# KBody: Balanced monocular whole-body estimation

Nikolaos Zioulis[1], James F. O'Brien*[1,2]
[1] Klothed Technologies Inc., [2] UC Berkeley
https://klothed.github.io/KBody

## Abstract

*KBody is a method for fitting a low-dimensional body model to an image. It follows a predict-and-optimize approach, relying on data-driven model estimates for the constraints that will be used to solve for the body's parameters. Compared to other approaches, it introduces virtual joints to identify higher quality correspondences and disentangles the optimization between the pose and shape parameters to achieve a more balanced result in terms of pose and shape capturing capacity, as well as pixel alignment.*

Author's preprint version. Published in CVPR 2023 6th Workshop on Computer Vision for Fashion, Art, and Design (CVFAD).

## 1. Introduction

Estimating the parameters of a low-dimensional human body is a cornerstone for human-centric applications such as virtual try-on based e-commerce [25]. However, for consumer-facing products that necessitate monocular inputs, it is a highly ill-posed problem that remains elusive due to the challenges arising from the problem formulation itself and the limitations of available constraints.

Estimating the human body from a single image corresponds to estimating the articulation parameters $\boldsymbol{\theta} \in \mathbb{SO}(3)^P$, the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^B$ and the global transform $\mathbf{T} = \left[\begin{smallmatrix}\mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1\end{smallmatrix}\right]$. These parameters reconstruct a human mesh $(\mathbf{V}, \mathbf{F}) = \mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T})$ via the body function $\mathcal{H}$. Two dominant classes of approach exist. The first is fitting the body by minimizing an objective function [7, 33]:

$$\underset{\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{T}}{\operatorname{argmin}} \mathcal{E}_{data} + \mathcal{E}_{prior}, \tag{1}$$

that includes a data fitting term, $\mathcal{E}_{data}$, and $\mathcal{E}_{prior}$, an important prior regularization term to prevent degenerate solutions and regularize the ill-posed problem. The constraints involved in the data term most typically include $2D$ keypoints [33], that are typically inferred by a data-driven method [8]. While the prior term helps, $2D$ keypoints usually lead to solutions that suffer from monocular ambiguity, producing poor results from a $3D$ accuracy perspective. The second class of approach consists of data-driven methods that encode a learned prior in the parameters, $\chi$, of a neural network, $f$, and perform monocular inference:

$$(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}, \boldsymbol{\pi}) = f_\chi(\mathbf{I}), \tag{2}$$

with $\boldsymbol{\pi}$ being the – typically weak perspective / orthographic – projection parameters that best explain the image content using the estimated parameters. As the neural network function $f_\chi$ is supervised, it preserves $3D$ awareness but usually suffers from predictions with poor pixel alignment.

Another challenge that also hinders high-quality pixel alignment is the conflict between the pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$, that are entangled through $\mathcal{H}$. Early works [10, 33] focused on the difficult problem of pose capturing foremost, with proper shape being an unaccomplished side-objective. Yet

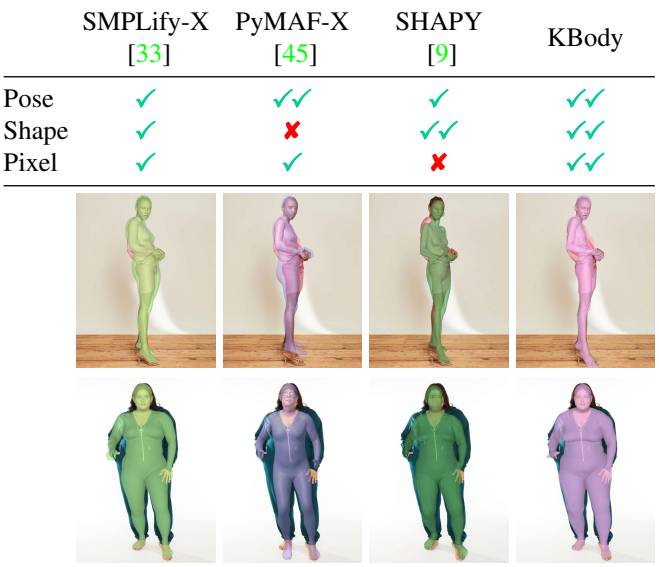


**Figure 1.** Flexible, pixel aligned, accurate body pose and shape capture is the challenging, yet ultimate goal of monocular expressive body fitting. KBody improves the balance between all 3 traits using a *predict-and-optimize* approach.

*Corresponding author: james@getklothed.com

as progress was made, it became evident that inaccurate shape was hindering further advances, and thus, more recent works [9, 11] started focusing on higher quality shape capture, but seemingly, at the cost of poorer pose estimation. Overall, achieving pixel-aligned estimates that are metrically correct (in world scale, not up to an unknown scale factor), and doing so robustly for a wide range of inputs remains a significant challenge.

In this work, we present a balanced whole-body monocular fitting method. We improve fitting quality by introducing virtual joints, adapted to fit the estimated data, and allowing for smooth interplay with silhouette constraints, expressed as an asymmetric distance field. We additionally show how disentangling the optimization process allows for improved joint shape and pose estimates.

## 2. Related Work

Estimating parametric human models from images is a rapidly evolving area forming a complex landscape of data, models, and training strategies, as discussed in a recent survey [32] and benchmark [41] papers. Several parametric human body models, including STAR [30], GHUM [1,44] and most recently SUPR [31] have been released, but we will focus on the expressive variant of SMPL, SMPL-X [33].

Pioneering the transition from keypoint estimation to full-body estimation involved the direct regression of low-dimensional body parameters from a single image [18]. The method was supervised using keypoint annotations and thus, end-to-end training was achieved after also regressing the camera parameters that would project the articulated body joints to correct positions. Regularization was applied in the form of a discriminator for the estimated pose and shape, so as to match a realistic distribution made available as a corpus of fit human scans. Various extensions were later proposed, integrating inverse kinematics [26], topological priors [29], and external camera estimation [21]. While the latter two approaches use silhouettes in their training schemes, they remain an intermediate representation for skeletonization [29] or they include clothing layers [21].

Initial efforts only regressed pure body parameters (*i.e.* SMPL), which unfortunately disregards details like hands and faces. ExPose [10] included regressing parameters for the hands and face. FrankMoCap [35] built an efficient system, achieving real-time rates. ExPose was extended to PIXIE [13], which had separate experts for the body, hands and face that were optimally combined to improve results. More recently, PyMAF-X [45] builds on the iterative nature of these models (*e.g.* [10, 18]) but instead of using global features at a single scale, PyMAF-X uses a pyramid of features, including finer-grained ones, achieving higher quality pixel alignment than other approaches.

Taking another direction, SHAPY [9] focuses on shape estimation using model agency annotations for shape measurements. Having been trained with this supervision, it is capable of regressing metric-scale shapes. SHAPY's pose estimation performance is below PyMAF-X, but its capacity to output metric-scale shapes heavily compensates.

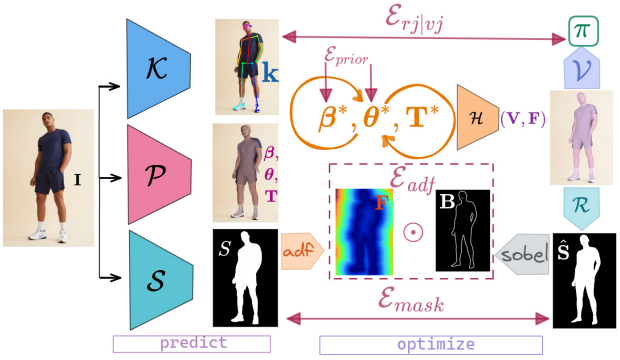


**Figure 2.** KBody body fitting. Keypoints $\mathbf{k}$ and silhouette $\mathbf{S}$ are predicted from the respective models $\mathcal{K}$ and $\mathcal{S}$. An initial state $\beta, \theta, \mathbf{T}$ predicted by $\mathcal{P}$ is iteratively optimized to fit $\mathbf{k}$ and $\mathbf{S}$ using the rendering $\mathcal{R}$, virtual joint $\mathcal{V}$, and projection $\pi$ functions.

SMPLify [7] is the seminal work that fits the SMPL body to a single image, showing the effectiveness of having priors for both the pose and shape alike. SMPLify was later extended to use annotated silhouettes in its iterative optimization scheme, with the goal of improving dataset annotations [24]. Using an L1 silhouette objective allowed for capturing human performances in video [15] using differentiable soft-rasterization [28, 34], and improved results when combined with a differentiable ray-tracer [27] and part-based masks [3]. In a follow up work, it was extended to SMPLify-X [33], adding details like hands and face, as well as a learned prior, VPoser [33]. Similarly, to improve shape capturing for use within forensic contexts [40], an L2 mask loss was added into the optimization scheme through a differentiable renderer [19].

While orthogonal improvements like better priors (*e.g.* Pose-NDF [42]) can improve fitting performance, results ultimately heavily rely on the constraints $\mathbf{k}$ and (optionally) $\mathbf{S}$. Another important component is the initialization of the optimization which can significantly affect convergence due to the ill-posedness of monocular fit.

## 3. Approach

Similar to prior approaches, we minimize Eq. (1) to fit a body model to image-domain constraints, using the same prior terms as [33], but with disentangling the optimization, adding virtual joints, and a silhouette-based objective:

$$\mathcal{E}_{data} = \underbrace{\lambda_k(\mathcal{E}_{rj} + \mathcal{E}_{vj})}_{\text{keypoints}} + \underbrace{\lambda_m \mathcal{E}_{mask} + \lambda_d \mathcal{E}_{adf}}_{\text{silhouette}}, \quad (3)$$

where $\mathcal{E}_{rj|vj} = \varrho(\mathbf{k}, \boldsymbol{\pi}(\mathbf{j}_{rj|vj}))$ is the Geman-McClure penalty function [14] for the regular, $\mathbf{j}_{rj}$, and virtual, $\mathbf{j}_{vj}$, joints, matching them to the corresponding keypoints $\mathbf{k}$ via the projection function $\boldsymbol{\pi}$ of given camera model. $\mathcal{E}_{mask} = \sum^{\Omega} ||\mathbf{S} - \hat{\mathbf{S}}||_1$ is an L1 silhouette overlay term defined on the image domain $\Omega$, between an inferred silhouette $\mathbf{S}$ and the body model's silhouette $\hat{\mathbf{S}} = \mathcal{R}(\mathbf{V}, \mathbf{F})$ rendered via a differentiable renderer $\mathcal{R}$. An overview is shown in Fig. 2.

**Disentangled Optimization (DO).** Prior monocular human body fitting works perform a staged optimization of Eq.(1), where each stage adds a layer of complexity in the optimization (e.g. details like fingers), and also anneals the constraints' [7, 33] weights across stages. Initial estimates of global parameters $\mathbf{T}$ have also been included as a first stage [7,33], but sensitivity to localisation of the torso joints has led to alternatives [24]. To partly address sensitivity to initialization, a data-driven initial estimate is used, offering a good initial starting point.

However, all prior work up to now optimize both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ simultaneously at each iteration $i$ of each stage $s$: $(\boldsymbol{\beta}^s_{i+1}, \boldsymbol{\theta}^s_{i+1}) = (\boldsymbol{\beta}^s_i + \Delta\boldsymbol{\beta} s_i, \boldsymbol{\theta}^s_i + \Delta\boldsymbol{\theta}^s_i)$. These two sets of parameters are entangled by the human body function $\mathcal{H}$ that allows for their joint optimization. While this is effective with a $3D$ objective that is conditioned on the same domain where the function $\mathcal{H}$ exists, it is much less effective in the monocular $2D$ case that comes with inherent $3D$ ambiguities. As a result, optimization is dominated by the pose updates $\Delta\boldsymbol{\theta}$. This imbalance is evident in both keypoint-only optimization approaches [33] as well as data-driven models trained with only keypoint losses [10, 22, 45]. Both tend to produce shape coefficients biased towards the zero mean vector. More recent shape-aware approaches either optimize in $3D$ [11] or use $3D$ losses during training [9].

Seeking to improve our optimization loop, we separate the parameter updates of the shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ components in an alternating fashion for stage $s$: $(\boldsymbol{\beta}^s_i, \boldsymbol{\theta}^s_{i+1}) = (\boldsymbol{\beta}^s_{i-1} + \Delta\boldsymbol{\beta}^s_{i-1}, \boldsymbol{\theta}^s_i + \Delta\boldsymbol{\theta}^s_i)$. Similar to block coordinate optimization, the shape $\boldsymbol{\beta}$ parameters are only updated in even iterations $i$, while the pose parameters $\boldsymbol{\theta}$ are only updated in odd iterations $i+1$. This method exhibits significantly better joint optimization of these parameters even in the highly ill-posed monocular case. However, as this approach suffers from local minima, it can only be introduced later in the optimization process.

**Virtual Joints (VJ).** An iterative fitting approach crucially relies on high quality correspondences. Defining proper joint locations on the body to match the keypoint estimates has troubled past approaches, with the hip joints ignored from the optimization [33], or regressed via empirically defined and manually created joint regressor functions [22]. However, the location of the keypoints $\mathbf{k}$ are typically inferred from a data-driven model which aggregates numerous annotations and thus, includes their biases as well. Recent works that acknowledge this have resorted to learning a joint regressor for a specific dataset [16] which comes with new challenges like properly constraining the joints' locations inside the human body.

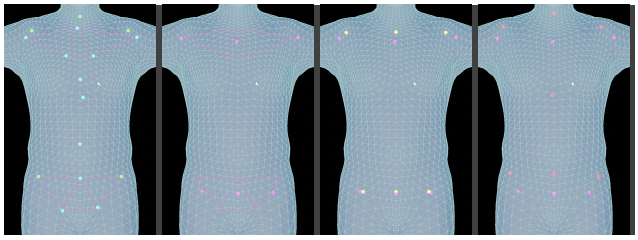

**Figure 3.** From left to right: **i)** the SMPL-X torso with the barycentric parameterization comprising the triangles formed by raw and manually picked [22] joints, **ii)** our best-estimated virtual joints, and their comparison with **iii)** manually picked openpose joints [4, 5] and **iv)** the learned regressor joints fit to Human3.6M [16].

Our approach also seeks to identify better matching locations, but not for a specific dataset, instead matching the inference distribution of a pre-trained $2D$ keypoint estimator. We introduce the concept of *virtual joints* $\mathbf{j}_{vj} = \mathcal{V}(\boldsymbol{b}, \mathbf{j}_s)$, by parameterizing joint locations as a linear combination of weights $\boldsymbol{b}$ and pre-defined (empirically or anthropomorphically) joint subsets $\mathbf{j}_s, s \in [1, \dots, S]$. More specifically, we focus on the more ambiguous torso joints, which carry a two-fold importance, **i)** they are high in the kinematic chain, and thus, highly influential of the articulated body fit, and **ii)** they are highly dependent on human shape, and thus, are necessary to avoid cross data-term conflicts between the keypoint and the silhouette terms.

Virtual joint localisation is restricted to planes formed by joint triangles (*i.e.* $S = 3$), illustrated in Fig. 3, using a barycentric formulation for the virtual joints. This allows for the reduction of the number of weights $\boldsymbol{b}$ to 2 (or 1 for joints that need to lie on one of the triangle's altitudes) by exploiting $\sum_{b \in \boldsymbol{b}} b = 1$. While this relies on a non-holding rigidity assumption for the joints subset, albeit relaxed in the torso area, the goal is to better localize joints matching those inferred by a $2D$ estimation model, which itself exhibits limited expressivity at the torso. Finding the best matching locations is an one-off process that involves fitting a variety of pre-defined poses to inferred keypoints and identifying the best performing weights using a performance indicator.

**Asymmetric Distance Fields (ADF).** Silhouette-based representations have long been used in parametric model fitting approaches [2, 39], and have lately appeared in both optimization or single-shot approaches [6, 15, 17, 24, 29, 37, 38, 40]. They are usually coupled with differentiable rendering [19, 34] and L1/2 losses. This loss is inefficient, suffering from an irregular loss landscape and the lack of

directional information for parameter updates [29].

To supplement the L1 mask loss we use $\mathcal{E}_{adf} = \sum^{\Omega} \mathbf{B} \odot \mathbf{F}$ summed over the pixel domain $\Omega$ which is minimized when the two silhouette boundaries align, with $\mathbf{B}$ being the boundary of $\mathbf{S}$ and $\mathbf{F}$ the asymmetric distance field. The latter is defined as:

$$\mathbf{F} = \lambda_o D(\mathbf{S}) \odot \bar{\mathbf{S}} + \lambda_i D(\bar{\mathbf{S}}) \odot \mathbf{S}, \quad (4)$$

with $D(\cdot)$ being the distance field function and $\bar{(\cdot)}$ denoting pixel-wise binary inversion. The asymmetry derives from the different inner ($\lambda_i$) and outer ($\lambda_o$) distance field weights.

## 4. Results

We refer to the approach depicted in Fig. 2 as KBody and implement it using SMPL-X [33] ($\mathcal{H}$), OpenPose [8] ($\mathcal{K}$), MODNet [20] ($\mathcal{S}$, producing $\mathbf{S}$ after binary thresholding the estimated matte at $0.85$), and ExPose [10] ($\mathcal{P}$). The objective is optimized with L-BFGS [43] for 30 iterations per stage. Similar to prior work we perform annealed optimization with the early stages using stronger regularization to make the objective function more convex, and then progressively reduce the regularization term weights and increase the data terms of the details (hands, face). Differentiable rendering $\mathcal{R}$ is implemented using high-performance rasterization [23]. We use the pose prior and regularization terms from SMPLify-X [33], but relax the latter's weights due to the better initialization and the silhouette constraints.

First we validate the effectiveness of the virtual joint localization by running only 2 stages of fitting to $\mathcal{K}$ after initializing with $\mathcal{P}$, without involving $\mathcal{S}$ for a fair comparison, and with only the second stage optimizing the details. We use the EHF [33] dataset to run a hierarchical and empirically defined search to identify the parameters $\boldsymbol{b}$. Performance is measured with the indicator $i = (1 - IoU) \times RMSE$, aggregating the keypoints' RMSE and the body's IoU using the service generated masks. We the compare against other approaches fitting to EHF in Tab. 1. Performance is assessed via procrustes-aligned vertex-to-vertex error on the SMPL-X body's vertices (PA-V2V-X) [33]. As also shown in Pose-NDF [42] and the first 3 rows of Tab. 1, simply optimizing the initial estimates of a data-driven model does not necessarily lead to improved fits. Using better priors [12, 42] slightly improves results over the baseline single-shot data-driven estimate [33], while a manually selected joint regressor [4, 5] (Fig. 3 **iii**) does not result in improved fits. The virtual joints produce the most significant gain, showcasing the importance of higher quality correspondences between the estimated keypoints used as constraints and the body's joints. It should be noted that, apart from the last 2 rows, bad joint-to-keypoint correspondences (*e.g.* hips) are ignored during optimization.

Next we evaluate the full KBody approach by adding the silhouette constraints using $\mathcal{S}$, a DO stage, and a final stage for detail (hands, face) capture. Two experiments are presented, on EHF and SSP3D [36] focusing on pose and shape performance respectively. Pixel-based IoU and use the PA-V2V and PVE-T-SC [36] metrics are used. Both experiments use the SMPL meshes to calculate metrics instead of the SMPL-X ones to reduce the effect of the densely sampled head via pre-calculated mesh-to-mesh vertex transfer maps [30]. Tab. 2 shows that how KBody outperforms the other methods with respect to pose estimation (EHF), and shape capturing (SSP3D) capacity. Evidently, KBody produces the best pixel alignment and also an ablation shows the benefit of disentangled optimization for shape capture.

| Initialization | Optimization | Joints | Prior | PA-V2V-X↓ |
|---|---|---|---|---|
| ✘ | SMPLify-X [33] | $\mathbf{j}_{rj}$ | VPoser [33] | 60.3 $mm$ |
| ExPose [10] | ✘ | $\mathbf{j}_{rj}$ | ✘ | 54.8 $mm$ |
| ExPose [10] | SMPLify-X [33] | $\mathbf{j}_{rj}$ | VPoser [33] | 67.2 $mm$ |
| ✘ | SMPLify-X [33] | $\mathbf{j}_{rj}$ | PoseNDF [42] | 57.4 $mm$ |
| ExPose [10] | SMPLify-X [33] | $\mathbf{j}_{rj}$ | PoseNDF [42] | 53.8 $mm$ |
| ExPose [10] | SMPLify-X [33] | $\mathbf{j}_{rj}$ | GAN-S [12] | 54.1 $mm$ |
| ExPose [10] | SMPLify-X [33] | $\mathbf{j}_{op}$ [4, 5] | VPoser [33] | 57.5 $mm$ |
| ExPose [10] | SMPLify-X [33] | $\mathbf{j}_{rj\|vj}$ | VPoser [33] | **49.3** $mm$ |

**Table 1.** Virtual joints improvement analysis on EHF [33]. The columns indicate parameter initialization and optimization, which joints are optimized, and with which pose prior.

| Method | EHF [33] | | SSP3D [36] | |
|---|---|---|---|---|
| | PA-V2V↓ | IoU↑ | PVE-T-SC↓ | IoU↑ |
| ExPose [10] | 71.7 $mm$ | 84.72% | 33.0 $mm$ | 71.00% |
| SMPLify-X [33] | 95.9 $mm$ | 81.46% | 33.9 $mm$ | 76.60% |
| PyMAF-X [45] | 66.6 $mm$ | 85.57% | 30.6 $mm$ | 75.87% |
| SHAPY [9] | 71.1 $mm$ | 81.29% | 29.3 $mm$ | 72.65% |
| KBody (w/o DO) | - | - | 28.1 $mm$ | 77.87% |
| KBody | **64.2** $mm$ | **87.72%** | **25.6** $mm$ | **80.35%** |

**Table 2.** Results on the the EHF [33] & SSP3D [36] datasets.

Finally, KBody's efficacy is qualitatively illustrated in Fig. 1 using images collected online. For these representative examples, KBody provides more balanced solutions, capturing pose and shape in high-quality for both heavy and lighter subjects, while also achieving good pixel alignment. An extended set of 112 randomly selected in-the-wild examples can be found in our supplemental material.

## 5. Conclusion

While the conflicts between pose and shape performance as well as world scale outputs and image alignment remain to be solved, we believe KBody is a step towards more balanced performance. However, relying on externally estimated constraints limits applicability in situations where the constraint models under-perform. Still, improving $2D$ estimation models is more practical than acquiring $3D$ data for

supervision [11] or a wide-range of images and corresponding measurements [9].

## References


[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5461–5470, 2021. 2

[2] Alexandru O. Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[3] Munkhtulga Battogtokh and Rita Borgo. Simple Techniques for a Novel Human Body Pose Optimisation Using Differentiable Inverse Rendering. *Eurographics 2022-Short Papers*, pages 65–684, 2022. 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 3, 4

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020. 3, 4

[6] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision*, pages 3–19. Springer, 2018. 3

[7] 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 *European conference on computer vision*, pages 561–578. Springer, 2016. 1, 2, 3

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 4

[9] 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, 2, 3, 4, 5

[10] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020. 1, 2, 3, 4

[11] Enric Corona, Gerard Pons-Moll, Guillem Alenyà, and Francesc Moreno-Noguer. Learned Vertex Descent: A New Direction for 3D Human Model Fitting. *arXiv preprint arXiv:2205.06254*, 2022. 2, 3, 5

[12] Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10997–11005, 2022. 4

[13] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 2

[14] Stuart Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 4:5–21, 1987. 3

[15] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 2, 3

[16] Eric Hedlin, Helge Rhodin, and Kwang Moo Yi. A Simple Method to Boost Human Pose Estimation Accuracy by Correcting the Joint Regressor for the Human3.6m Dataset. *arXiv preprint arXiv:2205.00076*, 2022. 3

[17] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 3

[18] 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. 2

[19] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 2, 3

[20] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1140–1147, 2022. 4

[21] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 2

[22] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 3

[23] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 4

[24] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2, 3

[25] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-Resolution Virtual Try-On

with Misalignment and Occlusion-Handled Conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 1

[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 2

[27] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2

[28] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 2

[29] Ramesha Rakesh Mugaludi, Jogendra Nath Kundu, Varun Jampani, et al. Aligning silhouette topology for self-adaptive 3D human pose recovery. *Advances in Neural Information Processing Systems*, 34:4582–4593, 2021. 2, 3, 4

[30] Ahmed AA Osman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer, 2020. 2, 4

[31] Ahmed AA Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. SUPR: A Sparse Unified Part-Based Human Representation. *arXiv preprint arXiv:2210.13861*, 2022. 2

[32] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and Analyzing 3D Human Pose and Shape Estimation Beyond Algorithms. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[33] 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, 2, 3, 4

[34] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 2, 3

[35] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMoCap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 2

[36] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020. 4

[37] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11219–11229, 2021. 3

[38] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16094–16104, 2021. 3

[39] Cristian Sminchisescu and Alexandru C Telea. Human pose estimation from silhouettes. a consistent approach using distance level sets. In *10th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG'02)*, volume 10, 2002. 3

[40] Neerja Thakkar, Georgios Pavlakos, and Hany Farid. The Reliability of Forensic Body-Shape Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 44–52, 2022. 2, 3

[41] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 2

[42] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 2, 4

[43] Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999. 4

[44] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[45] 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. *arXiv preprint arXiv:2207.06400*, 2022. 1, 2, 3, 4