Supplementary Material for: Human Mesh Recovery from Multiple Shots

Georgios Pavlakos, Jitendra Malik, Angjoo Kanazawa
University of California, Berkeley

This Supplementary Material provides more details about our approach that were not included in the main manuscript due to space constraints. Here, the supplementary material includes additional quantitative and qualitative evaluation, details about our processing on AVA, as well as more details about our method and evaluation.

1. Additional quantitative evaluation

Smoothness terms: The proposed multi-shot optimization incorporates a temporal smoothness regularization both on the pose parameters and on the 3D joints. Here, we provide a more fine-grained ablative, where we investigate the effect of each term in the optimization. We report the results in Table 1, where we observe that using both terms together improves upon using each term independently, which justifies the existence of the two terms during the optimization.

Detailed results on Partial Humans dataset: In Table 2 of the main manuscript, we provided results on the Partial Humans dataset [26]. This dataset includes four subsets (*i.e.*, VLOG, YouCookII, Instructions and Cross-Task), but for compactness, we reported the mean PCKh values, averaged across the four subsets. Here, we provide the detailed performance on each subset for the different methods/versions in Table 2.

Additional results for Transformer on 3DPW: In the main manuscript, we investigated the performance of different temporal encoding architectures on multi-shot movie data, and observed the benefit of the transformer encoder compared to convolutional or recurrent architectures when training/testing on movie sequences (Table 3 of the main manuscript). Here, we provide further results on the more conventional setting of monocular sequences, specifically on the popular 3DPW dataset [29]. For this comparison, we focus on the architecture of the temporal encoder, so we use the public code from VIBE [13], adopting all their training details, but using our proposed transformer architecture for the temporal encoder. The results of Table 3 indicate that our transformer-based temporal encoder achieves results that are on par with the recurrent encoder of VIBE. This demonstrates that our choice of a transformer-based temporal encoder while being a more appropriate choice when training/testing with movie sequences, which remains the main focus of this work, it is also on par with the recent state-of-the-art for conventional tasks (*i.e.*, monocular sequences from 3DPW). Simultaneously, this acts as further evidence that the improved performance on AVA is primarily an effect of the transformer being a more appropriate choice for movies (*i.e.*, not because it is a stronger architecture in general).

Optimization	H3.6M ↓	AVA ↑
Multi shot (no parameter smoothness)	63.5	49.0
Multi shot (no joints smoothness)	61.5	47.5
Multi shot (full)	**59.2**	**55.2**

Table 1. **The effect of different smoothness terms on multi-shot optimization.** We show PA-MPJPE (Human3.6M) and cross-shot PCK at $\alpha = 0.1$ (AVA). The combined application of smoothness terms on the joints and the pose parameters during our multi-shot optimization improves results compared to independent application of each term alone.

	PCKh on Partial Humans Uncropped ↑			
Method	VLOG	YouCook	Instr	Cross
HMR [11]	81.2	93.6	86.9	92.7
GraphCMR [16]	65.7	80.1	77.5	79.3
SPIN [15]	73.4	85.1	85.6	85.5
Partial Humans* [26]	68.7	95.4	77.9	91.1
ProHMR [17]	88.2	98.4	92.5	97.5
PARE [14]	**91.3**	96.4	93.5	96.3
HMR$^+$	86.9	96.8	92.4	96.2
+ AVA (2D keypoints)	88.1	98.3	92.0	97.3
+ AVA (single frame optim)	87.8	98.3	92.1	97.4
+ AVA (multi shot optim)	90.3	**98.9**	**94.1**	**98.2**

Table 2. **Detailed results on Partial Humans.** Note * operates in the harder setting, which uses the entire image as an input while others operate on cropped bounding boxes.

2. Additional qualitative results

We show more qualitative results in Figures 1 and 2. Figure 1 extends Figure 7 of the main manuscript providing examples from typical cases where the multi-shot optimization can improve compared to naive reconstruction without

Model	MPJPE	PA-MPJPE	PVE	accel error
VIBE	56.5	93.5	113.4	27.1
t-HMMR	55.6	94.3	112.9	29.7

Table 3. **Quantitative evaluation on 3DPW.** When compared on conventional monocular benchmarks, transformer is on par with state-of-the-art recurrent networks.

leveraging the shot continuity. Figure 2 provides results on the Partial Humans dataset [26], and compares our baseline HMR$^+$ model with our final model trained on frames from AVA with pseudo-ground truth from multi-shot optimization. The results of Figure 2 indicate that training with AVA frames using our pseudo-ground truth can improve accuracy and robustness on challenging examples, even when tested outside AVA.

Failure cases: The most common failure for the multi-shot optimization happens in cases with limited visibility of the person, *e.g.*, a head shot with only 2 or 3 detected keypoints. In fact, in 12% of the shot changes used for the evaluation on AVA, no keypoints are detected [4] for one of the two frames on the shot boundary. This means that the optimization cannot leverage the second shot. In these cases, predictive models, like the proposed transformer, could take advantage of the image cues (pixels) even in the absence of keypoint detections. Yet, extreme close-up shots (less than 5% of the body visible) is a challenge for predictive models too. In Figure 3, we present some representative examples with very limited visibility where the single frame regression model is producing incorrect reconstructions.

3. AVA dataset

A large part of our experiments is performed using the AVA dataset. After the preprocessing described in Section 3 of the main manuscript, we generate 6.7k tracklets. For each tracklet, we apply our multi-shot optimization which is used for pseudo ground-truth generation. Other forms of 2D detections could be applicable for the optimization (e.g., potentially silhouettes [18] or dense correspondences [7]), but we use 2D keypoints which tend to be the more robust and are used as the main source of supervision by recent work on human mesh recovery, *e.g.*, [2, 6, 10–13, 15, 16, 22–24, 26, 28]. Since only 2D keypoints are used, it is possible that the recovered body shape is not very accurate, but this is common issue in all of the above human mesh recovery approaches, where the regressed shape is often close to the mean human shape. Instead, our focus here is primarily on the pose estimation part, which is also reflected in the metrics we use for evaluation.

Test set verification: For the test set, we perform manual verification on the recovered tracklets to eliminate errors in shot change detection, tracklet Re-ID, or 2D keypoints. This manual cleaning returns a set of 2.3k instances with shot changes where the same person is visible in both frames (before and after the shot change). The reported cross-shot PCK metric is computed on these 2.3k pairs of images. For all results, we use the value $\alpha = 0.1$ to compute cross-shot PCK results. We highlight that we do not use the pseudo ground truth from multi-shot optimization for our evaluation, since this can still be inaccurate.

4. Method details

Preprocessing details: For our preprocessing, we use the Re-ID network of [8] to compute affinities between pairs of bounding boxes of humans. To avoid connecting identities in different scenes, we weight the affinity of each pair by the temporal distance (i.e., frames) between the two bounding boxes. If the affinity of a pair is larger than a threshold, we attribute the people in the corresponding bounding boxes to the same identity. For each bounding box, we also check if a set of keypoint detections from OpenPose is included in this box. If such detections exist, they are also associated with the corresponding bounding box, and are used as input to the multi-shot optimization.

Multi-shot optimization details: For our multi-shot optimization, we use the prediction of our baseline HMR$^+$ model as an initial estimate of the optimization for each frame. If the torso keypoints are not detected by the keypoint detector, we use the projection of the torso joints from the regressed model estimate as pseudo targets in the optimization. This helps to constrain the torso position and orientation in cases of extreme truncation. The multi-shot optimization lasts for one stage. This is equivalent to the last stage of the 4-stage SMPLify [3] optimization. We keep the same weights for the data term E_{proj} and the prior terms E_{prior}. More specifically, with E_{proj} we refer to the same data term as SMPLify, while with E_{prior}, we refer to the two body pose priors (GMM and angle prior) and the body shape prior from SMPLify [3]. The weights of the smoothness terms are set to $1e + 7$. For the optimization, the body root is considering to be in the origin before applying the rotation R_{gl}^T. The translation is applied afterwards, as part of the camera transformation. Consistency for camera parameters and body shape is enforced (focal length is constant for all cameras; we heavily penalize changes in the body shape parameters).

Novelty: From a technical perspective, our multi-shot optimization is similar to previous optimization approaches, *e.g.*, [3, 23, 25], but our contribution lies in 1) demonstrating that with some modification (*i.e.*, inference in the canonical frame, instead of the camera frame), we can adapt a canonical optimization example (SMPLify) to the multi-shot scenario, so that we can benefit from multi-shot continuity, and 2) demonstrating that with this novel insight of multi-shot

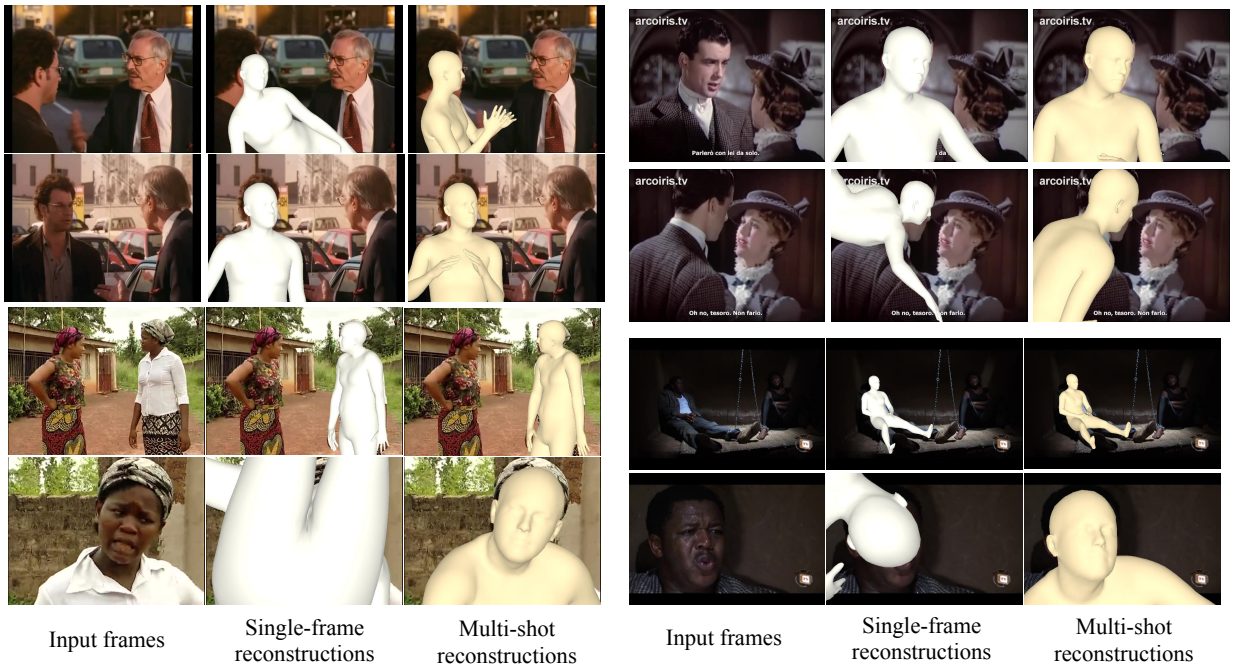

Figure 1. **Qualitative effect of our multi-shot optimization.** Extension of Figure 6 of the main manuscript. Although the single frame optimization baseline fails on more challenging frames with heavy truncation, our multi-shot optimization leverages information from the less ambiguous frame across the shot boundary to get a more accurate 3D reconstruction overall.

continuity, we can also also benefit in downstream tasks (training direct prediction models for human mesh recovery from image/video on movie sequences).

Pseudo-ground truth quality: As we discussed also in the main manuscript, our multi-shot reconstructions can include some incorrect 3D meshes. Depending on the quality required for each application, one could consider ways to manually, or automatically clean the reconstructions (*e.g.*, check for mesh interpenetrations, or evaluate consistency with other 2D cues, for example silhouettes). However, following other approaches that use optimization routines to generate pseudo-ground truth for training, *e.g.*, [2, 15], we do not perform elaborate filtering of the reconstructed data, and instead demonstrate their effectiveness by achieving improvements in downstream training tasks.

Body model choice: For our experiments, we use SMPL [21], since it is more widely adopted in the related literature for single/multi frame mesh recovery [11–13, 15]. Our optimization code also supports SMPL-X [23], however, the downstream methods using it are limited (*e.g.*, [5]), so we focus on SMPL to perform our analysis.

Regression model choice: As the basic component of our single-frame regression, we use the HMR model [11], which is widely adopted in the literature [2, 6, 15, 24, 27]. Most of our ablations also focus on models that share these design choices, *i.e.*, [11, 15, 16, 26]. However, for completeness, we have also included results from the most recent variations for human mesh recovery [14, 17].

Training details: For our single frame regression model, we use the same architecture as the original HMR model [11]. Following SPIN [15], the regression target for the SMPL pose parameters is expressed in the continuous 6D representation introduced by Zhou *et al*. [32]. To train our baseline model, we use images from Human3.6M [9], COCO [19] and MPII [1]. To benefit from supervision with SMPL parameters, in the case of Human3.6M we use MoSh [20] ground truth, while for COCO and MPII, we use SPIN [15] pseudo-ground truth. To train our final model, we also incorporate frames from AVA with pseudo-ground truth from multi-shot oprimization in the training. The batch size is equal to 64, learning rate is $1e-4$, and training lasts for 1.2M iterations. The weight for the keypoint reprojection loss L_{2D} is equal to 1, while the weight for the parameter losses L_{smpl} is equal to 0.1.

For our temporal t-HMMR model, following HMMR [12], the temporal receptive field is set to 13 frames. For the temporal encoder we use one transformer encoder layer with a single attention head. The batch size is set to 128 (128 subsequences of length equal to 13 frames), learning rate is set to $1e-4$ and training lasts for 100k iterations. The weights for the keypoint reprojection and parameter losses are the same with the HMR training. For

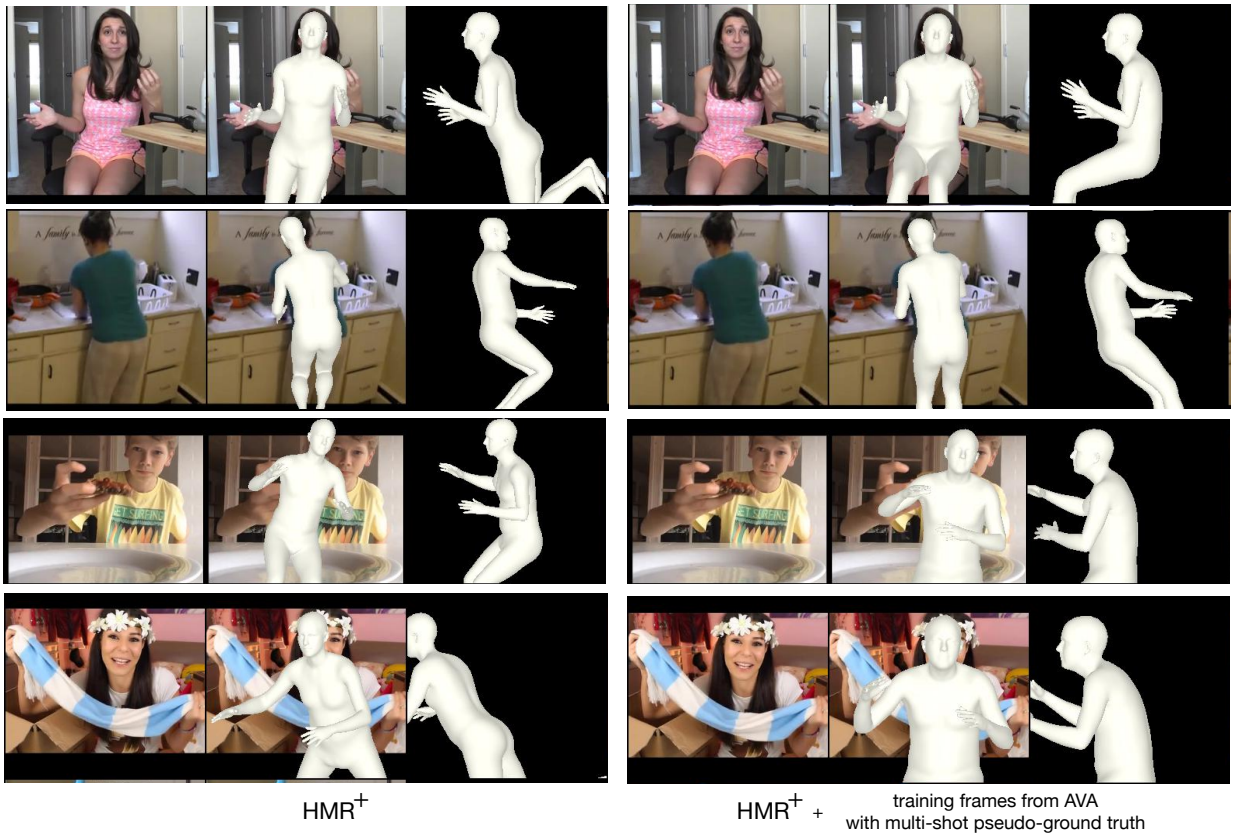

Figure 2. **Qualitative effect of training with data from AVA.** We present results on the Partial Humans dataset [26]. The inclusion of frames from AVA with multi-shot pseudo-ground truth can improve the accuracy and robustness on challenging examples, compared to our baseline HMR$^+$ model, even when it is tested outside AVA.

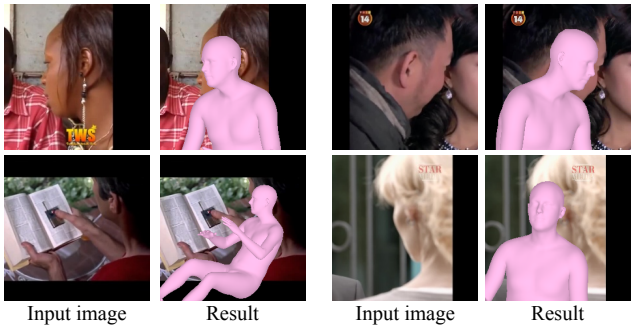

Figure 3. **Failure cases of our single-frame model.** Extreme close-up shots (*i.e.*, when less than 5% of the body is visible) is a huge challenge. Here we demonstrate some failures of our single frame human mesh regression model.

the temporal consistency losses, we use weight 0.015 for 3D keypoints $L^t_{\text{sm joint}}$ and 0.05 for parameter consistency $L^t_{\text{sm param}}$. Again, our training data combine sequences from standard benchmarks (*e.g.*, Human3.6M [9]) and our reconstructed AVA sequences. Also, for the non-transformer baselines, we explored two different strategies to deal with the missing frames (*e.g.*, due to b-rolls) in our AVA sequences; a) concatenate all frames together, ignoring missing frames and b) use a blank element (all zeros) to indicate a missing frame. Since the second strategy worked the best, we adopted it to report results.

5. Evaluation details

Cross-shot PCK: For the computation of the cross-shot PCK metric, we consider frames t and $t+1$, before and after the shot change, as well as the corresponding predictions for both frames, *i.e.*, the canonical shape, global orientation and camera parameters. For the evaluation, we project the canonical shape of frame t on frame $t+1$, using the global orientation and camera parameters estimated for frame $t+1$. Then, the typical PCK metric can be computed between the projected joints of the mesh and the detected keypoints for frame $t+1$. This procedure is also repeated in the opposite direction (projecting the canonical shape of frame $t+1$ on frame t using the camera estimated for frame t).

Evaluation metrics: In the main manuscript, we pro-

vide evaluations using a variety of metrics, depending on the ground truth information we have. Here, for completeness, we provide pointers to the definition of these metrics: **PA-MPJPE**: With MPJPE [9], we refer to the Mean Per Joint Position Error, *i.e.*, the Euclidean distance between ground truth and predicted 3D joints, averaged over all joints. With PA-MPJPE, we refer to this metric, when we can use Procrustes alignment to align the predicted 3D joints with the ground truth 3D skeleton before computing the per joint error. Detailed definition of this metric is provided in [31] (authors refer to it as "reconstruction error"). **PCKh**: PCK [30] refers to the percentage of correctly localized keypoints. For each predicted 2D keypoint, its distance from the ground truth keypoint is computed. Then, we consider as "correctly localized" the set of keypoints for which this distance is smaller than a specific threshold value and compute the percentage of this keypoints compared to all keypoints to report PCK. In the case of PCKh [1], for the threshold value, we use 50% of the head segment length.

Human3.6M evaluation setting: In Table 1 of the main manuscript we provide an evaluation on Human3.6M [9]. For this evaluation, we synthesized shot changes by using consecutive frames coming from different camera viewpoints. Following the usual evaluation, we use all actions from users S9 and S11, we introduce one shot change every second and we apply the three different optimization strategies (single frame, single shot, multi shot). In all cases, we use 2D keypoint detections from OpenPose [4] for the fitting. Eventually, the 3D reconstruction accuracy is estimated using the standard PA-MPJPE metric.

Partial Humans evaluation setting: Following previous work [11, 15, 16], our single frame model uses as input cropped bounding boxes of humans. Since the approach of [26] uses the full image as input, we evaluate their network using this setting (Table 2 of the main manuscript). We also experimented using the cropped bounding box as input, but this returned inferior results.

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. 3, 5

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 2, 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. 2

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *PAMI*, 2019. 2, 5

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

[6] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 2, 3

[7] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 2

[8] Qingqiu Huang, Wentao Liu, and Dahua Lin. Person search in videos with one portrait through visual and temporal links. In *ECCV*, 2018. 2

[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. 3, 4, 5

[10] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2021. 2

[11] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 5

[12] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 2, 3

[13] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2, 3

[14] 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

[15] 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, 2, 3, 5

[16] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 1, 2, 3, 5

[17] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 1, 3

[18] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 2

[19] 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. 3

[20] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 3

[21] 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. 3

[22] Gyeongsik Moon and Kyoung Mu Lee. I2L-meshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 2

[23] 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. 2, 3

[24] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 2, 3

[25] 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. 2

[26] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 1, 2, 3, 4, 5

[27] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3D human recovery in the wild. In *ICCV*, 2019. 3

[28] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 2

[29] 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. 1

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

[31] 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

[32] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3