# 3D Human Pose Estimation via Intuitive Physics
## **Supplemental Material**

## 1. MoCap Yoga Dataset (MoYo)

We capture a trained yoga professional in a MoCap studio with 54 Vicon Vantage V16 infrared cameras capable of tracking body markers as small as 3mm in diameter. The Vicon system was synchronized with 8 RGB cameras recording at 4112x3008 resolution and a Zebris FDM pressure measurement mat. The pressure mat offers a sensor resolution of 1.4sensors/cm$^2$ and can capture pressure in 10-1200 kPa range. Ground-truth SMPL-X [22] parameters are recovered from the MoCap data using MoSh++ [17]. A total of 200 yoga sequences were recorded at 30fps. The yoga poses we selected include all poses in the Yoga-82 dataset [25] as well as their variations. The T-SNE [24] plot in Fig. S.1 shows that the poses contained in MoYo are highly diverse and cover areas in the space of human poses not well represented in existing datasets [8, 17, 19, 21].

To compute a reference CoM, we use the commercially available tool, *Plug-in Gait (PiG)* from Vicon. PiG requires a-priori known anthropometric measurements (e.g. height, weight, shoulder offset, knee width, etc) and computes: (1) bone joints from a known marker topology, (2) per-bone mass as a proportion of body mass, (3) per-bone CoM as a proportion of each bone's length, and (4) whole-body CoM as a weighted average of per-bone CoMs. In contrast, our pCoM does not require anthropometric measurements and takes into account the full 3D body shape.

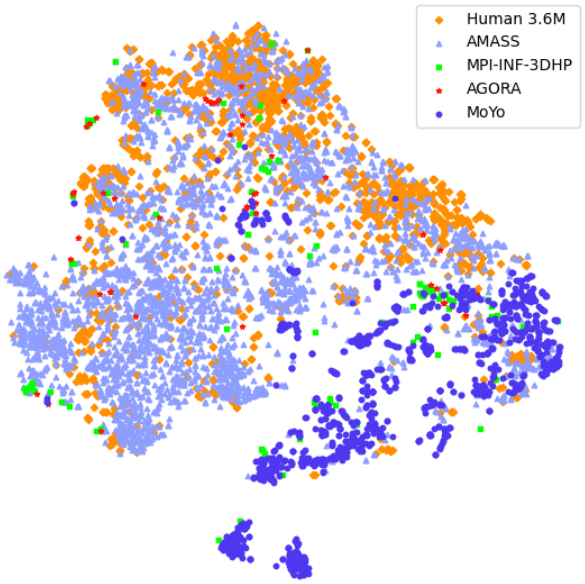


Figure S.1. The distribution of poses in MoYo and existing MoCap datasets are visualized after T-SNE dimension reduction.

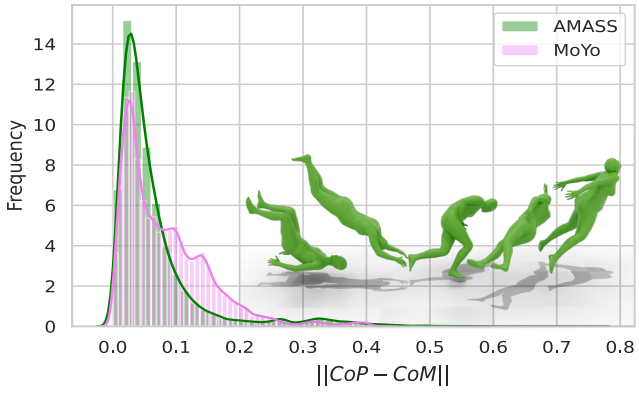


Figure S.2. Distribution of $\mathcal{L}_{\text{stability}}$ in AMASS and MoYo. Both peak at $\sim 0$, motivating using an $L_2$ formulation. Bottom right: Unstable long-tail poses from AMASS.

## 2. Method

### 2.1. Stability Loss

The suggested classic definition uses a binary stability criterion, i.e., the CoM "just" projects either inside or outside the BoS. This is discontinuous with sparse gradients.

Since CoP lies inside BoS, our L2 loss is a "soft" version that approximates the classic definition, but has two key benefits: (1) it is continuous and fully differentiable, and, (2) it informs about the *degree* of instability. The distribution of $\mathcal{L}_{\text{stability}}$ in Fig. S.2 for both AMASS and MoYo datasets peak at $\sim 0$, motivating using an $L_2$ formulation.

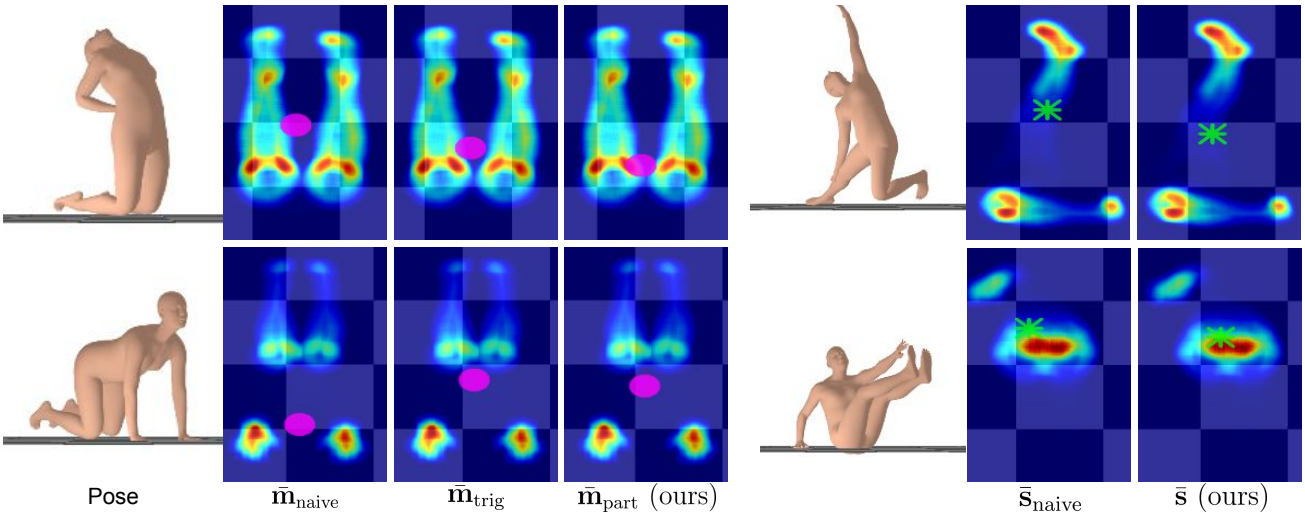


Figure S.3. Gravity-projections of different formulations of CoM (shown with pink) and CoP (shown with green) are shown with estimated pressure maps. Our proposed pCoM captures more accurate body mass distribution because it takes into account part-specific mass contributions. Similarly, our CoP leverages the pressure maps rather than binary contact.

### 2.2. Elements of Stability Analysis: Alternative formulations

Computation of the "Center of Mass", CoM, must be efficient and differentiable. The CoM could be naively approximated as the mean vertex position of a mesh:

$$\bar{\mathbf{m}}_{\text{naive}} = \frac{1}{N_V} \sum_{i=1}^{N_V} \boldsymbol{v}_i. \tag{S.1}$$

However, the SMPL and the SMPL-X body models have a non-uniform vertex distribution across the surface. There are a disproportionate number of vertices on the face and hands compared to the body. For instance, roughly half of SMPL-X's vertices lie on the head. Consequently, $\bar{\mathbf{m}}_{\text{naive}}$ is dominated by face and hand vertices.

A better formulation is the mean of uniformly sampled surface points:

$$\bar{\mathbf{m}}^{u}_{\text{naive}} = \frac{1}{N_U} \sum_{i=1}^{N_U} \boldsymbol{v}_i. \tag{S.2}$$

Another formulation computes the average of the mesh triangle face centroids weighted by the face area:

$$\bar{\mathbf{m}}_{\text{trig}} = \frac{\sum_{i=1}^{N_F} A_i \bar{F}_i}{\sum_{i=1}^{N_F} A_i}, \tag{S.3}$$

where $A_i$ denotes the area and $\bar{F}_i = \frac{1}{3}(\boldsymbol{v}_{i_1}^\top + \boldsymbol{v}_{i_2}^\top + \boldsymbol{v}_{i_3}^\top)$ the centroid of face $\boldsymbol{F}_i$. The problem with these approaches is that they assume that mass, $M$, is proportional to surface area, $S$, which is a poor approximation.

Our proposed pCoM formulation addresses this by (1) uniformly sampling vertices on the SMPL mesh and (2) taking part-specific mass contributions into account. Our pCoM computes mass from volume, $\mathcal{V}$, via the standard density equation, $M = \rho\mathcal{V}$. Tab. S.1 compares the CoM error across different formulations of CoM w.r.t. ground-truth CoM obtained using Vicon PiG. pCoM significantly outperforms all baselines. Figure S.3 shows an intuitive qualitative comparison between all formulations of CoM.

| | $\bar{\mathbf{m}}_{\text{naive}}$ | $\bar{\mathbf{m}}^{u}_{\text{naive}}$ | $\bar{\mathbf{m}}_{\text{trig}}$ | pCoM ($\bar{m}$) |
|---|---|---|---|---|
| CoM error ↓ | 264.1 mm | 68.5 mm | 70.0 mm | **53.3 mm** |

Table S.1. Comparison of various CoM formulations.

Similarly, for "Center of Pressure" (CoP), a simple heuristic used in previous works detects binary contact by thresholding body vertices using their Euclidean distance from the ground plane. However, such contact lacks information about the pressure distribution and assigns equal weight to all contact vertices. Moreover, binary contact is not differentiable and is therefore generally used at test-time [4, 26, 29, 30] or for data preprocessing [5, 31], not during training. In contrast, our CoP formulation is fully differentiable and takes the inferred pressure distribution of the body-floor contact into account. As shown in Fig. S.3, the naive CoP suffers from equally weighting all binary-contact whereas our CoP better represents the pressure profile of the body-ground contact.

### 2.3. Ablation of ground losses

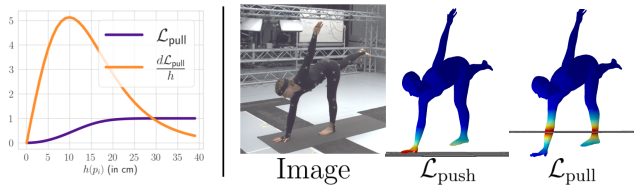


Figure S.4. *Left*: The gradient of $\mathcal{L}_{\text{pull}}$ decays gradually with $h(p_i)$; vertices with $h(p_i) \geq 20$cm contribute minimally to back-propagation. *Right*: Effect of $\mathcal{L}_{\text{push}}$ and $\mathcal{L}_{\text{pull}}$; high $\mathcal{L}_{\text{push}}$ results in inaccurate floor contact, high $\mathcal{L}_{\text{pull}}$ results in penetrations.

Instead of having a threshold to restrict $\mathcal{L}_{\text{pull}}$ only to vertices close to the ground, we chose a soft version of the loss to ensure full differentiability. However, as shown in Fig. S.4 (left), the loss gradient decays with height and vertices with $h(v_i) \geq 15$ cm contribute minimally during back-propagation. Further, we study the impact of $\mathcal{L}_{\text{push}}$ and $\mathcal{L}_{\text{pull}}$ in Tab. S.2 and Fig. S.4-right. The terms complement each other and are more effective when used jointly ($\mathcal{L}_{\text{ground}}$).

| **Method** | **MPJPE** ↓ | **PMPJPE** ↓ | **PVE** ↓ |
|---|---|---|---|
| HMR* [12] | 82.5 | 48.2 | 92.3 |
| HMR* [12]+$\mathcal{L}_{\text{push}}$ | 85.4 | 49.0 | 96.6 |
| HMR* [12]+$\mathcal{L}_{\text{pull}}$ | 88.0 | 48.8 | 99.4 |
| HMR* [12]+$\mathcal{L}_{\text{ground}}$ | 80.9 | 47.8 | 89.9 |

Table S.2. Ablation for $\mathcal{L}_{\text{push}}$ and $\mathcal{L}_{\text{pull}}$ on the RICH [7] dataset.

## 3. Experiments

We integrate our intuitive-physics terms in both an optimization- and a regression-based method for three reasons: (1) the community heavily uses both method types, (2) our terms generalize and benefit both types, despite their differences, and (3) our terms also work with different body models; SMPL-X (used by IPMAN-O) and SMPL (used by IPMAN-R).

### 3.1. IPMAN Implementation Details

#### 3.1.1 IPMAN-R.

Similarly to previous methods [9, 13, 14, 20], we take the widely used HMR [12] architecture to analyze the effect of adding our proposed IP terms. Note that, while HMR is not the most recent method, it is widely used as a backbone. As such, it provides a consistent foundation for evaluation and comparison. Our goal here is to isolate and evaluate the effect of adding intuitive physics. Such terms should then be readily applicable to other HPS regression frameworks.

The HMR regressor estimates the camera translation $\mathbf{t}^c$ and SMPL parameters (pose, global orientation, and shape) in the camera coordinates assuming $\mathbf{R}^c = \boldsymbol{I}_3$ and $\mathbf{t}^b = \mathbf{0}$. We initialize the HMR model using pretrained weights provided by SPIN [14] and finetune both IPMAN-R and HMR on the same datasets; namely RICH [7], Human3.6M [8], MPI-INF-3DHP [19], COCO [15], MPII [2] and LSP [10, 11]. In the main paper, we call the baseline as HMR* which uses the same training datasets and hyperparameters as IPMAN-R, albeit with the exception of the proposed IP terms. We follow the same training schedule, data augmentation and hyperparameters as SPIN [14] but do not use in-the-loop optimization. We use the Adam optimizer with learning rate of $5e^{-5}$ and finetuning takes 3 epochs ($\sim$ 8 hours) on a Nvidia Tesla V100 GPU.

We set the hyperparameters $\alpha = 100$, $\gamma = 10$ for the per-vertex pressure $\rho_i$, $\alpha_1 = 1.0$, $\alpha_2 = 0.15$ for the $\mathcal{L}_{\text{pull}}$ term and $\beta_1 = 10.0$, $\beta_2 = 0.15$ for the $\mathcal{L}_{\text{push}}$ term. The loss weights are empirically determined to be $\lambda_s = 0.01$ and $\lambda_g = 0.01$. We borrow the same configuration as [14] for all remaining loss weights, namely $\lambda_{\text{2D}}$, $\lambda_{\text{3D}}$ and $\lambda_{\text{SMPL}}$.

RICH [7] contains sequences with an uneven ground-plane. For training IPMAN-R, we therefore sample a subset of the RICH dataset where subjects mainly interact with an even ground plane (see Tab. S.3). In the Train/Val sequences, we use camera 0 for validation and cameras 1-5 for training.

| Train/Val | Test |
|---|---|
| 'Pavallion_000_yoga2' | 'Pavallion_002_yoga1' |
| 'Pavallion_000_yoga1' | 'Pavallion_013_yoga2' |
| 'Pavallion_006_yoga1' | 'ParkingLot2_014_pushup2' |
| 'Pavallion_018_yoga1' | 'ParkingLot1_005_pushup1' |

Table S.3. Training, validation and test sequences in the RICH dataset containing an even ground.

#### 3.1.2 IPMAN-O.

For IPMAN-O, we extend the baseline optimization-based method SMPLify-XMC [20]. We use the same configuration as SMPLify-XMC and only add extra hyperparameters for the proposed IP terms. Both methods are initialized with the same presented pose from the MoYo dataset. We extract 2D keypoints from images using MediaPipe [16].

Same as IPMAN-R, we set the hyperparameters $\alpha = 70$, $\gamma = 10$ for the per-vertex pressure $\rho_i$, $\alpha_1 = 1.0$, $\alpha_2 = 0.15$ for the $\mathcal{L}_{\text{pull}}$ term and $\beta_1 = 10.0$, $\beta_2 = 0.15$ for $\mathcal{L}_{\text{push}}$ term. The loss weights are empirically determined to be $\lambda_s = 10000$ and $\lambda_g = 10000$.

### 3.2. Evaluation Metrics

#### 3.2.1 BoS Error (BoSE) calculation.

Recall that the "Base of Support" (BoS) is defined by the convex hull of the contact regions. Since computing this

can be computationally inefficient, we reformulate the BoSE computation to test if projection of the CoM, $g(\bar{\mathbf{m}}_{\text{part}})$, on the ground plane can be represented as a convex combination of the gravity-projected contact vertices $C$. To this end, we solve the linear equation system via standard linear programming using interior point methods [1]:

$$\min_{\mathbf{a}} \quad \|\mathbf{a}^\top C - \bar{\mathbf{m}}_{\text{part}}\| \tag{S.4}$$

$$\text{s.t.} \quad a_i \in \mathbf{a} \geq 0 \tag{S.5}$$

$$\sum a_i = 1 \tag{S.6}$$

where $\mathbf{a}^\top C = a_1\boldsymbol{c}_1 + \cdots + a_n\boldsymbol{c}_n$ for the points $\boldsymbol{c}_i$ in $C$. If the system has a solution, $g(\bar{\mathbf{m}}) \in \mathcal{C}(C)$ holds, otherwise $g(\bar{\mathbf{m}})$ is not in the convex hull of $C$, i.e. $g(\bar{\mathbf{m}}) \notin \mathcal{C}(C)$.

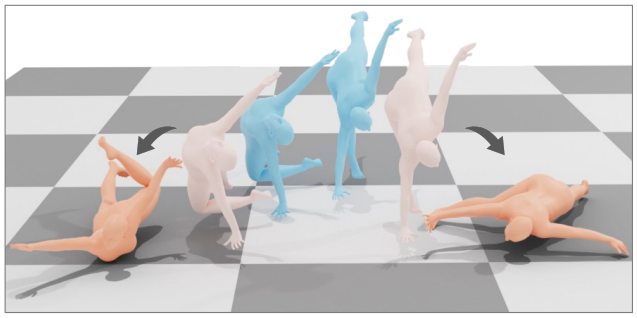

Figure S.5. Stability evaluation using the "Bullet" physics engine. Meshes produced by the baseline method [20] (in orange) topple but IPMAN-O's meshes (in cyan) remain stable after physics simulation.

### 3.3. Qualitative Results

Figures S.6 and S.7 show supplemental qualitative results for IPMAN-R and IPMAN-O, respectively.

## 4. Stability Evaluation via Physics Simulation

Current physics engines are incompatible with HPS methods, as they approximate SMPL bodies with rigid convex hulls and are non-differentiable. However, using them for posthoc stability evaluation of the estimated meshes is possible. Specifically, we evaluate IPMAN-O and SMPLify-XMC [20] by first, using V-HACD convex decomposition [18] of the estimated body meshes and then by simulating physics as in [6,23] via the "Bullet" physics engine [3]. We measure the displacement of the human mesh after 100 physics simulation steps; a small displacement denotes a stable pose and vice versa. IPMAN-O produces 14.8% more stable bodies than the baseline [20]; see Fig. S.5.

## 5. Evaluation of Biomechanical Elements

We use the pressure field defined in Eqn. 2 of the main paper to compute per-point pressure on the SMPL mesh. With this, the pressure heatmap is estimated by summing the per-point pressure projected to the ground-plane. Note that we recover relative pressure as we do not assume availability of ground-truth body mass or anthropometric measurements.

To measure the overlap of the inferred pressure heatmap w.r.t. the ground-truth, we compute the intersection-over-union (IOU) between the two. However, the ZEBRIS pressure sensor captures pressure measurements in the range 10-1200 KPa. Depending upon the contact area and the weight of the subject, some poses may fall outside this range. For instance, a person lying-down only exerts 1-5 kPa of pressure on the ground. To account for this, we tune the sensitivity of our pressure field for every pose and report mean of the best per-sample IOU.

We measure accuracy of our CoP by simply computing the Euclidean distance w.r.t. ground-truth. We call this as CoP error. Again, we report mean of the best CoP error after tuning the sensitivity of our inferred pressure field.

The CoM error is similar to the CoP error, albiet in 3D. It measures the Euclidean distance between the estimated and ground-truth CoM recovered from Vicon Plug-in Gait. Table S.4 presents summary results showing that our inferred pressure, CoP and CoM agrees with the ground-truth.

| | **Pressure** | | **CoM** |
|---|---|---|---|
| | **mIOU** | **CoP error (mm)** | **CoM error (mm)** |
| IPMAN (Ours) | 0.32 | 57.3 | 53.3 |

Table S.4. Quantitative evaluations of our estimated pressure, CoP and CoM w.r.t. ground-truth in MoYo.

## 6. IPMAN-O* (Extension of SMPLify-X).

To further explore the effect of our intuitive-physics terms, we extend the optimization method SMPLify-X [22] and name this IPMAN-O* (note that this is different from the main paper's IPMAN-O that extends SMPLify-XMC). We fit the SMPL-X body model to 2D image keypoints starting from mean pose and shape while exploiting the ground-truth ground plane. Adapted from SMPLify-X [22], we minimize the objective

$$\begin{aligned} E(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{t}^c) = & E_{J2D} + E_\theta + \lambda_\beta E_\beta + \lambda_\psi E_\psi + \\ & \lambda_\alpha E_\alpha + \lambda_\mathcal{C} E_\mathcal{C} + \\ & \lambda_s E_{\text{stability}} + \lambda_g E_{\text{ground}}. \end{aligned} \tag{S.7}$$

The energy term $E_{J2D}$ denotes the 2D re-projection error whereas the remaining terms $E_\theta = \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{\theta_h} E_{\theta_h}$ represent various priors for body, face, and hand pose. $E_\beta$, $E_\psi$, $E_\alpha$ and $E_\mathcal{C}$ are prior terms for body shape, expression, extreme bending and self-penetration (see [22] for details). $E_S$ and $E_G$ are the stability and ground contact losses. The results in Tab. S.5 show a clear improvement.

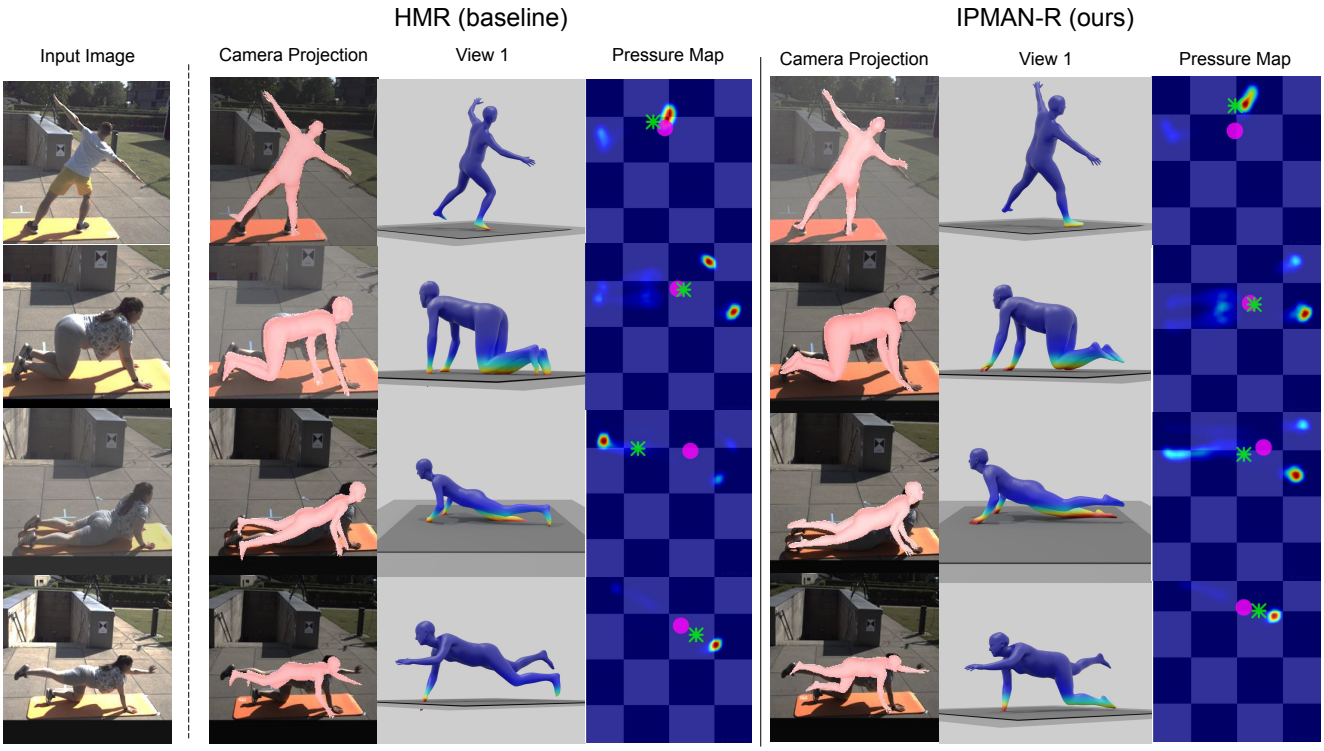


Figure S.6. Additional qualitative evaluation of IPMAN-R on RICH. The first column shows the input images of a subject doing various sports poses. The second and third block of columns show the results of HMR (baseline) and IPMAN-R, respectively. In each block, the first image shows the estimated mesh overlaid on the image. The next three images show different views of the estimated mesh in the world frame. The green sphere illustrates the CoM.

| Method | RICH [7] | | |
|---|---|---|---|
| | MPJPE ↓ | PVE ↓ | BoSE (%) ↑ |
| SMPLify-X [22] | 268.6 | 228.3 | 96.9 |
| IPMAN-O* (Ours) | **240.9** | **217.1** | **98.0** |

Table S.5. IPMAN-O* compared to the optimization method of [22] on RICH [7].

| Method | 3DPW [27] | |
|---|---|---|
| | MPJPE ↓ | PMPJPE ↓ |
| SPIN [14] | 97.2 | 59.6 |
| IPMAN-R (Ours) | **96.8** | **57.1** |

Table S.6. IPMAN-R compared to the regression method of [14] on 3DPW [27].

Note that SMPLify-X estimates the body's global orientation $\mathbf{R}^b$ and the camera translation $\mathbf{t}^c$, while camera rotation $\mathbf{R}^c$ and body translation $\mathbf{t}^b$ remain zero. In order to apply our IP terms, we use the ground-truth camera rotation $\mathbf{R}^c_w$ and translation $\mathbf{t}^c_w$ to transform the estimated mesh from camera to world coordinates. We empirically find that applying the IP terms to the final stage of optimization in SMPLify-X gives more accurate results than applying them to all stages. We hypothesize that this could be due to having a better body initialization before applying the IP terms.

## 7. Evaluation on 3DPW

3DPW [27] is an outdoor dataset containing pseudo ground-truth SMPL and camera parameters recovered using IMU sensors attached to the actors. As also noted in [28], we find that the ground plane in 3DPW is inconsistent. In fact, two subjects in the same scene can be supported by different ground-planes in the world coordinates. Additionally, 3DPW primarily contains dynamic poses like walking, climbing stairs, parkour, etc. Due to these reasons, 3DPW does not satisfy the core assumptions of IPMAN. Nevertheless, we report results on 3DPW to show that the IP terms do not degrade performance for such datasets; in fact, we see a slight improvement in performance as illustrated in Table S.6. This makes IPMAN applicable to everyday motion without needing special care.

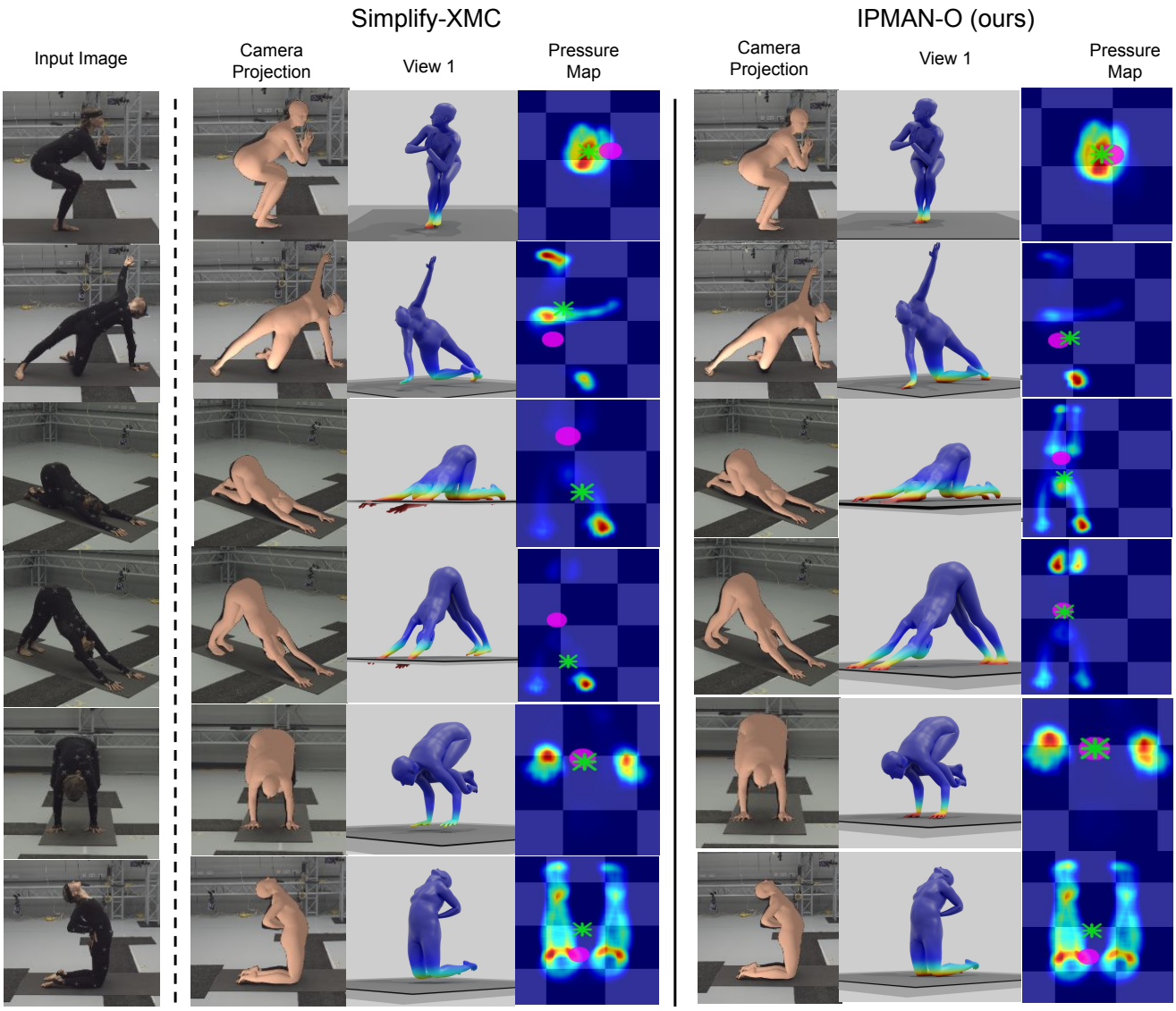


Figure S.7. Qualitative evaluation of IPMAN-O on MoYo. In the first column, the input images of a subject doing yoga poses. The second and third blocks show the results of the SMPLify-XMC and IPMAN-O respectively. In each block, the first and second column show the estimated mesh projected into the image and from a second view. The last images show the pressure map with the CoM (in pink) and the CoP (in green).

## References


[1] 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. 4

[2] 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. 3

[3] Bullet real-time physics simulation. https://pybullet.org. 4

[4] 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. 2

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

[6] 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. 4

[7] 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. 3, 5

[8] 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. 1, 3

[9] 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

[10] 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. 3

[11] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1465–1472, 2011. 3

[12] 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

[13] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021. 3

[14] 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

[15] 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. 3

[16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2019. 3

[17] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 1

[18] Khaled Mamou, E Lengyel, and A Peters. Volumetric hierarchical approximate convex decomposition. In *Game Engine Gems 3*, pages 141–158. AK Peters, 2016. 4

[19] 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. 1, 3

[20] 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. 3, 4

[21] 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. 1

[22] 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. 1, 4, 5

[23] 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. 4

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008. 1

[25] Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, and Shanmuganathan Raman. Yoga-82: A new dataset for fine-grained classification of human poses. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 1038–1039, 2020. 1

[26] 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. 2

[27] 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

[28] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11028–11039, 2022. 5

[29] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes – the importance of multiple scene constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 2

[30] 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. 2

[31] 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. 2