CVPR #13584

Supplementary Material

0.1. Addition related work

The literature [22, 24] on recovering 3D human representations from RGB images is vast. Techniques fall broadly into two categories. Parametric methods [3, 16] characterize the human body in terms of a parametric model. Model parameters defining body pose and shape are then estimated from images via direct optimization [20, 25, 26] or regression with deep networks [5, 6, 11, 12]. Non-parametric methods directly regress a 3D body representation from images using convolutional neural networks [13], transformers [15], intermediate representations [18] or implicit functions [7, 21]. On the other hand, the development of human shape estimation erupt 3D digital huaman research applications [4, 9] and the parametric models [16] from images and videos have attracted increasing attention. Optimization-based methods [20] detect 2D features corresponding to the whole body and fit the SMPL-X model. However, they suffer from slow speed and are ultimately limited by the quality of the 2D keypoint detectors. Hence, learning-based models are proposed. Due to the highly complex multi-stage pipelines, the reconstructed results inevitably generated unnatural articulation mesh and implausible 3D wrist rotations. [14] proposes the first ViT-based backbone [8] to relieve the issues in previous approaches. This provides a promising and concise way to leverage the scaling-up model for two-stage body measurements. However, there is a scarcity of benchmark datasets for comparing body measurements, and few researchers are exploring the integration of additional data toward generalizable and accurate body measurement results.

0.2. Detailed about datasets

We describe the datasets mentioned in the main paper. Note that all these are public academic datasets, each holding a license. We follow the common practice of using them in our non-commercial research and refer readers to their policies to ensure personal information protection.

Lidar dataset: We utilize a high-performance commercial Mojave Sensor, available at an affordable price, for scanning subjects aged between 18 and 24 years old. This choice allows us to conveniently conduct tests on LiDAR data. The output of this sensor is a distance and amplitude file. The distance image can be converted to a point cloud (x,y,z values for each distance point) by leveraging the sensor lens parameters that are stored on the device. Furthermore, each subject was measured by a skilled anthropologist, providing chest, waist, hip, wrist, and shoulder width as the GT measurement values for our experiments. RGB images were captured in a well-lit, indoor setup, with subjects standing in A-pose. Note that we never require subjects to wear tight-fitting clothing. Capture distance varied from 2.5-3.5 meters.

AGORA [19] is a synthetic dataset, rendered with high-quality human scans and realistic 3D scenes. It consists of 4240 textured human scans with diverse poses and appearances, each fitted with accurate SMPL-X annotations. There are 14K training images and 3K test images, and 173K instances.

3DPW [23] is the first in-the-wild dataset with a considerable amount of data, captured with a moving phone camera and IMU sensors. It features accurate SMPL annotations and 60 video sequences captured in diverse environments. We follow the official definition of train, val, and test splits.

Human3.6M [10] is a studio-based 3D motion capture dataset including 3.6M human poses and corresponding images captured by a high-speed motion capture system. In this paper, we use the annotation generated by NeuralAnnot, which fits the SMPL-X to the GT 2D joints and includes a total of 312.2K annotated data.

MPI-INF-3DHP [17] is captured with a multi-camera markerless motion capture system in constrained indoor and complex outdoor scenes. It records 8 actors performing 8 activities from 14 camera views. We use the annotations generated by NeuralAnnot, which fits the SMPL-X to the GT 2D joints and includes a total of 939,847 annotated data.

MPII [1] is a widely used in-the-wild dataset that offers a diverse collection of approximately 25K images. Each image within the dataset contains one or more instances, resulting in a total of over 40K annotated people instances. Among the 40K samples, 28K samples are used for training, while the remaining samples are reserved for testing. We use the annotations generated by NeuralAnnot, which fits the SMPL-X to the GT 2D joints and includes a total of ~28.9K annotated data.

0.3. Motivation about focusing network

Traditional fine-tuning methods require modifying the top layer of the network to adapt differences in label spaces and losses, which can disrupt the pretrained features and diminish the network's reusability. In contrast, our focusing network employs bypass network to extract various guidance features. This modification preserves the pretrained features for consistent performance and facilitates efficient model sharing. On the other hand, the traditional two-stage methods of reconstructing before measuring have limitations in generalizing to different scene categories. Additionally, the reconstruction quality lacks reliability under extreme viewing angles, making accurate measurement across various scenarios challenging. We are the first to solve the above problem by utilizing large-scale models for estimating anthropometric measurements. Moreover, we introduce a bypass network to fine-tune the output of the large model, an innovation not present in previous methods.

0.4. Details process about using Mojave Sensor

The output of this sensor for a subject is a distance and amplitude file. The distance image can be converted to a point cloud (x,y,z values for each distance point) by leveraging the sensor lens parameters that are stored on the device. By using the focal length and optical center of the lens a coefficient for x, y, and z can be calculated at each pixel location. These arrays of coefficients are stored on the device referred to as the pixel rays. After querying this information from the sensor, each value in the distance image can be multiplied by the three coefficients to obtain its x,y, and z coordinates in 3D space.

0.5. Training details

Following previous work, we use the following typical datasets for training focusing network, i.e., Human3.6M [10], MPI-INF-3DHP [17], MPII [2] and so on. Additionally, we provided 30 sets of samples, each containing frontal and profile images along with additional information. During the training of the bypass network, we utilized both our collected dataset and publicly available datasets, as described in Sec.5.1 of the main paper. Additionally, our Fashion-body dataset can be continually enhanced with diverse scenarios and individuals to meet the needs of a broader range of human reconstruction and measurement tasks. We considered the fairness of these baselines for comparison. The datasets used for HMR-BMViT and 4D-BMViT training are consistent with the bypass network, but NeuralAnthro could not support such a large amount of data for training.

0.6. Fixed hyper-parameter for loss function

In this part, we evaluate the effect of the loss design parameters on the measurement performance. We re-trained the proposed method on our Fashion-body dataset using fixed values for the parameters in the loss function introduced in main paper. We chose four different fixed values: 0.1,0.2,0.4,0.8. The validation error is presented in Tab. 1. The result shows that the larger value of the α parameter multiplies the measurement parts the smaller error of measurement results.

α	Chest	Waist	Hip	Wrist	Shoulder width
0.1	6.02	9.57	10.99	1.07	4.34
0.2	5.87	9.04	10.73	1.00	4.24
0.4	4.31	5.21	7.21	0.56	2.73
0.8	3.07	3.91	6.95	0.41	2.03

Table 1. **The MAE error of five body part in hyper-parameter tuning.**

0.7. Discussion

The generalization ability of the proposed approach, such as its robustness to background changes and variations in lighting conditions. Actually, our Fashion-body dataset consists of examples collected from diverse angles, lighting conditions, and backgrounds. Our experiments present the body measurement outcomes and MAE of our approach across multiple datasets, including our Fashion-body dataset, which validate the robustness of our approach.
Applications. We propose tailored measurement pipelines and scanner selections for diverse anthropometric applications: medicine, fashion, fitness, and entertainment. Our model can estimate key parameters of the human body from a simple RGB image captured by a camera or smartphone, making it applicable across various domains such as health, fashion, and entertainment. In the health domain, our model enables users to monitor changes in body measurements over fixed intervals (e.g., weekly or monthly) by capturing images, aiding in the formulation of fitness and diet plans. In the fashion domain, our model's outputs can assist users in selecting the most suitable clothing sizes in online shopping scenarios. In the entertainment domain, our model serves as a valuable tool for virtual character creation and clothing rendering, ensuring the physical realism and coherence of virtual scenes.
Limitations and Future works. Due to constraints related to the specialized equipment and cost required for dataset creation, our model is trained only on single-view human reconstruction, disregarding the potential benefits of information redundancy from multiple views. Utilizing multiple views to assist in human reconstruction can mitigate the impact of certain viewpoints or inappropriate poses on the model. In the future, we anticipate generating multi-view human reconstruction datasets. Additionally, considering the inference speed of the model is crucial for practical scenarios. Therefore, while ensuring model performance, we aim to enhance the inference speed.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 1

[4] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 1

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer, 2020. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2718–2728, 2022. 1

[7] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11875–11885, 2021. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 1

[9] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 1

[10] 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. 1, 2

[11] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1

[12] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 1

[13] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 1

[14] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 1

[15] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 1

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1

[17] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 1, 2

[18] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 1

[19] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 1

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1

[21] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1

[22] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 1

[23] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 1

[24] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 1

[25] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 1

CVPR
#13584

[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14484–14493, 2021. 1