Reconstructing Humans with a Biomechanically Accurate Skeleton

Yan Xia[1,2] Xiaowei Zhou[2] Etienne Vouga[1] Qixing Huang[1] Georgios Pavlakos[1]
[1]The University of Texas at Austin [2]Zhejiang University

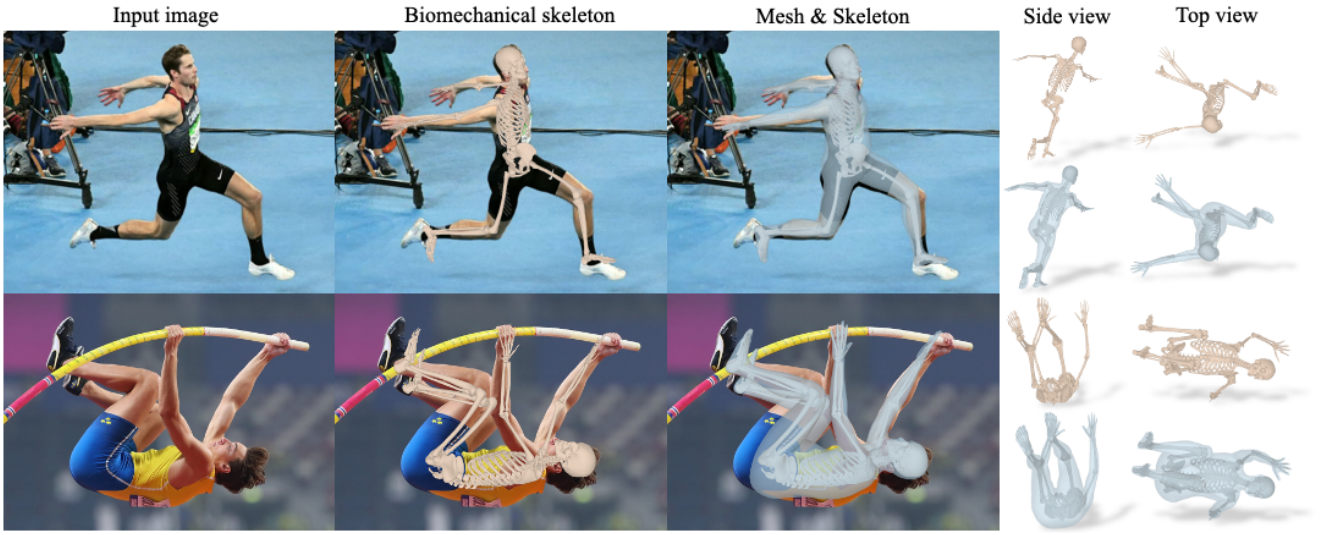

Figure 1. **Human Skeleton and Mesh Recovery (HSMR)**. We propose an approach that recovers the biomechanical skeleton and the surface mesh of a human from a single image. We adopt a recent biomechanical model, SKEL [24] and train a transformer to estimate the parameters of the model. We encourage the reader to see the skeleton and surface reconstructions in our project page.

Abstract

In this paper, we introduce a method for reconstructing 3D humans from a single image using a biomechanically accurate skeleton model. To achieve this, we train a transformer that takes an image as input and estimates the parameters of the model. Due to the lack of training data for this task, we build a pipeline to produce pseudo ground truth model parameters for single images and implement a training procedure that iteratively refines these pseudo labels. Compared to state-of-the-art methods for 3D human mesh recovery, our model achieves competitive performance on standard benchmarks, while it significantly outperforms them in settings with extreme 3D poses and viewpoints. Additionally, we show that previous reconstruction methods frequently violate joint angle limits, leading to unnatural rotations. In contrast, our approach leverages the biomechanically plausible degrees of freedom making more realistic joint rotation estimates. We validate our approach across multiple human pose estimation benchmarks. We make the code, models and data available at: https://isshikihugh.github.io/HSMR/

1. Introduction

In recent years, there has been remarkable progress in 3D human pose estimation, with proposed methods reaching the potential that computer vision researchers envisioned by finding applications in diverse fields such as robotics [12, 29, 44, 45], graphics and animation [57, 65], and AR/VR [51]. However, there are fields where these techniques would seemingly be a perfect fit, yet the adoption has been notably limited. Biomechanics is one such example. For biomechanics, the requirements are stricter: we need methods that estimate parameters compatible with biomechanical skeletons, respect joint limits, ensure physically plausible motion and return high accuracy estimates. Unfortunately, most state-of-the-art methods do not satisfy these constraints, and extensive post-processing is required [39, 54]. Our goal is to move towards bridging this gap by proposing an approach that can generate predictions aligned with a biomechanically accurate skeleton model.

Currently, the progress in the field of 3D human pose estimation has largely been driven by the use of parametric body models, like SMPL [33], SMPL-X [42] and GHUM [60]. These models provide a compact parameter-

ization, enabling the direct regression of model parameters from an input image [14, 22, 25, 27]. While these parametric models offer plausible surface representations, their skeleton design is not anatomically accurate. For instance, the kinematic tree does not align with the actual skeletal structure of the human body [24]. Moreover, the joints are represented as ball (socket) joints, introducing additional degrees of freedom. This modeling choice can lead to the prediction of unnatural joint angles, resulting in outputs that are incompatible with biomechanical applications and simulations [10]. Consequently, recent advances in 3D human pose estimation have not yet been fully leveraged by biomechanics, which rely on more anatomically accurate models.

The introduction of the SKEL model [24] marked a significant step forward by integrating a biomechanical skeleton with the SMPL surface mesh. This combination enables compatibility with biomechanical simulation environments, allowing for more anatomically realistic modeling. However, to fully harness advancements in computer vision, it is essential to develop methods that can accurately estimate the parameters of this model directly from image inputs.

Towards this goal, we propose **HSMR** (**H**uman **S**keleton and **M**esh **R**ecovery), a method for reconstructing humans with a biomechanically accurate skeleton from a single image. Our method leverages the recently introduced SKEL model [24] and adopts a transformer-based network [11, 61] to regress the SKEL parameters from image input. One key challenge is that there is no dataset of images with corresponding SKEL parameters, that could be used for training. To address this, we perform an initial optimization [24] to convert the SMPL (pseudo) ground truth of existing datasets to SKEL pseudo ground truth. While this is a reasonable starting point, the offline conversion is not perfect and can introduce annotation errors. To ensure high-quality data, we propose an iterative refinement routine during training, which progressively improves the SKEL pseudo ground truth, enabling us to train a more accurate and reliable model. This refinement is in the spirit of SPIN [27] – we optimize the SKEL model to align with the ground truth 2D body keypoints, while using the HSMR estimate as an initialization of the optimization. The result of this fitting is used as pseudo ground truth for future training iterations.

We carefully benchmark HSMR across multiple datasets. Despite starting without any ground truth SKEL parameters, we demonstrate that our approach matches the performance of state-of-the-art methods, when evaluated on the traditional metrics for 2D/3D joints accuracy. More importantly, our model has a clear advantage in cases with extreme poses and viewpoints (*i.e.*, yoga postures from the MOYO dataset [52]), which often lie outside the distribution of standard training data. This result indicates that the biomechanical skeleton model can be helpful at regularizing the estimated pose. Furthermore, we show that previous methods based on SMPL parameter regression frequently yield unnatural joint rotations, due to SMPL's simplified skeleton modeling, which provides more degrees of freedom than a realistic biomechanical model.

To summarize, our contributions are:

- We present HSMR, which is, to the best of our knowledge, the first end-to-end approach that can reconstruct humans in 3D from a single image by estimating the parameters of a biomechanical skeleton model, SKEL [24].
- Starting without any paired dataset of images and SKEL ground truth, we show how to generate data to train our model. Additionally, we incorporate a procedure to iteratively refine the quality of the pseudo ground truth.
- We demonstrate that our approach can match the performance of the most closely related state-of-the-art method that regresses SMPL parameters [14], while achieving clear improvements specifically for more challenging cases with extreme poses and viewpoints.
- We highlight the limitations of methods regressing parameters of simpler body models (*i.e.*, SMPL), and show how they tend to predict unnatural rotations for the body joints, leading to biomechanically inaccurate results.

2. Related Work

Human Body Models. A lot of the recent progress in pose estimation can be attributed to the access to simple, yet realistic models of the human body. SCAPE [3] was one of the seminal works in this space and was learned in a data-driven way from 3D scans of humans. The SMPL model [33] and follow-up work [37, 56, 60] streamlined and simplified the body model design making it compatible with traditional graphics pipelines. A number of extensions of SMPL improved the modeling capabilities, by introducing articulated hands [46], facial expressions [42], and deformations for the feet [38]. Although these models increased the detail and realism of the reconstructed surface, the skeleton maintained the simplistic design of SMPL, representing each body joint with a ball (socket) joint. This skeleton design was eventually improved by the SKEL model [24], which adopts most design principles from SMPL, but rigs the surface mesh using a biomechanically accurate skeleton. For our method, we adopt the SKEL model and estimate its parameters using a single image as input.

3D Human Pose Estimation. Earlier work on 3D human pose estimation was representing the human body with simplistic stick figures [35, 40, 50]. Since the introduction of the SMPL model [33], there has been a shift towards approaches that reconstruct the full body surface by estimating the parameters of the SMPL model. Although the initial approaches relied on iterative optimization [6, 42], currently most methods are based on deep learning and regress the SMPL parameters in a feedforward manner [22, 41]. HMR [22] was a seminal work in this direction that esti-

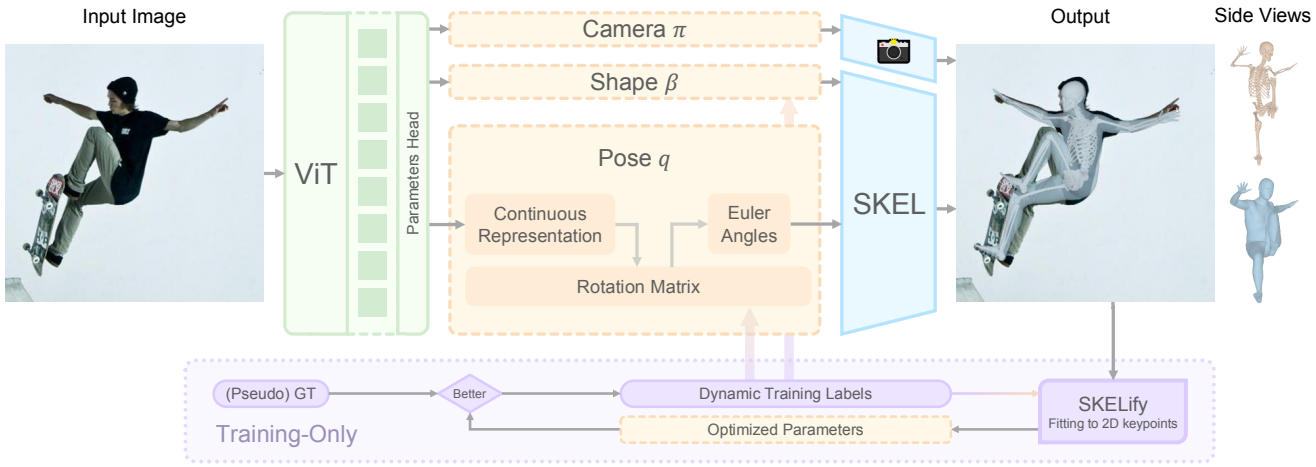

Figure 2. **Overview of our HSMR approach.** A key design choice of HSMR is the adoption of the SKEL parametric body model [24] which uses a biomechanically accurate skeleton. We employ a transformer-based architecture that takes as input a single image of a person and estimates the pose q and shape parameters β of SKEL, as well as the camera π. During training, we iteratively update the pseudo ground truth we use to supervise our model, aiming to improve its quality. For this, we optimize the HSMR estimate to align with the ground-truth 2D keypoints (SKELify). The output parameters of the optimization are used in future training iterations as supervision target.

mates SMPL parameters from a single image with a CNN in an end-to-end manner. Since then, different designs for the architecture of the network have been proposed [25, 26]. However, most key principles from HMR are still adopted by recent works [5, 14], even when other parametric models are used, like MANO [46] for hand reconstruction [43] or SMPL-X [42] for expressive reconstruction [7, 9]. One update of recent works [7, 14] is the adoption of Visual Transformers [11, 61] instead of the previous CNN designs [17, 49]. Following good practices, we also adopt a transformer-based neural network for SKEL regression.

In parallel with the investigation of architecture design for human mesh recovery, other works focused on the data for training. SPIN [27] proposed an optimization in-the-loop to create pseudo ground truth SMPL parameters for the training images. EFT [21] and CLIFF [30] followed a similar practice with an improved optimization. In our work, we face the problem that there is no existing image dataset with SKEL ground truth, so we describe how to get an initial dataset with SKEL pseudo ground truth and then iteratively refine these parameters to improve their quality during training. Besides the data quality, recent work has emphasized the importance of large scale data for training this kind of models [7, 14, 47]. We follow these good practices and we train using the large scale data of HMR2.0 [14].

Pose Estimation Meets Biomechanics. The most common use of human pose estimation methods in biomechanics is in the form of 2D keypoint detectors [8, 61] that can provide reliable 3D poses after triangulation from multiple views [39, 54]. Lin *et al*. [32] proposed an approach to regress the biomechanical model parameters by using input images from two views, while Bittner *et al*. [4] use video input. In contrast to them, we address the problem in its more challenging, single-image setting. Jiang *et al*. [19] use biomechanical constraints for more accurate 3D pose estimation, but their work adopts the SMPL model, making the output incompatible with biomechanical simulations [10]. Moreover, there is progress with the datasets for biomechanics. Werling *et al*. [58] introduced the AddBiomechanics dataset, a large scale collection of biomechanics data. This has the potential of acting similarly to the popular AMASS dataset [34] enabling training of pose and motion priors. More recently, Gozlan *et al*. [15] introduced a benchmark, OpenCapBench, for evaluating human pose estimation methods under physiological constraints. The benchmark was not available at the time of submission, but it could be useful for evaluating HSMR and future work.

3. Technical approach

In this section, we describe our technical approach for reconstructing humans using a biomechanically accurate skeleton model. First, we provide some preliminaries regarding the SKEL model [24] (Section 3.1), and then we present our HSMR model for Human Skeleton and Mesh Recovery (Section 3.2). We focus on the architecture, the procedure for training data generation, and the iterative refinement of the pseudo ground truth during training.

3.1. Preliminaries

SKEL Model. The SKEL model [24] is a parametric body model that combines the popular SMPL model [33] with a biomechanical skeleton model, BSM. Specifically, SKEL defines a function $\mathcal{S}(q, \beta)$ that takes as input parameters for pose ($q \in \mathbb{R}^{46}$) and shape ($\beta \in \mathbb{R}^{10}$), and outputs a skin mesh $M \in \mathbb{R}^{3\times N}$ with $N = 6890$ vertices and a skeleton

mesh S. The surface mesh shares the same topology with SMPL, so we can apply a regressor W to get the locations of the 3D joints $X = WM$. The shape space of SKEL, and the shape parameters β are the same with SMPL. However, there is a key difference for the pose representation. Previous models in the SMPL family [33, 42, 46] have treated every articulation joint as a ball (socket) joint with three degrees of freedom. In contrast to that, SKEL carefully designs the kinematic parameters according to the real human biomechanical structure and only models the realistic degrees of freedom. As a result, the pose parameters q are lower dimensional – 46 for SKEL, compared to 72 for SMPL. Each pose parameter corresponds to a single degree of freedom and is represented as an Euler angle. This allows us to associate each parameter with its explicit joint rotation limits. For example, the knee has one degree of freedom with limits of 0° extension and 135° flexion.

3.2. Human Skeleton and Mesh Recovery

Architecture. For our architecture, we follow best practices from the human mesh recovery literature [14, 22]. We start with a ViT backbone [11, 61], which takes as input an RGB image I of a person. A transformer head at the end of the network regresses the parameters of the SKEL model.

In terms of the model output, we regress the camera π, the shape parameters β, and the pose parameters q. Unlike the SMPL family of models, SKEL represents the pose parameters q with Euler angles. Although this representation is intuitive, we find that Euler angles can be challenging as a regression target (Section 4). Instead, we adopt the continuous rotation representation [64] for the pose parameters. Initially, the output of the network is in the form of this continuous representation, q_{cont}. We first convert the parameters to the rotation matrix representation, q_{mat}, using Gram–Schmidt [64]. The q_{mat} representation is where we apply our parameter loss. Then, we can convert the parameters to the Euler angle representation, q_{Euler}, which is compatible with the input of the SKEL model. Eventually, the losses on the SKEL parameters are:

$$\mathcal{L}_q = ||q_{\text{mat}} - q^*_{\text{mat}}||_2^2 \text{ and } \mathcal{L}_\beta = ||\beta - \beta^*||_2^2. \quad (1)$$

Here, q^*_{mat} and β^* are the ground truth pose and shape parameters, respectively. Besides the parameter losses $\mathcal{L}_q$ and $\mathcal{L}_\beta$ (which are applied only when the labels are available), we also apply losses on the 3D and 2D keypoints:

$$\mathcal{L}_{\text{kp3D}} = ||X - X^*||_1 \text{ and } \mathcal{L}_{\text{kp2D}} = ||\pi(X) - x^*||_1. \quad (2)$$

Training Data Generation. One key obstacle in training our HSMR model is that there are no image datasets with SKEL annotations. To address this, we propose to leverage existing image datasets with SMPL (pseudo) ground truth and convert them to SKEL parameters. This conversion is possible because the two models share the same topology for the surface mesh. This allows us to optimize the SKEL parameters, such that the SKEL mesh aligns with the target SMPL mesh [24]. Through this procedure, we can acquire some initial pseudo ground truth SKEL parameters for the datasets typically used for human mesh recovery.

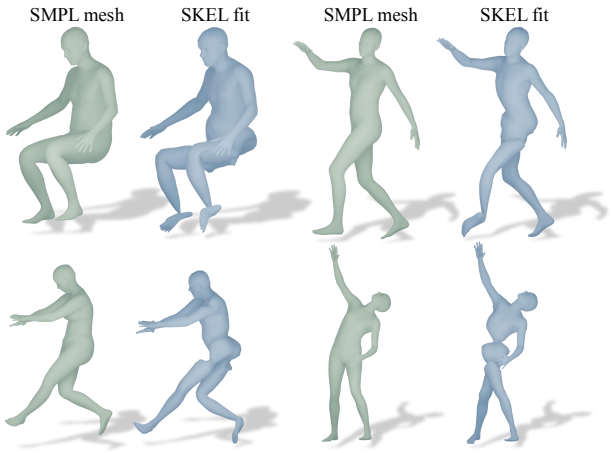

Figure 3. **Failure cases of SMPL-to-SKEL conversion.** While we can technically fit SKEL to an instance of the SMPL model, this conversion can often lead to problematic SKEL results. Here, we visualize SMPL meshes (light green), and the SKEL meshes we get when we try to fit the SKEL model to the SMPL mesh (light blue). For the fitting, we use the optimization code of [24].

Training with Pseudo-Label Refinement. Although the SMPL-to-SKEL conversion gives us a reasonable starting point, it is an imperfect procedure with frequent failure cases (Figure 3). This type of local minima are common in similar iterative optimization problems [6, 42]. If we aim to improve the accuracy of HSMR, we need to improve the quality of the pseudo ground truth we use for training.

To achieve this, we propose an iterative procedure that gradually updates the quality of the pseudo ground truth SKEL parameters for each example. This is inspired by previous work on pseudo ground truth refinement [21, 27]. More specifically, for each image I of a person, given a network estimate $q^{\text{reg}}, \beta^{\text{reg}}$, we refine the parameters iteratively, such that they align with the 2D keypoints x^* of the person on the image [6, 42]. The optimized estimates of the pose and shape parameters, q^*, β^* are used as more accurate pseudo ground truth for supervising the network.

For this iterative optimization, we propose an equivalent of SMPLify [6] for SKEL, which we call SKELify. The optimization is mainly guided by the 2D keypoints x^*. Specifically, we introduce a reprojection objective, E_{kp2D}, aiming to align the projection of the 3D joints with the 2D keypoints. This objective is similar to the second part of Equation 2, with the addition of a robustifier [13] as in [6]. To regularize the shape and pose parameters we add shape and pose priors. The shape prior is inherited from SMPL, *i.e.*,

Methods	COCO		LSP-Extended		PoseTrack		3DPW		Human3.6M		MOYO	
	@0.05↑	@0.1↑	@0.05↑	@0.1↑	@0.05↑	@0.1↑	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
PARE [25]	0.72	0.91	0.27	0.60	0.79	0.93	82.0	50.9	76.8	50.6	165.6	117.1
CLIFF [30]	0.64	0.88	0.32	0.66	0.75	0.92	– *	– *	47.1	32.7	154.6	109.3
HybrIK [28]	0.61	0.80	0.37	0.69	0.81	0.94	**80.0**	**48.8**	54.4	34.5	140.1	93.2
PLIKS [48]	0.62	0.90	0.26	0.66	0.74	0.94	– *	– *	**47.0**	34.5	132.6	91.8
HMR2.0 [14]	**0.86**	**0.96**	**0.53**	**0.82**	**0.90**	**0.98**	81.3	54.3	50.0	**32.4**	123.3	90.4
HSMR	0.85+0.01	**0.96**+0	0.51+0.02	0.81+0.01	**0.90**+0	**0.98**+0	81.5+0.2	54.8+0.5	50.4+0.4	32.9+0.5	**104.5**-18.8	**79.6**-10.8

Table 1. **Comparison with state-of-the-art approaches that regress SMPL parameters.** The primary baseline for HSMR is the HMR2.0 network [14], since it is the closest to our design, in terms of architecture and training data We report PCK @0.05 & @0.1 for the 2D datasets (COCO, LSP-Extended, PoseTrack) and MPJPE & PA-MPJPE for the 3D datasets (3DPW, Human3.6M, MOYO). Even though we adopt the SKEL model which is less flexible and we start without any initial ground truth for training, we are able to match the performance of HMR2.0 on most datasets - with up to 0.5mm difference. More importantly, we outperform HMR2.0 by a big gap of more than 10mm on the challenging MOYO dataset that includes extreme poses and viewpoints. In the table, we explicitly report the differences in evaluation metrics between our HSMR network and HMR2.0. *: trains on 3DPW.

$E_{\texttt{shape}}(\beta) = \|\beta\|^2$. For the pose parameters, however, we do not have an existing pose prior for SKEL. Instead, we leverage the known limits of natural rotation for each joint. For example, let us assume that for a pose parameter q_i, the lower limit is l_i and the upper limit is u_i, *i.e.*, $q_i \in [l_i, u_i]$. In this case, we can add a term:

$$E_{\texttt{pose}}(q) = \sum_i \exp(l_i - q_i) + \exp(q_i - u_i), \qquad (3)$$

which strongly penalizes rotations that exceed the known joint limits. If for a specific parameter there is no explicit limit, we can omit it from the calculation of the objective.

In the end, we sum the three objectives, $E_{\texttt{kp2D}}(q, \beta)$, $E_{\texttt{shape}}(\beta)$ and $E_{\texttt{pose}}(q)$ and solve for the optimal SKEL parameters, q^*, β^*. These parameters are used as pseudo ground truth to train the network. Unlike [27], this refinement is not happening in every training iteration, but we execute it periodically in batch mode for efficiency reasons. We refer to the SuppMat for more implementation details.

4. Experiments

4.1. Datasets and Metrics

We train HSMR using the training data from HMR2.0 [14], which include images from Human3.6M [18], MPI-INF-3DHP [36], COCO [31], MPII [1], AI Challenger [59], AVA [16] and InstaVariety [23]. We preprocess the data to convert the SMPL (pseudo) ground truth of HMR2.0 to SKEL parameters, as we describe in Section 3.2.

We evaluate our approach on multiple datasets for human pose estimation. Some of them provide 3D annotations, *i.e.*, Human3.6M [18], 3DPW [55] and MOYO [52], while others only include 2D annotations, *i.e.*, COCO [31], PoseTrack [2] and LSP Extended [20]. Accordingly, we report Percentage of Correct Keypoints (PCK) [62] at different thresholds as metrics for 2D pose accuracy, and Mean Per Joint Position Error (MPJPE) [18], Mean Per Vertex Position Error (MPVPE) [42] plus their Procrustes Alignment version PA-MPJPE [22, 63] and PA-MPVPE as 3D pose accuracy metrics. Moreover, we evaluate the results of different methods in terms of violation of the joint limits. Specifically, we focus on knees and elbows and report the frequency of violation for different angle thresholds. Please see the the SuppMat for more details.

	PARE	CLIFF	HybrIK	PLIKS	HMR2.0	HSMR
MPVPE↓	174.5	155.7	143.6	136.7	142.2	**120.1**
PA-MPVPE↓	121.9	110.6	94.4	94.8	103.4	**90.7**

Table 2. **Evaluation of the surface reconstruction accuracy.** We report MPVPE and PA-MPVPE on the MOYO dataset.

4.2. Comparison with methods for SMPL recovery

We build HSMR using best practices from the methods that regress SMPL parameters. More specifically, HMR2.0 [14] is closer to our design, so this is the primary baseline we compare against. In Table 1, we compare the performance of HSMR and HMR2.0 on various datasets. For context, we also include other state-of-the-art methods for SMPL reconstruction [25, 28, 30, 48]. In addition, in Table 2, we also present results on MOYO for per-vertex errors.

We observe that for most datasets, HSMR achieves results that are almost identical to HMR2.0, with the metrics in 3DPW and Human3.6M having a difference of *up to 0.5mm*. This is important, because even though we operate with a less flexible model (SKEL) and we started our investigation without any initial ground truth for training, we were able to actually match the performance of HMR2.0. Moreover, we observe that simultaneously we achieve a huge improvement of *more than 10mm* on the MOYO dataset [52]. The observations are similar for the surface-based evaluation (Table 2). This is significant, because MOYO includes challenging extreme poses (yoga poses) and viewpoints. We believe that this could be attributed to the stronger pose regularization that the biomechanical skeleton can impose, since it only allows the realistic degrees of freedom. In fact, in Section 4.4, we verify

Methods	COCO		LSP-Extended		PoseTrack		3DPW		Human3.6M		MOYO	
	@0.05↑	@0.1↑	@0.05↑	@0.1↑	@0.05↑	@0.1↑	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
HMR2.0 [14]	**0.86**	**0.96**	**0.53**	**0.82**	**0.90**	**0.98**	81.3	**54.3**	**50.0**	**32.4**	123.3	90.4
HMR2.0 + SKEL fit	0.78	0.95	0.49	0.79	**0.90**	**0.98**	**81.0**	54.4	53.6	34.1	130.5	93.7
HSMR	0.85	**0.96**	0.51	0.81	**0.90**	**0.98**	81.5	54.8	50.4	32.9	**104.5**	**79.6**

Table 3. **Comparison with baseline for SKEL recovery.** We start from the SMPL prediction of HMR2.0 [14] and we fit the SKEL model to it with terative optimization [24]. This baseline corresponds to the "HMR2.0 + SKEL fit" row. We observe that this two-stage baseline for SKEL recovery performs worse than HSMR, while it is also significantly slower (3 minutes for a single frame).

that the various networks regressing SMPL parameters are indeed suffering from frequent violations of the joint limits.

4.3. Baseline for SKEL recovery

Besides comparing with methods for SMPL-based reconstruction, we also consider an optimization-based baseline for SKEL reconstruction. This was introduced by [24] and it is the same with the approach we use for our pseudo ground truth generation (Section 3.2). For the comparison, we run HMR2.0 to get SMPL parameters and we fit SKEL to the SMPL mesh with the optimization approach. The full results are presented in Table 3. Although in some cases the SKEL fit is comparable with the HMR2.0 output (*e.g.*, PoseTrack and 3DPW), in most cases there is a clear degradation in the quality (*i.e.*, COCO, LSP-Extended, Human3.6M and MOYO). Additionally, the fitting procedure is computationally expensive, requiring 3 minutes per frame. This means that our end-to-end HSMR approach is not only more accurate, but also much faster than the SKEL fitting.

4.4. Biomechanically-sound reconstruction

Besides evaluating the 2D/3D pose accuracy of the different mesh recovery approaches, we also investigate the biomechanical validity of their outputs. As discussed in Section 3.1, SKEL only considers the realistic degrees of freedom for each joint, whereas SMPL models each joint with a ball (socket) joint, which endows three degrees of freedom for each joint. In this subsection, we investigate whether methods that regress SMPL parameters actually predict unnatural joint rotations. We focus our attention specifically on the elbow and the knee joints. We consider various thresholds (*i.e.*, 10°, 20°, 30°) and report the frequency that each method exceeds this threshold (*i.e.*, rotation violation). The complete results for MOYO are presented in Table 4. As we can see, the violations are more frequent than we might have expected and they happen for all the methods that regress SMPL parameters. These results are an indication that these methods might return poses with low 3D joint position errors that rotate the body parts in unnatural ways. We visualize some interesting failure cases in Figure 4. We believe this observation points to a clear direction for future improvement of the approaches for human mesh recovery.

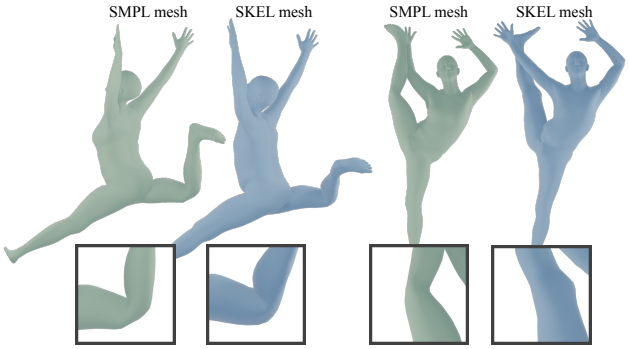

Figure 4. **Examples of unnatural joint rotation for SMPL.** SMPL represents the knee with a ball (socket) joint. This allows mesh recovery methods like HMR2.0 [14] to generate invalid rotations. We visualize examples from HMR2.0 (light green) where the knee is bend in unnatural ways. In comparison, the HSMR output (light blue) respects the biomechanical constraints.

4.5. Ablation study

Finally, we evaluate some key design decisions of our pipeline. More specifically, we investigate the choice of regression target for the pose parameters. We compare using the continuous rotation representation [64] as an alternative to the Euler angles (which is the native representation for SKEL). Moreover, we assess the importance of iterative refinement of the SKEL pseudo ground truth that we employ during training. For this evaluation, we perform a smaller scale ablation using a ViT-B backbone [61] for our network.

We present the detailed results of this ablation in Table 5. As we see, regressing the Euler angles directly produces a clear drop in performance, justifying the use of the continuous rotation representation for SKEL parameter regression.

Moreover, if we train without the iterative refinement of the labels, the performance decreases for most datasets, particularly for the 3D metrics (for the 2D metrics, the difference is small, because the refinement does not affect the quality of the 2D pseudo ground truth). These results confirm the importance of both design choices.

4.6. Qualitative evaluation

In Figure 5, we provide more qualitative results of our approach. We show reprojections on the image, as well as side and top views. We visualize both the (transparent) surface mesh and the skeleton output. HSMR performs well for a variety of poses, and viewpoints. Also, in Figure 6 we show

Figure 5. **Qualitative evaluation of HSMR.** For each input 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.

Methods	violation > 10° ↓				violation > 20° ↓				violation > 30° ↓			
	left elbow	right elbow	left knee	right knee	left elbow	right elbow	left knee	right knee	left elbow	right elbow	left knee	right knee
PARE [25]	36.4%	42.4%	20.0%	23.2%	14.6%	15.4%	3.2%	3.8%	5.5%	4.8%	0.3%	0.4%
CLIFF [30]	34.2%	33.0%	28.3%	31.0%	13.0%	12.4%	4.8%	4.5%	5.2%	5.2%	0.5%	0.3%
HybrIK	58.7%	60.9%	52.9%	48.6%	29.4%	34.6%	30.7%	27.0%	16.4%	21.0%	20.0%	17.5%
PLIKS	41.6%	44.7%	47.4%	43.8%	17.9%	22.7%	18.2%	17.6%	8.3%	11.4%	8.5%	8.5%
HMR2.0 [14]	47.6%	44.3%	45.7%	56.4%	19.8%	19.6%	6.4%	11.6%	8.5%	8.8%	1.0%	1.6%
HSMR	**0.0%**	**0.0%**	**3.9%**	**4.5%**	**0.0%**	**0.0%**	**0.2%**	**0.5%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**

Table 4. **Frequency of unnatural rotations for mesh recovery approaches.** We investigate how often each approach returns 3D bodies with unnatural joint rotations. We experiment on MOYO [52] and report the frequency that the unnatural rotation exceeds different thresholds (10°, 20° or 30°) for the elbow and the knee joints. Methods that regress SMPL parameters violate the joint limits frequently. Instead, our HSMR method avoids severe violations because it relies on SKEL which models only the realistic degrees of freedom.

Models	COCO		LSP-Extended		PoseTrack		3DPW		Human3.6M		MOYO	
	@0.05↑	@0.1↑	@0.05↑	@0.1↑	@0.05↑	@0.1↑	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
HSMR (ViT-B)	**0.79**	**0.94**	**0.38**	**0.70**	**0.86**	**0.96**	**76.7**	**50.0**	**49.8**	**37.1**	**124.0**	**92.6**
HSMR (ViT-B) w/ Euler angles	0.75	0.93	0.31	0.64	0.82	0.95	81.6	52.1	55.6	41.3	137.1	104.3
HSMR (ViT-B) w/o pseudo GT refinement	0.75	0.93	0.37	**0.70**	0.84	**0.96**	81.1	51.1	52.0	38.1	126.5	96.2

Table 5. **Ablation study on design choices.** We benchmark our proposed model and ablate two design choices. First, we change the regression target from the continuous representation [64] to the native Euler angles of SKEL. This has a negative effect across the board. Then, we experiment without the pseudo ground truth refinement process. This also has a negative impact particularly on the 3D metrics.

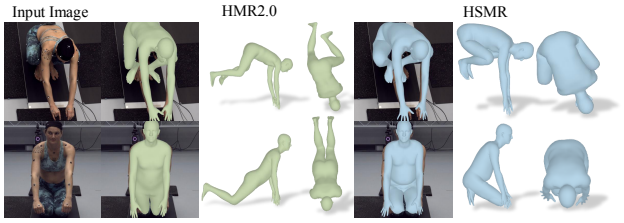

Figure 6. **Qualitative comparison with HMR2.0 on MOYO.** For each example we show the input image and results for HMR2.0 and HSMR. Although the interpretation in the input view is reasonable for both methods, HSMR achieves more accurate 3D reconstruction on the challenging poses and viewpoints of MOYO.

a comparison with HMR2.0 on images from the MOYO dataset. The qualitative improvements achieved by HSMR align with the MOYO quantitative results of Table 1. Finally, in Figure 7, we present some failure cases of HSMR.

5. Summary

In this paper, we presented an approach for reconstructing humans in 3D using a biomechanically accurate model, SKEL. We design a network that takes a single image as input and estimates the parameter of the SKEL model. To achieve that, we curate existing datasets with pseudo ground truth SKEL parameters and use them to train our model. In terms of 3D body pose estimation, our approach matches the performance of the state-of-the-art human mesh recovery methods while also outperforming them on cases with challenging poses and uncommon viewpoints. Moreover, we demonstrate how previous approaches for SMPL regression are failing to respect the biomechanical constraints, leading to serious violations of the joint angle limits. We

Figure 7. **Failure cases of our method.** HSMR often fails in cases with motion blur extreme poses and rare viewpoints.

hope that our work will help close the gap between vision-based methods for human pose estimation and the high precision required for biomechanical analysis.

Limitations and future work. One of the limitations of HSMR is the exclusive use of pseudo ground truth for training. Although our iterative refinement improves the pseudo ground truth quality, the network could benefit from more precise 3D labels. Moreover, we observe some inevitable jitter in our temporal reconstructions. We believe that follow-up work could address the recovery of smooth SKEL motions. Finally, future work could consider incorporating our estimates in a biomechanical simulation environment [10] to encourage physically-plausible motion [53].

Acknowledgements: E.V. was supported by CMMI-2310666. X.Z. was supported by Zhejiang Provincial Natural Science Foundation of China (No. LR25F020003) and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. Q.H. was supported by NSF IIS-2047677, NSF IIS-2413161, and Gifts from Adobe and Google. G.P. was supported by Gifts from Google and Adobe.

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. 5

[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. 5

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 2

[4] Marian Bittner, Wei-Tse Yang, Xucong Zhang, Ajay Seth, Jan van Gemert, and Frans CT van der Helm. Towards single camera human 3D-kinematics. *Sensors*, 23(1):341, 2022. 3

[5] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 3

[6] 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. 2, 4

[7] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. *NeurIPS*, 2024. 3

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 3

[9] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 3

[10] Scott L Delp, Frank C Anderson, Allison S Arnold, Peter Loan, Ayman Habib, Chand T John, Eran Guendelman, and Darryl G Thelen. OpenSim: open-source software to create and analyze dynamic simulations of movement. *IEEE transactions on biomedical engineering*, 54(11):1940–1950, 2007. 2, 3, 8

[11] 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. 2, 3, 4

[12] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *CoRL*, 2024. 1

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

[14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 3, 4, 5, 6, 8

[15] Yoni Gozlan, Antoine Falisse, Scott Uhlrich, Anthony Gatti, Michael Black, and Akshay Chaudhari. OpenCapBench: A benchmark to bridge pose estimation and biomechanics. In *WACV*, 2024. 3

[16] 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. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[18] 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. 5

[19] Jiaxi Jiang, Paul Streli, Xuejing Luo, Christoph Gebhardt, and Christian Holz. MANIKIN: Biomechanically accurate neural inverse kinematics for human motion estimation. In *ECCV*, 2024. 3

[20] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 5

[21] 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. 3, 4

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 4, 5

[23] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 5

[24] 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, 2, 3, 4, 6

[25] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2, 3, 5, 8

[26] 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 *ICCV*, 2021. 3

[27] 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. 2, 3, 4, 5

[28] 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 *CVPR*, 2021. 5

[29] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. OKAMI: Teaching humanoid robots manipulation skills through single video imitation. In *CoRL*, 2024. 1

[30] 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*, 2022. 3, 5, 8

[31] 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. 5

[32] Zhi-Yi Lin, Bofan Lyu, Judith Cueto Fernandez, Eline Van Der Kruk, Ajay Seth, and Xucong Zhang. 3D kinematics estimation from video with a biomechanical model and synthetic training data. In *CVPRW*, 2024. 3

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 1, 2, 3, 4

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3

[35] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 2

[36] 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. 5

[37] Ahmed AA Osman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *ECCV*, 2020. 2

[38] Ahmed AA Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. SUPR: A sparse unified part-based human representation. In *ECCV*, 2022. 2

[39] David Pagnon, Mathieu Domalain, and Lionel Reveret. Pose2Sim: An open-source python package for multiview markerless kinematics. *Journal of Open Source Software*, 7 (77):4362, 2022. 1, 3

[40] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 2

[41] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 2

[42] 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. 1, 2, 3, 4, 5

[43] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 3

[44] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018. 1

[45] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. In *NeurIPS*, 2024. 1

[46] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 2, 3, 4

[47] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *NeurIPS*, 2024. 3

[48] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. PLIKS: A pseudo-linear inverse kinematic solver for 3D human body estimation. In *CVPR*, 2023. 5

[49] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[50] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2

[51] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D human pose from an HMD camera. In *ICCV*, 2019. 1

[52] 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. 2, 5, 8

[53] Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. MultiPhys: multi-person physics-aware 3D motion estimation. In *CVPR*, 2024. 8

[54] Scott D Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. OpenCap: Human movement dynamics from smartphone videos. *PLoS computational biology*, 19(10):e1011462, 2023. 1, 3

[55] 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. 5

[56] Haoyang Wang, Riza Alp Güler, Iasonas Kokkinos, George Papandreou, and Stefanos Zafeiriou. BLSM: A bone-level skinned model of the human mesh. In *ECCV*, 2020. 2

[57] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 1

[58] Keenon Werling, Nicholas A Bianco, Michael Raitor, Jon Stingel, Jennifer L Hicks, Steven H Collins, Scott L Delp, and C Karen Liu. AddBiomechanics: Automating model scaling, inverse kinematics, and inverse dynamics from human motion data through sequential optimization. *Plos one*, 18(11):e0295152, 2023. 3

[59] 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. 5

[60] 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 *CVPR*, 2020. 1, 2

[61] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 2, 3, 4, 6

[62] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2012. 5

[63] 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. 5

[64] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4, 6, 8

[65] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *ECCV*, 2024. 1