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

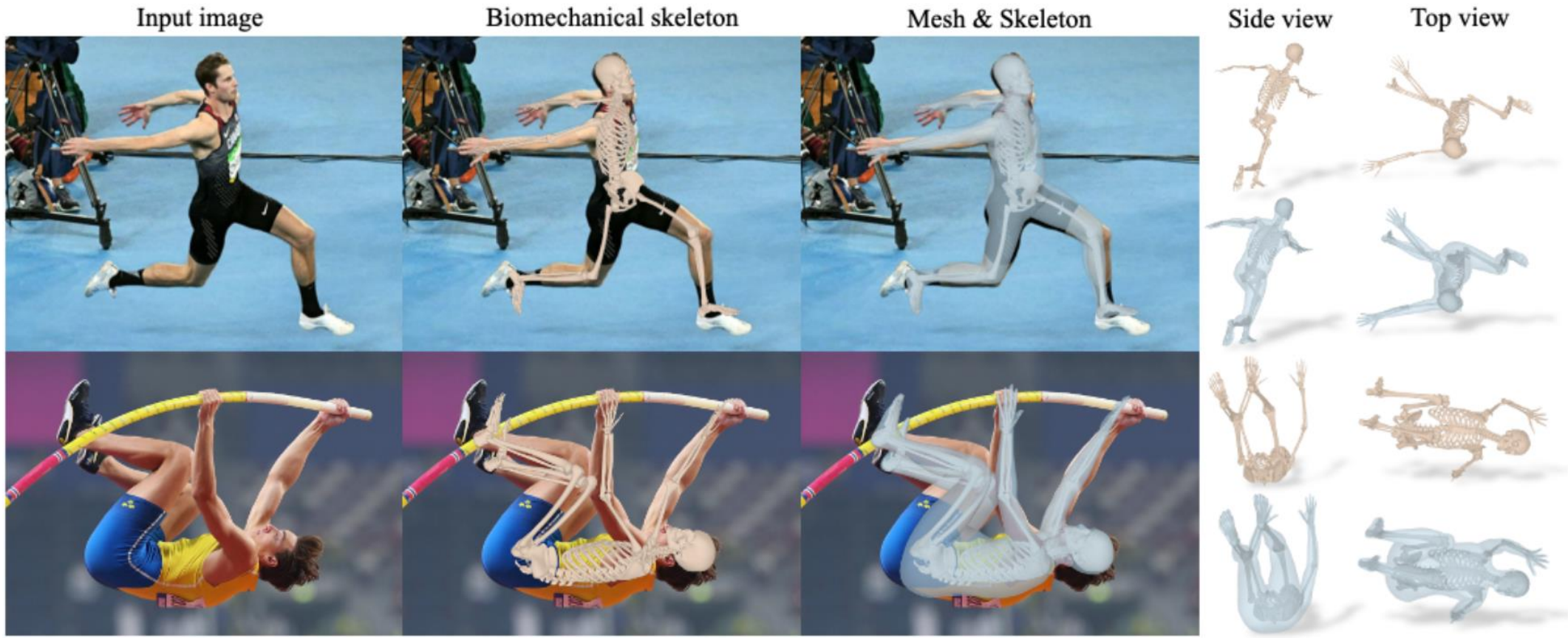


Figure 1. **Human Skeleton and Mesh Recovery (HSMR)**. We propose an approach that recovers the biomechanical skeleton and the surface mesh of a human from a single image. We adopt a recent biomechanical model, SKEL [24] and train a transformer to estimate the parameters of the model. We encourage the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

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.

From Skin to Skeleton:

Towards Biomechanically Accurate 3D Digital Humans

Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, Michael J. Black

SIGGRAPH ASIA 2023

[Paper] [Supplementary] [Code]

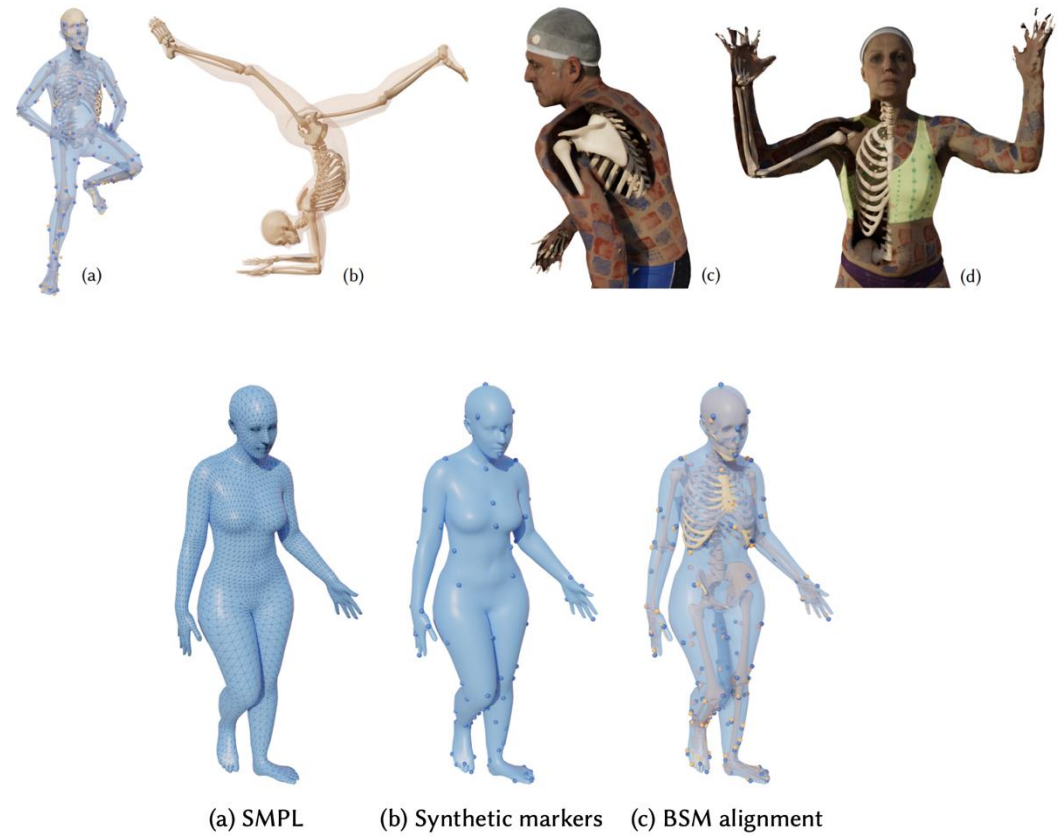


Fig. 2. Creation of the paired skeleton and body dataset. Given a SMPL motion sequence (a), we generate synthetic markers (b), and fit a biomechanical model to the makers using AddBiomechanics [Werling et al. 2022] (c).

Preliminary – SKEL model

SMAL Model

$\bar{T} \in \mathbb{R}^{3N}$ A mesh template in the zero pose.

$W \in \mathbb{R}^{N \times K}$ Blend weights

$N = 6890$ $K = 23$ $|\vec{\beta}| = 10$ $|\vec{\theta}| = 72 = 24 * 3$

shape pose

Key difference

Limitation:

Treat every articulation joint as a ball (socket) joint with three degrees of freedom.

SMAL and SKEL share shape space, mesh topology and kinematic tree.

Continuous rotation representation

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

SKEL carefully designs the kinematic parameters according to the real human biomechanical structure and only models the realistic degrees of freedom.

Each pose parameter corresponds to a single degree of freedom and is represented as an Euler angle.

wrists

24 joints, 10 have 3 DoF, 12 have 1 DoF, 2 have 2 DoF.
3 DoF: 6 values per joints (3x3 rotation matrix)
1 DoF: 2 values per joints (2x2 rotation matrix)
2 DoF: Directly regress the Euler angles.
Pose regression target: 88 = 6x10 + 2x12 + 2x2

Overview

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

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

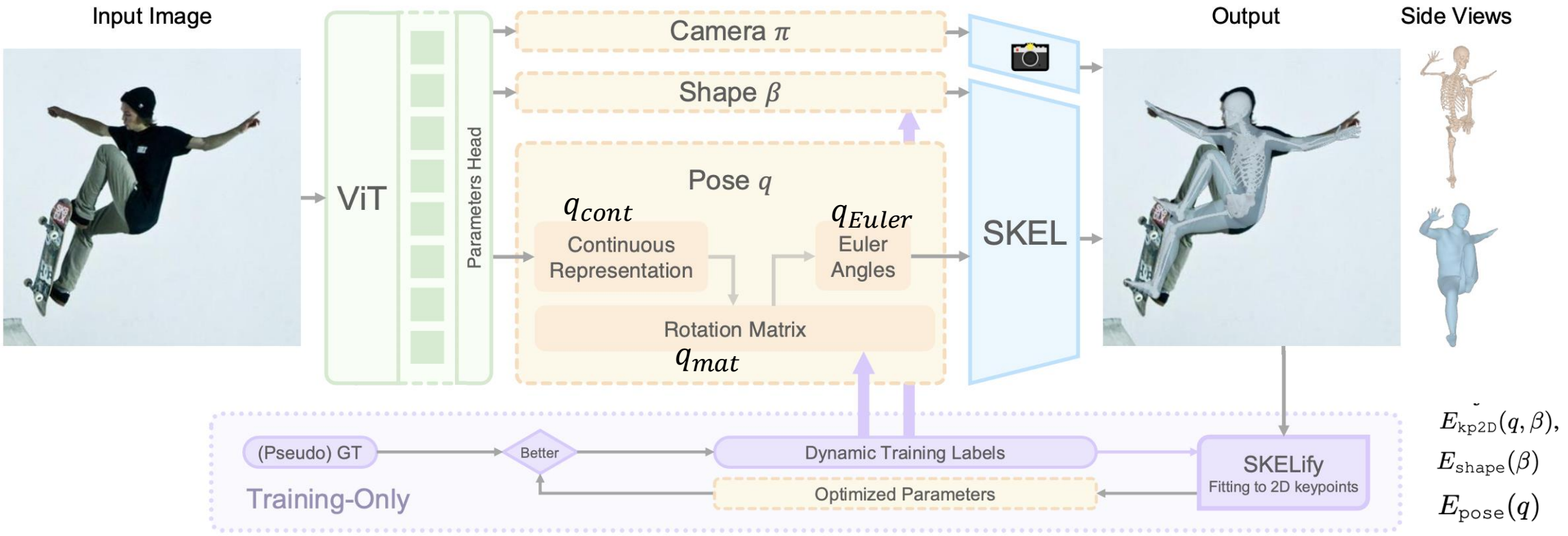


Figure 2. **Overview of our HSMR approach.** A key design choice of HSMR is the adoption of the SKEL parametric body model [24] which uses a biomechanically accurate skeleton. We employ a transformer-based architecture that takes as input a single image of a person and estimates the pose q and shape 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.

Training Data Generation

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

Initial pseudo GT:

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.

Quality Control (initial filtering stage)

ters 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.

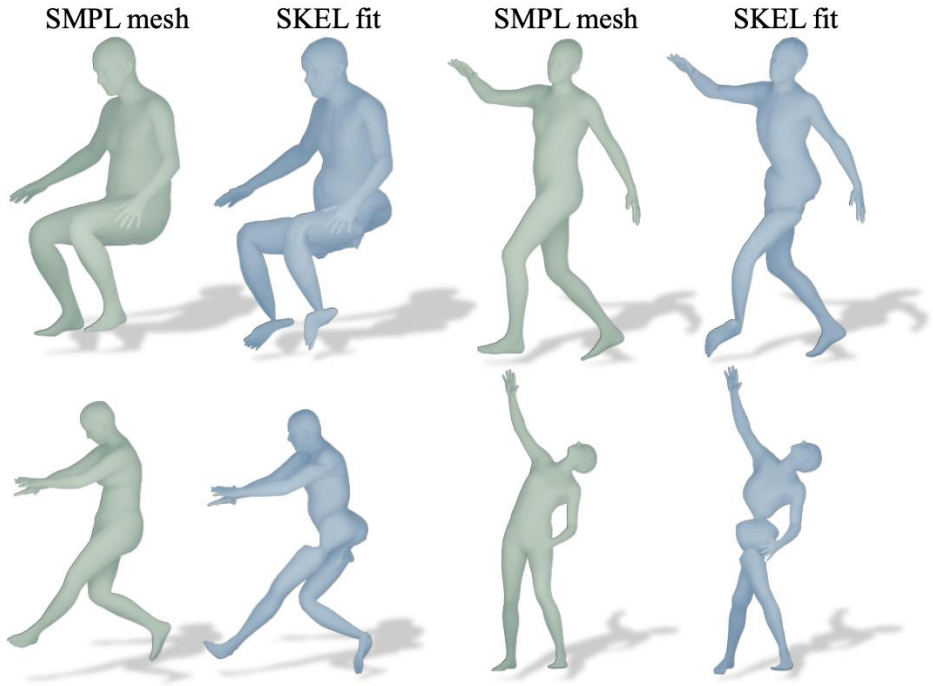


Figure 3. **Failure cases of SMPL-to-SKEL conversion.** While we can technically fit SKEL to an instance of the SMPL model, this conversion can often lead to problematic SKEL results. Here, we visualize SMPL meshes (light green), and the SKEL meshes we get when we try to fit the SKEL model to the SMPL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

maxPVE Check

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

$$\text{maxPVE} > 6\text{cm}.$$

Filtered images can potentially obtain pseudo ground truth SKEL parameters during the iterative refinement

Training with Pseudo-Label Refinement

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.*,

Shape Prior

$$E_{\text{shape}}(\beta) = \|\beta\|^2.$$

Pose Prior

$E_{\text{shape}}(\beta) = \|\beta\|^2$. 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:

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

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.

Training with Pseudo-Label 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.

Experiments

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.

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

	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**

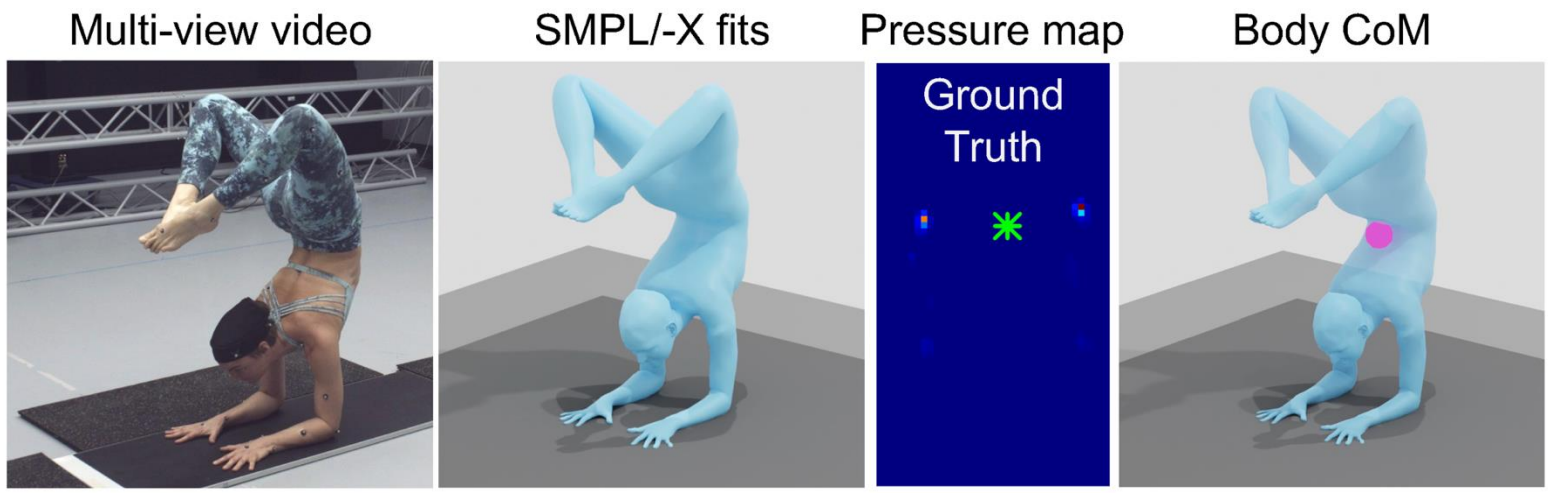
Table 2. **Evaluation of the surface reconstruction accuracy.** We report MPVPE and PA-MPVPE on the MOYO dataset.

Experiments

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).

The MoYo Dataset contains



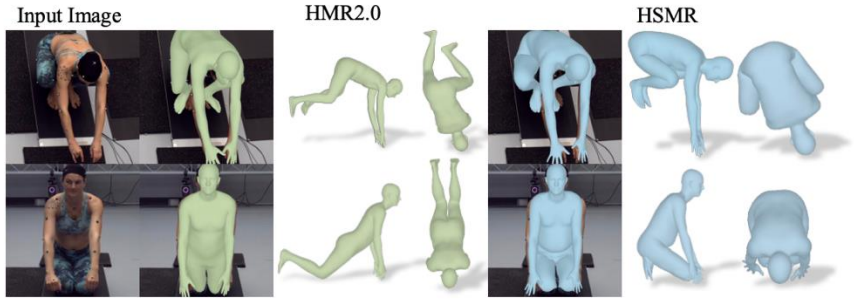


Figure 6. **Qualitative comparison with HMR2.0 on MOYO.** For each example we show the input image and results for HMR2.0 and HSMR. Although the interpretation in the input view is reasonable for both methods, HSMR achieves more accurate 3D reconstruction on the challenging poses and viewpoints of MOYO.

Experiments

Biomechanically-sound reconstruction

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.

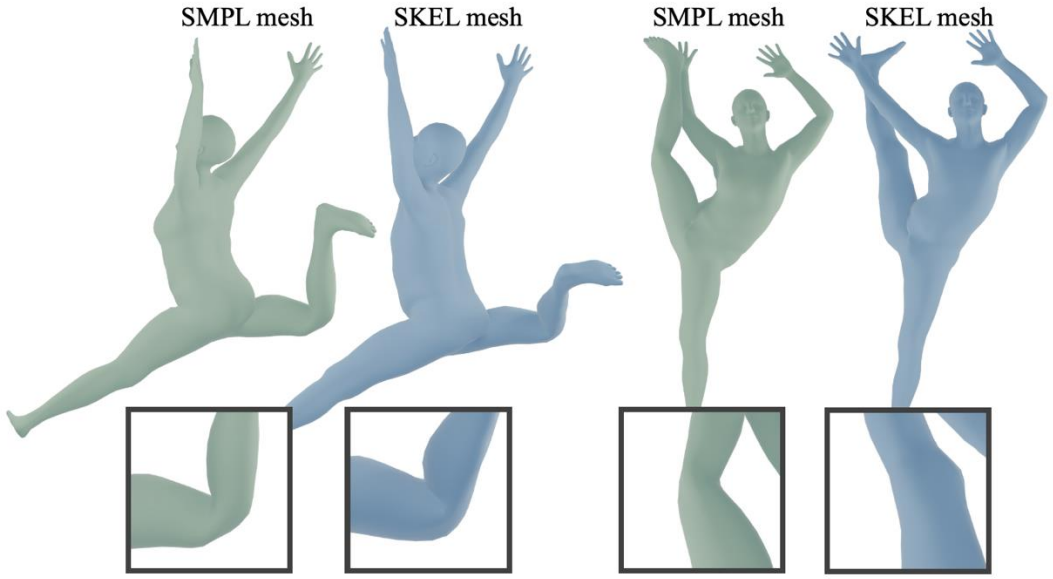


Figure 4. **Examples of unnatural joint rotation for SMPL.**

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.

Experiments

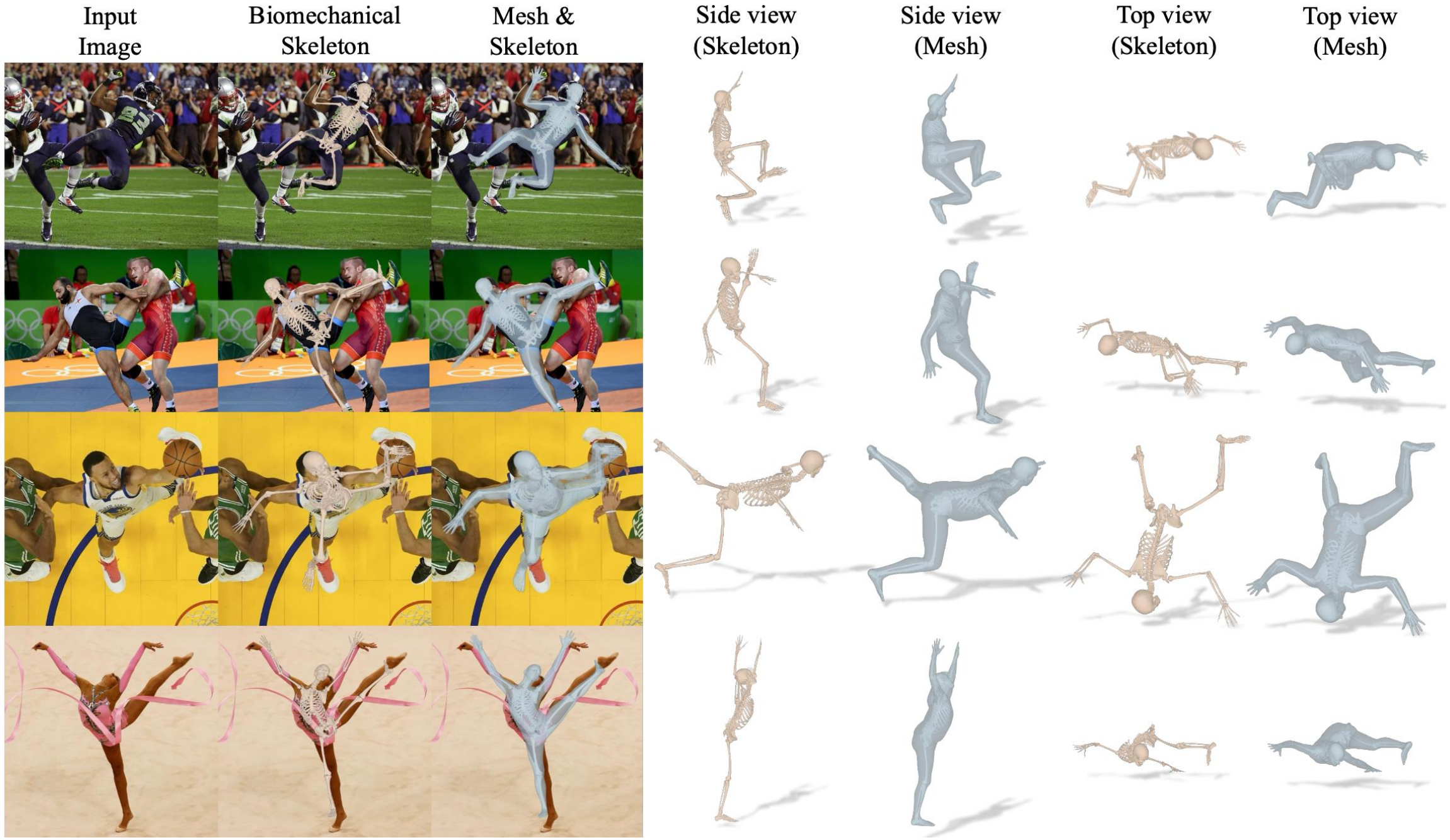
Models	COCO		LSP-Extended		PoseTrack		3DPW		Human3.6M		MOYO	
	@0.05↑	@0.1↑	@0.05↑	@0.1↑	@0.05↑	@0.1↑	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
HSMR (ViT-B)	**0.79**	**0.94**	**0.38**	**0.70**	**0.86**	**0.96**	**76.7**	**50.0**	**49.8**	**37.1**	**124.0**	**92.6**
HSMR (ViT-B) w/ Euler angles	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

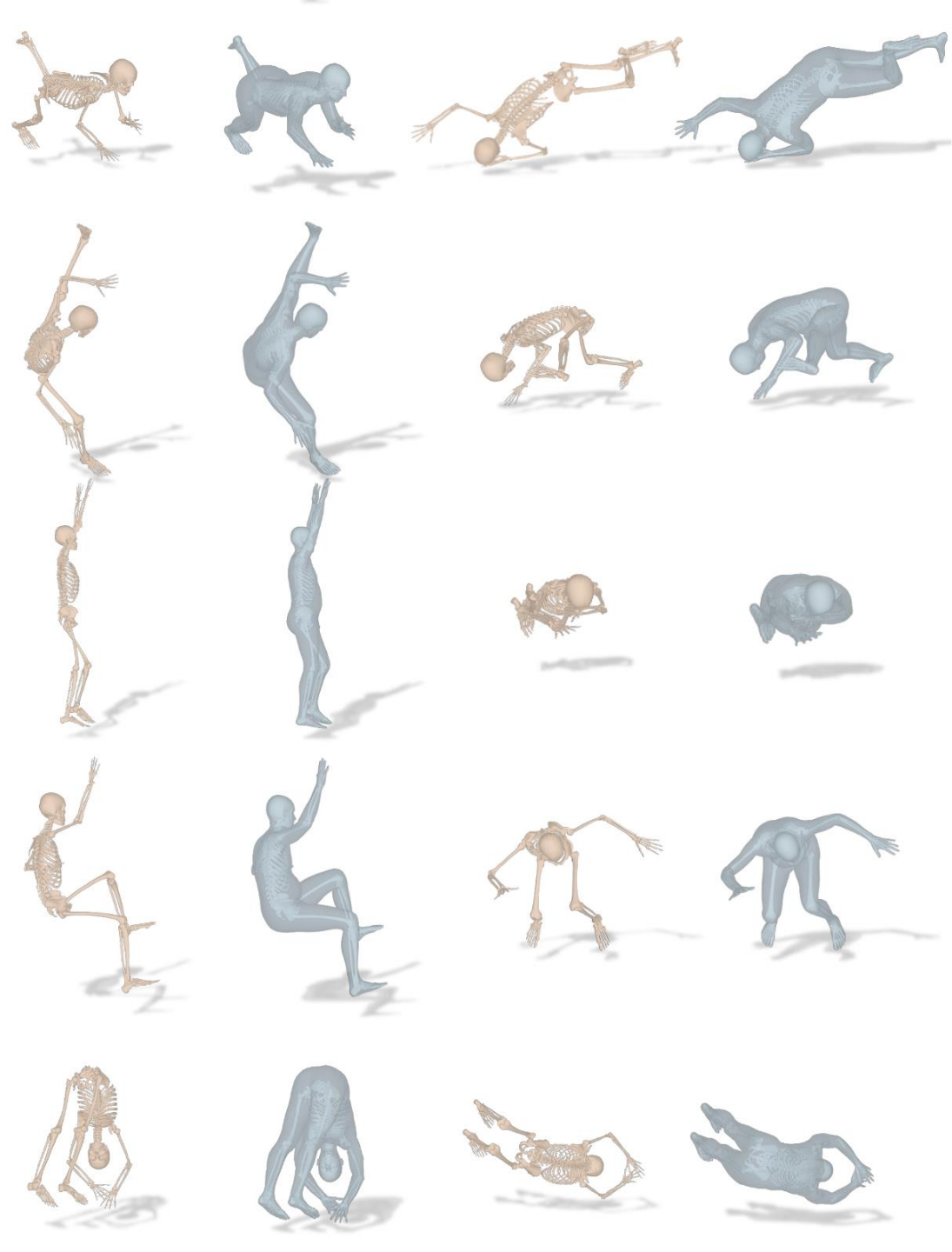
Table 5. **Ablation study on design choices.** We benchmark our proposed model and ablate two design choices. First, we change the regression target from the continuous representation [64] to the native Euler angles of SKEL. This has a negative effect across the board. Then, we experiment without the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

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.


Input Image
Biomechanical Skeleton
Mesh & Skeleton
Side view (Skeleton)
Side view (Mesh)
Top view (Skeleton)
Top view (Mesh)



Limitation

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].

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