E}_U$ (with $U = 3$) introduced in Section 3.2, with $N_{\text{sam}} = 1000$ samples. No iteration stands for choosing the best sample from the first generated batch (hence no resampling), while one iteration is the variant described in Sec. 3.2. This first iteration sharply reduces the MPJPE, while the benefit of the additional iterations is less pronounced. Based on these observations, we use only one re-sampling iteration with 1000 samples in the previous experiments. Finally, we ablate the confidence value-based pose merging in Eq. (7), setting N_{sam}=1000 and the number of iterations to 0. The measured MPJPE for with and without this confidence merging are 245.5 and 249.1, respectively.

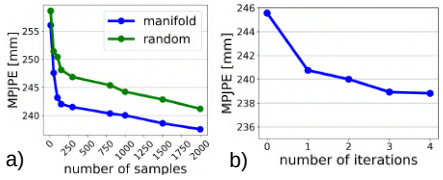

Fig. 3: (a) MPJPE [mm] comparison with different numbers of samples for the learned manifold sampling strategy vs. the naïve random sampling in the joint angle space of the kinematic skeleton.(b) MPJPE [mm] comparison with different numbers of iterations in the sampling strategy.

5.2 Qualitative Results

Figure 4 summarises the qualitative comparisons on GPA and PROX datasets. HULC produces more physically-plausible global 3D poses with mitigated collisions, whereas the other methods show body-environment penetrations. Even

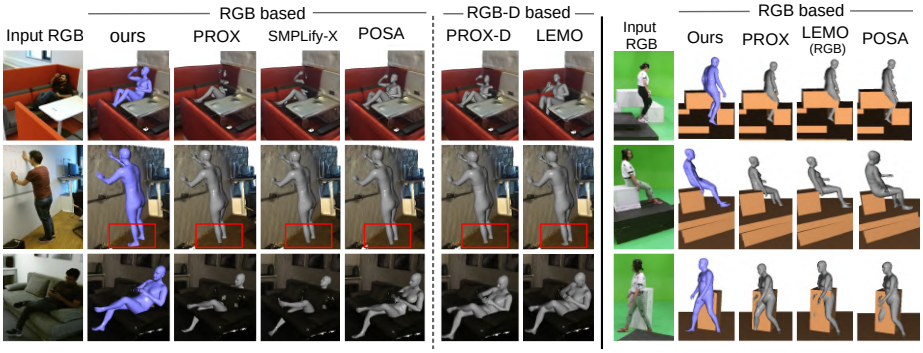

Fig. 4: The qualitative comparisons of our results with the related methods on PROX (left) and GPA dataset (right). Our RGB-based HULC shows fewer body-scene penetrations even when compared with RGB-D based methods; mind the red rectangles in the second row.

compared with the RGB(D) approaches, HULC mitigates collisions (mind the red rectangles). For more qualitative results, please refer to our video.

6 Concluding Remarks

Limitations. HULC requires the scene geometry aligned in a camera frame like other related works [11, 61, 12]. Also, HULC does not capture non-rigid deformations of scenes and bodies, although the body surface and some objects in the environment deform (*e.g.,* when sitting on a couch or lying in a bed). Moreover, since our algorithm relies on the initial root-relative pose obtained from an RGB-based MoCap algorithm, the subsequent steps can fail under severe occlusions. Although the estimated contact labels help to significantly reduce the 3D translation error, the estimated environment contacts contain observable false positives. These limitations can be tackled in the future.

Conclusion. We introduced *HULC*—the first RGB-based scene-aware MoCap algorithm that estimates and is guided by dense body-environment surface contact labels combined with a pose manifold sampling. HULC shows 60% smaller 3D-localisation errors compared to the previous methods. Furthermore, deep body-environment collisions are handled in hard manner in the pose manifold sampling-based optimisation, which significantly mitigates collisions with the scene. HULC shows the lowest collisions even compared with RGBD-based scene-aware methods.

Acknowledgements. The authors from MPII were supported by the ERC Consolidator Grant 4DRepLy (770784). We also acknowledge support from Valeo.

References

1. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision (ECCV) (2016)
2. Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision (ECCV) (2020)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
4. Charles, J., Pfister, T., Everingham, M., Zisserman, A.: Automatic and efficient human pose estimation for sign language videos. International Journal of Computer Vision **110**, 70–90 (10 2013)
5. Chen, C., Ramanan, D.: 3d human pose estimation = 2d pose estimation + matching. In: Computer Vision and Pattern Recognition (CVPR) (2017)
6. Choi, H., Moon, G., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Computer Vision and Pattern Recognition (CVPR) (2021)
7. Dabral, R., Shimada, S., Jain, A., Theobalt, C., Golyanik, V.: Gravity-aware monocular 3d human-object reconstruction. In: International Conference on Computer Vision (ICCV) (2021)
8. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Computer Vision and Pattern Recognition (CVPR) (2020)
9. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: Computer Vision and Pattern Recognition (CVPR) (2019)
10. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: International Conference on Computer Vision (ICCV) (2021)
11. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision (ICCV) (2019)
12. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: Computer Vision and Pattern Recognition (CVPR) (2021)
13. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2020)
14. John, V., Trucco, E., McKenna, S.: Markerless human motion capture using charting and manifold constrained particle swarm optimisation. In: British Machine Vision Conference (BMVC) (2010)
15. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
16. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Computer Vision and Pattern Recognition (CVPR) (2019)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014)

18. Knauer, C., Löffler, M., Scherfenberg, M., Wolle, T.: The directed hausdorff distance between imprecise point sets. In: International Symposium on Algorithms and Computation (ISAAC) (2009)
19. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Computer Vision and Pattern Recognition (CVPR) (2020)
20. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: International Conference on Computer Vision (ICCV) (2021)
21. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: International Conference on Computer Vision (ICCV) (2021)
22. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: International Conference on Computer Vision (ICCV) (2019)
23. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: International Conference on Computer Vision (ICCV) (2021)
24. Li, Z., Shimada, S., Schiele, B., Theobalt, C., Golyanik, V.: Mocapdeform: Monocular 3d human motion capture in deformable scenes. In: Arxiv (2022)
25. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3d motion and forces of person-object interactions from monocular video. In: Computer Vision and Pattern Recognition (CVPR) (2019)
26. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision (ICCV) (2019)
27. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: International Conference on Computer Vision (ICCV) (2017)
28. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: International Conference on 3D Vision (3DV) (2017)
29. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. ACM Transactions on Graphics (TOG) **39**(4) (2020)
30. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4) (2017)
31. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: Computer Vision and Pattern Recognition (CVPR) (2017)
32. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Computer Vision and Pattern Recognition (CVPR) (2021)
33. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV) (2016)
34. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2019)
35. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Computer Vision and Pattern Recognition (CVPR) (2017)

36. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Computer Vision and Pattern Recognition (CVPR) (2018)
37. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Computer Vision and Pattern Recognition (CVPR) (2019)
38. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: International Conference on Computer Vision (ICCV) (2021)
39. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video. In: European Conference on Computer Vision (ECCV) (2020)
40. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation learning for 3d human pose estimation. In: European Conference on Computer Vision (ECCV) (2018)
41. Saini, S., Rambli, D.R.B.A., Sulaiman, S.B., Zakaria, M.N.B.: Human pose tracking in low-dimensional subspace using manifold learning by charting. In: International Conference on Signal and Image Processing Applications (ICSIPA) (2013)
42. Saini, S., Rambli, D.R.B.A., Sulaiman, S.B., Zakaria, M.N.B., Rohkmah, S.: Markerless multi-view human motion tracking using manifold model learning by charting. Procedia Engineering **41**, 664–670 (2012)
43. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: International Conference on Computer Vision (ICCV) (2019)
44. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3d human pose estimation by generation and ordinal ranking. In: International Conference on Computer Vision (ICCV) (2019)
45. Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., Chen, B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Transactions on Graphics (TOG) **40**(1), 1–15 (2020)
46. Shimada, S., Golyanik, V., Xu, W., Pérez, P., Theobalt, C.: Neural monocular 3d human motion capture with physical awareness. ACM Transactions on Graphics (TOG) **40**(4) (aug 2021)
47. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: Physcap: Physically plausible monocular 3d motion capture in real time. ACM Transactions on Graphics **39**(6) (dec 2020)
48. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems (NIPS) (2015)
49. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: International Conference on Computer Vision (ICCV) (2019)
50. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured Prediction of 3D Human Pose with Deep Neural Networks. In: British Machine Vision Conference (BMVC) (2016)
51. Tomè, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2017)
52. Vicon blade. `https://www.vicon.com/`

53. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Computer Vision and Pattern Recognition (CVPR) (2021)
54. Wang, Z., Chen, L., Rathore, S., Shin, D., Fowlkes, C.: Geometric pose affordance: 3d human pose with scene constraints. In: Arxiv (2019)
55. Wang, Z., Shin, D., Fowlkes, C.: Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In: European Conference on Computer Vision Workshop (ECCVW) (2020)
56. Wei, X., Chai, J.: Videomocap: Modeling physically realistic human motion from monocular video sequences. ACM Transactions on Graphics (TOG) **29**(4) (2010)
57. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: Computer Vision and Pattern Recognition (CVPR) (2018)
58. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: Computer Vision and Pattern Recognition (CVPR) (2022)
59. Yuan, Y., Wei, S.E., Simon, T., Kitani, K., Saragih, J.: Simpoe: Simulated character control for 3d human pose estimation. In: Computer Vision and Pattern Recognition (CVPR) (2021)
60. Zanfir, A., Marinoiu, E., Sminchisescu, C.: 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) (2018)
61. Zhang, S., Zhang, Y., Bogo, F., Marc, P., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: International Conference on Computer Vision (ICCV) (Oct 2021)
62. Zhang, T., Huang, B., Wang, Y.: Object-occluded human shape and pose estimation from a single color image. In: Computer Vision and Pattern Recognition (CVPR) (2020)
63. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: A weakly-supervised approach. In: International Conference on Computer Vision (ICCV) (2017)
64. Zou, Y., Yang, J., Ceylan, D., Zhang, J., Perazzi, F., Huang, J.B.: Reducing footskate in human motion reconstruction with ground contact constraints. In: Winter Conference on Applications of Computer Vision (WACV) (2020)