Human Body Measurement Estimation with Adversarial Augmentation

Nataniel Ruiz[2†] Miriam Bellver[1] Timo Bolkart[1] Ambuj Arora[1]
Ming C. Lin[1] Javier Romero[3†] Raja Bala[1]
[1]Amazon [2]Boston University [3]Reality Labs Research
nruiz9@bu.edu {mbellver, timbolka, ambarora, minglinz, rajabl}@amazon.com

Abstract

We present a Body Measurement network (BMnet) for estimating 3D anthropomorphic measurements of the human body shape from silhouette images. Training of BMnet is performed on data from real human subjects, and augmented with a novel adversarial body simulator (ABS) that finds and synthesizes challenging body shapes. ABS is based on the skinned multiperson linear (SMPL) body model, and aims to maximize BMnet measurement prediction error with respect to latent SMPL shape parameters. ABS is fully differentiable with respect to these parameters, and trained end-to-end via backpropagation with BMnet in the loop. Experiments show that ABS effectively discovers adversarial examples, such as bodies with extreme body mass indices (BMI), consistent with the rarity of extreme-BMI bodies in BMnet's training set. Thus ABS is able to reveal gaps in training data and potential failures in predicting under-represented body shapes. Results show that training BMnet with ABS improves measurement prediction accuracy on real bodies by up to **10%**, *when compared to no augmentation or random body shape sampling. Furthermore, our method significantly outperforms SOTA measurement estimation methods by as much as* **3x**. *Finally, we release BodyM, the first challenging, large-scale dataset of photo silhouettes and body measurements of real human subjects, to further promote research in this area. Project website: https://adversarialbodysim.github.io.*

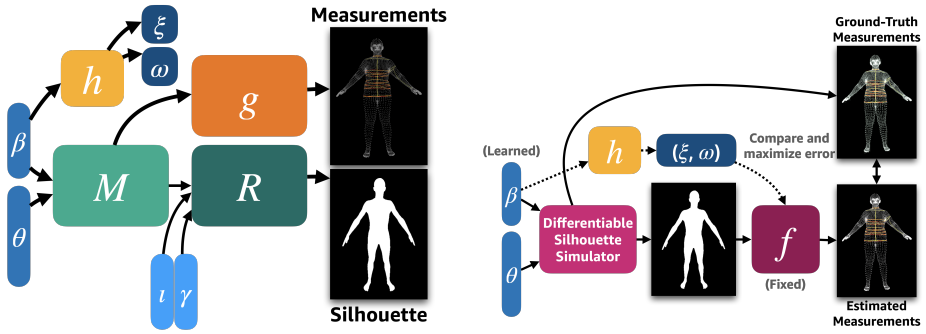
1. Introduction

Reconstruction of the 3D human body shape from images is an important problem in computer vision which has received much attention in the last few years [9, 14–17, 22, 23, 27, 31–33, 35, 39, 40, 45, 50, 54, 61, 62, 68, 84–86, 92, 93, 96]. However, 3D shape is not directly usable for applications where anthropomorphic body measurements are required. In healthcare, for example, measurements such as waist girth are a key indicator of body fat; while in the fashion industry, metric body measurements enable size recommendations and made-to-measure garments. Surprisingly, much less work has been published on directly estimating body measurements from images. This is the problem that we address in this paper. Note that body measurements can be viewed as a compact yet rich descriptor for 3D body shape. Indeed, previous work has shown that it is possible to accurately map a few body measurements to a 3D body mesh in a reference pose [57, 71].

Most existing body reconstruction methods do not incorporate knowledge of camera intrinsics or scale, and thus cannot guarantee metric accuracy (i.e. the distance between two points on the recovered mesh may not correspond to physical distances on a person's body) [34, 72, 80]. Furthermore, since these approaches have only been trained to generate a posed 3D avatar of a human, the body measurements have to be derived from the predicted mesh, which can limit resolution and accuracy. Finally, acquiring physical body measurements at scale is costly and time-consuming; hence, there is a dearth of training datasets pairing images with measurements of real humans. To circumvent this challenge, previous efforts have used synthetic data for training [19, 78], and evaluated on very small numbers (2-4) of human subjects [11, 19].

We present a method to predict body measurements from images that alleviates these shortcomings. We train a convolutional body measurement network (*BMnet*) to directly predict measurements from two silhouette images of a person's body. Silhouettes effectively convey body shape information, while preserving user privacy. To resolve scale ambiguity, we include height and weight as additional inputs to BMnet. We introduce a novel adversarial body simulator (*ABS*) that automatically discovers and synthesizes body shapes for which BMnet produces large prediction errors. ABS is fully differentiable with BMnet in-the-loop. It uncovers weaknesses in the model and gaps in the training data. For example, body shapes returned by ABS tend to be of predominantly high body-mass-index (BMI), consistent with the fact that these shapes are under-represented

†This research was performed while NR and JR were at Amazon.



(a) (b)

Figure 1: **(a) Differentiable silhouette simulator:** SMPL model M generates a body mesh from shape and pose parameters β and θ, which is passed to silhouette renderer R (parameterized by lighting ι and camera γ), and measurement extractor g. Regressor h generates height ξ and weight ω from β. **(b) Adversarial shape optimization:** The simulator renders silhouettes that are passed to BMnet (f) along with height ξ and weight ω to obtain measurement estimates, which are compared to ground truth measurements (also generated by the simulator). The error is maximized with respect to shape β under fixed pose θ.

in training. Fine-tuning BMnet with samples generated by ABS improves accuracy (up to *10%*) and robustness on real data, achieving state-of-art results. To train and evaluate BMnet, we introduce a new dataset, *BodyM*, comprising full-body silhouette images of 2,505 subjects in frontal and lateral poses, accompanied by height, weight, and 14 body measurements derived from 3D scans. To our knowledge, this is the first dataset that pairs photo silhouettes and body measurements for real humans at such a scale.

The main contributions of this work are:

- *BMnet*: A deep CNN to directly regress *physical body measurements* from 2 silhouettes, height and weight;
- *ABS*: A novel *differentiable simulator* for generating *adversarial body shapes* with BMnet in-the-loop, uncovering training gaps and improving BMnet performance on real data (up to **3x**);
- *BodyM*: A new *dataset for body measurement estimation* comprising silhouettes, height, weight and 14 physical body measurements for $2,505$ humans, publicly available for research purposes [1].

2. Related Work

Body reconstruction from RGB images: The literature on recovering 3D human representations from RGB images is vast; see [83] and [85] for excellent surveys. Techniques fall broadly into two categories. Parametric methods characterize the human body in terms of a parametric model such as SMPL{-X} [45, 53], Adam [30], SCAPE [3], STAR [51], or GHUM [88]. Model parameters defining body pose and shape are then estimated from images via direct optimization [10, 53, 87, 93], regression with deep networks [9, 16, 17, 27, 31–33, 40, 50, 61, 62, 94], or a combination of the two [34]. In contrast, non-parametric methods directly regress a 3D body representation from images using graph convolutional neural networks [14, 35], transformers [41], combinations of both [42], intermediate representations such as 1D heatmaps [49] or 2D depth maps [79], or with implicit functions [18, 68]. Recently, there have been successful explorations on probabilistic approaches for shape and pose estimation [36, 69–71].

Body reconstruction from silhouettes: Methods have been proposed to predict 3D body model parameters from binary human silhouette images [4, 5, 20, 55, 72]. Our approach is similar in flavor, but addresses a different task of predicting physical body measurements from silhouettes. Our constrained pose setting, height and weight inputs, and adversarial training scheme enable measurement prediction with state-of-art metric accuracy.

Body measurement estimation: Dibra et al. [19] reported the first attempt at using a CNN to recover a 3D body mesh and anthropomorphic measurements from silhouettes. The silhouettes are generated synthetically by rendering 3D meshes from the CAESAR (Civilian American and European Surface Anthropometry Resource) dataset [60] onto frontal and side views, and body measurements are derived as geodesic distances on 3D meshes. In contrast, our approach is trained on data from both real and synthetic humans, directly regresses measurements, and employs adversarial training for improved performance. Our approach is most closely related to the works of [78] and [90]. Yan et al. [90] use their BodyFit dataset to train a CNN to predict measurements from silhouette pairs. Smith et al. [78] proposed a multitask CNN to estimate body measurements, body mesh, and 3D pose from height, weight, two silhouette images and segmentation confidence maps. For training, they generate synthetic body shapes by sampling the SMPL shape space with multivariate Gaussian shape dis-

[1] https://adversarialbodysim.github.io

tributions and stochastic perturbations of body shapes from CAESAR. In contrast to both these methods, our approach seeks adversarial samples in the low performance regime of BMnet, enabling automatic discovery and mitigation of weaknesses in dataset and network in a principled manner.

Synthesis for training: With advances in simulation quality and realism, it has become increasingly common to train deep neural networks using synthetic data [21, 24, 40, 59, 63]. Recently, there have been attempts at learning to adapt distributions of generated synthetic data to improve model training [2, 6–8, 25, 46, 66, 74, 91]. These approaches focus on approximating a distribution that is either similar to the natural test distribution or that minimizes prediction error. Another flavor of approaches probes the weaknesses of machine learning models using synthetic data [28, 37, 38, 48, 56, 67, 77]. The works of [1, 76, 95] generate robust synthetic training data for object recognition and visual-question-answering by varying scene parameters such as pose and lighting, while preserving object characteristics. Shen et al. [75] tackle vehicle self-driving by introducing adversarial camera corruptions in training. In our work, we explore the impact of varying interpretable parameters that directly control human body shape.

Adversarial techniques: We take inspiration from the literature on adversarial attacks of neural networks [12, 26, 52, 81] and draw from ideas for improving network robustness by training on images that have undergone white-box adversarial attacks [47]. The main difference lies in the search space: previous works search the image space while we search the interpretable latent shape space of the body model. The works by [58, 65] find synthetic adversarial samples for faces using either a GAN or a face simulator. They are successful in finding interpretable attributes leading to false predictions; however, they do not incorporate this knowledge in training to improve predictions on real examples. In our work, we both discover adversarial samples and use them in training to improve body measurement estimation. Different from previous methods, we find adversarial bodies by searching the latent space of a body simulator comprising a pipeline of differentiable submodules, namely: a 3D body shape model, body measurement estimation network, height and weight regressors, and a renderer based on a soft rasterizer [43].

Datasets: Widely used human body datasets such as CAESAR [60] contain high volumes of 3D scans and body measurements; however these do not come with real images, which must therefore be simulated from the scans with a virtual camera. Recently Yan et al. [90] published the *BodyFit* dataset comprising over 4K body scans from which body measurements are computed, and silhouettes are simulated. They also present a small collection of photographs and tape measurements of 194 subjects. To resolve scale, they assume a fixed camera distance. Our BodyM is the first large-scale dataset comprising body measurements paired with silhouettes obtained by applying semantic segmentation on real photographs. To resolve scale, we store height and weight (easy to acquire) rather than assume fixed camera distance (hard to enforce in practice).

3. Method

We use the SMPL model [45] as our basis for adversarial body simulation. SMPL characterizes the human form in terms of a finite number of shape parameters β and pose parameters θ. Shape is modeled as a linear weighted combination of basis shapes (with weights β) derived from the CAESAR dataset, while pose is modeled as local 3D rotation angles θ on 24 skeleton joints. SMPL learns a regressor $M(\beta, \theta)$ for generating an articulated body mesh of 6890 vertices from specified shape and pose using *blend shapes*.

3.1. Body Measurement Estimation Network

BMnet takes as input either single or multi-view silhouette masks. For single-view, only a frontal segmentation mask is used. For multi-view, the model also leverages the lateral silhouette which provides crucial cues for accurate measurement in the chest and waist areas. Additionally, we use height and weight as input metadata. Height removes the ambiguity in scale when predicting measurements from subjects with variable distance to the camera, while weight provides important cues for body size and shape. Our multi-view measurement estimation network can be written as:

$$y = f_\psi(x_f, x_l, \xi, \omega), \tag{1}$$

where x_f and x_l are respectively the frontal and lateral silhouettes, (ξ, ω) are the height and the weight of the subject, and ψ represents network weights.

The network architecture comprises a MNASNet backbone [82] with a depth multiplier of 1 to extract features from the silhouettes. Each silhouette is of size 640×480 and the two views are concatenated spatially to form a 640×960 image. Constant-valued images of the same size representing height and weight are then concatenated depthwise to the silhouettes to produce an input tensor of dimension $3 \times 640 \times 960$ for the network. The resulting feature maps from MNASNet are fed into an MLP comprising a hidden layer of 128 neurons and 14 outputs corresponding to body measurements. Unlike previous approaches that attempt the highly ambiguous problem of predicting a high-dimensional body mesh and then subsequently computing the measurements from the mesh [19], we directly regress measurements, thus requiring a simpler architecture and obviating the need for storing 3D body mesh ground truth.

3.2. Adversarial Body Simulator

We present an *adversarial body simulator* (ABS) that searches the latent shape space of the SMPL model in order

to find body shapes that are challenging for BMnet. Given a set of shape and pose parameters (β, θ), we generate a SMPL body mesh $M(\beta, \theta)$. We then render a 2D silhouette image x of this body using a graphics renderer $R()$, given camera parameters γ and lighting conditions ι:

$$x = R(M(\beta, \theta), \iota, \gamma). \tag{2}$$

Combining Eq. 1 and 2 we arrive at an expression for measurements predicted by BMnet for a SMPL body as:

$$y = f_\psi(R(M(\beta, \theta), \iota, \gamma_f), R(M(\beta, \theta), \iota, \gamma_l), \xi, \omega), \tag{3}$$

where y is the vector of body measurements predicted by BMnet; γ_f are the frontal camera parameters, γ_l are the lateral camera parameters where the camera azimuth has been decreased by 90 degrees, and (ξ, ω) are the height and weight of the subject. The goal of adversarial simulation is to seek challenging inputs that result in high measurement prediction loss $L(y, y_{gt}) = ||y - y_{gt}||^2$ where y_{gt} are ground truth measurements:

$$\max_\beta [L(f_\psi(x_f(\beta), x_l(\beta), \xi, \omega), y_{gt})], \tag{4}$$

where, $x_f(\beta)$ and $x_l(\beta)$ are the frontal and profile renders with shape parameters β. We construct our setup so that loss L is differentiable with respect to shapes β, enabling the use of gradient back-propagation to find adversarial samples. We now investigate in detail the dependence of y and y_{gt} on β. Turning first to y in Eq. 3, the SMPL model M is linear and thus differentiable with respect to β. The renderer R is designed as a differentiable projection operator from the 3D body mesh to a 2D silhouette. First, the posed body is lit by a frontal diffuse point light and captured by a perspective camera pointed towards the body mesh. The 2D image is generated using a fully-differentiable soft silhouette rasterizer that aggregates mesh triangle contributions to each 2D pixel in a probabilistic manner [43]. While lighting is not critical for silhouette generation, we include it as part of a general RGB image generation framework.

Recall that height and weight (ξ, ω) are inputs to BMnet. These are not natural outputs of SMPL; however they are strongly correlated with body shape. We construct a differentiable 3-layer neural network regressor h that predicts height and weight ξ and ω from shape β. We train h in a supervised fashion on the CAESAR dataset, which contains subject height and weight as well as body mesh data. We fit a gender-neutral SMPL model with 10 shape parameters ($\beta \in R^{10}$) to the body meshes, providing tuples (β, ξ, ω) for training h. The choice of a gender-neutral model is based on our earlier findings that gender contributes minimal improvement to height/weight prediction, and the fact that gender must be determined either automatically (which is error prone) or by asking the user (not everyone shares or identifies with gender). Average prediction errors of h on independent test sets are within 1 cm and 1 kg respectively.

Next we turn to y_{gt}. For a given SMPL body mesh M the 14 body measurements are obtained by computing the lengths of curves traversing pre-specified vertex paths on the mesh. These curve lengths are computed by summing vertex-to-vertex distances along the path. This operation, denoted $y_{gt}(\beta) = g(M(\beta, \theta))$, is the same used to annotate the BodyM dataset (see Sec. 4). The fully differentiable silhouette renderer is shown in Figure 1a.

In order to sample adversarial bodies we optimize β by gradient ascent, backpropagating the gradient of the loss with respect to β:

$$\nabla_\beta L[f_\psi(x_f(\beta), x_l(\beta), \xi(\beta), \omega(\beta)), y_{\text{gt}}(\beta)], \tag{5}$$

where height and weight depend on β via $h()$, and y_{gt} depends on β via $g()$. For body shape analysis (Sec. 5.1) we fix pose θ to a canonical A-pose, while for training BMnet, we sample θ randomly from poses of real humans in the BodyM dataset. Henceforth we omit pose, camera and lighting parameters for brevity. The β are updated using the gradient ascent update rule:

$$\beta_{k+1} = \beta_k + \eta \nabla_\beta L[f_\psi(x_f(\beta), x_l(\beta), \xi(\beta), \omega(\beta)), y_{\text{gt}}(\beta)], \tag{6}$$

where η is a weight hyperparameter. Note that only β is updated, and the weights ψ of the model f are fixed. We illustrate the optimization in Figure 1b.

Adversarial augmentation: In order to train a network in an adversarial manner, we need to ensure our adversarial sampling selects diverse yet realistic body shapes. For this purpose, when doing adversarial augmentation, we initialize β by selecting at random shape parameters that have been fitted to real human bodies in the BodyM training set. We then optimize these shapes for k iterations following the update rule shown in Equation 6. This yields samples that are challenging, yet close to real body shapes. An alternate strategy would be to sample only around challenging examples in the training set; however we have found that excessive emphasis on hard examples causes BMnet to overfit on these and compromise mean performance.

We first pre-train BMnet on real examples from BodyM, and then fine-tune for 10 epochs using synthetic examples from the aforementioned augmentation. Training BMnet minimizes the L1 difference between regressed and target measurements. Synthetic bodies are not repeated over epochs, so that in 10 epochs the network sees roughly 10 times more data than the one using real data. Finally we perform another fine-tuning on the real BodyM data to bridge the synthetic-to-real domain gap. We note that synthetic silhouettes produced by the renderer R are noise-free, while silhouettes in BodyM are generated by segmenting real RGB photos, and thus contain realistic noise artifacts.

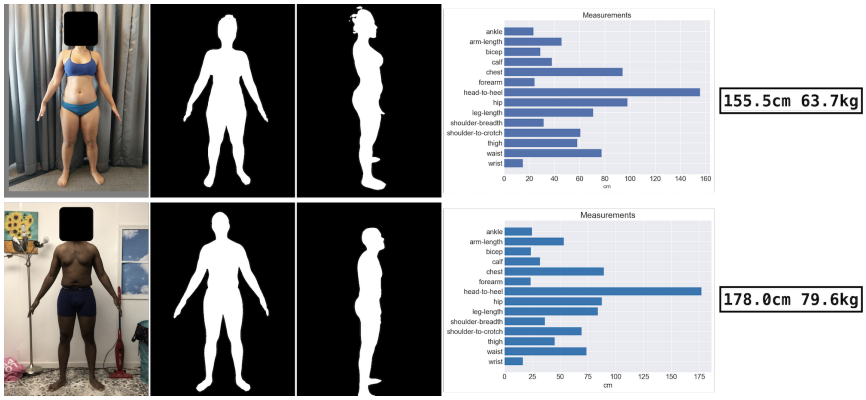


Figure 2: Example frontal color photograph, frontal and profile segmentation masks, body measurements and height/weight for different subjects in the BodyM Training Set (top) and the Test-B (bottom) datasets respectively.

Our augmentation strategy is inspired by adversarial training using pixel-level adversarial attacks [47], with some key differences: (1) we search through interpretable parameters of a simulator to find adversarial samples instead of modifying image pixels using high-frequency noise; (2) we use a gradient descent update instead of the quantized fast gradient sign update rule, since the latter leads to a coarse exploration of the landscape that is not suitable when searching for simulated adversarial examples in shape and pose space. Note that some additional training computational cost exists, but is in the order of 1%.

4. BodyM Dataset

Synthetic datasets used in previous body measurement work often lack the detail and diversity of real body shapes. To address this domain gap, we introduce *BodyM*, the first public dataset containing 8,978 frontal and lateral silhouette photos paired with height, weight and 14 body measurements for 2,505 real individuals. The ethnicity distribution of BodyM is: White 40%, Asian 30%, Black/African American 14%, American Indian or Alaska Native 1%, Other 15%; with 15% of the individuals also indicating Hispanic. The training-test breakdown is reported in Table 1 (top). Table 1 (bottom) reports gender and BMI statistics. We note that *BMI* $\in$ *18.5-25* and *BMI* $\in$ *25-30* are the dominant body shape categories. RGB photos were captured in a well-lit, indoor setup, with subjects standing in A-Pose wearing tight-fitting clothing, as shown in Figure 2. Capture distance varied between 5.5-6.5 feet. Silhouettes were obtained by applying semantic segmentation on RGB [13], thus exhibiting realistic segmentation artifacts not found in existing simulated datasets (e.g., Figure 2 top)3D scans of each subject were acquired with a Treedy photogrammetric scanner, registered to the SMPL mesh topology, and reposed to a canonical "A-pose". The following body measurements were then computed on the meshes using the procedure described in Sec. 3.2: *ankle girth, arm-length, bicep girth, calf girth, chest girth, forearm girth, head-to-heel length, hip girth, leg-length, shoulder-breadth, shoulder-to-crotch length, thigh girth, waist girth,* and *wrist girth.*

	Train	Test-A	Test-B
Subjects	2,018	87	400
Silhouettes	6,134	1,684	1,160

	Training Set		Test-A Set		Test-B Set	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
GENDER	60%	40%	52%	47%	39%	61%
BMI <18.5	0%	2%	1%	3%	1%	4%
BMI 18.5-25	28%	23%	33%	31%	18%	37%
BMI 25-30	25%	9%	15%	7%	13%	10%
BMI 30-40	6%	5%	2%	6%	7%	7%
BMI 40-50	0%	1%	0%	0%	1%	3%
BMI >=50	0%	0%	0%	0%	0%	0%

Table 1: BodyM dataset statistics (top), gender and BMI statistics for the BodyM (bottom).

For the training and Test-A sets, subjects were photographed and 3D-scanned by lab technicians. For the Test-B set, subjects were scanned in the lab, but photographed in a less-controlled environment with diverse camera orientations and lighting conditions, to simulate in-the-wild image capture. For privacy reasons, we do not release the original RGB images (not anyway needed by BMnet).

5. Experimental Results

For all experiments, unless noted, we train the baseline BMnet for 150k iterations on the BodyM training set using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 22. We select the best model using a validation set corresponding to 10% of the training data. The learning rate follows a multi-step schedule, whereby we reduce the learning rate at 75% and 88% of the training.

Metrics: We define measurement accuracy based on quantiles of absolute measurement errors. TP90 (TP75, TP50) metrics are defined by computing the 90th (75th,

50th) quantile cutoff for all 14 measurements, and reporting the mean of these values. Mean absolute error (MAE) is reported for selected experiments.

5.1. Adversarial Body Shape Analysis

We use ABS to reveal regimes of the body shape space where a pre-trained BMnet performs poorly. We initialize a 10-dimensional SMPL shape vector β to fall randomly within a small ball of radius 0.01 around the zero-vector. We then iteratively update β to maximize BMnet loss. The camera parameters γ are chosen to mimic the setup used to capture real images in BodyM. The lighting parameters ι represent a point illumination source that shines directly onto the subject from behind the camera, using only diffuse lighting, in order to avoid specular artifacts from corrupting the silhouette. While we fix pose, lighting and camera parameters as constants in this experiment, we note that our framework can be readily generalized to adversarial sampling of all these parameters.

For ABS we use adversarial sampling with a learning rate η of 0.1 and $k = 10$. Shape parameters are clamped in a $[-3, 3]$ range to prevent unrealistic body shapes. We compare samples generated using ABS to random body samples. Using this random sampling, we sample the shape space uniformly in the $[-3, 3]$ range. Our rationale is that in the absence of prior knowledge about f, the uniform distribution is the maximum entropy distribution, hence providing the strongest sampling baseline.

We show qualitative comparisons between randomly simulated bodies and adversarially simulated bodies in Figure 3, left. We observe that adversarial body shapes are of high BMI compared to random bodies. The mean measurement error of f for the adversarial bodies is also much higher than that for random bodies. We also note adversarial samples that are not of high BMI but with high measurement error (third sample in Fig. 3, left). Fig. 3, right, shows examples of real bodies, both random and samples with high error. We observe that the error for the hard samples is similar to that of the simulated adversarial samples. Furthermore, we can see that challenging samples in the real world are also of high BMI, similar to our simulated adversarial bodies. Aggregating this analysis over the entire population, the average BMI's for the random and adversarial body groups are 28.1 and 35.8 respectively; and the mean measurement errors in millimeters (mm) for the two groups are 34.8 and 92.2.

As another visualization, Figure 4 plots mean measurement error vs. BMI for adversarially sampled and random bodies. Error magnitudes are color-coded. We observe that the adversarially sampled population contains more bodies with higher error (red circles) and fewer bodies with low error (green circles). Furthermore, the adversarially sampled population contains many more samples with high BMI, which seem to directly contribute to higher mean error. In Figure 5 we visualize adversarial and random body sampling in the first two principal dimensions of the latent SMPL shape space. Again error magnitudes are color-coded. Adversarial (and high-error) bodies are largely concentrated in the negative quadrant of the shape space. For visual interpretation, Figure 6 shows that negative perturbations in β_1 and β_2 result in taller and wide bodies.

	Overall			Chest	Hip	Waist
	TP90	TP75	TP50	MAE	MAE	MAE
Single-View	41.91	29.13	17.09	33.95	31.03	31.93
Multi-View	39.02	26.55	14.85	28.66	28.29	27.32
Multi-View + Height	20.21	13.87	8.00	19.38	15.97	18.71
Multi-View + Weight	18.55	12.62	**7.20**	**15.22**	10.54	**13.69**
Multi-View + Height + Weight	**18.42**	**12.55**	7.34	15.92	**9.74**	15.44

Table 2: Ablations on single- vs. multi-view and height/weight inputs (errors in mm.) on BodyM TestA. Addition of a second view improves the accuracy of body measurements. Adding only the weight has stronger (positive) impact than adding only height. Robustness to outliers is improved when adding both height and weight.

	Overall			Chest	Hip	Waist
	TP90	TP75	TP50	MAE	MAE	MAE
Single-View (No Aug.)	19.10	13.00	7.64	19.18	11.53	16.12
Single-View (Random Aug.)	18.98	12.84	7.50	19.13	11.43	**15.76**
Single-View (Adv. Aug.)	**18.90**	**12.82**	**7.44**	**18.84**	**11.14**	15.78
Multi-View (No Aug.)	16.45	11.06	6.51	**14.40**	10.88	13.40
Multi-View (Random Aug.)	16.43	11.06	**6.48**	14.66	10.60	13.17
Multi-View (Adv. Aug.)	**16.00**	**10.00**	6.53	14.52	**10.00**	**13.00**
Multi-View (No Aug.)	26.52	17.64	10.04	24.60	19.55	21.75
Multi-View (Random Aug.)	26.13	17.35	9.90	23.09	18.87	22.44
Multi-View (Adv. Aug.)	**25.00**	**16.28**	**9.50**	**22.98**	**18.09**	**21.10**

Table 3: Ablations (on BodyM TestA) for synthetic data augmentation strategies. BMNet trained on the full training set (top two row blocks) and a reduced training set (bottom row block). Adversarial augmentation achieves lower errors (up to *10%*) over no augmentation or random sampling.

5.2. Ablation Studies

We highlight the impact of key elements of our body measurement estimation architecture in Table 2. First we study the effects of using one (frontal) input silhouette vs. two (frontal and lateral), as well as the effect of adding height and weight as metadata inputs to the network.

The addition of a second (lateral) view improves results by providing additional evidence not found in the frontal view. The addition of height and weight dramatically affect the network's ability to correctly predict measurements. Adding only the weight has stronger impact than adding only height. Robustness to outliers is improved when adding both height and weight, evidenced by the TP90 and TP75 metrics, although some specific measurements are less accurate than when only using one input.

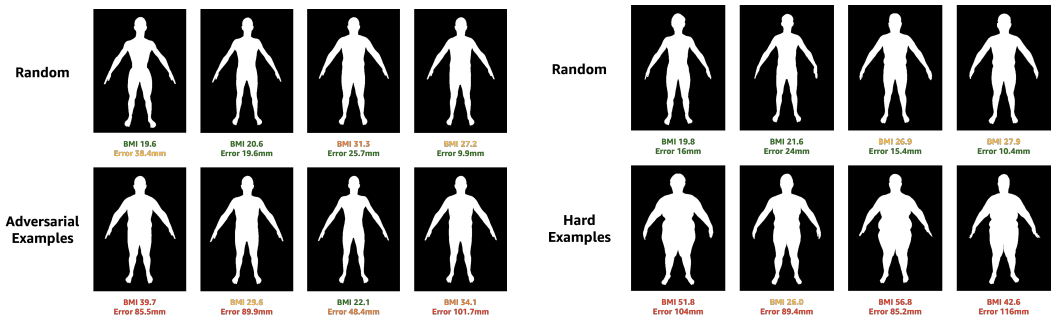


Figure 3: Comparison of randomly vs. adversarially simulated bodies (left). Comparison of random sampling vs. hard examples of real bodies with high body measurement estimation error (right).

	Ankle	Arm	Bicep	Calf	Chest	Forearm	H2H	Hip	Leg	S-B	S-to-C	Thigh	Waist	Wrist	Overall
Ours No Aug.	7.89	9.97	11.42	11.29	24.60	7.31	10.10	19.55	14.33	7.83	9.72	15.54	21.75	5.73	12.65
Ours Adv. Aug.	**7.59**	**9.91**	**11.26**	**10.88**	**22.98**	**7.16**	**9.19**	**18.09**	14.97	**7.67**	**9.30**	**14.11**	**21.10**	**5.52**	**12.12**

Table 4: Mean average individual measurement errors on Test-A using the reduced training set (bottom). *H2H stands for Head-to-Heel, S-B is Shoulder-Breadth and S-to-C is Shoulder-to-Crotch.*

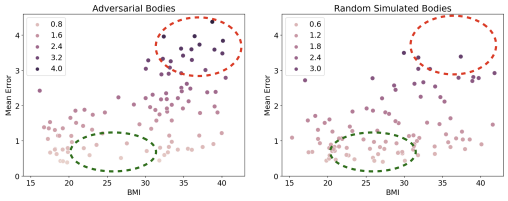


Figure 4: Prediction error vs. BMI for adversarial (left) and random (right) body sampling. The adversarial scheme selects more high-error and high-BMI samples (red ellipse) than the random sampling, more concentrated in low-error low-BMI areas (green ellipse).

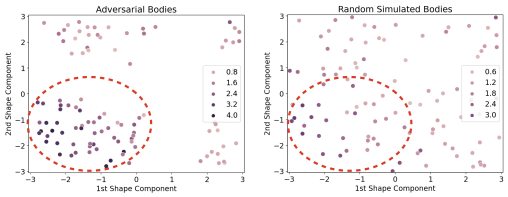


Figure 5: Distribution of adversarial (left) vs. random (right) sampling along the first two components of SMPL shape space. Adversarial bodies are more concentrated in the negative quadrant (red ellipse).

This could be attributed to slight noise in height and weight.

Next we evaluate the impact of augmenting real data samples with synthetic data drawn from different sampling strategies when training BMnet. We fit the SMPL model with 10 shape and 72 pose parameters to each real body in the BodyM training set. We then sample around these real parameters, in order to create synthetic augmentations with shape and pose in the vicinity of real bodies. We compare two methods of augmentation: random sampling and the proposed adversarial sampling around the BodyM shape parameters. Random sampling is performed uniformly within a hypercube with side length of 0.5 around the real shape parameters. For adversarial sampling, we initialize shape using the BodyM parameters and optimize using the Adam optimizer for 5 steps using a learning rate of 0.1.

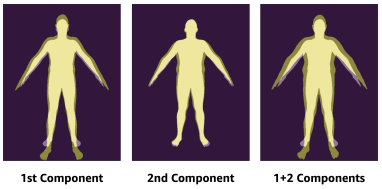


Figure 6: Changes in body shape (from yellow to green) from adding small negative perturbations to the 1st and 2nd shape components. We surmise that the network underperforms with taller, larger, and wider bodies.

We also evaluate a baseline network trained only on real BodyM data. All evaluations are performed on the independent BodyM Test-A set, and results are reported in Table 3. We observe that for the single-view scenario, random augmentation achieves slightly better results than no augmentation, and adversarial augmentation further improves results over random augmentation, with no additional data acquisition cost. This trend is consistent across aggregate TP metrics, most of the individual body measurements, and the multi-view scenario. We see relatively uniform improvements over different BMI categories.

We repeat the ablation study with a reduced training dataset of 1K randomly chosen real samples ($\sim$ 10% percent of the data used in the previous experiment). Training of BMnet thus weighs more heavily on synthetic augmentation. Results are shown in the lowest block of Table 3.

	Ankle	Arm-L	Bicep	Calf	Chest	Forearm	H2H	Hip	Leg-L	S-B	S-to-C	Thigh	Waist	Wrist	Overall
Dibra et al. [19]	2.0	2.7	3.3	3.3	7.2	2.3	4.0	*6.0*	2.8	2.9	2.9	4.9	8.1	2.0	3.78
Smith et al. [78]	2.1	**1.7**	2.7	2.3	4.7	1.9	**2.3**	3.0	**1.5**	1.9	**1.5**	2.4	4.8	2.5	2.72
Ours	**0.8**	1.9	**1.7**	**0.8**	**4.6**	**1.3**	3.6	***1.8***	2.1	**0.9**	1.9	**1.7**	**3.8**	**0.7**	**1.97**

Table 5: MAE (mm) for different methods on the simulated dataset from [78]. *H2H* stands for Head-to-Heel, *Arm-L* and *Leg-L* is arm/leg length, *S-B* is Shoulder-Breadth and S-to-C is *Shoulder-to-Crotch*. Ours achieves up to *70%* error reduction on real-body measurements.

	Neck	Chest	Waist	Pelvis	Wrist	Bicep	Forearm	Arm	Leg	Thigh	Calf	Ankle	Height	Shoulder	Overall
Yan et al. [90]	11.8	23.0	16.5	**13.3**	4.1	11.4	7.2	**7.6**	**9.2**	17.8	8.8	**5.4**	9.0	9.2	11.0
Ours	**11.0**	**15.2**	**15.7**	17.3	**3.8**	**4.7**	**3.9**	7.7	10.0	**7.5**	**8.6**	10.0	**13.4**	**7.1**	**9.7**
Yan et al. [90]	*14.6*	21.7	17.1	14.7	5.2	9.3	8.5	**6.4**	**6.5**	**11.6**	**9.2**	**6.1**	**8.6**	7.6	10.5
Ours	**4.4**	**9.1**	**10.8**	**7.7**	**5.2**	**3.9**	**5.3**	**6.4**	10.2	13.2	9.8	12.2	20.7	**6.5**	**9.0**

Table 6: MAE (mm) on Body-Fit [90] for male (top) and female (bot.) bodies with error reduction up to *70%* using ours.

	Overall			**Chest**	**Hip**	**Leg Length**	**Waist**
	TP90	TP75	TP50	MAE	MAE	MAE	MAE
SPIN [34]	81.10	57.33	33.96	74.45	65.41	35.81	77.39
STRAPS [72]	103.61	75.74	45.67	82.30	63.96	48.71	108.00
Sengupta et al. [69]	68.81	47.64	28.71	53.07	47.43	42.11	53.20
Ours (Single-View, No Metadata)	**41.91**	**29.13**	**17.09**	**33.95**	**31.03**	**25.80**	**31.93**

Table 7: Performance comparison for measurement estimation using different methods on the BodyM dataset.

Adversarial augmentation produces the best overall performance, with significant improvements on several individual measurements. This highlights the benefit of adversarial sampling in scenarios where real data is limited. Finally, in Table 4, we break down the performance of multi-view BMnet by individual measurements, with and without adversarial augmentation, in the reduced-data scenario. Adversarial augmentation produces a noticeable decrease in almost every individual body measurement error. Results for the Test-B set are included in Sup. Mat.

5.3. State-of-the-Art Comparisons

We compare our method with recent state-of-the-art body measurement approaches by Smith et al. [78] and Dibra et al. [19] on the simulated test set taken from [78]. Both these techniques tackle body measurement estimation under settings similar to ours, where inputs are silhouettes, pose is constrained, and body shape is highly variable. One difference is that our method and [78] directly incorporate height and weight inputs, while [19] infer height by assuming subjects are captured at a fixed distance. In Table 5, we evaluate our method with adversarial augmentation by ABS, and follow the testing protocols described in [78], comparing our method's performance directly with numbers reported in the respective references. Our method outperforms both alternatives in terms of overall mean error and 10 out of 14 individual measurements, often by a significant margin (up to 91% in the reduction of mean errors).

Table 6 compares our method with Yan et al. [90] on real human bodies in their *BodyFit* dataset. We train multi-view BMnet on *BodyFit* and adapt it to exclude height and weight inputs. Errors for their approach are taken directly from their paper. We outperform [90] on the majority of measurements, as well as the overall average error, demonstrating robustness of BMnet across different datasets.

Finally, for completeness, we compare our method with recent human mesh reconstruction (HMR) methods, SPIN [34], STRAPS [72], and Sengupta et al. [69] on BodyM Test-A. We compute body measurements from HMR by using the measuring function g on the predicted mesh. For fair comparison, we evaluate our technique with a single input and without height and weight metadata. Table 7 shows that our method reduces most errors by more than 35%. This error reduction is due to the fact that our network directly regresses measurements and that we have measurement supervision from the large training corpus in BodyM - we compare against methods with direct measurement prediction in Tables 5 and 6. Note in comparing Tables 5 and 7 that errors on real human data in BodyM are substantially higher than those on simulated data; demonstrating that BodyM provides a new challenging benchmark.

6. Conclusions

We propose BMnet to estimate body measurements from silhouettes, height and weight. The key contribution is a fully differentiable adversarial training scheme generating challenging bodies within the SMPL shape space, and revealing potential training gaps. When BMnet training is augmented with these adversarial shapes, measurement accuracy improves on real humans, producing new state-of-the-art results, particularly when real data is limited. We release BodyM, a new challenging large-scale dataset acquired with real human subjects to promote progress in body measurement research.

References

[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4845–4854, 2019. 3

[2] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. 3

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 2

[4] Alexandru O. Balan, Michael J. Black, Horst Haussecker, and Leonid Sigal. Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In *International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 2

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

[6] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 3

[7] Harkirat Singh Behl, Atilim Güneş Baydin, Ran Gal, Philip H. S. Torr, and Vibhav Vineet. Autosimulate: (quickly) learning synthetic data generation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 255–271.

[8] Binod Bhattarai, Seungryul Baek, Rumeysa Bodur, and Tae-Kyun Kim. Sampling strategies for gan synthetic data. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2303–2307, 2020. 3

[9] Benjamin Biggs, David Novotný, Sébastien Ehrhardt, Hanbyul Joo, Benjamin Graham, and Andrea Vedaldi. 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[10] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578. Springer, 2016. 2

[11] Jonathan Boisvert, Chang Shu, Stefanie Wuhrer, and Pengcheng Xi. Three-dimensional human shape inference from silhouettes: reconstruction and validation. *Machine Vision and Applications*, 24:145–157, 2013. 1

[12] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 3

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 5

[14] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *European Conference on Computer Vision (ECCV)*, volume 12352, pages 769–787, 2020. 1, 2

[15] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. 3DCrowdNet: 2D human pose-guided 3D crowd human pose and shape estimation in the wild. *arXiv preprint arXiv:2104.07300*, 2021.

[16] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

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

[18] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11875–11885, 2021. 2

[19] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *International Conference on 3D Vision (3DV)*, pages 108–117, 2016. 1, 2, 3, 8

[20] Endri Dibra, Himanshu Jain, Cengiz Oztireli, Remo Ziegler, and Markus Gross. Human shape from silhouettes using generative HKS descriptors and cross-modal neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[21] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3

[22] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. Revitalizing optimization for 3D human pose and shape estimation: A sparse constrained formulation. In *International Conference on Computer Vision (ICCV)*, pages 11457–11466, 2021. 1

[23] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 1

[24] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016. 3

[25] Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, S. M. Ali Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. In *International Conference on Machine Learning (ICML)*, 2018. 3

[26] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Inter-*

national Conference on Learning Representations (ICLR), 2015. 3

[27] Riza Alp Güler and Iasonas Kokkinos. Holopose: Holistic 3D human reconstruction in-the-wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10884–10894, 2019. 1, 2

[28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017. 3

[29] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52, 2021. 13

[30] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[32] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021.

[33] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021. 1, 2

[34] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 1, 2, 8

[35] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019. 1, 2

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

[37] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 2093–2102, 2018. 3

[38] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 2261–2268, 2019. 3

[39] Ren Li, Meng Zheng, Srikrishna Karanam, Terrence Chen, and Ziyan Wu. Everybody is unique: Towards unbiased human mesh recovery. *arXiv preprint arXiv:2107.06239*, 2021. 1

[40] Junbang Liang and Ming C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *International Conference on Computer Vision (ICCV)*, pages 4352–4362, 2019. 1, 2, 3

[41] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 2

[42] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *International Conference on Computer Vision (ICCV)*, pages 12939–12948, 2021. 2

[43] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3D reasoning. In *International Conference on Computer Vision (ICCV)*, pages 7708–7717, 2019. 3, 4

[44] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 13

[45] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1, 2, 3, 13

[46] Gilles Louppe, Joeri Hermans, and Kyle Cranmer. Adversarial variational optimization of non-differentiable simulators. In *International Conference on Artificial Intelligence and Statistics*, pages 1438–1447, 2019. 3

[47] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3, 5

[48] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 3

[49] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, pages 752–768, 2020. 2

[50] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, pages 484–494, 2018. 1, 2

[51] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. 2

[52] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Asia conference on computer and communications security*, pages 506–519. ACM, 2017. 3

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and

Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[54] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1485–1495, 2022. 1

[55] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 2

[56] Nicolas Pinto, James J DiCarlo, and David D Cox. Establishing good benchmarks and baselines for face recognition. In *Workshop on Faces In'Real-Life'Images: Detection, Alignment, and Recognition*, 2008. 3

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

[58] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision (ECCV)*, pages 19–37. Springer, 2020. 3

[59] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, pages 102–118. Springer, 2016. 3

[60] K. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 2, 3

[61] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[62] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision Workshops (ICCV-W)*, 2021. 1, 2

[63] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 3

[64] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge, 2018. 13

[65] Nataniel Ruiz, Adam Kortylewski, Weichao Qiu, Cihang Xie, Sarah Adel Bargal, Alan Yuille, and Stan Sclaroff. Simulated adversarial testing of face recognition models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4145–4155, 2022. 3

[66] Nataniel Ruiz, Samuel Schulter, and Manmohan Chandraker. Learning to simulate. In *International Conference on Learning Representations*, 2018. 3

[67] Nataniel Ruiz, Barry-John Theobald, Anurag Ranjan, Ahmed Hussein Abdelaziz, and Nicholas Apostoloff. Morphgan: One-shot face synthesis gan for detecting recognition bias. In *British Machine Vision Conference (BMVC)*, 2021. 3

[68] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 1, 2

[69] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 2, 8

[70] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16094–16104, 2021.

[71] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic estimation of 3D human shape and pose with a semantic local parametric model. In *British Machine Vision Conference (BMVC)*, page 419, 2021. 1, 2

[72] Akash Sengupta, Roberto Cipolla, and Ignas Budvytis. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 1, 2, 8

[73] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019. 13

[74] Yu Shen, Laura Zheng, Manli Shu, Weizi Li, Tom Goldstein, and Ming Lin. Gradient-free adversarial training against image corruption for learning-based steering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[75] Yu Shen, Laura Yu Zheng, Manli Shu, Weizi Li, Tom Goldstein, and Ming Lin. Improving generalization of learning-based steering using simulated adversarial examples. *https://arxiv.org/abs/2102.13262*, 2020. 3

[76] Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan Yuille. Identifying model weakness with adversarial examiner. In *Conference on Artificial Intelligence (AAAI)*, pages 11998–12006, 2020. 3

[77] Manli Shu, Yu Shen, Ming Lin, and Tom Goldstein. Adversarial differentiable data augmentation for autonomous systems. In *International Conference on Robotics and Automation*, 2021. 3

[78] Brandon M. Smith, Visesh Chari, Amit Agrawal, James M. Rehg, and Ram Sever. Towards accurate 3D human body reconstruction from silhouettes. *International Conference on 3D Vision (3DV)*, pages 279–288, 2019. 1, 2, 8

[79] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and accurate scans from an image in less than a second. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[80] Stephan Streuber, M. Alejandra Quiros-Ramirez,

Matthew Q. Hill, Carina A. Hahn, Silvia Zuffi, Alice O'Toole, and Michael J. Black. Body talk: Crowdshaping realistic 3D avatars with words. *Transactions on Graphics (TOG)*, 35(4), 2016. 1

[81] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 3

[82] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019. 3

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

[84] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3Dhuman shape and pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 13033–13042, 2021. 1

[85] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding (CVIU)*, 210:103225, 2021. 2

[86] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13211–13220, 2022. 1

[87] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10974, 2019. 2

[88] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: generative 3D human shape and articulated pose models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[89] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, 2012. 13

[90] S. Yan, J. Wirta, and J.-K. Kämäräinen. Silhouette body measurement benchmarks. In *International Conference on Pattern Recognition (ICPR)*, Milano, Italy (online), 2020. 2, 3, 8

[91] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *European Conference on Computer Vision (ECCV)*, pages 775–791. Springer, 2020. 3

[92] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11038–11049, 2022. 1

[93] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14484–14493, 2021. 1, 2

[94] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. THUNDR: Transformer-based 3D human reconstruction with markers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12971–12980, 2021. 2

[95] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4302–4311, 2019. 3

[96] 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 *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021. 1

A. Appendix

A.1. Body Shape Analysis

We include additional analysis of body shapes generated by ABS vs. random sampling. Fig. 7 plots weight vs. height for adversarial and random sampling. Consistent with analysis in Sec. 5.1 of the paper, we note that adversarial bodies are more densely concentrated in regions of greater weight and height. Fig. 8 plots the 3rd vs 4th SMPL [45] shape components. Unlike the first two components, these two variables do not distinctly separate ABS from random sampling, indicating that rare body shapes manifest themselves more strongly along some shape components than others.

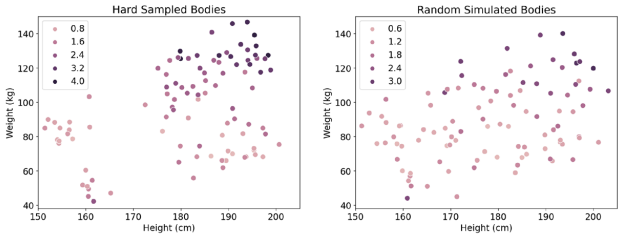


Figure 7: Distribution of adversarial (left) vs. random (right) sampling in terms of weight vs. height. Adversarial bodies are more concentrated in regions of higher weight and height.

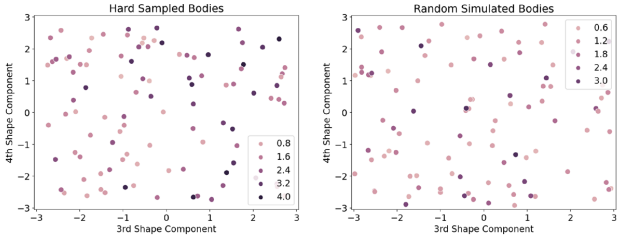


Figure 8: Distribution of adversarial (left) vs. random (right) sampling in terms of 4th vs 3rd principal component (β) of SMPL shape. Adversarial and random sampling are not clearly distinct along these two dimensions.

A.2. Additional Experimental Evaluation

We report additional results comparing our adversarial augmentation with an augmentation-free baseline for training BMnet. In Table 8 we show results on the minimal-clothing TestB subset. The findings are consistent: adversarial augmentation improves results in terms of both overall and individual metrics. In Table 9 we show mean errors for the individual body measurements from TestA. Adversarial sampling improves accuracy for 11 out of 14 measurements. The same analysis is performed on the reduced training scenario on TestA ($\sim$ 10% of the full training set) in the main manuscript. The gains from adversarial augmentation are even stronger in this scenario.

A.3. Limitations and Future Work

Similar to other adversarial training techniques, our method incurs a, small but real (order of 1%), computational overhead to achieve improved accuracy. Techniques such as "Adversarial Training for Free!" [73] may be explored to reduce training time and data storage. Our adversarial synthesizer currently does not account for environmental variations that affect the input silhouettes, such as camera characteristics, human pose variations, and segmentation noise. Adversarial sampling incorporating these dimensions is a fruitful future investigation. While some methods perform test-time optimization [29], the focus is usually on pose optimization rather than shape. Further improvement of all comparative methods in our work using test time optimization is interesting, but beyond the scope of this work.

A.4. Societal Impact

1. Our system predicts intimate attributes about a person (i.e. body measurements) from photo silhouettes. These attributes are considered confidential, as they can be linked to one's health, personal lifestyle, and choices. It is therefore important that such a pipeline is protected from access by unqualified authorities who could generate and misuse confidential body information. 2. Computer vision research in the fashion domain has been supported by datasets that are heavily biased to thin and tall body shapes. This is owed to the preponderance of photos of models and celebrities from which these datasets are sourced [44, 64, 89]. Consequently, networks that estimate body shape and measurements, and generate body avatars for virtual try-on experiences, tend to produce larger errors for body shapes that deviate from societal beauty standards. Our research aims to increase inclusivity in body shape by discovering body shapes that are rare with respect to available datasets. However, a purely computational approach to countering dataset bias may also introduce other unfavorable biases; hence it is important to check for alignment between the body shape distributions generated by our method and realistic shape distributions in a given demographic. In our work, we attempt to address this issue by performing adversarial perturbations around real body shapes in BodyM.

A.5. Ethics Statement

While we present research and datasets on human body measurement estimation, we take all precautions to respect the privacy of all individuals who have contributed to our data and research. Our collected human body dataset comprises silhouettes, height, weight and body measurements which do not reveal subject identity. The outputs of our

	Overall			Chest	Hip	Leg Length	Waist
	TP90	TP75	TP50	MAE	MAE	MAE	MAE
Single-View (No Aug.)	23.32	15.43	8.74	22.74	16.64	13.72	20.88
Single-View (Adv. Aug.)	**23.24**	**15.43**	**8.55**	**22.67**	**16.41**	**13.58**	**20.78**

Table 8: Comparison of No Augmentation (No Aug.) versus Adversarial Augmentation for BMnet on the minimal clothing subdivision of TestB (errors in mm).

	Ours Adv. Aug.	Ours No Aug.
Ankle	5.56	**5.48**
Arm Length	**7.07**	7.26
Bicep	**6.36**	6.50
Calf	**7.89**	7.95
Chest	**18.84**	19.18
Forearm	**5.18**	5.36
Head-to-Heel	**8.89**	9.11
Hip	**11.34**	11.53
Leg-Length	**11.24**	11.39
Shoulder-Breadth	6.05	**5.95**
Shoulder-to-Crotch	8.90	**8.85**
Thigh	**11.15**	11.16
Waist	**15.78**	16.12
Wrist	**4.31**	4.35
Mean Error	**9.19**	9.30

Table 9: Mean measurement error comparison of Adversarial Augmentation (Adv. Aug.) versus No Augmentation (No Aug.) for training BMnet on TestA (errors in mm).

adversarial body simulator are synthetic. All subjects have given written consent for the capture and release of the data.

A.6. Reproducibility

The BodyM dataset is publicly available at https://adversarialbodysim.github.io to enable reproducibility of our method and further research in this area.