UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Leveraging anthropometric measurements to improve human mesh estimation and ensure consistent body shapes

Katja Ludwig, Julian Lorenz, Daniel Kienzle, Tuan Bui, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Ludwig, Katja, Julian Lorenz, Daniel Kienzle, Tuan Bui, and Rainer Lienhart. 2025. "Leveraging anthropometric measurements to improve human mesh estimation and ensure consistent body shapes." In *2025 IEEE/CVF International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 11-12 June 2025, Nashville, TN, USA*, 5862–71. Piscataway, NJ: IEEE. https://doi.org/10.1109/CVPRW67362.2025.00585.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Deutsches Urheberrecht
Weitere Informationen finden Sie unter: / For more information see:
https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/

Leveraging Anthropometric Measurements to Improve Human Mesh Estimation and Ensure Consistent Body Shapes

Katja Ludwig, Julian Lorenz, Daniel Kienzle, Tuan Bui & Rainer Lienhart
Chair for Machine Learning & Computer Vision, University of Augsburg, Germany
{firstname.lastname}@uni-a.de

Abstract

The basic body shape (i.e., the body shape in T-pose) of a person does not change within a single video. However, most SOTA human mesh estimation (HME) models output a slightly different, thus inconsistent basic body shape for each video frame. Furthermore, we find that SOTA 3D human pose estimation (HPE) models outperform HME models regarding the precision of the estimated 3D keypoint positions. We solve the problem of inconsistent body shapes by leveraging anthropometric measurements like taken by tailors from humans. We create a model called A2B that converts given anthropometric measurements to basic body shape parameters of human mesh models. We obtain superior and consistent human meshes by combining the A2B model results with the keypoints of 3D HPE models using inverse kinematics. We evaluate our approach on challenging datasets like ASPset or fit3D, where we can lower the MPJPE by over 30 mm compared to SOTA HME models. Further, replacing estimates of the body shape parameters from existing HME models with A2B results not only increases the performance of these HME models, but also guarantees consistent body shapes.

1. Introduction

Creating an accurate 3D human mesh from monocular images or videos creates new opportunities in fields like 3D animation, gaming, fashion, sports, etc. In many of these application fields, videos are of main interest. While applying HME to videos, analyses of results of SOTA HME models show that the basic body shape of the meshes of the same person differs from frame to frame. [1] Worse, an analysis of currently used 3D mesh and pose datasets reveals the same inconsistencies in the provided ground truth (GT) data. For a precise body posture analysis, as it is necessary in many sports disciplines, an exact model of the athlete's body shape is required. Therefore, most professional athletes are measured anthropometrically during performance assessments today. Moreover, the body shape of an actor performing motions for 3D animations needs to be consistent as the basic body shapes does not change during performances. Thus, the changing body shapes of HME models for the same person are highly unwanted and simply wrong.

Figure 1. Two qualitative examples from the ASPset sports dataset. The result from a SOTA HME model, SMPLer-X [3], is shown on the left, the result from our model on the right, respectively. GT joints and estimated joints are color-coded. Corresponding joints are connected.

Our work aims to create a single perfectly fitting basic body shape for each human and reuse it for all videos with this person. Measuring the human body has already been done for centuries to fit suits or dresses perfectly to a specific body shape. In many applications, measuring the person in action beforehand would add only a marginal overhead, but improves the results dramatically. For this reason

[1] The body shape in a given pose is usually modeled by a basic body shape (given in T-pose) plus an additional pose-dependent deformation. We call the basic body shape just body shape in the following, as the pose-specific correction is computed from the pose and does not need to be estimated.

we propose to use these measurements. Body shape parameters of common human mesh models like SMPL-X [27] are not human interpretable. Therefore, it is not possible to obtain the perfect body shape parameters by anthropometric measurements. Hence, we train a machine learning model (called *A2B*, anthropometric measurements to body shape) to translate those measurements into body shape parameters for HME. With this approach, measuring a person once creates the body shape that can be used for all frames in all evaluation videos.

HME models are performing well on everyday data. However, in more challenging scenarios like sports, their performance is inferior to fine-tuned SOTA 3D HPE models. 3D HPE models only predict 3D keypoints resulting in a stick-figure pose, whereas HME models output a posed mesh including the human's surface. Due to the lack of GT meshes, HME models cannot be trained on datasets with solely 3D keypoint annotations. The usage of synthetic data is emerging in the field, but is not applicable to challenging or specific scenarios like sports. In this paper, we propose a solution to that problem. With our A2B model and anthropometric measurements, we can now create the body shape parameters of humans needed for HME. We further apply inverse kinematics (IK) to produce the rotations that are missing in the 3D stick-figure model that is created by 3D HPE models. Together with our A2B body shape, we are able to generate human meshes that have a consistent body shape and a precise pose. The main goal of this paper is to **estimate the best possible human mesh with a consistent body shape by adding the marginal overhead of measuring humans**. Since our main focus is sports, this overhead is negligible, as professional athletes are commonly measured anyway. We show qualitative examples of our model and a SOTA HME model, SMPLer-X [3], in Figure 1. Our approach is generally applicable to any HME problem. We choose sports datasets to validate our proposed approach, since the poses in sports are challenging for existent HME models and athletes are measured. Their performance is currently not good enough to use them in performance assessments of athletes, which we try to change. Our contributions can be summarized as follows:

- We reveal inconsistencies in the GT data of ASPset [25] and fit3D [12]. The body shape of a single person varies mistakenly in the GT.
- We create and evaluate different models to convert between anthropometric measurements and SMPL-X body shape parameters for all genders. We call them *A2B*.
- We analyze and compare the performance of existing HME models on ASPset and fit3D. Replacing the estimated body shape parameters (and keeping the pose) of each HME model with A2B body shape parameters increases the performance of all models.
- With fine-tuned SOTA 2D and 3D HPE models [10, 37], IK, anthropometric measurements, and our A2B model, we estimate accurate human meshes with a consistent body shape. We show that this approach achieves superior results to SOTA HME models, although still evaluated on the inconsistent GT.
- Our models and code for our approach are publicly available: https://github.com/kaulquappe23/a2b_human_mesh

2. Related Work

Human Mesh Estimation (HME) is an active area of research. Body models like SMPL [20] and its successor SMPL-X [27] are broadly used. Their advantage is that they decouple human pose and shape. The pose parameters θ give the rotations of the joints relative to the parent joint. The shape parameters β model the basic body shape. At first, a mesh is created with a linear mapping from β parameters to a T-shaped pose. Next, some pose-specific shape deformations are applied, and then the mesh is rotated at the joints according to the θ parameters.

The first HME model that estimates SMPL-X meshes from images, SMPLify-X, was introduced along SMPL-X [27]. It detects 2D image features and then fits an SMPL-X model to these. To achieve that, they incorporate a pose prior trained on a large motion capture dataset and an interpenetration test. A more recent model for HME is Multi-HMR [1]. It predicts 2D heatmaps for person centers and based on that the human mesh with a human prediction head. OSX [19] is a HME model using a component aware Transformer that is composed of a global body encoder and local decoders for face and hands. SMPLer-X [3] is introduced as a generalist foundation model for HME trained on a large amount of datasets mainly using vision transformers. There are many other HME models, some focussing more on whole-body HME [7, 11, 23], others on multi-person HME - either with a two stage approach using a person detector and a single person human mesh estimator [6, 13, 29], or a single stage approach estimating the meshes of all persons at once [30, 36, 38].

Choutas et al. [8] observed that existing HME models focus more on the body pose than the shape, although the shape is equally important for many applications. They propose SHAPY, a model that uses anthropometric and linguistic attributes to create accurate body shapes. Moreover, Sarkar et al. [32] introduce SoY, which contains specific loss functions to enhance the body shape accuracy. AnthroNet [28] propose a new body model that is learned with an end-to-end trainable pipeline. It takes anthropometric measurements as an input to learn a mesh model that accurately captures shapes of humans, but this model is different from the commonly used SMPL-X model. We use the common SMPL-X body model and decouple the estimation of the shape from the estimation of the pose. Sengupta et

al. [34] estimate anthropometric measurements from images and use a linear layer to convert them to body shape parameters. However, their model operates on single images and their measurement to shape conversion is different from ours and only available for a single gender. We further ensure the consistency of the body shape over time.

In the last years, 3D HME approaches leveraged Inverse kinematics (IK) to enhance their results. HybrIK [17] transforms 3D joint coordinates to relative body-part rotations for 3D HME by using a twist-and-swing decomposition. HybrIK-X [18] further enhances HybrIK with expressive face and hands. Cha et al. [4] leverage IK to tackle the challenge of person-to-person occlusions in images with interacting persons. PLIKS [35] (Pseudo-Linear Inverse Kinematic Solver) approaches HME by analytically reconstructing the human model via 2D pixel-aligned vertices in an IK-like manner.

Although HME is an active area of research, it is yet not common in computer vision for sports. Due to high velocities and a great variation of poses, sports is a challenging scenario for all kinds of human pose and shape estimation. The fit3D dataset [12] is a dataset which consists of videos from gym sports exercises with repetitions and is annotated with human meshes. AIFit [12] is a tool trained on fit3D which can reconstruct 3D human poses, reliably segment exercise repetitions, and identify the deviations between standards learned from trainers, and the execution of a trainee. Other sports datasets only consist of 3D joint annotations, like ASPset [25] or SportsPose [15]. SportsCap [5] is an approach for simultaneously capturing 3D human motions and understanding fine-grained actions from monocular challenging sports videos.

3. Errors in 3D Human Shape Ground Truth

Each person has a specific basic body shape that does not change over a short time period. Therefore, the SMPL-X body model decouples the human pose encoded by θ parameters from the basic body shape encoded by β parameters. Deformations to the basic body shape that are caused by the current pose are modeled separately. Therefore, it makes sense to assign a single set of shape parameters β to a person for a given short time period such as a recorded action to describe his/her shape. Further, there are lengths that can be calculated from 3D joints that should never change, since individual bones of humans are rigid and should not be deformed by different poses. Our approach enforces a single set of shape parameters per person and immutable bone lengths.

As a first step, we analyze if the GT data of our used datasets fulfills these properties. In this paper, we use ASPset [25] and fit3D [12], since both datasets consist of videos with fast changing poses and 3D GT. Results for the Human3.6M [16] and MPI-INF-3DHP [22] datasets are presented in the supplementary. For ASPset, we analyze bone lengths, since it has only GT annotations for 3D joints. For fit3D, GT SMPL-X β parameters are available, hence we can analyze the β parameters directly and further the derived anthropometric measurements. These values are the output of our deterministic B2A function: It generates a standard T-pose with the given β parameters and computes 36 anthropometric measurements from the resulting mesh. Results of our GT analysis for a subset of the anthropometric values are shown in Table 1. We can see that the GT itself is not consistent. The deviations are larger for ASPset. Although we have GT SMPL-X meshes for fit3D, every β parameter of a single person has a standard deviation of 0.64 on average.[2] This is a relevant flaw in the GT shape annotation, since based on the model, the GT shape should be consistent for each human. Nevertheless, we use the given inconsistent GT for our evaluations for comparability with related work and as we have no good means to correct them. The reader should keep this in mind. Nevertheless, we want to encourage future research in the field of 3D human pose and mesh data collection to try to eliminate these flaws in the provided GT.

ASPset				fit3D			
Measure	σ	r. σ	r. range	Measure	σ	r. σ	r. range
head	0.91	5.98%	57.91%	head	0.73	2.73%	17.52%
hip width	1.71	9.48%	85.46%	hip circ.	0.87	0.84%	8.17 %
forearm	1.99	8.37%	92.04%	forearm	0.34	1.40%	9.24%
upper arm	1.72	6.29%	66.35%	arm	0.76	1.51%	9.42%
lower leg	1.44	3.60%	41.36%	lower leg	0.52	1.31%	13.80%
thigh	1.65	4.23 %	35.46%	thigh	0.43	1.17%	11.74%
				height	1.60	0.94%	8.69%
				β param.	0.64		

Table 1. GT data analysis for ASPset (left) and fit3D (right): Standard deviation σ, relative standard deviation $\frac{\sigma}{avg}$ and relative range $\frac{\max - \min}{avg}$ of anthropometric measurements. Standard deviations are given in cm, but not for the β parameters. The values are averaged between left and right body parts and between all persons used for evaluations in Section 5. The β parameter standard deviation is averaged over all β parameters.

4. From Measurements to Body Shape

Humans have been measured for centuries [9]. Tailors know exactly which measurements to take for perfectly fitting a suit or dress to the body shape of a customer. In sports, it is already common practice that professional athletes are measured for precise performance assessments. Measuring a human is easy and well understood. In contrast, the parameters of the body shape for human mesh models like SMPL-X [27] are not humanly interpretable. The β parameters describe the principal components of the human body shape with typically around 10 to 16 values and are the re-

[2] Averaged standard deviation means (in the whole paper) that the standard deviation is calculated per person, and the mean of the resulting standard deviations is calculated afterwards.

sult of a PCA executed on the human meshes of a training dataset while learning the SMPL-X model. Fixing all β parameters despite one and looking at the results lets human observers get a notion of what this parameter might mean, but in total, the β parameters and their interactions are not well interpretable. Therefore, we want to leverage the well established technique of measuring humans to create precise body shape parameters for the commonly used SMPL-X human mesh model. We call our approach to convert from 36 Anthropometric measurements to Body shape parameters *A2B*. Since there is no known relation between anthropometric measurements (AMs) and β parameters, our aim is to learn this mapping. The reverse direction, *B2A*, is a deterministic function of the human mesh, as the AMs can be measured from the mesh.

4.1. Data Generation

We select 36 anthropometric measurements for our models based on the selections of AnthroNet [28] and an anthropometry study of the U.S. army [14]. They can be categorized into 23 lengths and 13 circumferences. Apart from the bone lenghts like arm length, thigh length, etc., this includes also detailed measurements like shoulder width, front torso height, lateral neck length, waist circumference, calf circumference, etc. A visualization and precise description of all AMs can be found in the supplementary.

Many existing datasets provide a wide range of different poses, but most incorporate the same humans. For learning a conversion model from AMs to β parameters, we need a lot of samples for different humans, no matter the pose. With given shape parameters, we can use the B2A function to compute the AMs. Recall, B2A is a deterministic function measuring the AMs from meshes in T-pose.

Because many different body shapes are required for the learning process, we use the AGORA [26] dataset. It consists of 1447 male and 1588 female subjects. We are not able to use the larger dataset from AnthroNet [28], since it uses its own mesh model and the authors did not publish their conversion to the SMPL-X model, which we want to use as it is most commonly used in research. Although comparably large, 1447/1588 subjects is still a little amount of data to learn a model. Hence, we analyze the β parameters in the AGORA dataset with the aim to randomly sample more data with realistic body shapes. Histograms (see Figure 2) of the occurring β parameters show that their distribution roughly follows a normal distribution. Therefore, we train our models with randomly sampled data according to these distributions, either assuming a normal distribution fitted to the histograms or a uniform distribution with the same minimum and maximum values as in the data analysis. This means that we sample each β parameter according to the selected distribution, create the mesh according to the sampled values and derive the AMs with B2A. With this

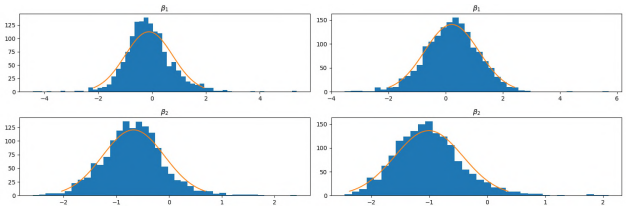

Figure 2. Histograms and fitted normal distribution (orange) for the first two β parameters for all male (left) and female (right) subjects of the AGORA [26] dataset.

strategy, we can create a dataset with as many subjects as we need. As we do not expect the analyzed AGORA data to cover the full range of human body shapes, we also train with extended distributions, meaning that we increase the standard deviation σ to $\alpha_n\sigma$ in the case of a normal distribution or stretch the interval by a factor α_u in case of a uniform distribution.

4.2. Models

We use the same number of β parameters for each gender as used in the AGORA dataset, meaning 11 for male, 10 for female, and 16 for neutral subjects. With 36 AMs as input values and $10 - 16$ output values for our A2B models, the dimensionality of the data is low. Therefore, we experiment with Support Vector Regression (SVR) and with small neural networks (NN). We split the AGORA dataset in an 80% train, 15% test and 5% validation subset. For SVR, we additionally randomly sample $10,000$ subjects for training. We use a hyperparameter search based on the validation split to determine the optimal settings, which leads us to a radial basis function kernel, an error margin of $\epsilon = 0.012$ and a regularization constant of $C = 3791$. For the NNs, we randomly sample new data in each iteration. The hyperparameter search for the NNs results in a model with 4 layers, 330 neurons per layer, tanh as activation function, and Xavier Glorot as initialization. We use mean squared error on the model output (the β parameters) as training loss.

4.3. Results

We train each model (NN and SVR) for each gender and with different dataset variants: We train solely on the AGORA train split, as well as on uniformly and normally distributed randomly sampled data according to the data analysis, and we further extend the range of the data as described in Section 4.1 with $\alpha_n = \alpha_u = 1.5$. The results are displayed in Table 2. We evaluate the performance of our models in two ways. At first, we calculate the error of the predicted and GT β parameters. Second, we calculate the mean deviation of the AMs of the meshes from the predicted and GT β parameters (A). Therefore, this evaluation can further be seen as a kind of cycle consistency evaluation of A2B (our learned model) and B2A (the deterministic measuring function). We provide a visualization of the

		⌀ error of β			⌀ error of A [in mm]		
	train data	m	f	n	m	f	n
NN	AGORA	9.11	13.9	24.0	0.814	0.934	1.459
NN	norm.	2.62	4.34	18.0	0.356	0.392	1.711
NN	norm. ext.	1.87	3.69	14.8	0.248	0.285	1.384
NN	unif.	5.08	**1.25**	**2.81**	**0.243**	**0.268**	0.623
NN	unif. ext.	**1.61**	3.20	8.37	0.274	0.419	**0.381**
SVR	AGORA	2.56	16.1	3.82	1.659	5.195	2.557
SVR	norm.	4.08	17.8	59.0	2.975	4.303	14.63
SVR	norm. ext.	0.210	4.60	6.27	0.280	1.090	2.211
SVR	unif.	0.0396	0.0350	**0.162**	0.124	0.284	0.214
SVR	unif. ext.	**0.0252**	**0.0193**	0.306	**0.082**	**0.136**	**0.164**

Table 2. Results of our A2B models on the test split of the AGORA dataset. The first block (β) shows the error if we take the GT β parameters, derive 36 anthropometric measurements (B2A), input them into the A2B models and evaluate the MSE of the predicted β parameters in the scale 10^{-3}. The second block (A) calculates B2A from the predicted β parameters and evaluates the mean difference between the GT and predicted AMs (all 36) in mm. Results are given for m(ale), f(emale), and n(eutral) models. Further visualizations are in the supplementary.

evaluation process and evaluation results of the real-world SSP-3D dataset [33] in the supplementary. The anthropometric error is our main metric as these values reflect the desired body shape given as an input by the user and are further interpretable. The β parameters are somehow arbitrary in their scale. For all genders and SVR, using an extended uniformly sampled dataset works best. For the NNs, a uniformly sampled dataset works best for male (m) and female (f) genders and an extended normally sampled dataset for the neutral (n) meshes. The results for the neutral model are worse in general, especially in the case of the NNs, which might be due to the fact that the neutral model needs to express a more diverse range of body shapes. Furthermore, the SVR achieves better results for all genders. Thus, we use these models for all datasets, without any fine-tuning or adaptation to specific datasets.

5. Leveraging A2B Model Results for HME

Now that we have trained the A2B models, we can use them to generate precise body shape parameters upfront and reuse them for every evaluation of a specific person. In the next section, we describe how the A2B results can be used to improve existing HME models (see Section 5.1). Further, **we introduce a new approach to HME** (see Section 5.2). We leverage the good performance of a sequence-based 2D-to-3D uplifting HPE model and convert the 3D stick-figure poses to human meshes with the help of our A2B models. With this approach, we achieve superior results compared to existing HME models. However, we want to emphasize that our approach is not exactly comparable to existing ones since it uses the additional information of anthropometric measurements. Since the performance of existing HME approaches is not good enough to be used for sports analyses, our **main goal is to achieve the best possible performance with marginal additional information**. As professional athletes are measured anyway, this results in actual no overhead in these use cases.

We evaluate all models on the ASPset [25] 3D human pose dataset. It consists of various different sports motion clips performed by different subjects, recorded from three camera perspectives. We evaluate on the test set, which contains two subjects and 30 videos for each subject. In the test set, only one camera perspective is public, so we evaluate on this perspective. Evaluating SMPL-X meshes for ASPset is non-trivial. Regressing standard SMPL-X joints from SMPL-X meshes is built-in, but for all other keypoint definitions it is necessary to define a custom regressor. Since there is no regressor available for ASPset, we create a custom SMPL-X regressor [31].

We further evaluate on fit3D [12], since this is the only sports dataset with public SMPL-X annotations. We evaluate the meshes and the SMPL-X joints since they are available. We select a subset of 37 SMPL-X joints. Since our focus is mainly on the body and not on the hands and face, we remove a lot of these joints and consider only the main body pose for MVE calculation. A list of the selected joints can be found in the supplementary. Hence, we achieve a fair comparison with this evaluation scheme. For both datasets, we do not have access to the athletes to measure them. Therefore, we simulate athlete measuring by measuring the GT meshes. Details can be found in the supplementary. Since there is no GT available for the official test set evaluation on the evaluation server of fit3D, we split the official training dataset into a training, validation, and test set for our evaluations. We perform a leave-one-out cross validation and average the results.

Sports datasets differ from most commonly used everyday activity datasets in the aspect that the poses are more diverse and the motions are faster, which makes sports datasets more difficult. In some cases, the poses are so difficult that some models do not detect a human at all. This makes a fair evaluation hard, since the standard MPJPE metric takes the mean of the joint position errors. Assuming a default pose for all frames where no person is detected would result in a very high error that shifts the mean enormously. Hence, we report the MPJPE only on the frames where persons are detected. Since mostly difficult frames are omitted, this will result in a slightly easier setting for methods that find fewer persons, but we include the number of missing frames in our results for comparison.

5.1. Improving HME Model Results

A major problem for HME based analyses is a varying basic body shape within a single video. Existing HME models output different β parameters for each frame. Exemplarily,

we show the standard deviation of the body height of one subject in Table 3. Recall that these measurements and β parameters are based on a T-pose mesh, hence varying poses have no influence on measuring and β parameters. Using β parameters generated with A2B models solves this problem. The necessary 36 measurements are either measured from the human directly, or averaged from the provided GT mesh (fit3D) or IK applied to the GT poses (ASPset). We call these measurements pseudo GT and include more details in the supplementary. We choose this process to simulate real measurements which exist for most professional athletes. We combine existing HME models with the body shape estimated by our A2B models by replacing the estimated β parameters with the ones predicted by the A2B models. We select three recent well performing models on the AGORA dataset (SMPLer-X [3], OSX [19], Multi-HMR [1]), and the first HME method developed by the SMPL-X authors, SMPLify [27]. Since SMPLer-X is trained on the official training data of fit3D, an evaluation with this model is not meaningful, and we omit it here. Moreover, SMPLify-X is not SOTA anymore and achieved the worst results for ASPset. Therefore we omit it, too.

The first evaluation contains the original result from the respective model, and evaluations where the pose from the model is kept, but the β parameters are replaced with the A2B body shape parameters with pseudo GT input. Results are displayed in Table 3. The results for the MVE of the meshes for fit3D are included in Table 5. We can see that for all models, replacing the estimated β parameters by β parameters from our A2B models with pseudo GT input leads to an improvement. For one model, the gendered meshes outperform the neutral ones and for all other models, the neutral meshes perform best. We use the correct gender (male or female) of the subject in the gendered results. Interestingly, the NN outperforms the SVR for all neutral experiments, although the SVRs achieved better results on the AGORA dataset evaluation. The reason could be that AGORA is a synthetic dataset and does not reflect reality. SMPLer-X achieves the best results for ASPset and Multi-HMR for fit3D, both with a significant margin. OSX performs worse on fit3D than on ASPset, but Multi-HMR performs better by a large margin and surpasses OSX. All methods benefit from our A2B β parameters based on pseudo GT with MPJPE improvements from 11 mm to 3 mm regarding both datasets and MVE improvements of approx. 8 mm regarding fit3D.

Although it is not our main goal, we further evaluate the capabilities of a fixed body shape without available GT measurements to ensure consistent body shapes in the case that no measurements are available. The simplest approach is to use the median of the β parameters across all frames of the respective model. However, the β parameters have no real meaning. Therefore, we compare this approach to taking the median of the anthropometric measurements of the generated meshes and then converting them to β parameters via the A2B models. Results are displayed in Table 4. For SMPLer-X and OSX, using the median β parameters lead to equal or even worse results on ASPset. Regarding ASPset, using our A2B models increases the performance of all models slightly. Switching from the neutral output that these models all have to a gendered model works best in most of these cases, but the neutral A2B models also lead to a marginal improvement. Regarding fit3D, using the median β parameters already enhances the MPJPE and MVE results. Using β parameters from an A2B model leads to the same improvement for both metrics, OSX achieves the best results with SVR and the neutral model, Multi-HMR with NN and the neutral model.

Anthropometric measurements can further be used to easily convert between neutral and gendered (male or female) models. In contrast, β parameters are not transferable between models of different genders. Therefore, until now, the conversion could only be achieved by minimizing the MVE between meshes of different genders in an iterative process. We can now use the B2A function to obtain measurements for a mesh of one gender and apply the A2B model of the other gender to these anthropometric measurements in order to get the corresponding β parameters for

	Model	orig.	σ	NN g	SVR g	NN n	SVR n	no r. ↓
ASPset	SMPLer-X	86.0	2.9	78.9	78.5	**78.3**	78.5	0.11%
	OSX	92.3	0.2	89.6	**89.3**	89.4	89.6	0.10%
	Multi-HMR	102.5	3.6	100.0	100.3	**99.3**	99.5	0.44%
	SMPLify-X	138.2	13.0	127.7	127.4	**126.8**	126.9	0.02%
fit3D	OSX	94.2	3.9	88.9	88.6	**87.1**	87.2	3.45%
	Multi-HMR	74.6	3.3	69.6	69.6	**68.0**	68.4	1.54%

Table 3. MPJPE results in mm for existing models on the test splits of ASPset (top) and fit3D (bottom). The second column (*orig.*) contains the original results, the other columns results with replaced β parameters from our **A2B models with pseudo GT anthropometric measurements** as input and either gendered (g) or neutral (n) meshes, and the percentage of frames with no result (no r.). The σ column displays the mean standard deviation of the body height per subject in cm for the original results, while all A2B body shapes have $\sigma = 0$.

	Model	orig.	median	NN g	SVR g	NN n	SVR n
ASPset	SMPLer-X	86.0	86.0	85.9	**85.7**	86.0	86.0
	OSX	92.3	92.4	92.4	**92.2**	92.3	92.4
	Multi-HMR	102.5	**102.0**	102.6	103.0	**102.1**	102.2
	SMPLify-X	138.2	133.6	133.8	**133.5**	133.6	133.5
fit3D	OSX	94.2	**93.0**	95.0	94.8	**93.0**	**93.0**
	Multi-HMR	74.6	**73.9**	75.8	76.1	**73.9**	74.1

Table 4. MPJPE results in mm for existing models on the test split of the ASPset (top) and fit3D (bottom) datasets. The second column contains the original results, the other columns results with replaced β parameters. Either the median β parameters are used or the results from our **A2B models with median anthropometric measurements from the respective model** as input.

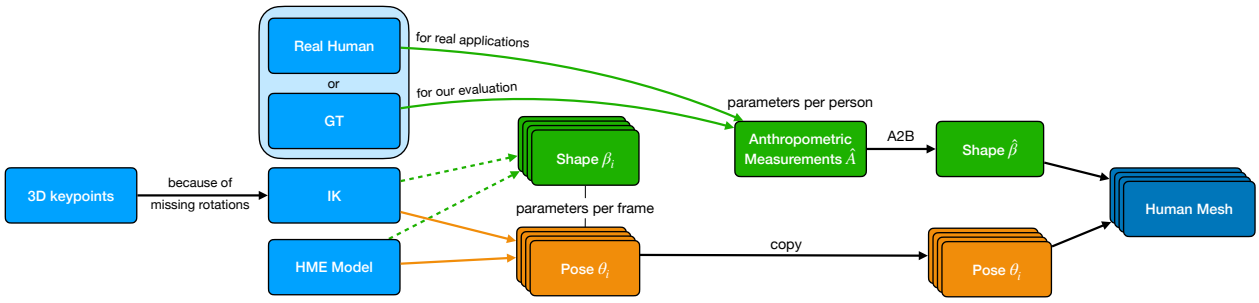

Figure 3. Overview of our inference pipeline. The pose and shape parameters are obtained either from IK applied to UU results (Sec. 5.2.2) or from an HME model (Sec. 5.1). In real applications, the anthropometric measurements will be taken directly from the humans. For our evaluations, we use the GT shape parameters and further experiment with the shape parameters of the respective model (IK or HME).

this gender.

5.2. HME with Sequence Based 3D HPE and A2B

All evaluated HME models are working image-wise. In contrast, SOTA 3D HPE models take a long sequence of 2D poses as an input, which helps to capture movements precisely. The models are called uplifting models, since they lift 2D pose sequences to 3D pose sequences. We use the efficient SOTA 3D HPE model *uplift and upsample* (UU) [10] to estimate the 3D poses on videos. To estimate the required 2D poses from the video frames, we use ViTPose [37], a SOTA 2D pose estimation model. It is important to note that UU operates on pose sequences instead of single frames like the HME models in Section 5.1 and can leverage the information of neighboring frames to estimate a more sophisticated pose. Since we have GT 3D joints available, we can fine-tune the models (ViTpose for 2D HPE and UU for 3D HPE) on our data. This is also necessary to adapt the model to the dataset specific joint definitions since many 3D HPE models like UU are pretrained on datasets like Human3.6M [16], but those joint definitions do not match ASPset nor fit3D. We fine-tune both 2D and 3D HPE models on the training subsets. On the test subsets, UU achieves an MPJPE of 63.85 mm on ASPset and an MPJPE of 29.60 mm on fit3D, which is better than the best existing HME model for both datasets (see Section 5.1). However, UU only outputs 3D joints, no meshes. Moreover, a stick-figure 3D pose is not sufficient to model the pose parameters θ of the SMPL-X mesh, since some rotations are missing. Hence, it is impossible to calculate the necessary rotation parameters directly from the UU result.

5.2.1. Inverse Kinematics for Full Pose Estimation

Because we need the rotation parameters, we use the well established approach of inverse kinematics (IK) with a pose prior to obtain the missing rotations by fitting an SMPL-X mesh to the 3D joint locations estimated by UU. IK outputs the best SMPL-X parameters (β and θ) that fit the mesh to the given 3D joint locations. Details can be found in the supplementary.

5.2.2. Experiments and Results

We evaluate different experiments in Table 5. For comparison, we mention the UU 3D HPE performance (first rows for each dataset in Tab. 5). These results correspond to stick-figure poses and not the required meshes. Therefore, they are not directly comparable to the other results.

Our main approach is shown in the second rows in Table 5. We evaluate the results of IK applied to the UU joint locations with original, median, and pseudo GT based A2B β parameters. The real-world scenario corresponds to the following. GT measurements can be measured from the athlete directly and the 3D pose can be estimated with UU and IK. The β parameters are estimated with the A2B models. A visualization of this pipeline can be found in Figure 3. We include the best result(s) from existing HME models in the respective last rows for comparison and provide qualitative results in Figure 1. **Our approach outperforms the best existing HME model for both datasets by a large margin.** But we need to mention that our approach is not directly comparable to the original results of the HME models, since our approach needs the additional information of measurements, but they already exist in our scenario. However, our approach still outperforms the existing HME models even when they use the same measurements and A2B results as our model does. Further, our model provides results for all frames, which is not the case for the other HME models.

We analyze the results of the building blocks of our model in detail. Applying IK to the UU results deteriorates the UU results by nearly 4 mm for ASPset and 5 mm for fit3D (see Tab. 5, first and second rows, column *orig.*), but this step is necessary since the UU result is only a stick-figure pose and not sufficient for our purpose. Moreover, these results are still better than the best existing HME model (see last rows in Tab. 5).

Next, we replace the inconsistent β parameters with the results from our A2B models. This is especially helpful for our approach since IK produces body shapes with high inconsistencies, as shown by the larger standard deviation of the body height compared to other HME models. For

DS	*inconsistent shape* pose	orig.	σ	measure	*consistent shape (ours)* NN g	SVR g	NN n	SVR n	median	σ	no r. ↓
ASPset	UU	63.9		no mesh							
	IK-UU	67.5	3.0	GT	56.4	56.6	**55.2**	**55.2**	-	0.0	0.0%
	IK-UU	67.5	3.0	IK-UU	66.9	66.6	67.3	67.1	67.2	0.0	0.0%
	SMPLer-X	86.0	2.9	GT	78.9	78.5	78.3	78.5	-	0.0	0.11%
fit3D	UU	34.3	-	no mesh							
	IK-UU	38.5 / 46.3	8.7	GT	41.2 / 47.5	41.3 / 46.9	38.8 / **45.3**	**38.7** / **45.3**	-	0.0	0.0%
	IK-UU	38.5 / 46.3	8.7	IK-UU	42.6 / 51.2	41.6 / 48.6	39.8 / 47.8	39.8 / 47.8	39.9 / 47.8	0.0	0.0%
	Multi-HMR	74.6 / 76.1	3.3	GT	69.6 / 67.8	69.6 / 67.6	68.0 / 68.0	68.4 / 68.8	-	0.0	1.54%
	Multi-HMR	74.6 / 76.1	3.3	Multi-HMR	75.8 / 77.3	76.1 / 76.7	73.9 / 75.6	74.1 / 75.8	73.9 / 75.5	0.0	1.54%
	OSX	94.2 / 89.0	3.9	GT	88.9 / 83.4	88.6 / 82.3	87.1 / 81.1	87.2 / 81.1	-	0.0	3.45%
	OSX	94.2 / 89.0	3.9	OSX	95.0 / 91.7	94.8 / 90.2	93.0 / 87.7	93.0 / 87.6	93.0 / 87.6	0.0	3.45%

Table 5. MPJPE and MVE results in mm on the test splits of ASPset (top) and fit3D (bottom) of our approach compared to the respective best HME model(s). For fit3D, we calculate the MVE, since we have GT meshes available. We display it as the second value in every column. The *pose* column indicates the origin of the pose. The *orig* column contains the result as it is estimated from the method indicated in the *pose* column (with inconsistent body shapes). The right block contains the results with the originally estimated β parameters replaced by consistent ones. The *measurements* column indicates which anthropometric measurements are used for the A2B computation (which β parameters are used for the median computation) whose results are the replacement β parameters in the last five columns. We highlight the overall best results for *estimated* meshes with *consistent shapes* in bold and underline the best (MPJPE and MVE) results in each line. We further add the mean standard deviation of the body height and the percentage of frames with no result as in Table 3.

ASPset, using pseudo GT AMs results in a large improvement of over 12 mm. Remarkably, this result surpasses even the original UU result by 8 mm. It seems that incorporating a clearly defined mesh helps to fix some typical errors of UU and enhance its result in case of ASPset. In general, the error on fit3D is much lower for UU based approaches. The reason might be that it consists of much more data, such that we can fine-tune UU for a longer time. Further, the videos are recorded in a lab in comparison to the in-the-wild videos of ASPset. The lab environment is very similar to the Human3.6M dataset [16], which serves as a training dataset for most recent HME models. Therefore, the results of ASPset are more relevant for future applications of our approach, where we assume only a few available 3D annotations and in-the-wild recordings. For fit3D, applying the A2B body shapes from pseudo GT AMs leads to a slight decrease in performance of 0.2 mm. Inconsistent shapes in the GT (see Section 3) are likely to cause this behavior. Still, our approach using a 3D HPE model and IK outperforms all existing HME models, no matter if the original inconsistent or the consistent body shapes from A2B are used.

Regarding the gendered meshes, we observe that the performance is slightly better for male than for female subjects. fit3D consists of two female and six male subjects. The best score of 40.2 mm for the male subjects is achieved with the SVR model. For the female subjects, the best score is 42.1 mm with the NN model.

As described in Section 5.1, we further evaluate the capabilities of a **consistent shape without available GT AMs**. The naive approach is to use the median of the estimated inconsistent β parameters (Tab. 5, column *median*). Another approach is to use the meshes created by IK applied to the UU results, compute the AMs with B2A, calculate the median AMs and convert them to β parameters via the A2B models. Results are displayed in Table 5, row three. For ASPset, using fixed body shape parameters from A2B models based on the measurements from UU results achieves a slightly better score than the results with inconsistent body shapes. For fit3D, the MPJPE increases by 0.9 mm, but the A2B model results are a slightly better alternative for consistent body shapes compared to the median β parameters.

Further, our approach can be used to generate pseudo GT meshes for datasets with only 3D keypoint annotations. We can use these pseudo GT meshes to fine-tune HME models and increase their performance regarding the estimated keypoints for the specific dataset. However, these results are still worse than the results of our approach. We present the results in the supplementary.

6. Conclusion

We address the problem of inconsistent estimated basic body shapes of humans in videos. We analyze the GT data of 3D pose and mesh datasets and find inconsistencies in their annotations. Then, we propose a family of learned *A2B* models to convert 36 anthropometric measurements to SMPL-X β parameters. This can be used to measure a human once (as it is established practice for athletes, our main focus) and use the resulting shape of the A2B model for all evaluations. With this strategy, the body shape is accurate and consistent per person. Evaluations show that using IK on the results of a SOTA 3D HPE model to estimate the mesh pose combined with our A2B model's shape parameters leads to superior and consistent results compared to existing HME models. Moreover, HME models also benefit from our approach. Replacing their estimated shape parameters with the A2B shape parameters leads to an improvement of their score and consistent body shapes. However, our approach based on 3D HPE still outperforms these scores.

References

[1] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. *arXiv preprint arXiv:2402.14654*, 2024. 2, 6

[2] David Bojanic. Smpl-anthropometry. https://github.com/DavidBoja/SMPL-Anthropometry/, 2023. 1

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6

[4] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *European Conference on Computer Vision*, pages 660–677. Springer, 2022. 3

[5] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 3

[6] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 2

[7] 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. 2

[8] 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. 2

[9] Luc Doyon, Thomas Faure, Montserrat Sanz, Joan Daura, Laura Cassard, and Francesco d'Errico. A 39,600-year-old leather punch board from canyars, gavà, spain. *Science Advances*, 9(15):eadg0834, 2023. 3

[10] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2903–2913, 2023. 2, 7

[11] 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

[12] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[13] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 2

[14] Claire C Gordon, Cynthia L Blackwell, Bruce Bradtmiller, Joseph L Parham, Patricia Barrientos, Stephen P Paquette, Brian D Corner, Jeremy M Carson, Joseph C Venezia, Belva M Rockwell, et al. 2012 anthropometric survey of us army personnel: Methods and summary statistics. *Army Natick Soldier Research Development and Engineering Center MA, Tech. Rep*, 2014. 4, 1

[15] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5219–5228, 2023. 3

[16] 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, 2014. 3, 7, 8, 1, 2

[17] 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. 3

[18] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 3

[19] 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. 2, 6

[20] 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. 2

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 4

[22] 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. 3, 1, 2

[23] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 2308–2317, 2022. 2

[24] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 2

[25] Aiden Nibali, Joshua Millward, Zhen He, and Stuart Morgan. Aspset: An outdoor sports pose video dataset with 3d keypoint annotations. *Image and Vision Computing*, 111: 104196, 2021. 2, 3, 5

[26] 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. 4, 3

[27] 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. 2, 3, 6, 4

[28] Francesco Picetti, Shrinath Deshpande, Jonathan Leban, Soroosh Shahtalebi, Jay Patel, Peifeng Jing, Chunpu Wang, Charles Metze III, Cameron Sun, Cera Laidlaw, et al. Anthronet: Conditional generation of humans via anthropometrics. *arXiv preprint arXiv:2309.03812*, 2023. 2, 4, 1

[29] Zhongwei Qiu, Qiansheng Yang, Jian Wang, and Dongmei Fu. Dynamic graph reasoning for multi-person 3d pose estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3521–3529, 2022. 2

[30] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21254–21263, 2023. 2

[31] Alessandro Russo. Domain analysis of end-to-end recovery of human shape and pose. https://github.com/russoale/hmr2.0, 2020. 5

[32] Rohan Sarkar, Achal Dave, Gerard Medioni, and Benjamin Biggs. Shape of you: Precise 3d shape estimations for diverse body types. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3520–3524, 2023. 2

[33] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 5, 3

[34] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic estimation of 3d human shape and pose with a semantic local parametric model. *arXiv preprint arXiv:2111.15404*, 2021. 3

[35] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 574–584, 2023. 3

[36] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 2

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 7

[38] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11446–11456, 2021. 2