Efficient Model-free Anthropometry from Depth Data

Thomas Probst[1], Andrea Fossati[1], Mathieu Salzmann[2], and Luc Van Gool[1]

[1]Computer Vision Lab, ETH Zurich, Switzerland
[2]CVLab, EPFL Lausanne, Switzerland

Abstract

Existing depth-based approaches to predicting anthropometric measurements, such as body height, arm span and hip circumference, either directly compute the measurements on 3D point clouds, and thus are sensitive to noise, or fit a model to the observed depth values, which typically is time-consuming. In this paper, we rely on the intuition that, to predict a specific anthropometric measurement, one does not need to have detailed information about the entire body shape. We therefore introduce an approach to anthropometry based on a random regression forest trained from local depth cues. The local predictions are then accumulated into one global, image-level anthropometric measurement prediction. We introduce a forest refinement scheme, whose objective function directly relies on both the image-level prediction, as well as on the local predictions' reliability. The resulting approach has the advantage of being both computationally highly efficient and accurate.

1. Introduction

Accurately and rapidly estimating anthropometric measurements, e.g., body height, hip circumference or shoulder breadth, could have a high impact in many applications, such as soft-biometrics for person re-identification, medical health diagnosis, and online garment shopping. Until recently, existing methods and commercial software [19, 35, 40] typically required high-quality 3D scans as input, which limited their applicability. Thanks to the growing availability of low-cost, real-time depth sensors, automatic, computer-based anthropometry has now the potential to become much more generally accessible. However, this would require developing efficient algorithms that have low hardware requirements and are robust to noise and occlusions.

Existing approaches to depth- or 3D-based anthropometry can be roughly classified into two categories. The first one consists of methods that directly estimate the measurements from the input point clouds [32, 6, 33, 5, 39]. These techniques typically rely on heuristics, and require the subject to stay in a reference pose and be completely visible by the sensor. As a consequence, they are quite sensitive to pose variation, sensor noise and occlusions. By contrast, the second category of methods perform anthropometry by first fitting a model to the depth data, and then estimating the measurements on the fitted model [12, 42, 13, 47, 43, 23, 40]. While this makes these approaches more robust to noise, the model-fitting step is usually time-consuming, and thus ill-suited for interactive and real-time applications, or for environments where hardware cost and power consumption are critical.

Here, we propose to overcome these limitations by relying on the following intuition: To predict a specific anthropometric measurement, one typically does not require observing the detailed shape of the complete human body; only portions of the body are sufficient. Following this intuition, we introduce an approach based on random regression forests, where we estimate an anthropometric measurement by accumulating the predictions obtained from local evidence. The benefits of our approach are twofold: First, by relying on local predictions, it is robust to noise and occlusions; second, it circumvents the expensive model-fitting procedure, and thus enables very efficient inference.

More specifically, given a depth image, our approach extracts depth features from local regions, each of which predicts the anthropometric measurement of interest using a random regression forest. In other words, we have access to multiple predictions of the anthropometric measurement, which we can then accumulate to improve robustness to noise. In principle, we are not truly interested in the local predictions, but in the global one. A standard random forest, however, is typically trained to maximize the accuracy of the individual predictions. We therefore introduce a new forest refinement procedure that optimizes a global prediction score based on the accumulated local predictions. Furthermore, we show how the reliability estimation of the local predictions can also be refined and employed in our framework to further improve prediction accuracy.

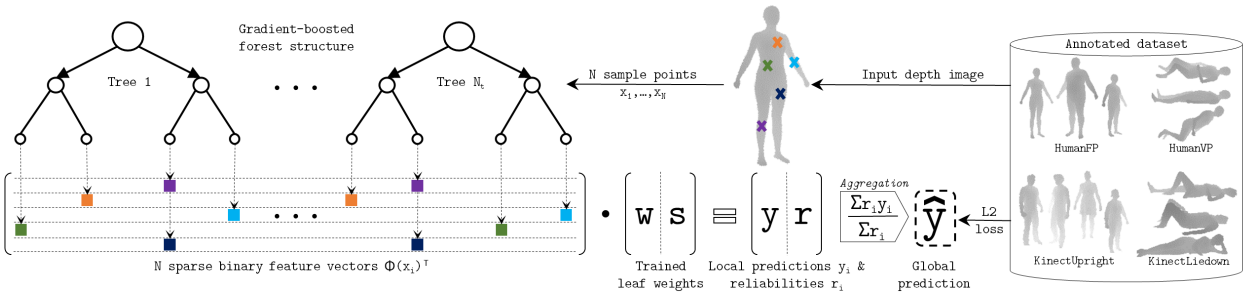


Figure 1. **Method Overview.** Based on RF activation patterns $\Phi(x_i)$ of randomly sampled points x_i, we predict local estimates y_i and their reliability r_i, by introducing two sets of leaf weights (w, s). These weights are trained jointly to minimize the loss on the aggregated estimate $\hat{y}$.

We demonstrate the effectiveness of our approach at accurately predicting diverse anthropometric measurements from real and synthetic data representing different human body shapes and poses, depicted by Fig. 1, and different levels of occlusion. Our experiments evidence the benefits of our method over several baselines in terms of both speed and accuracy. Furthermore, we show that our global prediction and reliability refinement strategy outperforms standard random regression forests and existing refinement schemes [30]. Finally, we demonstrate that our approach extends effortlessly, yet with state-of-the-art speed and accuracy, to performing anthropometry from silhouettes.

2. Related Work

Existing methods for depth-based anthropometry can be roughly grouped into those that directly estimate the measurements from a point cloud, and those that first fit a model to the observations and obtain the measurements from this fitted model. We now review these two types of approaches.

When one is provided with a high-quality 3D scan of the human body, anthropometric measurements can be extracted directly from the 3D point cloud. This is typically the strategy employed by commercial software and several other techniques [19, 35, 37]. In practice, however, such perfect data is difficult and expensive to obtain. One therefore has to deal with the noisy measurements provided by consumer depth cameras, such as the Microsoft Kinect. To this end, [32] proposed to fuse the information acquired from multiple Kinect sensors. Similarly, [6, 39] acquire multiple frames of a person in a fixed pose with a Kinect and combine these observations to obtain the complete human body shape. In both cases, heuristics are then employed to directly compute the anthropometric measurements from the resulting point cloud. There is a vast literature on multi-view 3D and 4D shape reconstruction from multiple cameras [14, 16, 17, 46], multiple depth sensors or depth sequences [33, 5, 21, 20, 44, 2, 45]. In principle, any of these methods could be used to produce a high-quality point cloud, from which anthropometric measurements can be computed. Unfortunately these methods, and those mentioned above, rely on having multiple views of the person, which limits their applicability. Furthermore, directly extracting the anthropometric measurements from the resulting point clouds will be sensitive to noise and occlusion.

By contrast, model-based techniques first attempt to fit a model to the observed data, which makes them more robust to noise. In this context, the SCAPE model [1] is the most popular one, but other models [12, 11, 4, 24, 48, 18] have been investigated, or could potentially be used for this purpose. In particular, [13] and [47] follow a procedure based on the Iterative Closest Point algorithm to fit a SCAPE model to observed point clouds. In [43] and [23], the SCAPE model is employed to obtain the human body shape under the garments worn by the subject. More directly related to anthropometry, [42] optimize the SCAPE parameters to jointly maximize the overlap of the projected model with the RGB-silhouette and minimize the distance between corresponding points on the model and on an input range image. The anthropometric measurements are then obtained from the SCAPE parameters via linear regression. In [40], anthropometric measurements are predicted from global and local geodesic features extracted from a BlendSCAPE model fitted to the data. While effective at handling noise, fitting a model to the observed data is typically time consuming. Therefore, the above-mentioned techniques are ill-suited for applications where run-time efficiency is a requirement. By contrast, [41] exploits the KinectSDK [36] to obtain a human skeleton from which they extract some body measurements and infer others by hand-crafted heuristics and regression. This method, however, requires the full skeleton to be visible in a reference pose, and relies on an accurate detection of each joint. Furthermore, it is ill-suited to predict certain measurements, such as girth and foot size, which limits its applicability.

In short, the model-free approaches are typically sensitive to noise and often rely on multiple views, whereas the model-based methods are computationally expensive. In this paper, we overcome these limitations by relying on the assumption that, to predict a particular anthropometric measurement, one does not need to have a complete, detailed

view of the human body. We therefore rely on the collective power of multiple local predictions, which lets us achieve versatility, efficiency and robustness to noise and occlusion.

3. Method

We now introduce our model-free approach to depth-based anthropometry. We first propose to make use of random regression forests to predict anthropometric measurements from local depth features. We then introduce a novel forest refinement strategy that intuitively models the relationship between local evidence and optimizes for the global, image-level estimate. Importantly, our refinement strategy is not specific to depth-based anthropometry, but applies generally to problems where one seeks to aggregate multiple local estimates into a global prediction.

3.1. Random Forests for Anthropometry

Given a depth image, our goal is to predict a specific anthropometric measurement, such as the hip circumference, or the body height. To this end, we propose to make use of random forests, which have proven effective and efficient at diverse regression tasks [38, 34, 9, 29, 27, 15, 28].

For our purpose, directly working at the level of the complete image would be ill-suited, since it would be highly sensitive to the pose of the subject and to the viewpoint of the sensor. We therefore follow an approach based on local evidence. More specifically, we randomly sample N_s observable center pixels in each depth image, and make use of fast and depth-invariant features [34] to represent each such center. These features encode depth differences between the center pixel and other pixels located at randomly chosen offsets, which will be selected while learning the decision forest. The features are parametrized by the maximal offset range, which regulates the radius of spatial support around each point. We found a maximal offset of ± 125pxm to capture enough context so as to let the forest discover non-obvious correlations between the body parts. The depth differences are then thresholded at values τ, learned during training, thus yielding a binary decision.

This combination of random forest and depth-difference features has several beneficial properties. First, it can cope effortlessly with regression tasks on arbitrarily different input (depth map) and output domains (anthropometrics). Second, the features are inherently invariant to 3D translation. Third, only one path through the tree is explored at test-time, and the features can be computed efficiently "on the fly". These properties enable rapid training, good generalization, and highly efficient inference that can be performed on cheap hardware with low power consumption. Altogether, this makes our approach better-suited to our task than CNNs, which, to our knowledge, have not been applied to depth-based anthropometry.

Below, we discuss our learning procedure to obtain local, per-sample predictions. In Section 3.2, we then introduce our novel forest refinement strategy modeling the fact that we are not interested in local, patch-level predictions, but in a single, global, image-level one.

Learning the forest structure. Since our approach to predicting a global estimate from local samples is formulated as a forest refinement strategy, we first need to obtain a forest structure. The refinement then optimizes the leaf weights, while keeping the tree structure fixed. Learning the forest structure can be thought of as learning a feature extractor to describe a sample point.

To this end, we make use of gradient boosting and of the differential Gaussian information gain as an objective function [10]. We train the trees sequentially, with every tree becoming an expert on the error of its predecessor. The final prediction then has the form

$$\hat{y}_T^{\text{raw}}(\mathbf{x}) = \bar{y} + \sum_{t=1}^{T} \gamma_t r_t(\mathbf{x}) \ , \tag{1}$$

where $\bar{y}$ is the mean value over the training data, and $r_t(\mathbf{x})$ is the quantity predicted by the t^{th} tree to compensate for the remaining residual error. The weights γ_t determine the influence of each tree in the final prediction. Once tree t has been trained, this weight is optimized in a grid search manner, so as to minimize the prediction error

$$\gamma_t = \underset{\gamma}{\text{argmin}} \frac{1}{|\mathcal{P}|} \sum_{(\mathbf{x},y)\in\mathcal{P}} (y - \hat{y}_t(\mathbf{x},\gamma))^2 \ , \tag{2}$$

where $\mathcal{P}$ is the training set.

Reliability-based inference. To predict an anthropometric measurement for a test depth image, we first sample center pixels, extract the depth features at each center, and obtain the prediction of each center according to the random forest. Our goal, however, is to compute a single prediction for the entire image, not one per center pixel. To accumulate the local predictions, we therefore propose to make use of the notion of local reliability. In other words, we associate a reliability value r_i to each local prediction i.

For our *Raw Forest* baseline, we compute r_i as a function of leaf sample variance. Specifically, let us denote by $\phi(\mathbf{x}_i)$ the binary vector whose j^{th} position $\phi_j(\mathbf{x}_i)$ has value 1 if the local sample i has reached the corresponding leaf in the forest, and 0 otherwise[1].We define a local reliability value as

$$r_i^{\text{raw}} = \sum_{j|\phi_j(\mathbf{x}_i)=1} e^{-\sigma_j^2} \ , \tag{3}$$

where σ_j^2 is the variance of the training samples in leaf j.

[1]This vector concatenates the leaves of all the trees in the forest.

The most straightforward way to make use of our random forest then consists of computing the prediction of each local sample point using the raw boosted forest (Eq. 1), and obtain the global, image-level prediction using a mean-shift strategy, where each local prediction is weighted by its reliability value. In the next section, we introduce our novel forest refinement strategy, which, in contrast to this baseline, directly exploits the global prediction during training.

3.2. Reliability-Aware Global Forest Refinement

While effective, the training procedure described in Section 3.1 treats each local sample individually. Therefore, it ignores the fact that, ultimately, we are not interested in the local predictions, but truly in the global, image-level one. To address this issue, we introduce a forest refinement strategy that takes into account the fact that local, per-sample predictions (PSPs) differ in quality due to sample point locations and local occlusions, and aggregates PSPs of the same image into a global, image-level estimate.

Note that our refinement scheme differs fundamentally from the one proposed in [30]. While, in [30], 'global' indicates the set of all training samples, independently of the images they come from, here, it corresponds to the set of 'local' samples that are part of the same image. Furthermore, [30] does not exploit any notion of sample reliability, which, for our purpose, is crucial in the presence of occlusions. These differences required us to design a completely different optimization framework.

Overview (Fig. 1). To refine the forest, we discard the leaf weights of the "raw" forest, and equip each leaf node with two weights $(\mathbf{w}, \mathbf{s})$ that serve for *anthropometric* and *local reliability* prediction, respectively. In particular, we propose to re-use the RF structure $\phi(\mathbf{x}_i^I)$ in conjunction with these weight vectors to obtain new local predictions (Eq. 4) and reliabilities (Eq. 8).

We then introduce two energy terms to learn these weights: The first one, $E(\mathbf{w})$ in Eq. 6, penalizes the error of the aggregated per-image prediction $\bar{y}^I$ (Eq 5), which is intuitively modeled as the normalized sum of the PSPs weighted by their new per-sample reliability $\hat{r}_i^I$. The second energy term, $E_R(\mathbf{s})$ in Eq. 9, aims to learn to estimate the per-sample reliabilities $\hat{r}_i^I$ used for weighted aggregation (Eq. 5). To this end, we define the target reliability value as a function of the error each PSP $\hat{y}_i^I$ produces on its own in the current iteration (Eq. 10).

The energies $E(\mathbf{w})$ and $E_R(\mathbf{s})$ are coupled via the per-sample predictions $\hat{y}_i^I$ and reliabilities $\hat{r}_i^I$. We minimize the resulting non-convex energy using an alternating SGD strategy to jointly optimize the leaf weights $(\mathbf{w}, \mathbf{s})$.

Measurement prediction energy. Formally, let $\mathcal{P}_I = \{(\mathbf{x}_i^I, y_i^I)\}$ represent the set of samples belonging to image I. We model the refinement process as that of learning weights $\mathbf{w}$ generating a local prediction of the form

$$\hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I) = \mathbf{w}^T \phi(\mathbf{x}_i^I) \ . \tag{4}$$

Since, as mentioned above, $\phi(\cdot)$ is a binary vector whose length is equal to the total number of leaf nodes in the forest, this can be thought of as learning the prediction value of each leaf. To account for the entire image, we then define an image-level prediction as

$$\bar{y}^I(\mathbf{w}) = \frac{1}{\sum \hat{r}_i^I} \sum_{\{\mathbf{x}_i^I\}} \hat{r}_i^I \hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I) \ , \tag{5}$$

where $\hat{r}_i^I$ is the current prediction of the reliability value of sample i in image I (Eq. 8).

To make use of the global image prediction in our refinement procedure, we introduce an energy of the form

$$E(\mathbf{w}) = \frac{\lambda_w}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2}\sum_{I\in\mathcal{I}}\sum_{\{\mathbf{x}_i^I\}}(1-r^I)(y^I - \hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I))^2 \ , \tag{6}$$

where $\mathcal{I}$ is the set of all training images, y^I is the ground-truth anthropometric measurement for image I, and

$$r^I = e^{-(y^I - \bar{y}^I(\mathbf{w}))^2} \ . \tag{7}$$

We employ this *per-image* reliability to help the learning procedure focus on difficult images.

Reliability prediction energy. We further propose to follow a similar refinement procedure for the reliability values r_i^I. In particular, following the same idea as for the predictions themselves, we write the reliability value of sample i in image I as

$$\hat{r}_i^I(\mathbf{s}|\mathbf{x}_i^I) = \mathbf{s}^T \phi(\mathbf{x}_i^I) \ . \tag{8}$$

Here, the weights $\mathbf{s}$ can thus be interpreted as reliability values for the leaves, similarly to the $\exp(-\sigma_j^2)$ in Eq. 3. We can then write an additional energy term of the form

$$E_R(\mathbf{s}) = \frac{\lambda_s}{2}\|\mathbf{s}\|_2^2 + \frac{1}{2}\sum_{I\in\mathcal{I}}\sum_{\{\mathbf{x}_i^I\}}(r_i^I - \hat{r}_i^I(\mathbf{s}|\mathbf{x}_i^I))^2 \ , \tag{9}$$

where the target reliability measure is given by

$$r_i^I = e^{-(y_I - \hat{y}_i^I(\mathbf{w}|\mathbf{x}_i^I))^2} \ , \tag{10}$$

and is thus a function of the error each PSP $\hat{y}_i^I$ would produce on its own in the current iteration.

Coupled energy minimization. The per-sample predictions $\hat{y}_i^I$ and reliabilities $\hat{r}_i^I$ therefore couple the two energies in Eqs. 6 and 9. Our goal is to minimize the sum of these two energies with respect to $\mathbf{w}$ and $\mathbf{s}$. Since this is a non-convex optimization problem, we follow an alternating SGD procedure, where we iteratively fix $\mathbf{s}$ to update $\mathbf{w}$ and

vice-versa (for more details see Table 4 in the supplementary material). In each step, we update the variable by running a few stochastic gradient descent iterations. To exploit the sparsity of the indicator vectors $\phi(\mathbf{x})$, we make use of a lazy update strategy and update only the variables affected by the data term.

Inference for a new image. The prediction for a image is performed by first dropping the local samples down the forest trees to obtain the vectors $\phi(\mathbf{x}_i^*)$. The local predictions and reliability values are then computed according to Eqs. 4 and 8, respectively, using the weights $\mathbf{w}$ and $\mathbf{s}$ learned by our refinement procedure. We then compute the image-level prediction as before, by making use of mean-shift on the local predictions with the reliability values as weights.

4. Experiments

4.1. Datasets

To the best of our knowledge, there is no publicly available dataset for anthropometry. Therefore, to evaluate our method on a large set of different body shapes and poses, we introduce several synthetic and real datasets.

Our synthetic dataset was created from synthetically generated mesh data. The motivation behind this is that it allows us to easily annotate the ground-truth anthropometric measurements, and carefully evaluate the behavior of our method under different conditions, such as different body shapes, different poses, and different levels of occlusion. In particular, we employed the MPII Human Shape Model [24]. This rigged statistical shape model was created from the CAESAR dataset [31], which contains a wide variety of body shapes represented as 3D meshes. These meshes are in vertex correspondence, and annotations for body parts and joint positions are provided. We used this model to create two different datasets (see Fig. 1): one where all subjects have the same pose but different body shapes, and one where both the pose and body shape vary. We will refer to these datasets as *HumanFP* and *HumanVP*, respectively. In both cases, the data was obtained by sampling from the 4000 CAESAR-fitted body meshes of [24]. For *HumanFP*, we used the reference pose with additional small pose noise. For *HumanVP*, we randomly assigned a pose combination chosen from a set of 750 sub-poses of upper body (bending, torsion), arm (straight, angled, supporting head,...) and leg (straight, angled) to each one of these meshes. Some of the resulting meshes are shown in Fig. 1. We then obtained the ground-truth anthropometric measurements by using geodesic distances on the meshes and the joint position annotations. In particular, we measured the body height, the shoulder width, the leg length, the foot length, as well as a set of circumferences and thicknesses. Ultimately, our goal is to predict anthropometric measurements from depth data. Therefore, given all the meshes introduced above, we rendered depth images from 12 different viewpoints (random 22.5° steps around 2 rotation axes) using OpenGL. To this end, we employed a virtual camera that mimics the projection and noise properties of the Kinect sensor [22]. This resulted in a total of about 35 000 depth images for *HumanFP* and 50 000 for *HumanVP*. In our experiments, we partitioned these images into training and test sets based on the 4000 mesh models. We used 70% of the models for training and 30% for testing.

To evaluate our approach on real data, we recorded depth images of 20 subjects (aged 24-34, 20/80 female/male, 161-195cm height) wearing clothes in upright and lie-down poses using a Kinect sensor. To obtain ground-truth, we physically measured the anthropometric values of interest on these subjects. Some examples from the resulting datasets are shown in Fig. 1(c,d).

4.2. Setup

We trained two versions of our model, one for each of the two synthetic datasets described above. We then evaluated the resulting models on the corresponding synthetic test data and real Kinect data. In the case of *HumanFP*, where all the subjects have the same reference pose, we employed a forest with 8 trees, trained to depth 24. For *HumanVP*, which also contains different poses, we made use of 16 trees, thus reflecting the fact that this dataset is more challenging. At test time, we randomly sampled 512 center pixels in every image. With 16 trees, this results in a total runtime of 2.3 ms per image on the CPU (Hardware: Intel Core i7-4790K CPU 4.00GHz, 16Gb RAM). We found that increasing the number of sample points did not further improve accuracy. This number, however, can be reduced down to 128, which increases speed (0.6 ms per image) with only minor loss in accuracy. We trained one forest for each body measurement, although we empirically found that transferability of the trained forests between different measurements is very high: With only little loss in accuracy, we can re-use one single forest structure with different leaf weights per measurement (obtained by the refinement strategy) to speed up inference and reduce memory impact. The complete set of hyper-parameters is summarized in the supplementary material.

In our experiments, we compare our results to the following model-based baseline: We employed metric regression forests [27] to predict dense correspondences between the depth map pixels and the MPII HumanShape model [24]. We then used these correspondences to first align the mesh model with the observed data, and second to perform model fitting by gradient descent on the model shape parameters. After this initialization, we further refined the results by alternating alignment and fitting with ICP-based correspondences and optimized until convergence. To avoid local minima, we performed optimization starting from 3 different initial shapes and selected the solution with the best fit. Finally, we computed the body measurements from the re-

sulting mesh. We refer to this baseline as *[27]+ICP*.

4.3. Results

We report results on a variety of anthropometric body measurements, which, we believe, illustrates the generality of our approach. We measure prediction errors using mean absolute error (MAE) and error standard deviation (ESD).

Real data. We first compare our method to the model-based baseline on real Kinect data, which reflects the realistic application scenario. Our real datasets are unfortunately too small to directly train a regression model. Since our models were trained from synthetic data, this real data involves several challenges that have not been explicitly addressed at training time: clothing, reflections, shape model realism, and pose deviations. To account for this domain transfer, we increase our regularization to $\lambda_w = \lambda_s = 10^{-2}$. As the 3D mesh model is only an approximation to a real person, measurements should in principle be corrected. To this end, we estimate a constant offset value from the mean error, which we found to generalize well.

In Table 1, we compare our results with those of our model-based baseline. In short, our approach yields slightly more accurate predictions for much faster runtimes. On *KinectUpright*, we found the hip girth prediction to be the least reliable. We attribute this primarily to the effect of clothing, which makes the thighs appear thicker than they actually are. In general, the *KinectLiedown* scenario is more challenging for both our approach and the baseline, due to very challenging self occlusions. We conjecture that this introduces ambiguities between the arms and the torso, which affect our predictions. We observed that leg length is typically easier to estimate, due to the good leg visibility.

We also provide the errors and computation times reported by other papers on the same anthropometric measurements in Table 2. Although a comparison of these values with ours has to be treated with caution, since they have not been computed on the same data, it still gives an idea of the accuracy of our method and shows in particular the different orders of magnitude in the required processing time. Note that we were unable to find publicly available implementations of other methods, and that the results reported by these methods are on proprietary data. This data, however, typically depicts varying shapes and poses, but all related methods make use of multiple views [5,8,10] or sequences [7,13], and thus assume a controlled environment without any occlusions.

Synthetic data. We made use of our synthetic data to provide an extensive evaluation of the different components of our approach and of its robustness to occlusions. For these experiments, we set the regularizer weights to weaker values, i.e., $\lambda_w = \lambda_s = 10^{-3}$. In Table 3, we compare the results of our complete approach with those obtained with several baselines. A complete comparison of our approach with several baselines for varying levels of occlusions is provided in supplementary material.

Our first baseline is a standard random regression forest, as described in Section 3.1. Note that we significantly outperform this baseline on both *HumanFP* and *HumanVP*. A more interesting comparison can be done with the forest refinement procedure of [30], which still works on individual samples. Note that our approach consistently outperforms this baseline, thus showing the importance of working at image-level. Finally, to evaluate the influence of our reliability refinement step, we compare our approach with a baseline using our image-level prediction refinement scheme, but without exploiting our reliability values. Our complete approach reduces error deviation and still outperforms this baseline. However, the gap on *HumanVP* is quite small in several cases; our approach really yields noticeably higher accuracy on the shoulder width and weight.

In Table 4, we evaluate the robustness of our approach to occlusions. In particular, we included different levels of rectangular occlusions (RO), including none at all (NO) at training time. We created multiple (2 to 4) instances of each depth image, and incorporated these occlusions at random locations by overwriting regions of the image or cropping the image itself. For evaluation, we applied the same random procedure to the test depth maps while making sure a desired percentage of depth points are occluded. We further tested on more realistic occlusions, created using free-form occlusions shaped as humans and pieces of furniture (FO). The results in Table 4 show that, in both the single pose and the varying pose settings, while influenced by the occlusions, our approach consistently outperforms [30] and our approach without reliability (NR), which was the best-performing baseline according to our previous experiments. Note that our method also remains robust to free-form occlusions, even though they were not included in the training set. We can conclude that the concept of local reliability improves prediction accuracy in most cases of severe occlusion. In summary, we believe that our experiments evidence the importance of leveraging multiple local predictions, some of which will come from unoccluded data.

Our model trained from *HumanFP* corresponds to a fixed pose. In the following experiment, we evaluate the robustness of such a model to pose variations at test time. To this end, we employed the *DYNA* dataset [26].Starting from a fixed pose, the 'jumping_jacks' sequence lets us evaluate how much the predictions degrade as the pose varies in the following frames. This is illustrated in Fig. 5 for several anthropometric measurements. We apply each method frame-by-frame, without temporal filtering. Note that our method is robust to small pose variations, but the error of some measurements, such as hip girth and arm length, increases when the variations become too large. Other measurements, such as shoulder width and neck-to-hip distance, are more ro-

Dataset	Approach	Leg Length	Shoulder W.	Neck-Hip D.	Arm Length	Hip Girth	CPU Time
	Average	77	38	63	50	105	-
KinectUpright (trained on *HumanFP*)	[27]+ICP	2.6(3.0)	2.4(3.0)	3.2(3.8)	2.7(3.3)	4.0(4.9)	12
	Ours	2.5(2.8)	2.2(2.6)	2.7(3.1)	2.5(2.8)	3.9(4.7)	0.002
KinectLieDown (trained on *HumanVP*)	[27]+ICP	2.5(2.4)	3.1(2.8)	4.1(5.3)	3.6(4.5)	4.6(6.2)	12
	Ours	2.3(2.5)	2.9(2.4)	4.1(5.1)	3.4(3.9)	4.8(5.9)	0.002
		MAE (ESD) [cm]					s

Table 1. **Comparison of our approach with a model-based baseline on real data.** We report errors for different measurements, as well as runtimes. The average value for each measurement was taken from the MPII Human Shape [24] model pool. Note that our approach yields lower errors than the model-fitting baseline and is much faster.

Approach	Leg Length	Shoulder W.	Neck-Hip D.	Arm Length	Hip Girth	CPU Time
Model-based						
[43]	1.5	–	2.4	1.7	2.1	76
[42]	–	–	2.3	2.6	3.5	3900
Model-free						
[39]	2.1	1.5	2.5	3.0	3.8	354
[6]	–	–	–	–	3.2	20
[5]	3.1	1.0	2.1	2.3	2.6	826
[41]	2.9	–	–	2.6	8.4	–
Unit	MAE (ESD) [cm]					s

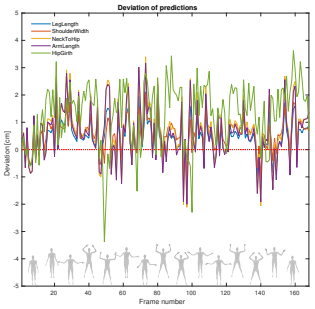
Table 2. **Errors and runtimes reported by other papers on proprietary datasets for the same measurements as in Table 1.** Note that these methods make use of depth sequences and/or multiple views to predict the measurements. Despite this, our errors, reported in Table 1 are comparable to theirs, and our runtimes much faster.

	Measurement	[27]+ICP	Raw Forest	[30]	Ours (NR)	Ours	Unit
HumanFP	Body Height	2.1(1.4)	1.9(2.4)	1.8(2.4)	1.1(1.5)	1.1(1.4)	cm
	Belly Thickness	2.0(1.7)	0.6(0.8)	0.4(0.6)	0.3(0.4)	0.3(0.3)	cm
	Shoulder Width	1.0(0.8)	0.4(0.5)	0.3(0.4)	0.2(0.2)	0.2(0.2)	cm
	Leg Length	0.5(0.4)	0.5(0.6)	0.5(0.6)	0.3(0.4)	0.3(0.4)	cm
	Foot Length	0.6(0.5)	0.3(0.4)	0.2(0.3)	0.2(0.2)	0.1(0.2)	cm
	Body Weight	8.5(7.1)	2.8(3.5)	1.8(2.6)	1.4(1.6)	1.1(1.5)	kg
HumanVP	Body Height	6.4(6.5)	4.0(5.2)	3.9(5.1)	2.9(3.5)	2.9(3.4)	cm
	Belly Thickness	3.3(2.8)	1.1(1.6)	1.1(1.6)	0.9(1.2)	0.9(1.2)	cm
	Shoulder Width	2.1(1.6)	0.8(1.1)	0.8(1.1)	0.7(0.9)	0.6(0.7)	cm
	Leg Length	1.6(1.6)	1.0(1.3)	1.0(1.3)	0.8(1.0)	0.8(0.9)	cm
	Foot Length	1.3(1.1)	0.6(0.8)	0.6(0.8)	0.5(0.6)	0.5(0.6)	cm
	Body Weight	14.0(12.6)	5.1(7.6)	5.2(7.7)	4.4(5.7)	4.2(5.2)	kg

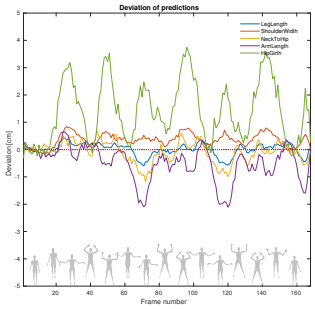
Table 3. **Evaluation on synthetic data.** We compare our approach without (NR) and with reliability measure with a model-fitting baseline, a raw random forest, sample-based refinement [30]. We report the MAE (ESD) in the no occlusion (NO) setting.

		[30]			Ours (NR)			Ours		
	Measurement	25% RO	50% RO	FO	25% RO	50% RO	FO	25% RO	50% RO	FO
HumanFP	Body Height	2.7(3.6)	3.7(4.8)	4.7(5.9)	1.9(2.5)	2.8(3.8)	3.9(5.0)	1.7(2.4)	2.4(3.4)	3.6(4.7)
	Belly Thickness	0.7(1.0)	1.0(1.5)	2.0(2.7)	0.5(0.7)	0.8(1.2)	1.7(2.4)	0.4(0.6)	0.7(1.1)	1.7(2.3)
	Shoulder Width	0.5(0.7)	0.7(1.0)	1.2(1.5)	0.3(0.5)	0.6(0.8)	1.0(1.3)	0.3(0.4)	0.5(0.7)	0.9(1.2)
	Leg Length	0.7(0.9)	0.9(1.2)	1.1(1.4)	0.5(0.7)	0.7(1.0)	0.9(1.2)	0.4(0.6)	0.6(0.8)	0.9(1.1)
	Foot Length	0.4(0.5)	0.6(0.7)	0.8(1.0)	0.3(0.4)	0.4(0.6)	0.7(0.9)	0.2(0.3)	0.4(0.5)	0.6(0.8)
	Body Weight	3.1(4.5)	4.8(6.8)	8.4(12.0)	2.2(3.1)	3.7(5.2)	7.2(10.1)	1.9(2.8)	3.3(4.9)	7.0(9.4)
HumanVP	Body Height	5.2(6.6)	6.0(7.6)	5.1(6.4)	4.3(5.4)	5.3(6.7)	4.2(5.3)	4.1(5.3)	5.2(6.6)	4.1(5.0)
	Belly Thickness	1.8(2.5)	2.3(3.0)	2.1(2.8)	1.5(1.9)	2.0(2.5)	1.6(2.2)	1.4(1.8)	1.9(2.5)	1.6(2.2)
	Shoulder Width	1.3(1.6)	1.5(1.9)	1.4(1.7)	1.2(1.5)	1.5(1.8)	1.2(1.6)	0.9(1.2)	1.3(1.6)	1.1(1.3)
	Leg Length	1.3(1.6)	1.5(1.8)	1.2(1.6)	1.2(1.5)	1.4(1.7)	1.1(1.4)	1.0(1.3)	1.3(1.6)	1.0(1.3)
	Foot Length	0.9(1.1)	1.1(1.3)	0.9(1.1)	0.7(0.9)	0.9(1.2)	0.7(0.9)	0.7(0.9)	0.9(1.1)	0.7(0.8)
	Body Weight	8.2(11.4)	10.3(14.0)	8.8(12.2)	7.2(8.5)	9.0(11.2)	6.2(8.8)	6.4(8.2)	8.5(11.1)	6.1(8.8)

Table 4. **Impact of occlusions.** We compare our approach without (NR) and with reliability measure with the sample-based refinement of[30]. We consider rectangular occlusions (RO) covering up to 25% and 50% of the image, as well as free-form occlusions (FO).



[27]+ICP



Ours

	[27]+ICP	Ours
Leg Length	0.78	0.22
Shoulder Width	0.82	0.40
Neck-Hip D.	1.24	0.39
Arm Length	1.16	0.77
Hip Girth	1.78	1.70

Standard Deviation [cm]

Table 5. **Pose stability on the DYNA [25] 'jumping_jacks' sequence.** We compare our approach with the model-based baseline. Note that both methods have been trained only on the reference pose (start frame). Altogether, both methods yield larger errors as the pose varies, but the results of our approach are smoother over time.

Approach	Leg Length	Shoulder W.	Neck-Hip D.	Arm Length	Hip Girth	Time
[3]	0.6(0.7)	0.6(0.7)	0.4(0.5)	1.5(2.1)	1.1(1.2)	216
[7]	0.9(0.6)	0.2(0.4)	0.3(0.5)	0.3(0.2)	0.4(0.4)	0.450 (GPU)
[8]	2.0(1.9)	0.6(0.6)	1.8(1.7)	1.3(1.2)	2.6(2.5)	0.30
Ours (frontal)	0.5(0.7)	0.4(0.5)	0.7(0.9)	0.8(1.1)	1.4(2.0)	0.002
Ours (trained frontal)	0.2(0.4)	0.1(0.3)	0.3(0.5)	0.4(0.7)	0.5(0.8)	0.002
Ours	0.5(0.7)	0.5(0.6)	0.8(1.0)	0.9(1.2)	2.0(2.8)	0.002
Ours (depth data)	0.5(0.6)	0.3(0.4)	0.6(0.6)	0.5(0.6)	1.3(1.2)	0.002
Unit	MAE (ESD) [cm]					s

Table 6. **Anthropometry from silhouette.** We compare our approach to state-of-the-art methods using synthetically generated silhouettes. Although the test datasets are not the same for all methods, all are derived from CAESAR [31] data and are therefore sampled from the same distribution. Our approach yields errors comparable to the baselines for faster runtimes, which demonstrates its generality.

bust to these pose variations. Compared to the model-based baseline, our method tends to produce slightly higher errors on poses far away from the training pose, but overall the results are temporally smoother.

Anthropometry from silhouette. Note that our approach is not limited to working with depth images. To illustrate this, we therefore evaluate its accuracy at predicting anthropometric measurements from *single* silhouette images, assuming known camera calibration and constant distance to the camera. We rendered noisy silhouette images from our *HumanFP* dataset and trained 16 trees for each body measurement. Here, the depth-difference features act as foreground/background comparisons. In Table 6, we compare our method with different state-of-the-art approaches, including a model fitting approach [3], a CNN-based method [7] and CCA-Forests [8]. Note that these baseline numbers were directly taken from their respective papers, and were obtained using different training and test data. In all cases, however, this data was obtained from the CAESAR data in a similar way as in this paper. Thus, assuming all methods used a representative set, this comparison remains meaningful. Although using only a single image from varying viewpoints in our test data, our approach yields accuracies that are on par with methods that use both a frontal as well as a lateral silhouette [3, 7, 8]. When using frontal test data only, our results further improve, particularly if our model was trained from frontal silhouettes as well. Importantly, our method is much faster than these baselines. Table 6 also shows the accuracy loss compared to using depth data. We believe that this shows the generality of our approach, and its potential to adapt to other applications.

5. Conclusions

We have introduced a highly efficient approach to inferring anthropometric measurements from depth data. To this end, we have exploited random regression forests to compute local predictions from depth features, and have proposed to accumulate these local predictions into a global, image-level one using sample reliability values. Furthermore, we have introduced a novel forest refinement strategy and developed a joint optimization framework for global prediction and sample reliability refinement. Finally, we also contribute several real and synthetic datasets, which include diverse scenarios, such as varying body shapes, poses and levels of occlusions. Our experiments have demonstrated that our approach consistently outperforms standard random forests and sample-based refinement procedures, as well as a model-based approach. Being computationally inexpensive, our approach has the potential to be applied for many tasks, such as soft-biometrics, person re-identification and ambulant medical diagnosis.

ACKNOWLEDGMENT

This work is funded by the EU Framework Seven project ReMeDi (Grant 610902).

References

[1] D. Anguelov, P. Srinivasan, S. Thrun, K. Daphne, J. Davis, and J. Rodgers. SCAPE: Shape Completion and Animation of People. *Lecture Notes in Computer Science*, 7729 LNCS(PART 2):133–147, 2013. 2

[2] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. *ICCV*, 2015. 2

[3] J. Boisvert, C. Shu, S. Wuhrer, and P. Xi. Three-dimensional human shape inference from silhouettes: Reconstruction and validation. *Machine Vision and Applications*, 2013. 8

[4] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[5] Y. Cui, W. Chang, and N. Tobias. KinectAvatar:Fully Automatic Body Capture Using a Single Kinect. In *ACCV Workshop on Color Depth Fusion in Computer Vision*, 2012. 1, 2, 7

[6] N.-L. Dao, T. Deng, and J. Cai. Fast and automatic body circular measurement based on a single kinect. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2014. 1, 2, 7

[7] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. HS-Nets : Estimating Human Body Shape from Silhouettes with Convolutional Neural Networks. In *International Conference on 3D Vision (3DV)*, 2016. 8

[8] E. Dibra, C. Oztireli, R. Ziegler, and M. Gross. Shape from Selfies : Human Body Shape Estimation using CCA Regression Forests. *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 2016. 8

[9] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. van Gool. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision (IJCV)*, 101(3):437–458, 2013. 3

[10] G. Fanelli, J. Gall, and L. V. Gool. Real time head pose estimation with random regression forests. *Cvpr 2011*, pages 617–624, 2011. 3

[11] O. Freifeld and M. J. Black. Lie bodies: A manifold representation of 3D human shape. *Lecture Notes in Computer Science*, 7572 LNCS(PART 1):1–14, 2012. 2

[12] N. Hasler and C. Stoll. A statistical model of human pose and body shape. *Eurographics*, 28(2):1–10, 2009. 1, 2

[13] T. Helten, A. Baak, G. Bharaj, M. Muller, H. P. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *International Conference on 3D Vision (3DV)*, 2013. 1, 2

[14] C. H. Huang, E. Boyer, B. d. C. Angonese, N. Navab, and S. Ilic. Toward User-specific Tracking by Detection of Human Shapes in Multi-Cameras. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[15] H. Y. Jung, S. Lee, D. Comp, and E. S. Eng. Random Tree Walk toward Instantaneous 3D Human Pose Estimation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[16] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless Motion Capture of Multiple Characters Using Multiview Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2720–2735, 2013. 2

[17] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2720–35, 2013. 2

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1—-248:16, oct 2015. 2

[19] L. Mckinnon and C. Istook. Comparative analysis of the image twin system and the 3T6 body scanner. *Journal of Textile and Apparel, Technology and Management*, 2001. 1, 2

[20] R. a. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[21] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. W. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 2

[22] C. V. Nguyen, S. Izadi, and D. Lovell. Modeling kinect sensor noise for improved 3D reconstruction and tracking. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, 2012. 5

[23] F. Perbet, S. Johnson, M.-T. Pham, and B. Stenger. Human Body Shape Estimation Using a Multi-resolution Manifold Forest. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[24] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building Statistical Shape Spaces for 3D Human Modeling. *Pattern Recognition*, 67:276–286, 2017. 2, 5, 7

[25] G. Pons-Moll, D. Fleet, and B. Rosenhahn. Posebits for Monocular Human Pose Estimation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 8

[26] G. Pons-Moll, R. Javier, N. Mahmood, and M. J. Black. Dyna: A Model of Dynamic Human Shape in Motion. *ACM Transactions on Graphics*, pages 1–14, 2015. 6

[27] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric Regression Forests for Human Pose Estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013. 3, 5, 6, 7, 8

[28] T. Probst, A. Fossati, and L. Van Gool. Combining Human Body Shape and Pose Estimation for Robust Upper Body Tracking Using a Depth Sensor. In G. Hua and H. Jégou, editors, *Computer Vision ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 285–301. Springer International Publishing, Cham, 2016. 3

[29] U. Rafi, J. Gall, and B. Leibe. A Semantic Occlusion Model for Human Pose Estimation from a Single Depth Image.

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2015. 3

[30] S. Ren, X. Cao, Y. Wei, and J. Sun. Global Refinement of Random Forest. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 4, 6, 7

[31] K. M. Robinette, H. Daanen, and E. Paquet. The CAESAR project: a 3-D surface anthropometry survey. In *International Conference on 3-D Digital Imaging and Modeling*, 1999. 5, 8

[32] M. Robinson and M. Parkinson. Estimating Anthropometry with Microsoft Kinect. *Proceedings of the 2nd International Digital Human Modeling Symposium*, 2013. 1, 2

[33] A. Shapiro, A. Feng, R. Wang, H. Li, M. Bolas, G. Medioni, and E. Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–2011, 2014. 1, 2

[34] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *Studies in Computational Intelligence*, 411:119–135, 2013. 3

[35] Software. Anthroscan (Human Solutions GmbH). *http://www.human-solutions.com/*. 1, 2

[36] Software. Microsoft Kinect SDK. *https://dev.windows.com/en-us/kinect*. 2

[37] Software. Styku. *http://www.styku.com/*. 2

[38] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 103–110, 2012. 3

[39] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3D Full Human Bodies using Kinects. *IEEE Transactions on Visualization & Computer Graphics*, 2012. 1, 2, 7

[40] A. Tsoli, M. Loper, and M. J. Black. Model-based Anthropometry : Predicting Measurements from 3D Human Scans in Multiple Poses. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 1, 2

[41] C. Velardo and J.-L. Dugelay. Building the space scale or how to weigh a person with no gravity. *2012 IEEE International Conference on Emerging Signal Processing Applications (ESPA)*, pages 67–70, 2012. 2, 7

[42] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D Body Scans from Noisy Image and Range Data. In *International Conference on Computer Vision (ICCV)*, 2011. 1, 2, 7

[43] H. Xu, Y. Yu, Y. Zhou, Y. Li, and S. Du. Measuring accurate body parameters of dressed humans with large-scale motion using a Kinect sensor. *Sensors*, 13(9):11362–11384, 2013. 1, 2, 7

[44] W. Xu, M. Salzmann, Y. Wang, and Y. Liu. Deformable 3D Fusion : From Partial Dynamic 3D Observations to Complete 4D Models. *International Conference on Computer Vision (ICCV)*, 2015. 2

[45] J. Yang, J.-S. Franco, F. Hetroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pages 1344795–4, 2016. 2

[46] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[47] Q. Zhang, B. Fu, and M. Ye. Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[48] S. Zuffi and M. J. Black. The Stitched Puppet : A Graphical Model of 3D Human Shape and Pose. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2