A Simple Strategy for Body Estimation from Partial-View Images

Yafei Mao, Xuelu Li, Brandon Smith, Jinjin Li, Raja Bala

Amazon Inc.

{yafeimao, xueluli, smithugh, jinjinli, rajabl}@amazon.com

Abstract

Virtual try-on and product personalization have become increasingly important in modern online shopping, highlighting the need for accurate body measurement estimation. Although previous research has advanced in estimating 3D body shapes from RGB images, the task is inherently ambiguous as the observed scale of human subjects in the images depends on two unknown factors: capture distance and body dimensions. This ambiguity is particularly pronounced in partial-view scenarios. To address this challenge, we propose a modular and simple height normalization solution. This solution relocates the subject skeleton to the desired position, thereby normalizing the scale and disentangling the relationship between the two variables. Our experimental results demonstrate that integrating this technique into state-of-the-art human mesh reconstruction models significantly enhances partial body measurement estimation. Additionally, we illustrate the applicability of this approach to multi-view settings, showcasing its versatility.

1. Introduction

In modern online shopping, obtaining customers' body measurements is crucial for delivering accurate virtual try-on experiences [27, 30]. A common strategy is to request full-body photos from customers, taken in a designated pose and wearing tight-fitting clothing, along with additional information such as height and/or garment sizes, which are then used to estimate measurements [12, 42]. However, requiring full-body captures can create friction for customers, as it necessitates positioning the camera at a distance of at least 5-6 feet, which may not always be feasible in confined private spaces, as depicted in Fig. 1. Consequently, customers may occasionally submit partial body (view) images from the knees upwards.

Previous work [14, 18, 21, 41] has attempted to predict body shapes from monocular images. While they can provide reasonable results from full-body inputs, their performance on partial views is unsatisfactory [31]. The difficulty stems from the ill-posed nature of estimating the distance to
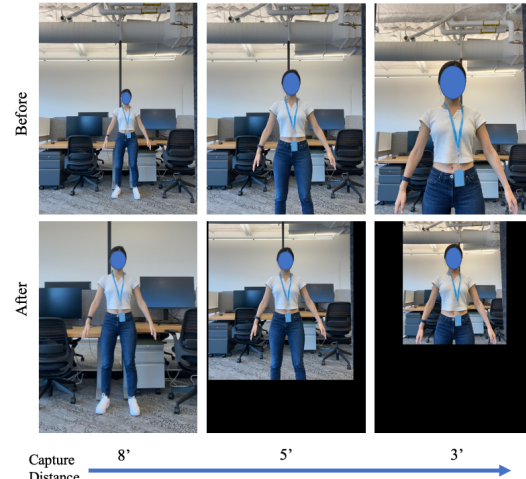

Figure 1. From left to right, the capture distance decreases from 8 to 3 feet. The height normalized versions are shown in the bottom row, where effective capture distance and scale are equalized.

the camera and the object's size jointly, as the same photo can correspond to multiple valid combinations of the two. When both parameters are unknown, ambiguity arises, such as the inability to determine if the person is close to the camera or has a large body size.

In this paper, we address, for the first time, the challenge of body size and capture distance ambiguity in 3D body measurement estimation from partial-view images. We introduce a novel method called *height normalization*, which aims to mitigate these issues by normalizing the scale across all images. Firstly, we estimate a reference 3D skeleton using the subject's gender, height, and the SMPL [25] model, and then project it to a designated location in the image. Next, utilizing an affine transformation computed from the target skeleton and the detected one, we resize and translate the subject in the images to the designated location. This normalization process allows the network to focus more on learning the characteristics of body dimensions regardless of the capture distance, leading to more accurate measurement predictions. We demonstrate through experiments that our method can be easily integrated into existing HMR
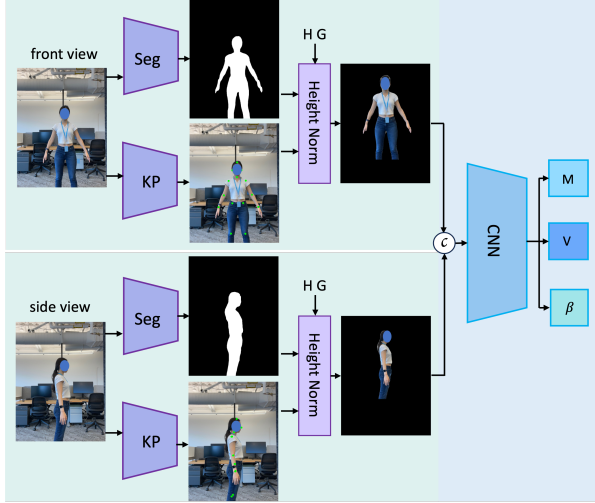
Figure 2. **Network architecture of BMN model** with height normalization. Segmentation masks and body landmarks are first predicted on each view using the respective models *Seg* and *KP*. Then, this information and the subject's height H and gender G data are passed to the height normalization module to transform the image. Finally, the two views are concatenated together and fed into the network to predict measurements M, mesh vertices V, and SMPL shape parameters β.

models, reducing body estimation errors by up to 2 inches on our real human dataset. Besides, we build a 2-view body estimation pipeline as shown in Fig. 2 to prove its applicability to multi-view settings.

Our key contributions are: (1) We identify and analyze the challenges related to capture distance and body size ambiguity in estimating body measurements from partial-view images, and propose a straightforward and modular solution. (2) We show that our approach can be easily integrated into existing monocular Human Mesh Recovery (HMR) models and extended to multi-view setups to enhance body measurement accuracy. (3) We conduct thorough experiments to validate the effectiveness of our approach.

2. Related Work

Current methods for obtaining body measurements from RGB images typically involve either regressing from the reconstructed mesh or directly from image features.

Human mesh reconstruction has received much attention over the years. While non-parametric methods directly regress a 3D body representation from images [5, 8, 19, 24, 32, 35], parametric methods rely on 3D body models like SMPL{-X} [25, 29] to estimate the shape and pose parameters via optimization [2, 29, 37, 39] or deep neural regression [1, 6, 7, 10, 15, 17, 22, 23, 31, 40]. SPIN [18] leverages a collaboration between regression and optimization-based approaches. EFT [14] augments existing 2D datasets with 3D pseudo-annotations, extreme crop
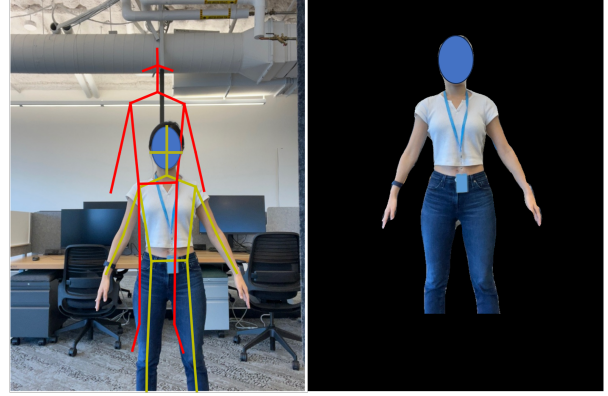

Figure 3. Height normalization involves transforming the skeleton of the person from its detected location (connected by yellow lines) to the designated location (connected by the red lines) through affine transformations.

augmentation and auxiliary inputs. PyMAF [41] rectifies predicted parameters through a mesh alignment feedback loop. HyBriK [21] improves body mesh estimation with inverse kinematics. While these methods achieve reasonable mesh alignment with subjects in full-body images, their performance degrades under partial visibility. Methods like [20, 33] predict body shape and/or pose distributions under occlusions. However, the complexity of their dedicated pipeline makes integration into other methods challenging.

Body measurement estimation models rely on silhouettes [9, 38] and/or synthetic data [34] to train deep networks to directly predict measurements. Many of these works are hindered by the lack of 3D real data, highlighting the necessity for an efficient data preprocessing strategy. Our proposed method can be easily integrated into the aforementioned pipelines, thus bridging the data gap with minimal effort.

3. Method

Estimating body measurements from partial-view images is challenging. As shown in the first row of Fig. 1, noticeably different-looking images may correspond to the same set of body shape parameters or measurements, posing a complex one-to-many learning problem for the network. Additionally, when dealing with partial-view images, directly estimating depth using camera and human height information is difficult, because the subject's height in the image cannot be easily obtained. To address these challenges, we propose a height normalization method in this section.

3.1. Height Normalization

Our height normalization strategy aims to position human subjects virtually in the scene at a target skeleton location (as illustrated by the red outline in Fig. 3). To achieve this, we first generate a reference 3D human mesh using an aver-

TP90 (inch)	H	BustGirth	WaistGirth	HipGirth	Overall$_{59}$
PyMAF [41]	✗	4.48	4.49	4.31	2.38
PyMAF [41]	✓	2.92 (**-1.56**)	3.67 (-0.82)	4.03 (-0.28)	1.74 (-0.64)
HybriK [21]	✗	4.09	4.88	5.57	2.58
HybriK [21]	✓	3.09 (-1.00)	3.71 (**-1.17**)	3.53 (**-2.04**)	1.65 (**-0.93**)
ETF [14]	✗	4.02	4.58	4.27	2.40
ETF [14]	✓	3.06 (-0.96)	3.68 (-0.90)	3.53 (-0.74)	1.66 (-0.74)
SPIN [18]	✗	3.82	4.51	4.17	2.41
SPIN [18]	✓	2.93 (-0.89)	3.79 (-0.72)	3.24 (-0.93)	1.67 (-0.74)

Table 1. Comparison of TP90 errors ($\downarrow$) in inches and their differences (in Green) on body measurements predicted by SOTA human mesh reconstruction models when trained with and without the height normalization (**H**). The largest difference is in **Bold**.

age SMPL [25] shape parameters derived from the subject's height and gender in an A-pose. Subsequently, we infer the skeletal joints with the SMPL model and render them onto the scene using a perspective camera. This camera is positioned at a distance of 70 inches away from the root of the SMPL mesh along the optical axis. This distance was empirically selected to ensure that individuals with a height of less than 78 inches can be fully projected into the image.

Next, we estimate the 2D body landmarks from the input image using a ResNet50-like keypoint detection model trained on synthetic data (other off-the-shelf keypoint detection models like OpenPose [3] can also be used, although we will not delve into them in our paper). The detection model outputs the coordinates and the visibility score of the joints, as depicted by the yellow outline in Fig. 3. From this information, we derive an affine mapping through least squares fit that aligns the subject's skeletal coordinates with those of the reference body. This transformation scales the partial-view image to a nominal scale and translates the body to center it in the image. We rely on the visible points to compute the transform. An example output of this process is illustrated in Fig. 1, showing the effective normalization of capture distances.

3.2. Data Augmentation Strategy

To train models capable of handling both fully visible and partially occluded bodies, we augment partial views by randomly cropping full-body images from the bottom at various occlusion levels. This process includes deliberately retaining some full-body images without any occlusion (occlusion level 0). The cropped images are then resized back to their original resolutions. The synthesized partial-view data undergoes height normalization before being fed into the network. Additionally, we use segmentation masks to suppress background pixels in the height-normalized images. This step ensures that irrelevant background information doesn't interfere with subsequent body measurement estimation. During test time, all images undergo height normalization directly without random resized cropping.

4. Results

4.1. Dataset and Metric

Public HMR datasets such as [13, 28, 36] do not include information about subjects' height, gender, or body measurements. Therefore, we collected a training dataset of 6,100 individuals with full-body front-view and left side-view images, along with their height and gender, 59 body measurements, and SMPL-aligned 3D scans. The 59 measurements cover various aspects such as girth, height, and length in areas like the bust, waist, hips, arms, legs, and neck, aiding precise body shape and size estimation. Additionally, we obtained at-home photos of 400 new subjects, excluding 3D scans, for our test set. Both training and test images were taken indoors under well-lit conditions, with subjects in an A-Pose wearing form-fitting attire.

Body estimation accuracy is reported in terms of the mean 90th percentile (TP90) errors in inches of the 59 body measurements across the test set subjects to provide a robust statistic of model performance. Additionally, TP90 errors for specific body dimensions are reported for a more granular understanding of the error distribution.

4.2. Using Height Normalization in HMR Models

We conducted fine-tuning experiments on state-of-the-art (SOTA) HMR models, including PyMAF [41], HybriK [21], SPIN [18], and EFT-HMR [14], both with and without height normalization.

As a standalone module, height normalization can easily serve as a pre-processing step for HMR models to improve body measurement accuracy. During training, all images are processed with height normalization before being passed to the network. Specifically, utilizing 6,100 frontal view images in our dataset, we first apply resized crop on the input image by uniformly sampling an occlusion level parameter α from between 0 and 0.4, mimicking scenarios from no occlusion (0) to truncation at the waist (0.4). Then we extract segmentation masks using DeepLab-V3 [4] and detect body landmarks. Subsequently, these partial-view data are height normalized and fed into the network. We incorporate an additional measurement head on top of the feature map from the last layer of the backbone to predict 59 body measurements. This head consists of two linear layers with ReLU activation, and training utilizes the L_2 loss function to compare ground truth measurements with predicted ones for supervision.

For all four HMR models, we use a batch size of 64 images, AdamW [26] optimizer with a learning rate of $1e-3$, and weight decay of $1e-4$, and trained for 25 epochs. We present the overall mean TP90 errors of the 59 measurements in inches, along with a detailed analysis of the TP90 errors for Bust Girth, Waist Girth, and Hip Girth. Results are reported based on the 400 images with varying degrees

of occlusion. The results are summarized in Table 1. It is evident that height normalization leads to a notable reduction in itemized errors, with decreases of up to 2.04 inches, and an overall error reduction ranging from 0.64 to 0.93 inches, thus validating its effectiveness.

4.3. Using Height Normalization in Multi-View

Next, we demonstrate that the proposed method can also be applied to multiple views. We developed a simple two-view Body Measurement Network (BMN) baseline model, incorporating the height normalization module separately for each view, as depicted in Fig. 2. We also tested sharing height normalization parameters between the two views and found consistent performance. The height-normalized two-view images are concatenated and forwarded into the CNN backbone and the three task heads for the prediction of body measurements, SMPL mesh vertices, and SMPL shape parameters, respectively. The loss function is a weighted combination of the L_1 loss of predicted vertices V, shape parameters β, and measurements M: $\mathcal{L}_{\text{total}} = w_1\mathcal{L}_{1V} + w_2\mathcal{L}_{1\beta} + w_3\mathcal{L}_{1M}$, where $[w_1, w_2, w_3] = [10, 1, 1]$.

Our BMN baseline uses the ImageNet pretrained ResNet50 [11] as the backbone and all the task heads comprise two fully connected layers with ReLU activation. The experiments are implemented using PyTorch and are conducted on an NVidia V100 GPU. The model is trained for 75 epochs with a batch size of 8, using the Adam [16] optimizer with an initial learning rate of $1e-4$. We reduce the learning rate to $0.88\times$ and $0.75\times$ at $112.5K$ and $132K$ steps respectively using the multistep scheduler.

We compare the BMN baseline model trained with and without height normalization using the 400-subject test set. The mean TP90 error of all 59 measurements and the TP90 error details of five key measurements are summarized in Tables 2 (a) and (b) respectively, with the best scores highlighted. In (a), height normalization outperforms the baseline across all BMI buckets, resulting in an overall improvement of 0.4 inches, indicating its effectiveness across diverse body shapes and sizes. In (b), height normalization leads to 0.34 to 1.27 inches lower errors in various body parts, demonstrating its effectiveness in disambiguating scale and facilitating training.

We also qualitatively compare a random test subject's body mesh predicted by the baseline w and w/o height normalization when the input is a partial-view photo. Our model produces overall better body shape fidelity, especially around the chest and waist areas, shown in Fig. 4.

4.4. Ablation Studies

Occlusion level. During training, we uniformly sample the occlusion level from the range of [0.0, 0.4] for each image. Now, we ablate the lower bound of the hyperparameter occlusion level α on HybriK [21] in Table 3 (a). It is evident

	Under	Normal	Overweight	ObeseI	ObeseII	Overall$_{59}$
BMN	1.67	1.44	1.38	1.77	1.97	1.52
BMN✓	**1.23**	**1.03**	**0.99**	**1.39**	**1.73**	**1.12**

(a)

	BustGirth	WaistGirth	HipGirth	BackWidth	Outseam
BMN	2.73	3.44	2.48	1.11	1.68
BMN✓	**1.78**	**2.17**	**1.68**	**0.77**	**1.24**

(b)

Table 2. TP90 errors ($\downarrow$) in inches for the BMN model w and w/o height normalization: (a) across BMI categories, and (b) for five key measurements. (✓) denotes with height normalization.
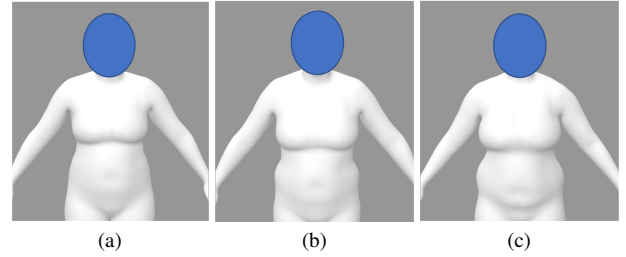

(a)   (b)   (c)

Figure 4. Qualitative comparison of the 3D body mesh predicted by BMN model trained (a) without and (b) with height normalization. (c) shows the SMPL-aligned groundtruth scan of the subject.

that our choice of 0.4 results in the smallest TP90 errors.

Background removal. We ablate the impact of background removal in Table 3 (b) and find that it reduces TP90 error by 0.09 inches, providing a marginal benefit.

α	0.5	0.4	0.3	0.2
	1.76	**1.65**	1.67	1.76

(a)

BR	✓	✗
	1.65	1.74

(b)

Table 3. Ablation studies on (a) lower bound of occlusion levels α and (b) background removal (BR). We show the overall TP90 errors ($\downarrow$) in inches for both experiments.

5. Conclusions

We introduce a height normalization technique to address body measurements and scale ambiguity in partial-view body estimation by adjusting the position of human subjects in images to a nominal distance. This method is straightforward yet effective, suitable for both single-view HMR models and multi-view setups. Our experimental results show that integrating this technique enhances body measurement accuracy. However, its applicability may be limited by the requirement for height and gender information. In future work, we plan to explore additional augmentation techniques, such as camera perspective distortion, to further enhance the realism of synthesized partial views.

References

[1] Benjamin Biggs, David Novotný, Sébastien Ehrhardt, Hanbyul Joo, Benjamin Graham, and Andrea Vedaldi. 3D multibodies: Fitting sets of plausible 3D human models to ambiguous image data. In *NIPS*, 2020. 2

[2] 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*, pages 561–578, 2016. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, volume 12352, pages 769–787, 2020. 2

[6] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *CVPR*, 2022. 2

[7] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, pages 20–40, 2020. 2

[8] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *CVPR*, pages 11875–11885, 2021. 2

[9] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. HS-Nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3DV*, pages 108–117, 2016. 2

[10] Riza Alp Güler and Iasonas Kokkinos. Holopose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[12] Denise Incandela. Walmart Levels Up Virtual Try-On for Apparel With Be Your Own Model Experience, https://corporate.walmart.com/news/2022/09/15/walmart-levels-up-virtual-try-on-for-apparel-with-be-your-own-model-experience. Accessed: April 12, 2024. 1

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 3

[14] 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*, pages 42–52, 2021. 1, 2, 3

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

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[17] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, pages 11127–11137, 2021. 2

[18] 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*, pages 2252–2261, 2019. 1, 2, 3

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 2

[20] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, pages 11605–11614, 2021. 2

[21] 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*, pages 3383–3393, 2021. 1, 2, 3, 4

[22] 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*, pages 590–606, 2022. 2

[23] Junbang Liang and Ming C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *ICCV*, pages 4352–4362, 2019. 2

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH ASIA*, 34(6):248:1–248:16, 2015. 1, 2, 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[27] Yafei Mao, Christopher Merkle, and Jan Allebach. A color image analysis tool to help users choose a makeup foundation color. *Electron. Imaging*, 34:373–1, 2022. 1

[28] 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*, pages 506–516. IEEE, 2017. 3

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[30] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: rapid creation of metrically accurate avatars from 3D measurements. *Transactions on Visualization and Computer Graphics*, 25(5):1887–1897, 2019. 1

[31] Chris Rockwell and David F Fouhey. Full-body awareness

from partial observations. In *ECCV*, pages 522–539, 2020. 1, 2

[32] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, pages 81–90, 2020. 2

[33] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *ICCV*, pages 11219–11229, 2021. 2

[34] Brandon M. Smith, Visesh Chari, Amit Agrawal, James M. Rehg, and Ram Sever. Towards accurate 3D human body reconstruction from silhouettes. *3DV*, pages 279–288, 2019. 2

[35] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and accurate scans from an image in less than a second. In *ICCV*, 2019. 2

[36] 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*, pages 601–617, 2018. 3

[37] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10965–10974, 2019. 2

[38] Song Yan, Johan Wirta, and Joni-Kristian Kämäräinen. Silhouette body measurement benchmarks. In *ICPR*, 2020. 2

[39] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *CVPR*, pages 14484–14493, 2021. 2

[40] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. THUNDR: Transformer-based 3D human reconstruction with markers. In *CVPR*, pages 12971–12980, 2021. 2

[41] 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*, pages 11446–11456, 2021. 1, 2, 3

[42] Zyler. See This Outfit On You, https://www.zyler.com. Accessed: April 12, 2024. 1