Supplementary Material for: "Humans in 4D: Reconstructing and Tracking Humans with Transformers"

Shubham Goel Georgios Pavlakos Jathushan Rajasegaran Angjoo Kanazawa* Jitendra Malik*
{shubham-goel, pavlakos, jathushan, kanazawa}@berkeley.edu, malik@eecs.berkeley.edu
University of California, Berkeley

We provide more details about HMR 2.0, *i.e.*, the architecture we use (Section S.1), the data (Section S.2) and the training pipeline (Section S.3). Furthermore, we describe the aspect of pose prediction (Section S.4) and we discuss the metrics we use for evaluation (Section S.5). Then, we discuss the experimental settings for tracking (Section S.6), and action recognition (Section S.7). Finally, we provide additional qualitative results (Section S.8).

S.1. HMR 2.0 architecture details

The architecture of our HMR 2.0 model is based on a ViT image encoder and a transformer decoder. We use a ViT-H/16 ("huge") pre-trained on the task of 2D keypoint localization [25]. It has 50 transformer layers, takes a 256×192 sized image as input, and outputs 16×12 image tokens, each of dimension 1280. Our transformer decoder is a standard transformer decoder architecture [23] with 6 layers, each containing multi-head self-attention, multi-head cross-attention, and feed-forward blocks, with layer normalization [2]. It has a 2048 hidden dimension, 8 (64-dim) heads for self- and cross-attention, and a hidden dimension of 1024 in the feed-forward MLP block. It operates on a single learnable 2048-dimensional SMPL query token as input and cross-attends to the 16×12 image tokens. Finally, a linear readout on the output token from the transformer decoder gives pose θ, shape β, and camera π.

S.2. Data details

In our training, we adopt the training data conventions of previous works [10], using images from Human3.6M [4], COCO [13], MPII [1] and MPI-INF-3DHP [18]. This forms the training set for the version we refer to as HMR 2.0a in the main manuscript. For the eventual HMR 2.0b version, we additionally generate pseudo-ground truth SMPL [14] fits for images from AVA [3], InstaVariety [6] and AI Challenger [24]. Since AVA and InstaVariety include videos, we collect frames by sampling at 1fps and 5fps respectively. For pseudo-ground truth generation, we use ViTDet [11] for bounding box detection and ViTPose [25] for keypoint detection, while fitting happens using ProHMR [10]. We discard detections with very few 2D detected keypoints (less than five) and low detection confidence (threshold 0.5). We also discard fits with unnatural body shapes (*i.e.*, body shape parameters outside $[-3, 3]$), unnatural body poses (computed using a per-joint histogram of poses on AMASS [17]), and large fitting errors (*i.e.*, which indicates that the reconstruction was not successful). For training our HMR 2.0b model, we sample with different probabilities from each dataset, *i.e.*, Human3.6M: 0.1, MPII: 0.1, MPI-INF-3DHP: 0.1, AVA: 0.15, AI Challenger: 0.15, InstaVariety: 0.2, COCO: 0.2.

S.3. Training details

We train our main model using 8 A100 GPUs with an effective batch size of $8 \times 48 = 384$. We use an AdamW optimizer [15] with a learning rate of 1e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1e-4. Training lasts for 1M iterations, which takes roughly six days. For our main model HMR 2.0b, we train the network end-to-end. However, for the HMR 2.0a variant, the ViT encoder remains frozen, allowing a larger effective batch size of $8 \times 512 = 4096$, learning rate of 1e-4, and fewer training iterations of 100K (*i.e.*, roughly equivalent number of epochs).

While training, we weigh the different losses. $\mathcal{L}_{\texttt{kp3D}}$, $\mathcal{L}_{\texttt{kp2D}}$, and $\mathcal{L}_{\texttt{adv}}$ have weights 0.05, 0.01, and 0.0005 respectively. The terms within $\mathcal{L}_{\texttt{smpl}}$ are also weighed differently, the θ and β terms weigh 0.001 and 0.0005 respectively.

S.4. Pose prediction

For the pose prediction model, we train a vanilla transformer model [23] from the tracklets obtained by [19]. Each tracklet at every time instance contains 3D pose and 3D location information, where the pose is parameterized by the SMPL model [14] and the location is represented as the translation in the camera frame. The transformer has 6 layers and 8 self-attention heads with a hidden dimension of 256. Each output token regresses the 3D pose and 3D loca-

tion of the person at the specified time-step. We train this model by randomly masking input pose tokens and applying the loss on the masked tokens. During inference, to predict a future 3D pose, we query the model by reading out from a future time-step, using a learned mask-token as input to that time-step. Similarly for amodal completion, we replace the missing detections with the learned mask-token and read out from the output at the corresponding time-step. The model is trained with a batch size of 64 sequences and a sequence length of 128 tokens. We use the AdamW optimizer [15] with a learning rate of 0.001 and $\beta_1 = 0.9, \beta_2 = 0.95$.

S.5. Metrics

For our evaluation, we use the metrics that are common in the literature:

3D Pose: We follow [5] and we use MPJPE and PA-MPJPE. MPJPE refers to Mean Per Joint Position Error and it is the average L2 error across all joint, after aligning with the root node. PA-MPJPE is similar but is computed after aligning the predicted pose with the ground-truth pose using Procrustes Alignment.

2D Pose: We use PCK as defined in [26]. This is the Percentage of Correctly localized Keypoints, where a keypoint is considered as correctly localized if its L2 distance from the ground-truth keypoint is less than a threshold t. We report results using different thresholds (@0.05 and @0.1 of image size).

Tracking: Following [20, 21], we use standard tracking metrics. This includes ID switches (IDs), MOTA [7], IDF1 [22], and HOTA [16].

Action Recognition: We report results using mAP metrics as defined in the AVA dataset [3]. We further provided a more fine-grained analysis reporting results on different action categories: actions that involve Object Manipulation (OM), actions that involve Person Interactions (PI), and actions that involve Person Movement (PM). The results in these categories are also reported using mAP.

S.6. Tracking with PHALP$'$

In the main manuscript, we compare different human mesh recovery systems on the downstream problem of tracking (Table 3 of the main manuscript). For this, we modify the PHALP approach [21], so that pose distance is computed on the SMPL space that all the models share. To make this comparison fair, we keep other variables similar to the original PHALP (*e.g.*, same appearance embedding). Note that this comparison is generous to baselines that do not model appearance themselves. Eventually, our final 4DHumans system uses a sampling-based appearance head and our new pose prediction, which lead to the state-of-the-art performance for tracking on PoseTrack (Table 4 of the main manuscript). To model appearance, we texture visible points on the mesh by projecting them onto the input image and sampling color from the corresponding pixels.

S.7. Action recognition

As an alternative way to assess the quality of 3D human reconstruction, we evaluate various human mesh recovery systems on the downstream task of action recognition on AVA (please refer to [19] for more details on the task definition). More specifically, we take the tracklets from [19], which were generated by running PHALP [21] on the Kinetics [8] and AVA [3] datasets. Then, we replace the poses from various human mesh recovery models (*i.e.*, PyMAF [28], PyMAF-X [27], PARE [9], CLIFF [12], HMAR [21], HMR 2.0) and evaluate their performance on the action recognition task. In this pose-only setting, the action recognition model has access only to the 3D poses (in the SMPL format) and 3D location and is trained to predict the action of each person. For a fair comparison and to achieve the best performance for each 3D pose regressor, we retrain the action recognition model specifically for each 3D pose method.

S.8. Additional qualitative results

We have already provided a lot of qualitative results of HMR 2.0, both in the main manuscript and in videos on the project webpage. Here, we provide additional results, including comparisons with our closest competitors (Figure S.1), and a demonstration of our results in a variety of challenging cases, including successes (Figure S.2) and failure cases (Figure S.3).

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.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] 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.

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

Figure S.1: **Qualitative comparison of our approach with state-of-the-art methods.** We compare HMR 2.0 with our closest competitors, PyMAF-X [27], PARE [9] and CLIFF [12]. For each example, we show the input image, and results from each method (including the frontal and a side view). HMR 2.0 is significantly more robust in a variety of settings, including images with unusual poses, unusual viewpoints and heavy person-person overlap.

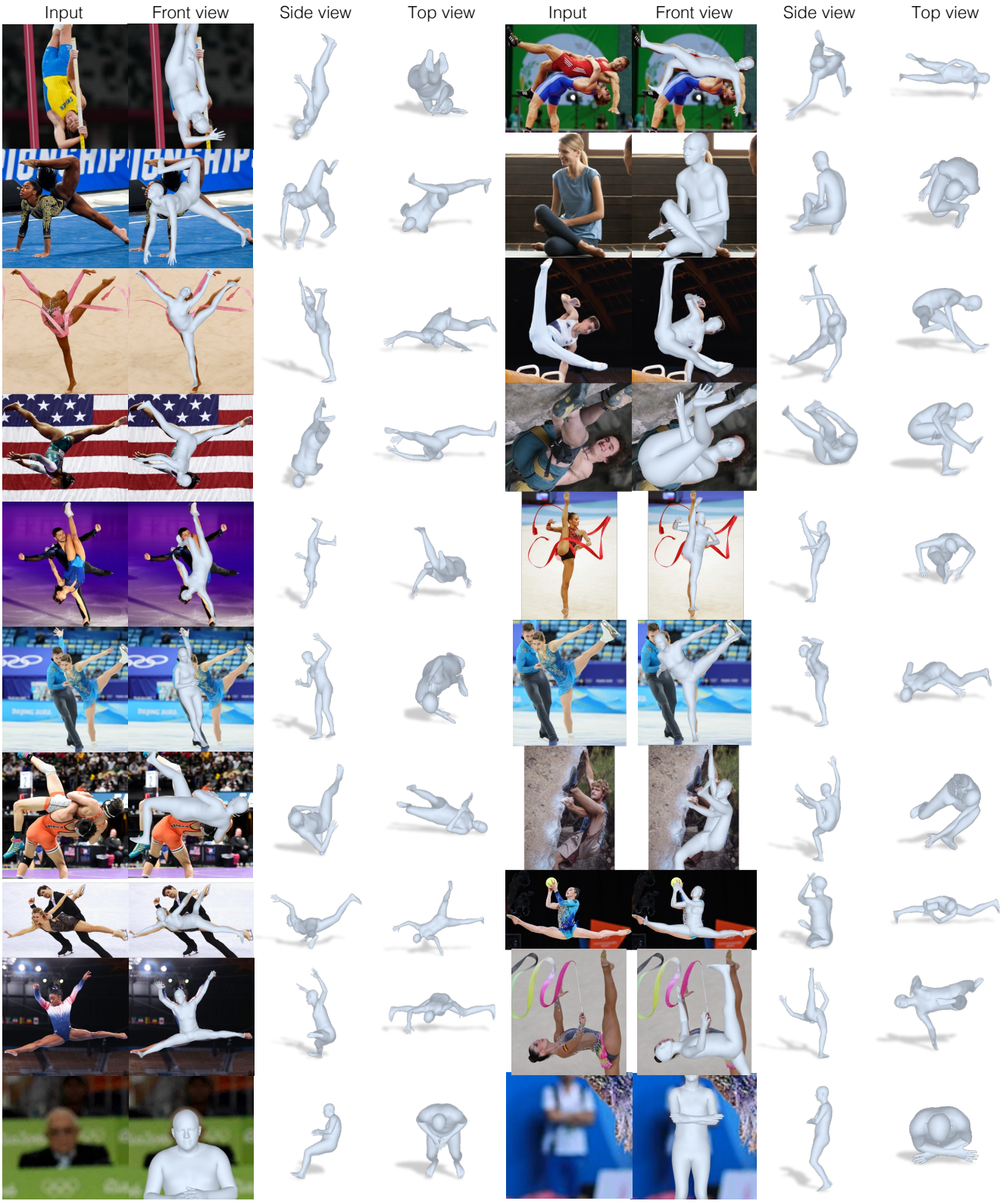

Figure S.2: **Qualitative results of our approach on challenging examples.** For each example we show the input image, the reconstruction overlay, a side view and the top view. The examples include unusual poses, unusual viewpoints, people in close interaction, extreme truncations and occlusions, as well as blurry images.

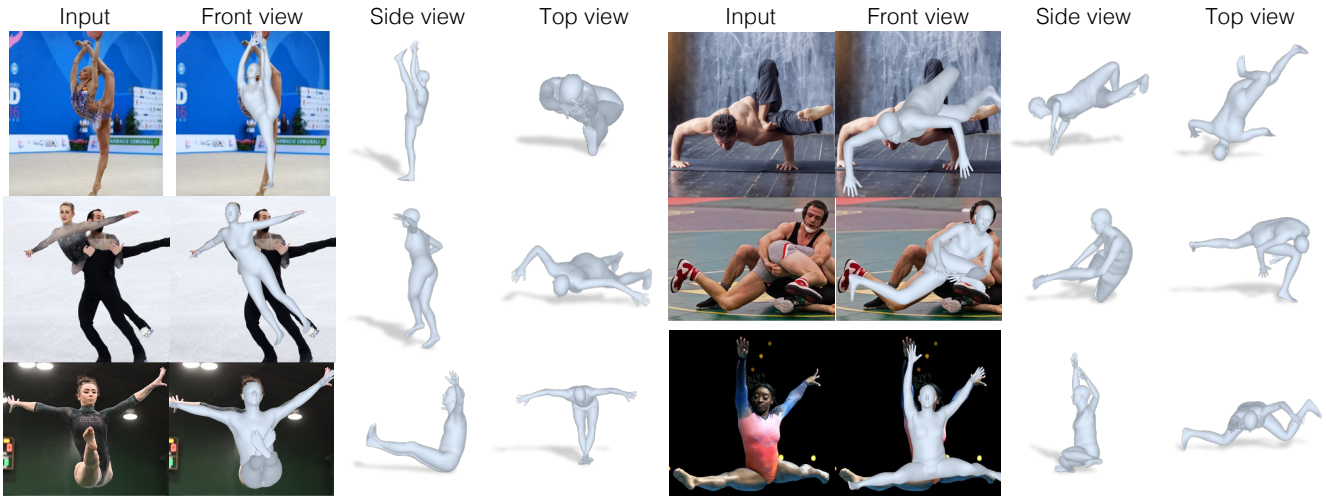

Figure S.3: **Failures of single frame 3D human reconstruction with HMR 2.0.** Despite the increased robustness of our method, we observe that HMR 2.0 occasionally recovers erroneous reconstructions in cases with very unusual articulation (first row), heavy person-person interaction (second row), and very challenging depth ordering for the different body parts (third row).

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

[7] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *PAMI*, 2008.

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

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

[10] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021.

[11] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022.

[12] 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.

[13] 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.

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[16] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *IJCV*, 2021.

[17] 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.

[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.

[19] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3D tracking and pose for human action recognition. In *CVPR*, 2023.

[20] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3D representations. In *NeurIPS*, 2021.

[21] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location and pose. In *CVPR*, 2022.

[22] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[24] 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.

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

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

[27] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *PAMI*, 2023.

[28] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021.