Reconstructing Humans with a Biomechanically Accurate Skeleton

Supplementary Material

In this Supplementary Material, we provide additional details and qualitative results that were not included in the main manuscript due to space constraints. The reader is also encouraged to watch our supplementary video for more temporal results of our approach.

S.1. More qualitative results

In Figure S.1 we provide more qualitative results of our approach. We choose a variety of scenarios, poses, viewpoints and activities, to demonstrate the robustness of our HSMR model. Additionally, the readers are encouraged to watch our supplementary video in our project page where one can appreciate the temporal consistency of our output.

S.2. Data generation

Initial SKEL parameter dataset. As we highlight in the main paper, there is no previous dataset of images with corresponding SKEL parameters. For this reason, we adopt existing image datasets with SMPL (pseudo) ground truth and we apply an offline optimization to convert the SMPL parameters to SKEL parameters. For this procedure, we follow the offline optimization proposed by SKEL [14]. The optimization aligns the location of the vertices and the joints for the two models while following multiple (four) stages to gradually improve fitting (*e.g.* avoiding scapula failure). For efficiency reasons, we modify the original code to a batch-wise version and set the batch size to 25k examples per batch. The remaining settings are kept the same with [14]. We apply the optimization to all the datasets used by HMR2.0 [7] and obtain the initial set of SKEL parameters. More specifically, we process images from Human3.6M [9], MPI-INF-3DHP [18], COCO [17], MPII [1], AI Challenger [22], AVA [8] and InstaVariety [13].

Moreover, unlike most past works using a neutral SMPL model [7, 12, 16], SKEL only provides a male and a female model. We adopt the male model for all our experiments and so our pseudo ground truth is compatible with the male SKEL model as well.

Finally, we note a SKEL-specific observation that affected our pipeline. For training pose estimation models (in 2D or 3D), we typically apply left-right flipping augmentation. For SMPL, we can produce the corresponding mirrored mesh (which is used for supervision), by applying a simple transformation to the pose parameters. However, for SKEL, we noticed that a similar transformation lead to a slight imperfection to the mirrored mesh, possibly due to an asymmetry to the SKEL model. To avoid using noisy supervision, we instead keep track of two sets of SKEL parameters for each image example – one for the original image and one for the mirrored version. This led to over 13M sets of SKEL pseudo ground truth parameters for the original images and an additional 13M for their mirrored versions.

Quality control. As discussed in the main manuscript, the procedure for converting SMPL parameters to SKEL parameters is not perfect. The optimization can occasionally get stuck in local minima and produce unlikely poses (see Figure 3 of the main manuscript). To avoid using some of these severe failures as supervision, we apply an initial filtering stage to discard the SKEL parameters for low quality SKEL fits. For this procedure, we use the Max Per Vertex Error (maxPVE) as measure of the quality. This is defined as the max position error across all the vertices V^i between the SMPL and SKEL surface mesh:

$$\text{maxPVE} = \max_{i<6890} ||V^i_{\text{SMPL}} - V^i_{\text{SKEL}}||_2^2. \quad (1)$$

This metric can quantify the worst part of the fitting, making it possible for us to strictly bound the quality of the data. In practice, we discard the SKEL parameters for examples where maxPVE $>$ 6cm. We note that we discard only the SKEL parameters for these examples, while the images (and 2D/3D keypoints) remain in our datasets and can potentially obtain pseudo ground truth SKEL parameters during the iterative refinement of our training. Besides the maxPVE check, we also adopt the other quality checks that HMR2.0 [7] performs to remove low quality fits. These include discarding a fit for examples that a) have a shape parameter with absolute value larger than 3, or b) have less than four keypoints with confidence larger than 0.

S.3. Training

Architecture. For our main model, we adopt the architecture of HMR2.0 [7]. We use a ViT-H backbone [5], which is initialized with weights from ViTPose [23]. After the backbone, a transformer head is used to regress the SKEL parameters. This is also similar to HMR2.0's transformer head. The only difference is that we regress a lower dimensional output for the pose representation (SKEL has 46 pose parameters, while SMPL has 72), so we adapt the output accordingly. Regarding the exact size of the output, see details in the next paragraph.

Rotation Representation. The SKEL model natively uses Euler angles for the pose parametrization. As we discussed in the main manuscript, our network does not regress Euler angles – instead, we use the continuous rotation representation [26] as the regression target. As is common in the literature [7, 11, 15, 16], we convert this representation to

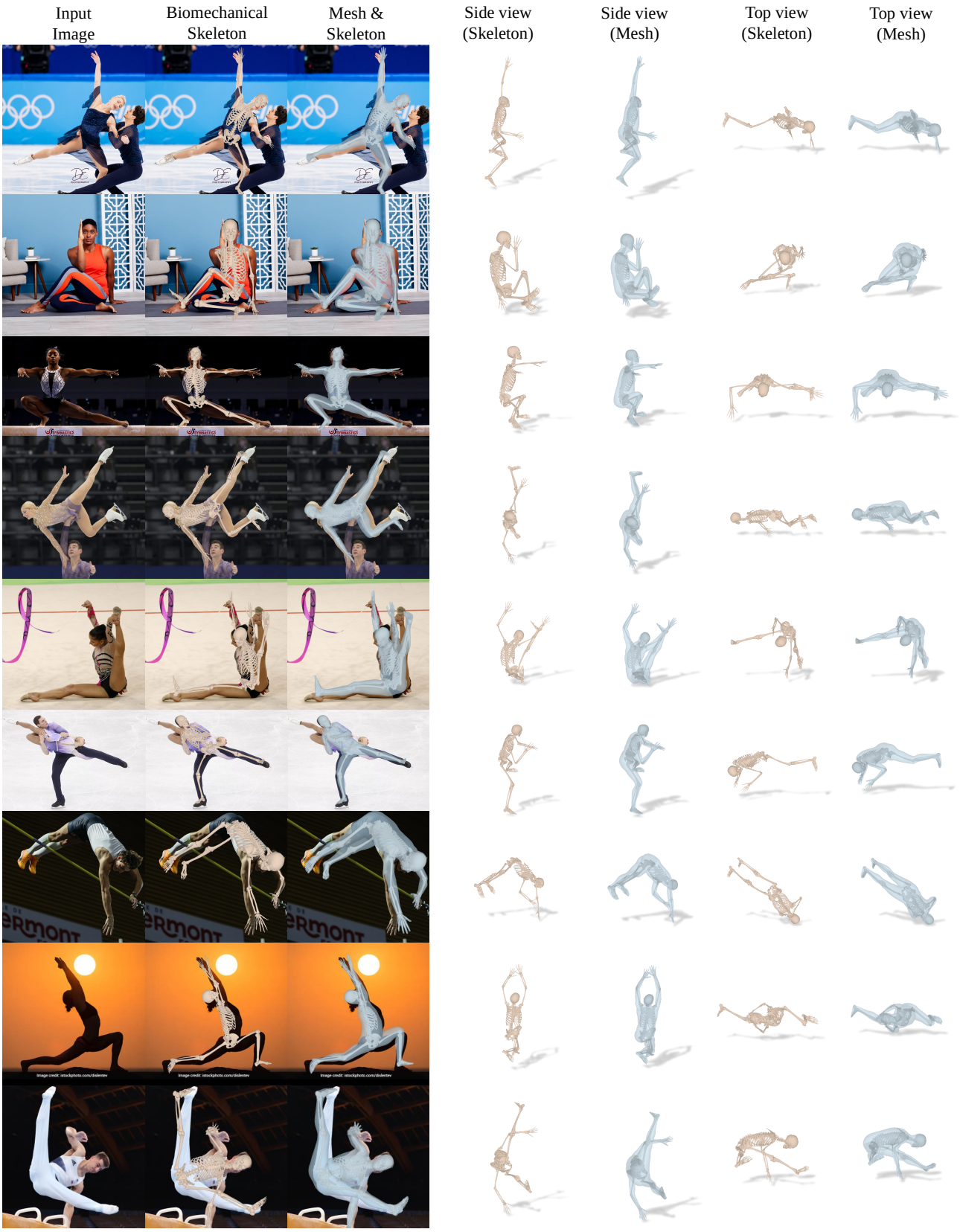

Figure S.1. **Additional qualitative results of HSMR.** This figure extends Figure 7 of the main manuscript. For each example we show: a) the input image, b) the overlay of SKEL in the input view, c) a side view, d) the top view. We visualize both the skeleton and the transparent mesh of the estimated SKEL model.

rotation matrices and the pose parameter loss is applied directly on the elements of the rotation matrix. Finally, we convert the rotation matrices to Euler angles, which is what we provide as input to SKEL.

Something that we need to highlight here is that for joints with three degrees of freedom, we need to regress six values per joint. This is the common 6D representation [7, 11, 15, 16], which gives us a 3×3 rotation matrix (after Gram-Schmidt) and corresponds to three Euler angles. Overall, 10 joints among the 24 SKEL joints have three rotational degrees of freedom. However, SKEL also has 12 joints with one rotational degree of freedom (*e.g.*, knees). In this case, we only need to regress two values per joint. This gives us the 2×2 rotation matrix for this joint, which corresponds to one Euler angle. Finally, the only joints with two degrees of freedom are the wrists. In this case, we simply regress the Euler angles directly. As a result, the length of our regression target is 88.

Hyperparameters. Most of our hyperparameters mirror the choices of HMR2.0 [7]. We use AdamW optimizer with learning rate equal to $1\text{e}{-5}$ and weight decay equal to $1\text{e}{-4}$. As for the loss weights, we set 0.05 for the 3D keypoints loss and 0.01 for the 2D keypoints loss. The parameter losses include the global orientation, the body pose and the shape. Their weights are 0.002, 0.001, 0.0005 respectively. We train the network on 8 A6000 GPUs with a batch size of 78 per GPU (effective batch size 624). We use half-precision (16 bits) for training. We train our model for 160k iterations.

S.3.1. Pseudo ground truth refinement

SKELify. To enable the refinement of the pseudo ground truth, we implement a fitting pipeline, similar to SMPLify [3], that will allow us to fit the SKEL model to 2D body keypoints. To be compatible with previous conventions, we call this SKELify. The SKELify objective for the 2D keypoints reprojection follows [3, 16, 19]:

$$E_{\texttt{kp2D}}(q, \beta) = \sum_i c_i \rho\big(\pi(X_i) - x_i^*\big). \tag{2}$$

Here, ρ is the Geman-McClure robustifier [6], and c_i is the confidence of the keypoint x_i^*. We already defined $E_{\texttt{shape}}(\beta)$ and $E_{\texttt{pose}}(q)$ in the main manuscript. The loss weights for normalized 2D keypoints loss, shape prior loss and pose prior loss are 1.0, 5.0^2, $(4.78 \times 0.17)^2$ respectively. For this iterative optimization, we use an LBFGS optimizer equipped with strong Wolfe line search.

Iterative refinement routine. We execute the SKELify optimization periodically during training. More specifically, we first warm up our network for 5k iterations. After the warmup, SKELify runs every 230 steps, and it will run the optimization on the latest 18k prediction results.

After the optimization, we compare the results of SKELify, q^*, β^*, with the ones that we maintain in our dictionary of pseudo ground truth SKEL parameters. If the SKELify results have improved keypoint reprojection, then we update the pseudo ground truth in our dictionary with the pseudo labels q^*, β^* acquired by SKELify.

S.4. Ablation

For the ablation experiment (Table 5 of the main paper), we perform a simpler setting of HSMR, to make it easier to run more experiments. First, we employ the ViT-B backbone [5] for HSMR (pretrained by ViTPose [23]). This allows us to increase the batch size (from 78 images per GPU to 300 images per GPU) and train the network on one GPU only. Additionally, we train on a subset of HSMR's training data. Specifically, following HMR2.0 [7], we choose Human3.6M [9], MPI-INF-3DHP [18], MPII [1], and COCO [17] for the ablation study. With the reduced dataset, we train each network for 60k iterations. The rest of the decision choices remain the same with the main network training, unless explicitly stated by the ablation – *i.e.*, "with Euler angles" for the first ablation setting or "without pseudo GT refinement" for the second ablation setting. In the ablation, we observe that the ViT-B baseline has better performance on the 3D metrics, but clearly lags behind the ViT-H version on the 2D metrics, so we choose ViT-H as the main backbone of our approach.

S.5. Volume estimation accuracy

To further evaluate the mesh reconstruction, we perform another experiment which considers the accuracy of the reconstructed volume. This volume is tightly connected to the body shape parameters, β. In general, methods that regress SMPL parameters tend to be less accurate in terms of the body shape β. We observe this qualitatively for HSMR too. As a sanity check, we compare HSMR with our main baseline, HMR2.0 [7], on volume estimation accuracy (in dm^3 or Liters) for datasets that provide ground truth meshes. For 3DPW [21], HMR2.0 has a volume error of 13.3 dm^3, while HSMR has 11.8 dm^3. Similarly, on MOYO [20], HMR2.0 has a volume error of 12.9 dm^3, while HSMR is again better at 4.2 dm^3. Although the sample is small (indoor datasets only include a few individuals), this indicates that HSMR is in a similar ballpark with HMR2.0.

S.6. Evaluation

For the evaluation, we adopt the same protocols with previous work [7] to be compatible with their evaluation. This includes reporting results on Human3.6M [9], 3DPW [21], COCO [17], PoseTrack [2] and LSP Extended [10]. The metrics are also consistent with previous work. For 3D metrics we report MPJPE and PA-MPJPE, as they are defined

in [9, 12, 25]. For the 2D metrics we report PCK at different thresholds as defined in [24].

The only dataset that is new to our evaluation is MOYO [20]. For MOYO, we report results on the whole validation subset (rouhgly 155k frames). We evaluate our results on the 3D joints using the 24 SMPL joints, and additionally on the SMPL mesh vertices. For the vertex-based evaluation, we consider the Mean Per Vertex Position Error (MPVPE), as defined in [4, 19], as well as PA-MPVPE, which is the Procrustes Alignment version of this metric. For the evaluation on MOYO, we use the improved SMPL fits (v2 of the dataset – check this GitHub issue for more details). Our original results were using the first version of the dataset, but we have updated all the tables with the recently released and improved SMPL ground truth.

Regarding the rotation violation, SKEL provides degrees of freedom across the realistic rotation axes only, so we evaluate whether the predicted body configuration from HSMR violates the known joint limits. For SMPL, we first convert the predicted axis angle rotations from the various methods to Euler angles. Then, we evaluate the violation along each axis and for each joint of interest (*i.e.*, knees, elbows) we select the maximum violation along these axes.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 3

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 3

[3] 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 *ECCV*, 2016. 3

[4] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 4

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1, 3

[6] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 4:5–21, 1987. 3

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 1, 3

[8] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 1

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 2013. 1, 3, 4

[10] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 3

[11] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2021. 1, 3

[12] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 4

[13] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 1

[14] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C Karen Liu, and Michael J Black. From skin to skeleton: Towards biomechanically accurate 3D digital humans. *ACM Transactions on Graphics (TOG)*, 42(6): 1–12, 2023. 1

[15] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 1, 3

[16] 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 *ICCV*, 2019. 1, 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 3

[18] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 1, 3

[19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 3, 4

[20] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *CVPR*, 2023. 3, 4

[21] 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 *ECCV*, 2018. 3

[22] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. AI Challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 1

[23] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 1, 3

[24] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2012. 4

[25] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior. *PAMI*, 2018. 4

[26] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 1