3D Human Pose Estimation via Intuitive Physics

Shashank Tripathi[1] Lea Müller[1] Chun-Hao P. Huang[1] Omid Taheri[1]
Michael J. Black[1] Dimitrios Tzionas[2*]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Amsterdam, the Netherlands

{stripathi, lmueller2, chuang2, otaheri, black}@tue.mpg.de d.tzionas@uva.nl

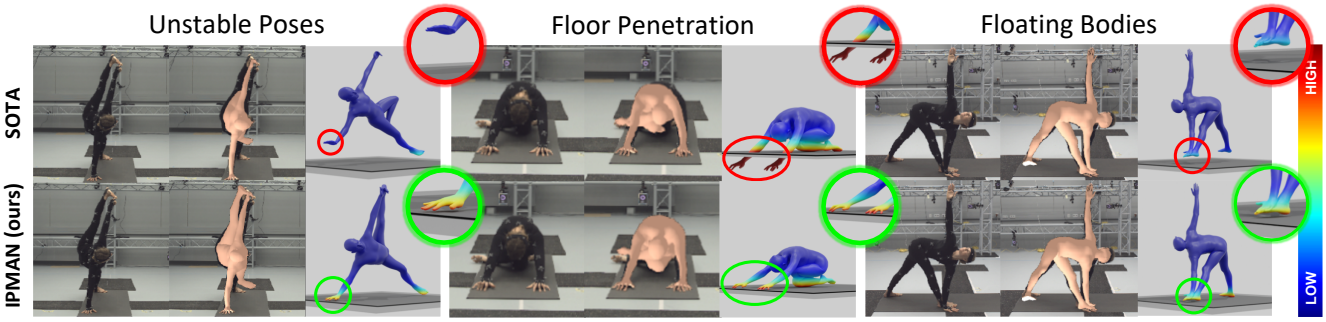

Figure 1. Estimating a 3D body from an image is ill-posed. A recent, representative, optimization method [59] produces bodies that are in unstable poses, penetrate the floor, or hover above it. In contrast, IPMAN estimates a 3D body that is physically *plausible*. To achieve this, IPMAN uses novel *intuitive-physics* (IP) terms that exploit inferred *pressure* heatmaps on the body, the *Center of Pressure* (CoP), and the body's *Center of Mass* (CoM). Body heatmap colors encode per-vertex pressure.

Abstract

Estimating 3D humans from images often produces implausible bodies that lean, float, or penetrate the floor. Such methods ignore the fact that bodies are typically supported by the scene. A physics engine can be used to enforce physical plausibility, but these are not differentiable, rely on unrealistic proxy bodies, and are difficult to integrate into existing optimization and learning frameworks. In contrast, we exploit novel intuitive-physics (IP) terms that can be inferred from a 3D SMPL body interacting with the scene. Inspired by biomechanics, we infer the pressure *heatmap on the body, the* Center of Pressure *(CoP) from the heatmap, and the SMPL body's* Center of Mass *(CoM). With these, we develop IPMAN, to estimate a 3D body from a color image in a "stable" configuration by encouraging plausible floor contact and overlapping CoP and CoM. Our IP terms are intuitive, easy to implement, fast to compute, differentiable, and can be integrated into existing optimization and regression methods. We evaluate IPMAN on standard datasets and MoYo, a new dataset with synchronized multi-view images, ground-truth 3D bodies with complex poses, body-floor contact, CoM and pressure. IPMAN produces more plausible results than the state of the art, improving accuracy for static poses, while not hurting dynamic ones. Code and data are available for research at* https://ipman.is.tue.mpg.de.

1. Introduction

To understand humans and their actions, computers need automatic methods to reconstruct the body in 3D. Typically, the problem entails estimating the 3D human pose and shape (HPS) from one or more color images. State-of-the-art (SOTA) methods [46, 51, 75, 102] have made rapid progress, estimating 3D humans that *align* well with image features in the camera view. Unfortunately, the camera view can be deceiving. When viewed from other directions, or when placed in a 3D scene, the estimated bodies are often physically implausible: they lean, hover, or penetrate the ground (see Fig. 1 top). This is because most SOTA methods reason about humans *in isolation*; they ignore that people move in a scene, interact with it, and receive physical support by contacting it. This is a *deal-breaker* for inherently 3D applications, such as biomechanics, augmented/virtual reality (AR/VR) and the "metaverse"; these need humans to be reconstructed faithfully and *physically plausibly* with respect to the scene. For this, we need a method that estimates the 3D human on a ground plane from a color image in a configuration that is *physically "stable"*.

This is naturally related to reasoning about physics and support. There exist many physics simulators [10, 30, 60] for games, movies, or industrial simulations, and using these for plausible HPS estimation is increasingly popular [66, 74, 96]. However, existing simulators come with two significant

* This work was mostly performed at MPI-IS.

problems: (1) They are typically non-differentiable *black boxes*, making them incompatible with existing optimization and learning frameworks. Consequently, most methods [64, 95, 96] use them with reinforcement learning to evaluate whether a certain input has the desired outcome, but with no ability to reason about how changing inputs affects the outputs. (2) They rely on an unrealistic proxy body model for computational efficiency; bodies are represented as groups of rigid 3D shape primitives. Such proxy models are crude approximations of human bodies, which, in reality, are much more complex and deform non-rigidly when they move and interact. Moreover, proxies need *a priori* known body dimensions that are kept fixed during simulation. Also, these proxies differ significantly from the 3D body models [41, 54, 92] used by SOTA HPS methods. Thus, current physics simulators are too limited for use in HPS.

What we need, instead, is a solution that is fully differentiable, uses a realistic body model, and seamlessly integrates physical reasoning into HPS methods (both optimization- and regression-based). To this end, instead of using full physics simulation, we introduce novel intuitive-physics (IP) terms that are simple, differentiable, and compatible with a body model like SMPL [54]. Specifically, we define terms that exploit an inferred *pressure* heatmap of the body on the ground plane, the *Center of Pressure* (CoP) that arises from the heatmap, and the SMPL body's *Center of Mass* (CoM) projected on the floor; see Fig. 2 for a visualization. Intuitively, bodies whose CoM lie close to their CoP are more *stable* than ones with a CoP that is further away (see Fig. 5); the former suggests a *static pose*, e.g. standing or holding a yoga pose, while the latter a *dynamic pose*, e.g., walking.

We use these intuitive-physics terms in two ways. First, we incorporate them in an objective function that extends SMPLify-XMC [59] to optimize for body poses that are stable. We also incorporate the same terms in the training loss for an HPS regressor, called IPMAN (Intuitive-Physics-based huMAN). In both formulations, the intuitive-physics terms encourage estimates of body shape and pose that have sufficient ground contact, while penalizing interpenetration and encouraging an overlap of the CoP and CoM.

Our intuitive-physics formulation is inspired by work in biomechanics [32, 33, 61], which characterizes the stability of humans in terms of relative positions between the CoP, the CoM, and the *Base of Support* (BoS). The BoS is defined as the convex hull of all contact regions on the floor (Fig. 2). Following past work [6, 71, 74], we use the "inverted pendulum" model [85, 86] for body balance; this considers poses as stable if the gravity-projected CoM onto the floor lies inside the BoS. Similar ideas are explored by Scott et al. [71] but they focus on predicting a foot pressure heatmap from 2D or 3D body joints. We go significantly further to exploit stability in training an HPS regressor. This requires two technical novelties.

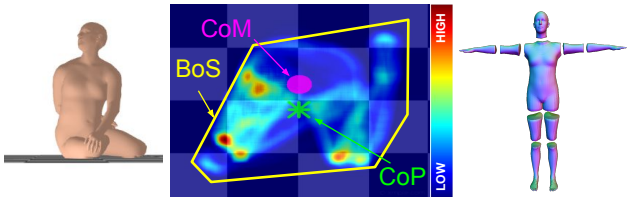

Figure 2. (1) A SMPL mesh sitting. (2) The inferred pressure map on the ground (color-coded heatmap), CoP (green), CoM (pink), and Base of Support (BoS, yellow polygon). (3) Segmentation of SMPL into $N_P = 10$ parts, used for computing CoM; see Sec. 3.2.

The first involves computing CoM. To this end, we uniformly sample points on SMPL's surface, and calculate each body part's volume. Then, we compute CoM as the average of all uniformly sampled points weighted by the corresponding part volumes. We denote this as pCoM, standing for "part-weighted CoM". Importantly, pCoM takes into account SMPL's shape, pose, and all blend shapes, while it is also computationally efficient and differentiable.

The second involves estimating CoP directly from the image, without access to a pressure sensor. Our key insight is that the soft tissues of human bodies deform under pressure, e.g., the buttocks deform when sitting. However, SMPL does not model this deformation; it *penetrates* the ground instead of deforming. We use the penetration depth as a proxy for pressure [68]; deeper penetration means higher pressure. With this, we estimate a pressure field on SMPL's mesh and compute the CoP as the pressure-weighted average of the surface points. Again this is differentiable.

For evaluation, we use a standard HPS benchmark (Human3.6M [37]), but also the RICH [35] dataset. However, these datasets have limited interactions with the floor. We thus capture a novel dataset, MoYo, of challenging yoga poses, with synchronized multi-view video, ground-truth SMPL-X [63] meshes, pressure sensor measurements, and body CoM. IPMAN, in both of its forms, and across all datasets, produces more accurate and stable 3D bodies than the state of the art. Importantly, we find that IPMAN improves accuracy for static poses, while not hurting dynamic ones. This makes IPMAN applicable to everyday motions.

To summarize: (1) We develop IPMAN, the first HPS method that integrates intuitive physics. (2) We infer biomechanical properties such as CoM, CoP and body pressure. (3) We define novel *intuitive-physics* terms that can be easily integrated into HPS methods. (4) We create MoYo, a dataset that uniquely has complex poses, multi-view video, and ground-truth bodies, pressure, and CoM. (5) We show that our IP terms improve HPS accuracy and physical plausibility. (6) Data and code are available for research.

2. Related Work

3D Human Pose and Shape (HPS) from images. Existing methods fall into two major categories: (1) non-

parametric methods that reconstruct a free-form body representation, e.g., joints [1, 56, 57] or vertices [52, 58, 100], and (2) parametric methods that use statistical body models [5, 25, 41, 54, 63, 92, 97]. The latter methods focus on various aspects, such as expressiveness [13, 18, 63, 69, 87], clothed bodies [15, 88, 91], videos [24, 45, 78, 99], and multi-person scenarios [38, 75, 103], to name a few.

Inference is done by either optimization or regression. Optimization-based methods [7, 16, 63, 87, 88] fit a body model to image evidence, such as joints [11], dense vertex correspondences [2] or 2D segmentation masks [23]. Regression-based methods [42, 44, 48, 51, 76, 102, 106, 109] use a loss similar to the objective function of optimization methods to train a network to infer body model parameters. Several methods combine optimization and regression in a training loop [47, 50, 59]. Recent methods [24, 40] fine-tune pre-trained networks at test time w.r.t. an image or a sequence, retaining flexibility (optimization) while being less sensitive to initialization (regression).

Despite their success, these methods reason about the human in "isolation", without taking the surrounding scene into account; see [77, 107] for a comprehensive review.

Contact-only scene constraints. A common way of using scene information is to consider body-scene *contact* [12, 17, 27, 28, 65, 84, 90, 94, 98, 104, 105, 110]. Yamamoto et al. [93] and others [19,27,70,98,104] ensure that estimated bodies have plausible scene contact. For videos, encouraging foot-ground contact reduces foot skating [36,65,72,105,110]. Weng et al. [84] use contact in estimating the pose and scale of scene objects, while Villegas et al. [80] preserve self- and ground contact for motion retargeting.

These methods typically take two steps: (1) detecting contact areas on the body and/or scene and (2) minimizing the distance between these. Surfaces are typically assumed to be in contact if their distance is below a threshold and their relative motion is small [27, 35, 98, 104].

Many methods only consider contact between the ground and the foot joints [66, 110] or other end-effectors [65]. In contrast, IPMAN uses the full 3D body surface and exploits this to compute the pressure, CoP and CoM. Unlike binary contact, this is differentiable, making the IP terms useful for training HPS regressors.

Physics-based scene constraints. Early work uses physics to estimate walking [8, 9] or full body motion [82]. Recent methods [21, 22, 66, 73, 74, 89, 96] regress 3D humans and then refine them through *physics-based optimization*. Physics is used for two primary reasons: (1) to regularise dynamics, reducing jitter [49, 66, 74, 96], and (2) to discourage interpenetration and encourage contact. Since contact events are discontinuous, the pipeline is either not end-to-end trainable or trained with reinforcement learning [64, 96]. Xie et al. [89] propose differentiable physics-inspired objectives based on a soft contact penalty, while DiffPhy [21] uses a differentiable physics simulator [31] during inference. Both methods apply the objectives in an optimization scheme, while IPMAN is applied to both optimization and regression. PhysCap [74] considers a pose as balanced, when the CoM is projected within the BoS. Rempe et al. [66] impose PD control on the pelvis, which they treat as a CoM. Scott et al. [71] regress foot pressure from 2D and 3D joints for stability analysis but do not use it to improve HPS.

All these methods use unrealistic bodies based on shape primitives. Some require known body dimensions [66, 74, 96] while others estimate body scale [49, 89]. In contrast, IPMAN computes CoM, CoP and BoS directly from the SMPL mesh. Clever et al. [14] and Luo et al. [55] estimate 3D body pose but from pressure measurements, not from images. Their task is fundamentally different from ours.

3. Method

3.1. Preliminaries

Given a color image, $\mathbf{I}$, we estimate the parameters of the camera and the SMPL body model [54].

Body model. SMPL maps pose, $\boldsymbol{\theta}$, and shape, $\boldsymbol{\beta}$, parameters to a 3D mesh, $\boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$. The pose parameters, $\boldsymbol{\theta} \in \mathbb{R}^{24\times 6}$, are rotations of SMPL's 24 joints in a 6D representation [108]. The shape parameters, $\boldsymbol{\beta} \in \mathbb{R}^{10}$, are the first 10 PCA coefficients of SMPL's shape space. The generated mesh $\boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ consists of $N_V = 6890$ vertices, $\boldsymbol{V} \in \mathbb{R}^{N_V \times 3}$, and $N_F = 13776$ faces, $\boldsymbol{F} \in \mathbb{R}^{N_F \times 3 \times 3}$.

Note that our regression method (IPMAN-R, Sec. 3.4.1) uses SMPL, while our optimization method (IPMAN-O, Sec. 3.4.2) uses SMPL-X [63], to match the models used by the baselines. For simplicity of exposition, we refer to both models as SMPL when the distinction is not important.

Camera. For the regression-based IPMAN-R, we follow the standard convention [42, 43, 47] and use a weak perspective camera with a 2D scale, s, translation, $\mathbf{t}^c = (t_x^c, t_y^c)$, fixed camera rotation, $\mathbf{R}^c = \boldsymbol{I}_3$, and a fixed focal length (f_x, f_y). The root-relative body orientation $\mathbf{R}^b$ is predicted by the neural network, but body translation stays fixed at $\mathbf{t}^b = \mathbf{0}$ as it is absorbed into the camera's translation.

For the optimization-based IPMAN-O, we follow Müller et al. [59] to use the full-perspective camera model and optimize the focal lengths (f_x, f_y), camera rotation $\mathbf{R}^c$ and camera translation $\mathbf{t}^c$. The principal point (o_x, o_y) is the center of the input image. $\mathbf{K}$ is the intrinsic matrix storing focal lengths and the principal point. We assume that the body rotation $\mathbf{R}^b$ and translation $\mathbf{t}^b$ are absorbed into the camera parameters, thus, they stay fixed as $\mathbf{R}^b = \boldsymbol{I}_3$ and $\mathbf{t}^b = \mathbf{0}$. Using the camera, we project a 3D point $\mathbf{X} \in \mathbb{R}^3$ to an image point $\mathbf{x} \in \mathbb{R}^2$ through $\mathbf{x} = \mathbf{K}(\mathbf{R}^c\mathbf{X} + \mathbf{t}^c)$.

Ground plane and gravity-projection. We assume that the gravity direction is perpendicular to the ground plane in the world coordinate system. Thus, for any arbitrary point in

3D space, $\boldsymbol{u} \in \mathbb{R}^3$, its *gravity-projected* point, $\boldsymbol{u'} = g(\boldsymbol{u}) \in \mathbb{R}^3$, is the projection of $\boldsymbol{u}$ along the plane normal $\boldsymbol{n}$ onto the ground plane, and $g(.)$ is the projection operator. The function $h(\boldsymbol{u})$ returns the signed "height" of a point $\boldsymbol{u}$ with respect to the ground; i.e., the signed distance from $\boldsymbol{u}$ to the ground plane along the gravity direction, where $h(\boldsymbol{u}) < 0$ if $\boldsymbol{u}$ is below the ground and $h(\boldsymbol{u}) > 0$ if $\boldsymbol{u}$ is above it.

3.2. Stability Analysis

We follow the biomechanics literature [32, 33, 61] and Scott et al. [71] to define three fundamental elements for stability analysis: We use the Newtonian definition for the "Center of Mass" (CoM); i.e., the mass-weighted average of particle positions. The "Center of Pressure" (CoP) is the ground-reaction force's point of application. The "Base of Support" (BoS) is the convex hull of all body-ground contacts. Below, we define intuitive-physics (IP) terms using the inferred CoM and CoP. BoS is only used for evaluation.

Body Center of Mass (CoM). We introduce a novel CoM formulation that is fully differentiable and considers the per-part mass contributions, dubbed as pCoM; see Sup. Mat. for alternative CoM definitions. To compute this, we first segment the template mesh into $N_P = 10$ parts $P_i \in \mathcal{P}$; see Fig. 2. We do this once offline, and keep the segmentation fixed during training and optimization. Assuming a shaped and posed SMPL body, the per-part volumes $\mathcal{V}^{P_i}$ are calculated by splitting the SMPL mesh into parts.

However, mesh splitting is a non-differentiable operation. Thus, it cannot be used for either training a regressor (IPMAN-R) or for optimization (IPMAN-O). Instead, we work with the full SMPL mesh and use differentiable *"close-translate-fill"* operations for each body part on the fly. First, for each part P, we extract boundary vertices $\mathcal{B}_P$ and add in the middle a *virtual* vertex $\boldsymbol{v}_g$, where $\boldsymbol{v}_g = \sum_{j \in \mathcal{B}_P} \boldsymbol{v}_j / |\mathcal{B}_P|$. Then, for the $\mathcal{B}_P$ and $\boldsymbol{v}_g$ vertices, we add virtual faces to *"close"* P and make it *watertight*. Next, we *"translate"* P such that the part centroid $\mathbf{c}_P = \sum_{j \in P} \boldsymbol{v}_j / |P|$ is at the origin. Finally, we *"fill"* the centered P with tetrahedrons by connecting the origin with each face vertex. Then, the part volume, $\mathcal{V}^{\mathcal{P}}$, is the sum of all tetrahedron volumes [101].

To create a uniform distribution of surface vertices, we uniformly sample $N_U = 20000$ surface points $\boldsymbol{V}_U \in \mathbb{R}^{N_U \times 3}$ on the template SMPL mesh using the Triangle Point Picking method [83]. Given $\boldsymbol{V}_U$ and the template SMPL mesh vertices $\boldsymbol{V}_T$, we follow [59], and analytically compute a sparse linear regressor $\mathbf{W} \in \mathbb{R}^{N_U \times N_V}$ such that $\boldsymbol{V}_U = \mathbf{W}\boldsymbol{V}_T$. During training and optimization, given an arbitrary shaped and posed mesh with vertices $\boldsymbol{V}$, we obtain uniformly-sampled mesh surface points as $\boldsymbol{V}_U = \mathbf{W}\boldsymbol{V}$. Each surface point, v_i, is assigned to the body part, P_{v_i}, corresponding to the face, $\boldsymbol{F}_{v_i}$, it was sampled from.

Finally, the part-weighted pCoM is computed as a volume-weighted mean of the mesh surface points:

$$\bar{\mathbf{m}} = \frac{\sum_{i=1}^{N_U} \mathcal{V}^{P_{v_i}} v_i}{\sum_{i=1}^{N_U} \mathcal{V}^{P_{v_i}}}, \tag{1}$$

where $\mathcal{V}^{P_{v_i}}$ is the volume of the part $P_{v_i} \in \mathcal{P}$ to which v_i is assigned. This formulation is fully differentiable and can be employed with any existing 3D HPS estimation method.

Note that computing CoM (or volume) from uniformly sampled surface points does not work (see Sup. Mat.) because it assumes that mass, M, is proportional to surface area, S. Instead, our pCoM computes mass from volume, $\mathcal{V}$, via the standard density equation, $M = \rho\mathcal{V}$, while our *close-translate-fill* operation computes the volume of deformable bodies in an efficient and differentiable manner.

Center of Pressure (CoP). Recovering a pressure heatmap from an image without using hardware, such as pressure sensors, is a highly ill-posed problem. However, stability analysis requires knowledge of the pressure exerted on the human body by the supporting surfaces, like the ground. Going beyond binary contact, Rogez et al. [68] estimate 3D forces by detecting intersecting vertices between hand and object meshes. Clever et al. [14] recover pressure maps by allowing articulated body models to deform a soft pressure-sensing virtual mattress in a physics simulation.

In contrast, we observe that, while real bodies interacting with rigid objects (e.g., the floor) deform under contact, SMPL does not model such soft-tissue deformations. Thus, the body mesh penetrates the contacting object surface and the amount of penetration can be a proxy for pressure; a deeper penetration implies higher pressure. With the height $h(v_i)$ (see Sec. 3.1) of a mesh surface point v_i with respect to the ground plane Π, we define a *pressure field* to compute the per-point pressure ρ_i as:

$$\rho_i = \begin{cases} 1 - \alpha h(v_i) & \text{if } h(v_i) < 0, \\ e^{-\gamma h(v_i)} & \text{if } h(v_i) \geq 0, \end{cases} \tag{2}$$

where α and γ are scalar hyperparameters set empirically. We approximate soft tissue via a "spring" model and "penetrating" pressure field using Hooke's Law. Some pressure is also assigned to points above the ground to allow tolerance for footwear, but this decays quickly. Finally, we compute the CoP, $\bar{\mathbf{s}}$, as

$$\bar{\mathbf{s}} = \frac{\sum_{i=1}^{N_U} \rho_i v_i}{\sum_{i=1}^{N_U} \rho_i}. \tag{3}$$

Again, note that this term is fully differentiable.

Base of Support (BoS). In biomechanics [34, 85], BoS is defined as the "supporting area" or the possible range of the CoP on the supporting surface. Here, we define BoS as the convex hull [67] of all gravity-projected body-ground contact points. In detail, we first determine all such contacts

by selecting the set of mesh surface points v_i close to the ground, and then gravity-project them onto the ground to obtain $C = \{g(v_i) \mid |h(v_i)| < \tau\}$. The BoS is then defined as the convex hull $\mathcal{C}$ of C.

3.3. Intuitive-Physics Losses

Stability loss. The *"inverted pendulum"* model of human balance [85, 86] considers the relationship between the CoM and BoS to determine stability. Simply put, for a given shape and pose, if the body CoM, projected on the gravity-aligned ground plane, lies within the BoS, the pose is considered *stable*. While this definition of stability is useful for evaluation, using it in a loss or energy function for 3D HPS estimation results in sparse gradients (see Sup. Mat.). Instead, we define the stability criterion as:

$$\mathcal{L}_{\text{stability}} = \|g(\bar{\mathbf{m}}) - g(\bar{\mathbf{s}})\|_2, \tag{4}$$

where $g(\bar{\mathbf{m}})$ and $g(\bar{\mathbf{s}})$ are the gravity-projected CoM and CoP, respectively.

Ground contact loss. As shown in Fig. 1, 3D HPS methods minimize the 2D joint reprojection error and do not consider the plausibility of body-ground contact. Ignoring this can result in interpenetrating or hovering meshes. Inspired by self-contact losses [19, 59] and hand-object contact losses [26, 29], we define two ground losses, namely pushing, $\mathcal{L}_{\text{push}}$, and pulling, $\mathcal{L}_{\text{pull}}$, that take into account the height, $h(v_i)$, of a vertex, v_i, with respect to the ground plane. For $h(v_i) < 0$, i.e., for vertices under the ground plane, $\mathcal{L}_{\text{push}}$ discourages body-ground penetrations. For $h(v_i) \geq 0$, i.e., for hovering meshes, $\mathcal{L}_{\text{pull}}$ encourages the vertices that lie close to the ground to "snap" into contact with it. Note that the losses are non-conflicting as they act on disjoint sets of vertices. Then, the ground contact loss is:

$$\mathcal{L}_{\text{ground}} = \mathcal{L}_{\text{pull}} + \mathcal{L}_{\text{push}}, \text{ with} \tag{5}$$

$$\mathcal{L}_{\text{pull}} = \alpha_1 \tanh\left(\frac{h(v_i)}{\alpha_2}\right)^2 \quad \text{if } h(v_i) \geq 0, \text{ and} \tag{6}$$

$$\mathcal{L}_{\text{push}} = \beta_1 \tanh\left(\frac{h(v_i)}{\beta_2}\right)^2 \quad \text{if } h(v_i) < 0. \tag{7}$$

3.4. IPMAN

We use our new IP losses for two tasks: (1) We extend HMR [42] to develop IPMAN-R, a regression-based HPS method. (2) We extend SMPLify-XMC [59] to develop IPMAN-O, an optimization-based method. Note that IPMAN-O uses a reference ground plane, while IPMAN-R uses the ground plane only for training but not at test time. It leverages the *known* ground in 3D datasets, and thus, does not require additional data beyond past HPS methods.

3.4.1 IPMAN-R

Most HPS methods are trained with a mix of direct supervision using 3D datasets [37, 56, 81] and 2D reprojection losses using image datasets [4, 39, 53]. The 3D losses, however, are calculated in the camera frame, ignoring scene information and physics. IPMAN-R extends HMR [42] with our intuitive-physics terms; see Fig. 3 for the architecture. For training, we use the *known* camera coordinates and the world ground plane in 3D datasets.

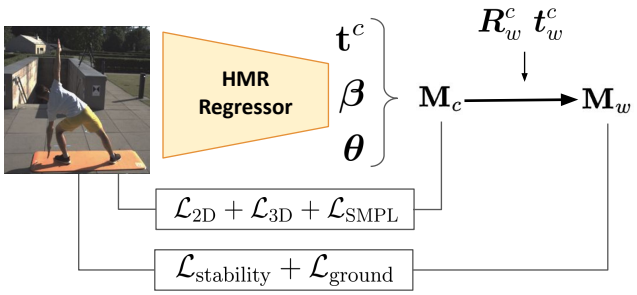

Figure 3. IPMAN-R architecture. First, the HMR regressor estimates camera translation and SMPL parameters for an input image. These parameters are used to generate the SMPL mesh in the camera frame, $\boldsymbol{M}_c$. To transform the mesh from camera into world coordinates ($\boldsymbol{M}_c \rightarrow \boldsymbol{M}_w$), IPMAN-R uses the ground-truth camera rotation, $\boldsymbol{R}_w^c$, and translation, $\boldsymbol{t}_w^c$. The IP losses, $\mathcal{L}_{\text{ground}}$ and $\mathcal{L}_{\text{stability}}$, are applied on the mesh in the world coordinate system.

As described in Sec. 3.1 (paragraph "Camera"), HMR infers the camera translation, $\mathbf{t}^c$, and SMPL parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, in the camera coordinates assuming $\mathbf{R}^c = \boldsymbol{I}_3$ and $\mathbf{t}^b = \mathbf{0}$. Ground truth 3D joints and SMPL parameters are used to supervise the inferred mesh $\boldsymbol{M}_c$ in the *camera frame*. However, 3D datasets also provide the ground, albeit in the world frame. To leverage the known ground, we transform the predicted body orientation, $\mathbf{R}^b$, to world coordinates using the ground-truth camera rotation, $\mathbf{R}_w^c$, as $\mathbf{R}_w^b = \mathbf{R}_w^{c\top}\mathbf{R}^b$. Then, we compute the body translation in world coordinates as $\mathbf{t}_w^b = -\mathbf{t}^c + \mathbf{t}_w^c$. With the predicted mesh and ground plane in world coordinates, we add the IP terms, $\mathcal{L}_{\text{stability}}$ and $\mathcal{L}_{\text{ground}}$, for HPS training as follows:

$$\mathcal{L}_{\text{IPMAN-R}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t}^c) = \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{\text{SMPL}}\mathcal{L}_{\text{SMPL}} + \lambda_{\text{s}}\mathcal{L}_{\text{stability}} + \lambda_{\text{g}}\mathcal{L}_{\text{ground}}, \tag{8}$$

where λ_{s} and λ_{g} are the weights for the respective IP terms. For training (data augmentation, hyperparameters, etc), we follow Kolotouros et al. [47]; for more details see Sup. Mat.

3.4.2 IPMAN-O

To fit SMPL-X to 2D image keypoints, SMPLify-XMC [59] initializes the fitting process by exploiting the self-contact and global-orientation of a known/presented 3D mesh. We posit that the presented pose contains further information, such as stability, pressure and contact with the ground-plane. IPMAN-O uses this insight to apply stability and ground

contact losses. The IPMAN-O objective is:

$$E_{\text{IPMAN-O}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Phi}) = E_{J2D} + \lambda_{\beta} E_{\beta} + \lambda_{\theta_h} E_{\theta_h} + \lambda_{\tilde{\theta}_b} E_{\tilde{\theta}_b} + \lambda_{\tilde{C}} E_{\tilde{C}} + \lambda_s E_{\text{stability}} + \lambda_g E_{\text{ground}}. \quad (9)$$

$\boldsymbol{\Phi}$ denotes the camera parameters: rotation $\mathbf{R}^c$, translation $\mathbf{t}^c$, and focal length, (f_x, f_y). E_{J2D} is a 2D joint loss, E_{β} and E_{θ_h} are L_2 body shape and hand pose priors. $E_{\tilde{\theta}_b}$ and $E_{\tilde{C}}$ are pose and contact terms w.r.t. the presented 3D pose and contact (see [59] for details). E_S and E_G are the stability and ground contact losses from Sec. 3.3. Since the estimated mesh is in the same coordinate system as the presented mesh and the ground-plane, we directly apply IP losses without any transformations. For details see Sup. Mat.

4. Experiments

4.1. Training and Evaluation Datasets

Human3.6M [37]. A dataset of 3D human keypoints and RGB images. The poses are limited in terms of challenging physics, focusing on common activities like walking, discussing, smoking, or taking photos.

RICH [35]. A dataset of videos with accurate marker-less motion-captured 3D bodies and 3D scans of scenes. The images are more natural than Human3.6M and Fit3D [20]. We consider sequences with meaningful body-ground interaction. For the list of sequences, see Sup. Mat.

Other datasets. Similar to [47], for training we use 3D keypoints from MPI-INF-3DHP [56] and 2D keypoints from image datasets such as COCO [53], MPII [4] and LSP [39].

4.1.1 MoCap Yoga (MoYo) Dataset

We capture a trained Yoga professional in 200 highly complex poses (see Fig. 4) using a synchronized MoCap system, pressure mat, and a multi-view RGB video system with 8 static, calibrated cameras; for details see Sup. Mat. The dataset contains $\sim$ 1.75M RGB frames in 4K resolution with ground-truth SMPL-X [63], pressure and CoM. Compared to the Fit3D [20] and PosePrior [1] datasets, MoYo is more challenging; it has extreme poses, strong self-occlusion, and significant body-ground and self-contact.

4.2. Evaluation Metrics

We use standard 3D HPS metrics: The Mean Per-Joint Position Error (MPJPE), its Procrustes Aligned version (PA-MPJPE), and the Per-Vertex Error (PVE) [62].

BoS Error (BoSE). To evaluate stability, we propose a new metric called BoS Error (BoSE). Following the definition of stability (Sec. 3.3) we define:

$$\text{BoSE} = \begin{cases} 1 & g(\bar{\mathbf{m}}) \in \mathcal{C}(C) \\ 0 & g(\bar{\mathbf{m}}) \notin \mathcal{C}(C) \end{cases} \quad (10)$$

where $\mathcal{C}(C)$ is the convex hull of the gravity-projected contact vertices for $\tau = 10$ cm. For efficiency reasons, we formulate this computation as the solution of a convex system via interior point linear programming [3]; see Sup. Mat.

4.3. IPMAN Evaluation

IPMAN-R. We evaluate our regressor, IPMAN-R, on RICH and H3.6M and summarize our results in Tab. 1. We refer to our regression baseline as HMR* which is HMR trained on the same datasets as IPMAN-R. Since we train with paired 3D datasets, we do not use HMR's discriminator during training. Both IP terms individually improve upon the baseline method. Their joint use, however, shows the largest improvement. For example, on RICH the MPJPE improves by 3.5mm and the PVE by 2.5mm. It is particularly interesting that IPMAN-R improves upon the baseline on H3.6M, a dataset with largely dynamic poses and little body-ground contact. We also significantly outperform ($\sim$ 12%) the MPJPE of optimization approaches that use the ground plane, Zou et al. [110] (69.9 mm) and Zanfir et al. [98] (69.0 mm), on H3.6M. Some video-based methods [49, 96] achieve better MPJPE (56.7 and 52.5 resp.) on H3.6M. However, they initialize with a stronger kinematic predictor [45, 50] and require video frames as input. Further, they use heuristics to estimate body weight and non-physical residual forces to *correct* for contact estimation errors. In contrast, IPMAN is a single-frame method, models complex full-body pressure and does not rely on approximate body weight to compute CoM. Qualitatively, Fig. 5 (top) shows that IPMAN-R's reconstructions are more stable and contain physically-plausible body-ground contact. While HMR is not SOTA, it is simple, isolating the benefits of our new IP formulation. These terms can also be added to methods with more modern backbones and architectures.

IPMAN-O. Our optimization method, IPMAN-O, also improves upon the baseline optimization method, SMPLify-XMC, on all evaluation metrics (see Tab. 2). We note that adding $L_{\text{stability}}$ independently improves the PVE, but not joint metrics (PA-MPJPE, MPJPE) and BoSE. This can be explained by the dependence of our IP terms on the relative position of the mesh surface to the ground-plane. Since joint metrics do not capture surfaces, they may get worse. Similar trends on joint metrics have been reported in the context of hand-object contact [29, 79] and body-scene contact [27]. We show qualitative results in Fig. 5 (bottom). While both SMPLify-XMC [59] and IPMAN-O achieve similar image projections, another view reveals that our results are more stable and physically plausible w.r.t. the ground.

4.4. Pressure, CoP and CoM Evaluation

We evaluate our estimated pressure, CoP and CoM against the MoYo ground truth. For pressure evaluation, we measure Intersection-over-Union (IoU) between our esti-

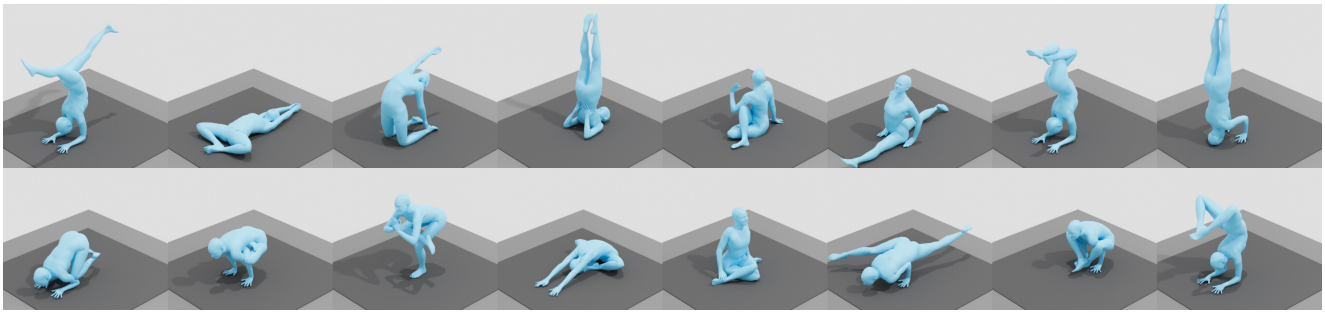

Figure 4. Representative examples illustrating the variation and complexity of 3D pose and body-ground contact in our new MoYo dataset.

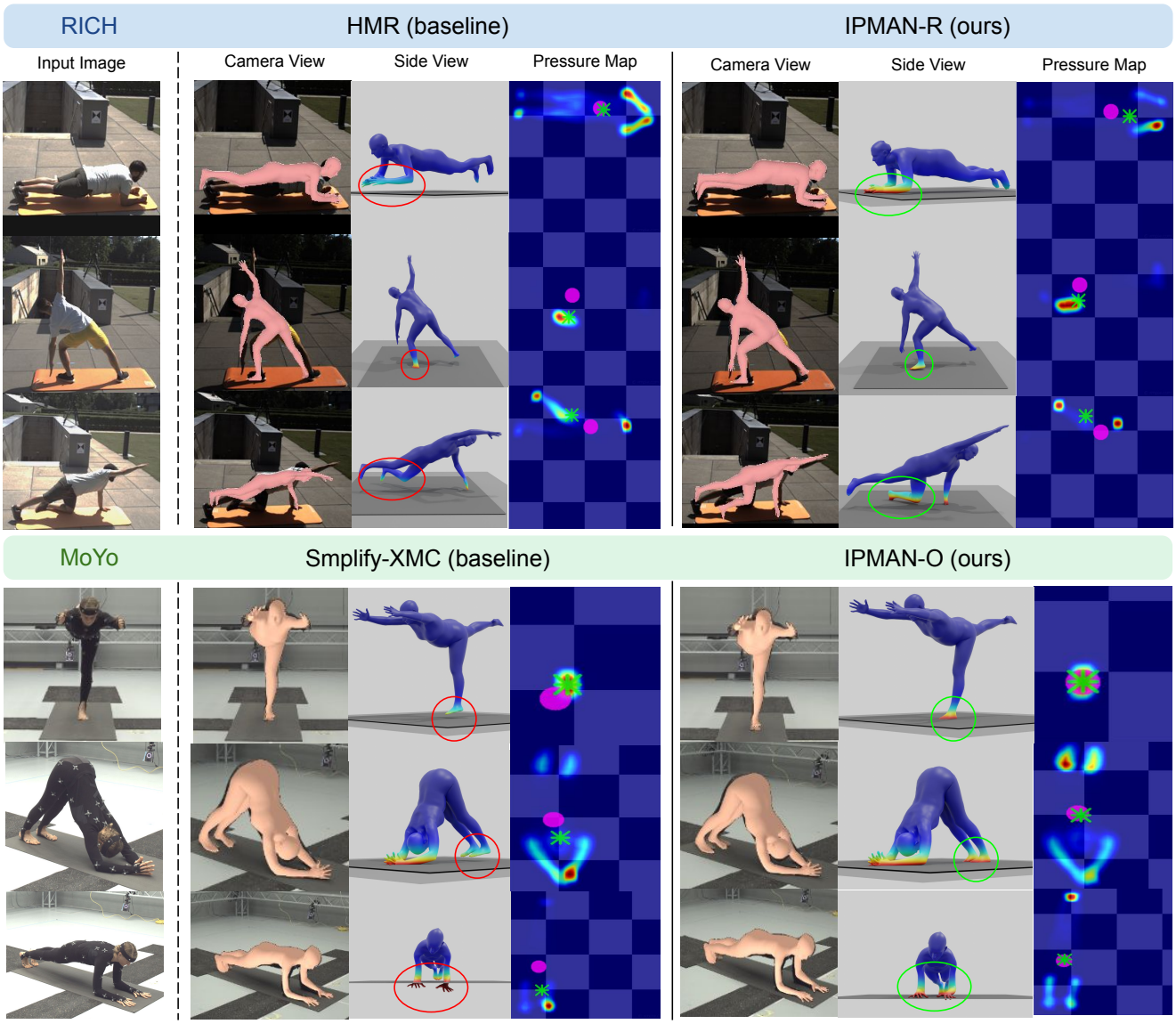

Figure 5. Qualitative evaluation of IPMAN-R and IPMAN-O on the RICH and MoYo datasets. The first column shows the input images of a subject doing various sports poses. The second and third block of columns show the baseline's and our results, respectively. In each block, the first image shows the estimated mesh overlayed on the image (camera view), the second image shows the estimated mesh in the world frame (side view), and the last image shows the estimated pressure map with the CoM (in pink) and the CoP (in green).

Method	RICH				Human3.6M	
	MPJPE ↓	PAMPJPE ↓	PVE ↓	BoSE (%) ↑	MPJPE ↓	PAMPJPE ↓
PhysCap [74]	-	-	-	-	113.0	68.9
DiffPhy [21]	-	-	-	-	81.7	55.6
Zou et al. [110]	-	-	-	-	69.9	-
Xie et al. [89]	-	-	-	-	68.1	-
VIBE [45]	-	-	-	-	61.3	43.1
Simpoe [96]	-	-	-	-	56.7	41.6
D&D [49]	-	-	-	-	**52.5**	**35.5**
HMR [42]	-	-	-	-	88.0	56.8
Zanfir et al. [98]	-	-	-	-	69.0	-
SPIN [47]	112.2	71.5	129.5	54.7	62.3	41.9
PARE [46]	107.0	73.1	125.0	74.4	-	-
CLIFF [51]	107.0	67.2	122.3	67.6	81.4	52.1
Finetuning on Human3.6M						
HMR* [42]	-	-	-	-	62.1	41.6
IPMAN-R (Ours)	-	-	-	-	**60.7 (-1.4)**	**41.1 (-0.5)**
Finetuning on all datasets						
HMR* [42]	82.5	48.3	92.4	62.0	61.6	41.9
HMR* [42]+$\mathcal{L}$ground	80.9	47.8	89.9	66.5	61.9	41.8
HMR* [42]+$\mathcal{L}$stability	81.0	**47.5 (-0.8)**	90.8	69.6	61.2	41.9
IPMAN-R (Ours)	**79.0 (-3.5)**	47.6	**89.9 (-2.5)**	**71.2 (+9.2)**	**60.6 (-1.0)**	**41.8 (-0.1)**

Table 1. Top to Bottom: Comparisons with video-based and single-frame regression methods. IPMAN-R outperforms the single-frame baselines across all benchmarks. * indicates training hyperparameters and datasets are identical to IPMAN-R. All units are in mm except BoSE. Bold denotes best results (per category), and parentheses show improvement over the baseline. **Zoom in**

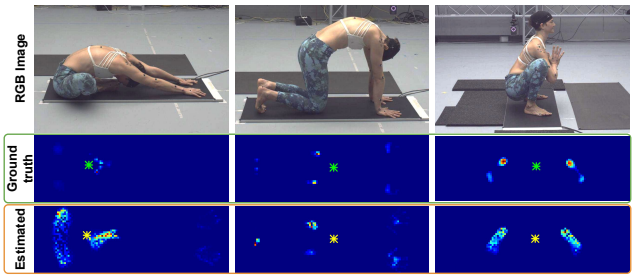

Figure 6. Qualitative comparison of estimated vs the ground-truth pressure. The ground-truth CoP is shown in green and the estimated CoP is shown in yellow. Pressure heatmap colors as per Fig. 2.

mated and ground-truth pressure heatmaps. We also compute the CoP error as the Euclidean distance between estimated and ground-truth CoP. We obtain an IoU of 0.32 and a CoP error of 57.3 mm. Figure 6 shows a qualitative visualization of the estimated pressure compared to the ground truth. For CoM evaluation, we find a 53.3 mm difference between our pCoM and the CoM computed by the commercial software, Vicon Plug-in Gait. Unlike Vicon's estimate, our pCoM does not require anthropometric measurements and takes into account the full 3D body shape. For details about the evaluation protocol and comparisons with alternative CoM formulations, see Sup. Mat.

Physics Simulation. To evaluate stability, we run a post-hoc physics simulation in "Bullet" [10] and measure the displacement of the estimated meshes; a small displacement denotes a stable pose. IPMAN-O produces 14.8% more stable bodies than the baseline [59]; for details see Sup. Mat.

Method	MoYo			
	MPJPE ↓	PAMPJPE ↓	PVE ↓	BoSE (%) ↑
SMPLify-XMC [59]	75.3	36.5	16.8	98.0
SMPLify-XMC [59]+$\mathcal{L}$ground	73.3	36.2	14.5	98.2
SMPLify-XMC [59]+$\mathcal{L}$stability	88.5	38.6	15.3	97.8
IPMAN-O (Ours)	**71.9 (-3.4)**	**34.3 (-2.2)**	**11.4 (-5.4)**	**98.6 (+0.5)**

Table 2. Evaluation of IPMAN-O and SMPLify-XMC [59] (optimization-based) on MoYo. Bold shows the best performance, and parentheses show the improvement over SMPLify-XMC.

5. Conclusion

Existing 3D HPS estimation methods recover SMPL meshes that align well with the input image, but are often physically implausible. To address this, we propose IPMAN, which incorporates *intuitive-physics* in 3D HPS estimation. Our IP terms encourage stable poses, promote realistic floor support, and reduce body-floor penetration. The IP terms exploit the interaction between the body CoM, CoP, and BoS – key elements used in stability analysis. To calculate the CoM of SMPL meshes, IPMAN uses on a novel formulation that takes part-specific mass contributions into account. Additionally, IPMAN estimates proxy *pressure* maps directly from images, which is useful in computing CoP. IPMAN is simple, differentiable, and compatible with both regression and optimization methods. IPMAN goes beyond previous physics-based methods to reason about arbitrary full-body contact with the ground. We show that IPMAN improves both regression and optimization baselines across all metrics on existing datasets and MoYo. MoYo uniquely comprises synchronized multi-view video, SMPL-X bodies in complex poses, and measurements for pressure maps and body CoM. Qualitative results show the effectiveness of IPMAN in recovering physically plausible meshes.

While IPMAN addresses body-floor contact, future work should incorporate general body-scene contact and diverse supporting surfaces by integrating 3D scene reconstruction. In this work, the proposed IP terms are designed to help static poses and we show that they do not hurt dynamic poses. However, the large body of biomechanical literature analyzing dynamic poses could be leveraged for activities like walking, jogging, running, etc. It would be interesting to extend IPMAN beyond single-person scenarios by exploiting the various physical constraints offered by multiple subjects.

Acknowledgements. We thank T. Alexiadis, T. McConnell, C. Gallatz, M. Höschle, S. Polikovsky, C. Mendoza, Y. Fincan, L. Sanchez and M. Safroshkin for data collection, G. Becherini for MoSh++, Z. Fang, V. Choutas and all of Perceiving Systems for fruitful discussions. This work was funded by the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and in part by the German Federal Ministry of Education and Research (BMBF), Tübingen AI Center, FKZ: 01IS18039B.

Disclosure. https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015. 3, 6

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 3

[3] Erling D. Andersen and Knud D. Andersen. The Mosek interior point optimizer for linear programming: An implementation of the homogeneous algorithm. In *High Performance Optimization*, 2000. 6

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 5, 6

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *Transactions on Graphics (TOG)*, 24:408–416, 2005. 3

[6] Michael Barnett-Cowan, Roland W. Fleming, Manish Singh, and Heinrich H. Bülthoff. Perceived object stability depends on multisensory estimates of gravity. *PLOS ONE*, 6(4):1–5, 2011. 2

[7] 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)*, volume 9909, pages 561–578, 2016. 3

[8] Marcus A. Brubaker, David J. Fleet, and Aaron Hertzmann. Physics-based person tracking using the anthropomorphic walker. *International Journal of Computer Vision (IJCV)*, 87(1–2):140–155, 2010. 3

[9] Marcus A. Brubaker, Leonid Sigal, and David J. Fleet. Estimating contact dynamics. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2389–2396, 2009. 3

[10] Bullet real-time physics simulation. https://pybullet.org. 1, 8

[11] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 3

[12] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. June 2023. 3

[13] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, volume 12355, pages 20–40, 2020. 3

[14] Henry M. Clever, Zackory M. Erickson, Ariel Kapusta, Greg Turk, C. Karen Liu, and Charles C. Kemp. Bodies at rest: 3D human pose and shape estimation from a pressure image using synthetic data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6214–6223, 2020. 3, 4

[15] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11875–11885, 2021. 3

[16] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. Revitalizing optimization for 3D human pose and shape estimation: A sparse constrained formulation. In *International Conference on Computer Vision (ICCV)*, pages 11437–11446, 2021. 3

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[18] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 3

[19] Mihai Fieraru, Mihai Zanfir, Teodor Alexandru Szente, Eduard Gabriel Bazavan, Vlad Olaru, and Cristian Sminchisescu. REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 3, 5

[20] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. AIFit: Automatic 3D human-interpretable feedback models for fitness training. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9919–9928, 2021. 6

[21] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3D human motion reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13180–13190, 2022. 3, 8

[22] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3D human pose from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13096–13105, 2022. 3

[23] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7450–7459, 2019. 3

[24] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10472–10481, 2021. 3

[25] Riza Alp Güler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10876–10886, 2019. 3

[26] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020. 5

[27] 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. 3, 6

[28] 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. 3

[29] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 5, 6

[30] Havok: Customizable, fully multithreaded, and highly optimized physics simulation. http://www.havok.com. 1

[31] Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S. Sukhatme. NeuralSim: Augmenting differentiable simulators with neural networks. In *International Conference on Robotics and Automation (ICRA)*, pages 9474–9481, 2021. 3

[32] At L. Hof. The equations of motion for a standing human reveal three mechanisms for balance. *Journal of Biomechanics*, 40(2):451–457, 2007. 2, 4

[33] At L. Hof. The "extrapolated center of mass" concept suggests a simple control of balance in walking. *Human movement science*, 27(1):112–125, 2008. 2, 4

[34] At L. Hof, M. G. J. Gazendam, and Sinke W. E. The condition for dynamic stability. *Journal of Biomechanics*, 38(1):1–8, 2005. 4

[35] Chun-Hao Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13264–13275, 2022. 2, 3, 6

[36] Leslie Ikemoto, Okan Arikan, and David Forsyth. Knowing when to put your foot down. In *Symposium on Interactive 3D Graphics (SI3D)*, page 49–53, 2006. 3

[37] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014. 2, 5, 6

[38] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5578–5587, 2020. 3

[39] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2010. 5, 6

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

[41] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2, 3

[42] 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. 3, 5, 8

[43] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. *Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, 2019. 3

[44] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1715, 2022. 3

[45] 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 5252–5262, 2020. 3, 6, 8

[46] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 1, 8

[47] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 3, 5, 6, 8

[48] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4496–4505, 2019. 3

[49] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&D: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 6, 8

[50] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 3, 6

[51] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, volume 13665, pages 590–606, 2022. 1, 3, 8

[52] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 3

[53] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014. 5, 6

[54] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2, 3

[55] Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma, Tomás Palacios, Antonio Torralba, and Wojciech Matusik. Intelligent carpet: Inferring 3D human pose from tactile signals. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11255–11265, 2021. 3

[56] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. *International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 3, 5, 6

[57] 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. *Transactions on Graphics (TOG)*, 36(4):44:1–44:14, 2017. 3

[58] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, volume 12352, pages 752–768, 2020. 3

[59] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021. 1, 2, 3, 4, 5, 6, 8

[60] NVIDIA PhysX: A scalable multi-platform physics simulation solution. https://developer.nvidia.com/physx-sdk. 1

[61] Yi-Chung Pai. Movement termination and stability in standing. *Exercise and sport sciences reviews*, 31(1):19–25, 2003. 2, 4

[62] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 6

[63] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. 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

[64] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3

[65] 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)*, pages 11468–11479, 2021. 3

[66] 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)*, volume 12350, pages 71–87, 2020. 1, 3

[67] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015. 4

[68] Grégory Rogez, James Steven Supancic, and Deva Ramanan. Understanding everyday hands in action from RGB-D images. In *International Conference on Computer Vision (ICCV)*, pages 3889–3897, 2015. 2, 4

[69] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 1749–1759, 2021. 3

[70] Nadine Rueegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. BITE: Beyond priors for improved three-D dog pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[71] Jesse Scott, Bharadwaj Ravichandran, Christopher Funk, Robert T Collins, and Yanxi Liu. From image to stability: Learning dynamics from human pose. In *European Conference on Computer Vision (ECCV)*, volume 12368, pages 536–554, 2020. 2, 3, 4

[72] 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. *Transactions on Graphics (TOG)*, 40(1):1:1–1:15, 2021. 3

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

[74] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *Transactions on Graphics (TOG)*, 39(6):235:1–235:16, 2020. 1, 2, 3, 8

[75] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *International Conference on Computer Vision (ICCV)*, pages 11179–11188, 2021. 1, 3

[76] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *International Conference on Computer Vision (ICCV)*, pages 5348–5357, 2019. 3

[77] Yating Tian, Hongwen Zhang, Yebin Liu, and limin Wang. Recovering 3D human mesh from monocular images: A survey. *arXiv:2203.01923*, 2022. 3

[78] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit K. Agrawal. PoseNet3D: Learning temporally consistent 3D human pose via knowledge distillation. In *International Conference on 3D Vision (3DV)*, pages 311–321, 2020. 3

[79] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, 118:172–193, 2016. 6

[80] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *International Conference on Computer Vision (ICCV)*, pages 9720–9729, 2021. 3

[81] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume 11214, pages 614–631, 2018. 5

[82] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Physical simulation for probabilistic motion tracking. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 3

[83] Eric W. Weisstein. Triangle point picking. https://mathworld.wolfram.com/TrianglePointPicking.html, 2014. From MathWorld – A Wolfram Web Resource. 4

[84] 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, 2020. 3

[85] David A. Winter. *A.B.C. (Anatomy, Biomechanics and Control) of balance during standing and walking*. Waterloo Biomechanics, 1995. 2, 4, 5

[86] David A. Winter. Human balance and posture control during standing and walking. *Gait & Posture*, 3(4):193–214, 1995. 2, 5

[87] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 3

[88] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. MonoClothCap: Towards temporally coherent clothing capture from monocular RGB video. In *International Conference on 3D Vision (3DV)*, pages 322–332, 2020. 3

[89] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*, pages 11532–11541, 2021. 3, 8

[90] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[91] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal Integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[92] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 2, 3

[93] Masanobu Yamamoto and Katsutoshi Yagishita. Scene constraints-aided tracking of human body. In *Computer Vision and Pattern Recognition (CVPR)*, pages 151–156, 2000. 3

[94] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[95] Ye Yuan and Kris Kitani. 3D ego-pose estimation via imitation learning. In *European Conference on Computer Vision (ECCV)*, volume 11220, pages 735–750, 2018. 2

[96] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7159–7169, 2021. 1, 2, 3, 6, 8

[97] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, pages 465–481, 2020. 3

[98] 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. 3, 6, 8

[99] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. SmoothNet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision (ECCV)*, volume 13665, pages 625–642, 2022. 3

[100] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[101] Cha Zhang and Tsuhan Chen. Efficient feature extraction for 2d/3d objects in mesh representation. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 3, pages 935–938. IEEE, 2001. 4

[102] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11426–11436, 2021. 1, 3

[103] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *Computer Vision and Pattern Recognition (CVPR)*, pages 546–556, 2021. 3

[104] 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)*, volume 12357, pages 34–51, 2020. 3

[105] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4D human body capture in 3D scenes. In *International Conference on Computer Vision (ICCV)*, pages 11343–11353, 2021. 3

[106] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7374–7383, 2020. 3

[107] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah.

Deep learning-based human pose estimation: A survey. *arXiv:2012.13392*, 2022. 3

[108] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 3

[109] 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)*, pages 4811–4822, 2021. 3

[110] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 459–468, 2020. 3, 6, 8