Learning Clinical Body Composition Metrics from 2D and 3D Optical Imaging

Isaac Yuheng Tian

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Brian Curless, Chair

John A. Shepherd

Linda Shapiro

Program Authorized to Offer Degree:

Computer Science & Engineering

University of Washington

Abstract

Learning Clinical Body Composition Metrics from 2D and 3D Optical Imaging

Isaac Y. Tian

Chair of the Supervisory Committee:
Brian Curless
Computer Science and Engineering

Accurate human body shape representation has many applications within computer graphics including 3D animation, virtual tailoring, ergonomic engineering, and virtual reality reconstruction. For clinical researchers working at the intersection of computer graphics, machine learning, and obesity-related epidemiology, computational modeling of human body shape presents a novel and accessible pathway to quantifying, classifying, and monitoring risk factors associated with premature mortality caused by metabolic syndrome. Total and regional body composition and body shape are strongly correlated with progression of metabolic syndrome as well as degenerative conditions such as sarcopenia and osteoporosis. Estimates of body composition from optically measured human body geometry are cheap, safe, and non-invasive relative to current reference methods that require exposure to ionizing radiation.

In this thesis, I present a body of work that thoroughly investigates predicting body composition from optical images of human body shape with clinically significant precision and accuracy. This thesis contributes the following:

1. Introduces a model that predicts 3D body shape and total and regional body composition metrics from monocular 2D images.

2. Develops a method that automatically standardizes 3D human body scans to watertight manifold mesh templates with consistent topology and anatomical correspondence and demonstrates the viability of this tool for constructing new shape and regression models that predict body composition with agnosticism towards input scanning devices.
3. Extends the automatic mesh templating method to create the first autoencoded shape model for a pediatric cohort paired with body composition prediction from shape parameters derived with unsupervised learning.
4. Performs a systematic review of deep 3D shape autoencoders for total human body geometry with the goal of identifying the current state of the art methods and architectures in reconstruction accuracy while also suggesting standards and best practices for future work in this research field.
5. Performs a novel estimation of body composition from nonlinear features extracted by a deep autoencoder with nonlinear Gaussian process regression and comprehensively compares marginal contributions of linear and nonlinear shape and regression algorithms against the linear baselines of prior works with systematic ablation studies.

Dedication

To my family, for their support and inspiration

To my friends, for their joy and companionship

To all my former teachers, for their guidance and motivation

To my advisors, for their patience and mentorship

To Sarah, for your love and grace

TABLE OF CONTENTS

1. Introduction

1.1 Motivation

Metabolic syndrome is a grouping of related health conditions associated with increased risk of heart attack, stroke, and other potentially fatal diseases such as cancer and diabetes [51]. Conditions such as high blood sugar, abnormal cholesterol levels, and high blood pressure contribute to increased risk of premature mortality. Excess body fat, especially when disproportionately distributed around the waist, is a symptom of metabolic syndrome that is quantified by total and regional body composition analysis [10]. Clinical measurements of body composition range from simple approximations such as waist-to-hip-ratio to detailed internal anatomy visualization and measurement with radiological imaging.

Many clinical studies have shown the importance of regional body composition as a predictor for metabolic disease risk and increased mortality even when controlling for total body variables such as weight and body mass index (BMI). A criterion method for body composition assessment is Dual-Energy X-ray absorptiometry (DXA), an imaging technique that is currently considered the gold standard for measurement of total and regional body composition in clinical trials and research studies because of its precision and accuracy. However, DXA is only available in specialized clinics and its use of ionizing radiation limits repetitive imaging.

The importance of body composition monitoring coupled with its high cost and low accessibility suggest a need for methods that can easily be used without access to a controlled clinical environment with cost prohibitive equipment and expertise to monitor the status of and changes in total and regional body composition compartments. Ideally, this technology would be

affordable to middle- and low-income individuals, who are the populations most likely to be adversely affected by high costs and low access due to the increased risk of metabolic disease among lower socioeconomic brackets, and accessible through hardware that is widely distributed and commonly available outside of specialized clinics. Such a method would allow for measurement of body composition "in the wild" and would enable the outsourcing of body composition tracking from the professional clinic to the domestic household. This large-scale broadening of accessibility to monitoring clinically important body metrics can enable participation in self-monitoring and population health data analysis at previously infeasible scales. Although the criterion methods for measuring body composition and metabolic risk factors reside mainly in radiology facilities, many mortality predicting body features have visually observable external effects on body shape [81,83,135]. Body shape as a predictive signal for disease and mortality risk management provides researchers and clinicians with a non-invasive, low cost alternative for assessing metabolic health. Surrogate models for radiological imaging can be trained using external body shape data captured with optical imaging and training transfer functions to reference variables measured by criterion clinical imaging methods.

1.2 Optical Imaging

Our work on this problem correlates body composition with optical imaging. Optical imaging is the class of image capture that measures the visible part of the electromagnetic spectrum to assemble its picture. This includes 2D images from standard digital photography and 3D geometry captured with structured light scanning. We can transform or simplify the captured

image data to make it compatible with specific analysis architectures, such as extracting a silhouette or fitting a standard template to 3D mesh.

Figure 1.1 Three 3D optical (3DO) scanners used for data collection in this thesis. From back to front: SizeStream SS20, Fit3D Proscanner 4.x, Styku S100.

Optical imaging is relatively cheap compared to medical radiology. 2D photography is ubiquitous with the widespread distribution of smartphones across a wide range of socioeconomic status. Even 3D optical (3DO) scanners such as Fit3D [117] are an order of magnitude cheaper than MRI or DXA alternatives. Optical imaging devices do not require certified and trained technicians to operate and do not pose the risk of radioactive injury. The accessibility and affordability of optical imaging make it an attractive alternative to radiology if meaningful medical outcomes can be predicted from their data. However, in exchange these imaging devices trade away the ability to image the internal structure of the body, as optical

imaging is reflective and non-penetrating. Models that predict those variables from optical data trained on associated radiological measurements must be created to bridge the information gap.

The primary focus of my work, in collaboration with medical researchers, is in modeling the relationship between optical images and gold-standard measurements derived from DXA. The works contained in this thesis were trained and tested on data from the Shape Up! Adults (NIH R01 DK109008) and Shape Up! Kids (NIH R01 DK111698) datasets. These two clinical trials recruited a cross-sectional sample of the United States population spanning the range of age, ethnicity, and body mass index (BMI) from three locales: San Francisco, CA, Honolulu, HI, and Baton Rouge, LA. Participants were scanned on three separate 3DO systems (Fit3D Proscanner 4.x, Fit3D Inc, Redwood City, CA, USA, Styku S100 4.1, Styku LLC, Los Angeles, CA, USA, and Size Stream SS20, Size Stream, Cary, NC, USA) shown in Fig. 1.1, were photographed with front and side view 2D digital photos, and were imaged with full-body DXA scans to record reference body composition metrics. This dataset collection gave us the novel ability to correlate 2D and 3D optical images with body composition variables measured with a clinically accepted gold-standard reference method (DXA). Body composition was first mapped to PCA shape parameters derived from 3DO scans in Ng et al. [83] using the templated fitting algorithm of Allen et al. [5] Although the scale of data acquisition of the Shape Up! Studies was novel and unprecedented, the number of individuals that were recruited over a 5-year study was only around 1000 total. Balancing model complexity to sufficiently express the relationship between optical image data and body composition while staying sparse enough to train on the limited amount of paired data is a fundamental constraint of our projects. Our initial efforts in body shape modeling and body composition prediction from shape models were built on linear algorithms such as principal component analysis (PCA) and ordinary least squares (OLS)

regression. These linear methods were effective for unsupervised shape model parameterization and body composition regression with the available dataset sizes.

1.3 Linear Modeling

Linear models can have relatively small parameter counts and train well with limited datasets. These are the first class of techniques we attempt on a new problem, dataset, or data transformation as they are conservative and easy to optimize. Their inherent advantages also double as their fundamental weakness, as a linear relationship often oversimplifies the true underlying function between inputs and outputs. In this thesis, the full pipeline for prediction of body composition targets from optical imaging is divided into two components: shape encoding, which converts an optical image into a shape feature vector, and body composition regression, which maps shape features to reference metrics determined by DXA. Both the shape encoding and the regression mapping modules can be trained with linear or nonlinear algorithms.

A linear shape model for optical images of human bodies assumes the variation in the observed dataset can be explained with a linear combination of descriptive features. Recent work on human shape modeling relied extensively on PCA where the linear shape features are determined via unsupervised learning on the mesh vertex distribution for a set of fixed topology 3D meshes, as opposed to *a priori* fixed assignments to body circumferences or anthropometric measurements [5]. An example of a PCA-based shape model is shown in Fig. 1.2. A linear regression model assumes a linear mapping function between shape features and continuous body composition target variables and estimates the function parameters via well-established algorithms such as ordinary least squares (OLS) or stepwise regression.

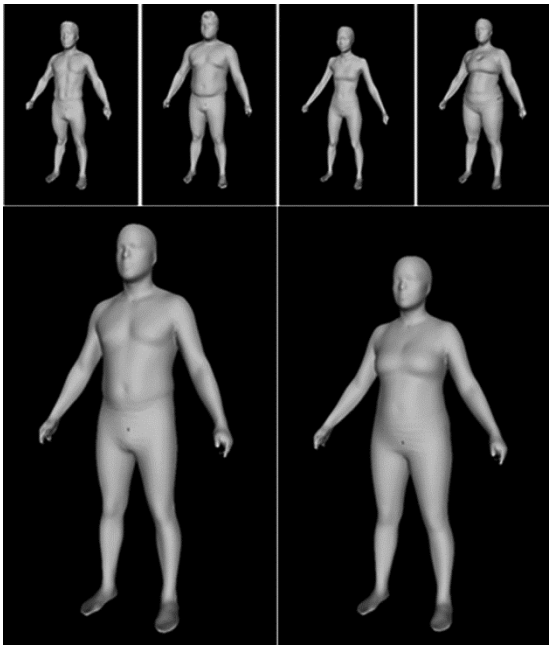


Figure 1.2 Top row shows two bodies of each sex sampled with different parameters **w**. Bottom row shows the mean shape **μ**, such that a new shape **s** = **Aw** + **μ** for some orthogonal PCA matrix **A**.

Much of the work in this thesis achieved increasingly accurate state-of-the-art results on body composition prediction using fully linear pipelines where both the shape model and the regression model were trained with linear algorithms. In Chapter 8 we show that hybrid linear-nonlinear and fully nonlinear shape model-regression pipelines can improve upon the best results established by all prior works using fully linear model pipelines.

Figure 1.3. PCA parameterization of a body shape (left) compared to the actual scan (right) with 457 components used for projection.

1.4 Nonlinear Modeling

A limitation of linear methods is the inability to fit nonlinear mapping functions between shape and body composition, as shown in Fig. 1.3. Nonlinear methods may better represent the observed variation in optical image data as well as the mapping to body composition metrics but come at the cost of more difficult training and optimization. Our datasets consisting of a few hundred examples of associated optical image and DXA body composition measurements are orders of magnitude smaller than standard datasets used for nonlinear feature extraction models such as deep convolutional neural nets.

Nonlinear models include all methods that do not simplify assumptions about the relationship between shape features and the body composition response variables to strictly linear functions. The most general of these methods include deep network methods, which generate features from optical image input and can map these features to body composition nonlinearly with multi-layer perceptrons (MLP). By the Universal Approximation Theorem,

these methods can in theory approximate any relationship provided enough training data. Our dataset is two orders of magnitude smaller than what is typically used to train such networks. We can circumvent this problem by constraining the model from a general function to a specific parameterized nonlinear function, or by leveraging transfer learning from a network trained on similar data such as MRI and CT images. Nonparametric regressions such as Gaussian Process Regression (GPR) sits between generalized deep NLPs and parameterized nonlinear methods such as kernel regressions, instead opting to probabilistically estimate the distribution of the true regression function conditioned on observed training data and a function shape specified by a distance function (also known as a kernel). We achieved higher precision and accuracy using GPR than prior work with linear methods or experimental trials with deep MLPs in Chapter 8.

We constrained the optical image estimation problem by subdividing it into two independently trained modules, a shape feature extractor and a feature-to-body composition regressor. An end-to-end prediction model from 2D or 3D optical imagery was attempted in Chapter 4.6.2 and again in Chapter 8. These experiments did not exceed the performance of methods that performed feature extraction and body composition regression separately. We leave investigations of fully end-to-end networks relating optical imagery of human body shape to body composition to future work.

1.5 Thesis and Impact

In this dissertation, we applied algorithms and methods from computer graphics, computer vision, and machine learning to investigate parameterized shape modeling of 3D human bodies for the purpose of body composition regression. We built our models on a clinically stratified cross-sectional sample of the US population (Shape Up! Adults,

NCT03637855 & Shape Up! Kids, NCT03706612) containing over 1,000 combined pairs of optical imagery (both 2D and 3D) and reference DXA measurements. Throughout the course of our work, we implemented both linear and nonlinear feature extraction methods such ranging from PCA to deep 3D graph convolutional networks. We tested a wide array of regression methods for deriving body composition predictions from extracted feature spaces, starting with simple linear regression, and moving to continuous stepwise regressions (Least Angle Regression), nonparameterized probabilistic regressions (Gaussian Process Regression), and deep multilayer perceptrons.

We summarize the related work in the field of 3D human shape modeling and body composition prediction from optical images in Chapter 2. We discuss a preliminary project for capturing 3DO human body scans using photometric stereo in Chapter 3. We chose to operate commercial optical scanners such as Fit3D in place of photometric stereo for usability and practicality reasons in later work; however, the method described in Chapter 3 may potentially yield better 3DO shape captures if pursued to a more advanced conclusion. We developed methods for predicting body composition from monocular 2D images in Chapter 4. Although similar work on body shape and pose estimation from monocular images existed in the field of computer vision, none were targeted with clinically relevant estimations and applications in design. In Chapter 5, we developed a method for automatically fitting a topologically consistent 3D manifold mesh to raw 3DO scanner inputs and developed new regression models that, unlike previous work, were agnostic to the scanning device used for data collection and exhibited robustness to slight pose variations. This work enabled rapid standardization of hundreds of additional unordered point clouds captured by 3DO scanners into manifold templates with identical topology, a process that previously took manually guided labor by a trained technician

that scaled linearly with the number of collected scans. Topologically consistent manifold meshes were essential for feature extraction and shape model construction, which were in turn instrumental for building accurate regression models to body composition. This automated pipeline enabled the creation of previously unavailable quantities and diversity of 3D training data for later work in Chapter 6, which investigated 3DO body composition prediction of pediatric populations aged 5-17.

Chapter 7 conducts a systematic review of the computer vision and machine learning literature for deep shape modeling of 3D human bodies using autoencoder methods. We picked a method from this review with favorable characteristics that covered perceived shortcomings in linear PCA shape modeling and tested its accuracy in 3D shape reconstruction and body composition prediction from extracted features in the following chapter. Chapter 8 used the combined auto-templated 3DO data generated by the method of Chapter 5 to create the largest and most complete multiple identity, consistently posed ensemble dataset with topologically consistent mesh fits to date. The large and diversified 3DO ensemble dataset in Chapter 8 was used in conjunction with a 40,000-member 4D scan database (DFAUST) [13] containing multiple continuously captured dynamic poses of 10 individuals to train deep feature extractors with the 3D convolutional autoencoder identified in Chapter 7. The dataset and its resulting features were used to comprehensively compare linear and nonlinear implementations of both 3D shape models and body composition regression methods in a controlled environment to determine the isolated effect of each permutation on the precision and accuracy of body composition estimation from 3D optical scans.

Our work provides future researchers working on clinical applications of 3D computer graphics and 2D computer vision algorithms with practical, portable, and usable models,

workflows, and methods for advancing the state of the art in body composition estimation from optical imagery. Our auto-templating method enabled the execution of multiple subsequent studies with data scales that would have been impractical for manual processing. Future adoption of our work could expedite topological standardization of future optical data collection trials while also providing validated pre-trained model initializations with cross-compatible data formats enforced by the templating algorithm. Finally, we identify multiple avenues of unresolved experimental directions that future work can expand upon, accelerated by the contributions present in both our experimental results and our computational models.

2. Related Work

Prior work on learning total and regional body composition from optical imaging is sparse due to the expense and difficulty of collecting paired optical and radiological training data for model learning. The literature on learning representations of optical images of human bodies absent medical considerations is much more developed. To bridge the gap between optical image learning and target medical variable prediction, we can apply models and techniques published on independently developed 3DO datasets without body composition targets to paired optical image datasets and learn transfer functions between optical image features and body composition. In this chapter, we give an overview of works either directly performing body composition prediction from optical imaging or relevant works that introduce methods and techniques that can translate to subtasks of the prediction pipeline. We will do an in-depth analysis of deep 3D autoencoders of human body shape in Chapter 7.

2.1 End-to-End Optical Image Body Composition Prediction

While models for learning non-medical metrics such as age from optical images are plentiful [53], predicting outcomes like body fat mass, percentage, or distribution is an underexplored problem due to the difficulty of collecting ground truth measurements. Many existing methods learn transfer functions to body composition from a selected set of features [28] often based on anthropometrics. Anthropometric features are not desirable target inputs as it is labor and expertise intensive to measure tens of circumferences in order to build a feature vector. Other techniques such as Affuso et al. [2] and Farina et al. [29] start with an optical image as input, but only use it to automatically extract estimates of anthropometric features. While these

methods automate the collection of anthropometric features, they risk oversimplifying the input image by reducing a fundamentally 3D structure to a sequence of 1D scalar measurements. Using anthropometric estimates as an intermediate feature for body composition prediction is analogous to using features derived from statistical parameterization as PCA or latent representation derived from an autoencoder. However, there is risk of introducing bias and error when choosing anthropometrics as an intermediate feature space. The choice of which subset of defined measurements to use and the precision and accuracy of anthropometric measurement affect the descriptive accuracy of the intermediate feature vector. It is also likely that the reported results of [2] and [29] are overfit as the published results are training accuracies and not held out data. For our work, we treat anthropometric prediction as a separate problem that is decoupled from body composition and prefer intermediate feature representations that are empirically calculated.

Multiple papers from the James Hanh Lab pursued parallel objectives to our study with different problem formulations. Lu & Zhao et al. [71] published a methodology for calibrating commodity depth sensors (Kinects) to capture 3DO images of human bodies, which was used in their subsequent work in place of commercial systems. In Lu & McQuade et al. [70], features like generated circumferences and their curvatures were defined on 50 irregular meshes of adult males and regressed to body fat percentage. This method performed comparably to our work in Chapter 5 but on a much more restricted data set. Lu, Hahn, & Zhang [68] predicted fat distribution heatmaps based on DXA scans by adding the nearest neighbor residual heatmap from the training set to a baseline test prediction. The baseline was generated with a single variable regression from body density to the first principal component of normalized DXA heatmaps. As density was measured with a BOD POD, this method was not a purely optical one.

In Wang & Lu et al. [125] the authors implemented a method more similar to their previous work in [70] and regressed visceral adipose tissue (VAT) from extracted features of a 3D scan. They performed dimensionality reduction on their features with functional PCA and predicted VAT with gaussian process regression (GPR). GPR is of particular interest to our future efforts as it provides a probabilistic model for estimating a nonlinear regression function without having to specify it a priori.

In Wang & Xue et al. [126], the authors trained a conditional generative adversarial network (cGAN) with 5-fold cross validation on 270 3D CT scans. This was an advancement on the method in [68] and sought to predict the visceral and subcutaneous adipose tissue distribution as 2D heatmaps from projected 2D body silhouette using CT slices as ground truth. Visceral and subcutaneous fat deposits had to be segmented manually. GANs may be a useful advancement in constructing parameterized shape, but predicting pixel-wise body composition maps is beyond the scope of this thesis.

Chhatkuli et al. [21] described a neural network that predicts percent body fat from optical front and side images associated with height, weight, gender, and age. This was a very similar construction to our work in section 5.1. They only reported validation results rather than test results, and the coefficient of determination (R2) and root mean squared error (RMSE) were worse than our equivalent models. Their ground truth measurements were taken with an InBody bioelectrical impedance scale instead of DXA.

Klarvquist et al. [56] developed a 2-view silhouette model based on Densenet-121 for compartmental fat prediction. They converted over 40,000 whole body MRI scans from the UK Biobank cohort into binary silhouette masks and trained the network to predict MRI derived values for visceral fat (VAT), abdominal subcutaneous fat (ASAT), and gluteofemoral fat

(GFAT). Relative to our best results, they achieved very high (0.885) R2 values for VAT, which was the only directly comparable metric we studied, but this value could be inflated due to the high sample count. All 40,000 images were fed through a network corresponding to its held-out fold during 5-fold cross validation training. The authors acknowledged that their method is only practical if it can be transferred to a cheaper imaging acquisition format such as 2D photography.

2.2 Feature Extraction from Optical Images of Human Bodies

Many works addressing feature extraction from 2D or 3D human images exist without consideration of target variable prediction for medical or clinical applications. It is possible methods from these previously unrelated studies could be adapted for predicting medical targets; feature extraction methods applied to an optical image dataset with paired medical imaging reference data can provide intermediate results from which to further train modular regression functions that are agnostic to the feature selection method. Another advantage of unsupervised feature extraction on optical images is the potential to discover more informative features than predetermined measurements such as anthropometric circumferences. Pathak et al. [85] trained a 3-layer neural network on the demographic and anthropometric data of the 1999-2006 NHANES DXA scans and achieved R^2 values greater than 0.92 for lean and fat mass prediction, but we were able to exceed 0.95 R^2 using unsupervised features in Chapter 8. A comprehensive review of shape encoding and feature extraction of total 3D human body shape is provided in Chapter 7.

Feature extraction in early work on human body shape was done with shape models parameterized by PCA, which is a linear autoencoder [5]. A conservative advancement on this

implementation is a neural network autoencoder such as non-linear PCA (NLPCA) which is described in Scholz et al. [105] as a 3-4-2-4-3 structured MLP. For a 3DO mesh, the 3 input and output channels could be interpreted as the vectors containing all x, y, and z coordinates of the mesh vertices. These methods treat meshes as a point cloud and are blind to the connectivity of mesh triangles. Autoencoders based on graph convolution such as Zhou et al. [146], Bouritsas et al. [15], or Hanocka et al. [46]. are more tailored for learning on meshes and exploit the additional information embedded in the edges. However, increasing the complexity of the input domain may also inflate the size of the network and the amount of data required for training.

Data availability is an ongoing challenge in past and present work on human body shape modeling and feature extraction. The training data from the Shape Up! Studies currently contain 742 unique individuals between the Adults and Kids datasets. The total mesh count is 2146 total scans representing two captures across three scanning devices, accounting for dropouts due to quality control exclusions. Human shape datasets prior to the Shape Up! clinical trials did not contain reference body composition measurements to serve as reference regression targets. Among those datasets, diversity of identity was often low; diversity of posing was the variable of choice used to augment the size of the dataset due to the relative difficulty of recruiting hundreds or thousands of unique participants. As a result, most prior works on 3D mesh feature extraction favored training and testing on multi-pose datasets rather than multi-identity data. Bouritsas et al. was trained on over 40k scans from the DFAUST [13] dataset, which is a multi-pose dataset that only contains 10 unique people. A few works leveraged existing multi-identity data to add more diversity of individuals into their model. Jiang et al. [54] trained an autoencoder that disentangles pose and shape as part of its decoder, using the 4,300-member CAESAR [97] dataset along with some multi-pose data to encode and train on about 5000 people. The authors transformed their

meshes into an ACAP (as-consistent-as-possible) feature from Gao et al. [36] to use as input into the encoder. This conversion was claimed to be more efficient for derivative computation and reconstruction in the autoencoder process. Datasets like Shape Up! stand to benefit little from the pose separation due to the largely standardized posing of the 3DO meshes. However, the ACAP feature parameterization is an example of a method that may be translatable to paired medical data such as Shape Up!, enabling us to learn more accurate and representative unsupervised features for downstream regression model inputs.

We discovered in Chapter 7 that feature extracting shape models such as Jiang et al. trained on previously available multi-identity data used a PCA derivative of the original CAESAR dataset published by Pishchulin et al. [90]. This rendition of the CAESAR data was low resolution and not identity preserving relative to the original detail of the 3DO scans. Thus, the learned latent representations of those works may be artificially restricted to a linear subspace despite training with a deep network model. Work presented in the following chapters address these limitations by assembling a high-resolution ensemble of all existing multi-identity data including Shape Up! and CAESAR and training feature extraction models using both PCA and deep 3D autoencoder networks.

2.3 Initial Work on Body Composition Prediction from 3DO features

Assuming we have a feature vector representation of a human body captured in an optical image, predicting a body composition value from this is a generic regression problem from a d x 1 dimensional vector. A wide array of regression methods both linear and nonlinear published in isolation of any consideration for body composition or human body shape learning can be applied to this generic problem after features have been defined. MLPs are the most general and

least conservative prediction method and require relatively large data volumes to train. In this case, there were 20,137 subjects. Nonlinear regression such as Gaussian Process Regression (GPR) used in [125] may be more stable on smaller training sets due to implicit regularizations to a prior distribution. Early work in body composition prediction from 3DO features has been based largely on linear models for both shape (PCA) and regression (linear ordinary least squares OLS).

Ng et al. (2016) [81] and Wong et al. (2019) [135] published initial investigations into prediction body composition from 3DO scans of human bodies for Shape Up! Adults and Shape Up! Kids data respectively. Unsupervised feature extraction was not used in these early works; regression models used predetermined features such as hip and waist circumferences derived from the 3D scan as input.

Ng et al. (2019) [83] published the first study training regression models from linear shape parameterizations of 3D human scans on the Shape Up! Adults dataset. PCA was used as the shape parameterization and stepwise regression was used for the regression model. The first 11 principal components (PCs) captured 95% of the shape variance in the dataset; regression models were trained on a mixture of anthropometric variables and the first 11 PCs. The 95% variance cutoff was chosen to enforce a sparse parameterization space in order to reduce overfitting. However, we show in our work in Chapter 8 that including a maximum number of shape parameters tended to benefit prediction accuracy and precision on test data without adverse effects on overfitting. We performed an equivalent study covering body composition from PCA features on Shape Up Kids! in Chapter 6.

Wong et al. (2021) [136] showed factoring out pose variations in the Shape Up! Adults dataset by registering 3DO scans to a skinned template and reposing them to a T-pose reduced

root-mean-square error (RMSE) in body composition prediction by up to 50% in males and around 30% in females. We complemented this work in Chapters 5 and 8 by building models incorporating more diverse pose variation into the training data and learning pose agnosticism with respect to body composition through data variation. Future work explicitly disentangling pose parameters from shape variation may produce models with even greater accuracy.

3. Full Body 3D Scanning and Reconstruction with Photometric Stereo

Our first project in 3DO shape modeling and body composition regression of human bodies aimed to improve hardware performance of 3D body scanners. Most 3D scanning systems, both then and now, incur substantial noise or distortion during the capture process. Some of this noise was caused by minute movements including breathing during the capture process, which could take up to a minute. A system that could capture high resolution detail of human shapes while also operating at very low latency could overcome the limitations of time-of-flight and structured light scanning systems. To explore this, we built a photometric stereo [137] rig for capturing multiple images of a human body in the same pose and camera orientation under controlled lighting conditions.

3.1 Methods

3.1.1 Photometric Stereo Formulation under Linear and Nonlinear Assumptions

In calibrated photometric stereo, the same still scene is captured under multiple known lighting conditions. The orientation of the surface at a pixel location (r, c) can be determined by solving a system of equations modeling the pixel intensity as a function of the normal vector $n(r, c)$ given at least three observations with varying known lighting for the three unknowns n_x, n_y, n_z. We modeled the reflectance function as the Phong [88] shading model for simplicity, shown in Fig. 3.1 as $I(z, \rho, \gamma)$ for depth z, diffuse albedo ρ and specular exponent γ.

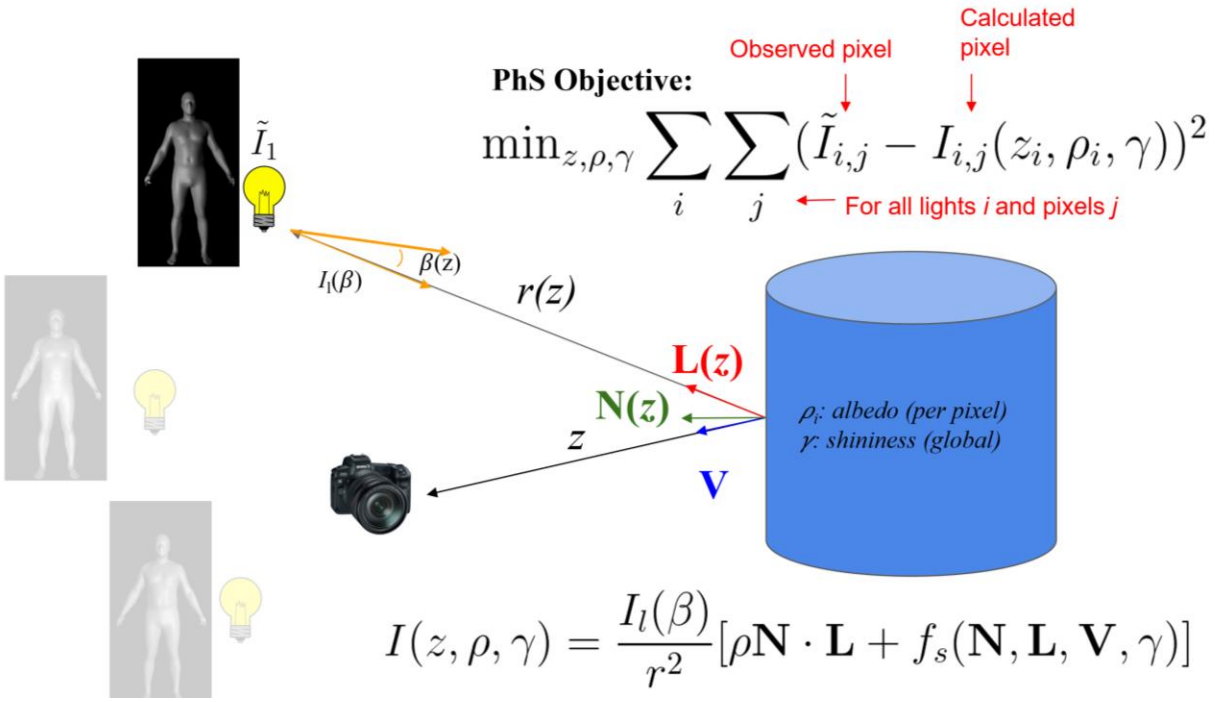


Figure 3.1. Diagram of photometric stereo with a single camera position and three known light positions. $\frac{I_l(\beta)}{r^2}$ is the intensity falloff attenuated by distance *r* and light angle β. The numerator is constant for ideal point lights. $\rho\mathbf{N} \cdot \mathbf{L}$ is the diffuse (Lambertian) term for albedo ρ and $f_s(N, L, V, \gamma)$ is the specular reflection term.

Under linear and orthographic assumptions, the normals $n(r, c)$ can be solved with a linear system of equations given *i* observations with known lighting direction and reflectance parameters. The depth $z(r, c)$ at each pixel can then be determined from $n(r, c)$ with the linear system of equations modeling normal vectors as forward differences between pixels as shown in Fig 3.2:

$$n_x + n_z(z_{x+1} - z_{x,y}) = 0$$

$$n_y + n_z(z_{x,y+1} - z_{x,y}) = 0$$

With special consideration at the boundaries:

$$-n_x + n_z\left(z_{x-1} - z_{x,y}\right) = 0$$

$$-n_y + n_z\left(z_{x,y-1} - z_{x,y}\right) = 0$$

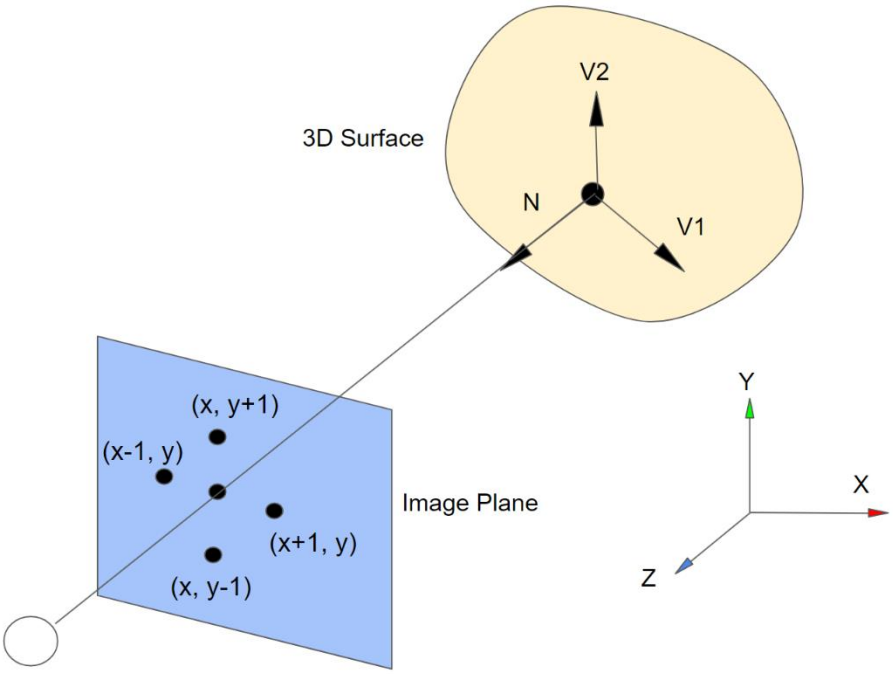


Figure 3.2 Photometric stereo depth integration diagram.

In practice, this solution creates artificially flattened and distorted images due to a lack of accounting for perspective projection. To account for perspective distortion, we transformed all positions and normal vectors by a perspective projection factor with known camera focal length *f* and distance from camera *d.* The resulting nonlinear constraints were optimized in Ceres solver. This also allowed us to treat nonlinear parameters such as the specular exponent as variables; in practice we set this to a constant 5 to speed up convergence.

3.1.2 Photography and Lighting Hardware

Photometric stereo typically assumes ideal point light behavior from all light sources with no global or indirect illumination contaminating the scene. Brighter lights also increase the dynamic range of the scene, providing better signal-to-noise ratio and enabling greater depth resolution in photometric stereo. In small scenes where geometry is measured on the scale of centimeters, small light emitting diodes (LEDs) approximate point light behavior with sufficient brightness. Full body human capture requires light sources to illuminate a subject up to two meters tall and up to three meters away. We acquired five Bolt VB-22 bare bulb remote flash devices each with 360W output to maximize the illumination of the target subject. Each remote flash was controlled with a PocketWizard Plus IV radio shutter transceiver.

The bulbs in the flash units had coiled emitters and produced randomized refraction artifacts due to interactions with the encasing glass (Fig. 3.3). This violated the isotropic point light assumption of our model. We mounted a spherical diffuser to the bulb, created with a manufacturer designed reflective metal snoot glued to a spherical or hemispherical filter.

Figure 3.3. Bolt VB-22 flash bulb. The coiled emitter and nonuniform glass refraction nullify the assumption of a point light model in photometric stereo. A uniform isotropic diffuser is necessary to filter this light source into a point-like distribution.

For the filter, we tested a ping-pong ball with the snoot connection cut off and a clear plastic ball with the exposed surface sanded into a frosted appearance. The ping-pong diffuser obstructed too much light even with one of its surfaces cut out. The sanded ball solution was clear on the side facing the bulb and preserved more illumination in the scene. (Fig. 3.4)

Figure 3.4. Left: Spherical diffuser made by cutting out one end of a ping-pong ball. Removing transmissive surfaces kept the light output higher. Right: diffuser made by sanding the exposed portion of a clear plastic ball into a frosted appearance.

We tested the behavior of the light by photographing its scatter pattern on a flat white surface and by photographing the light source directly (Fig. 3.5). At a distance of ~3 meters, this small spherical light source of about 4cm was acceptably close to a point light assumption. Lights were placed coplanar with the camera position at varying heights and horizontal translations to maximize shading differences observed in the image captures.

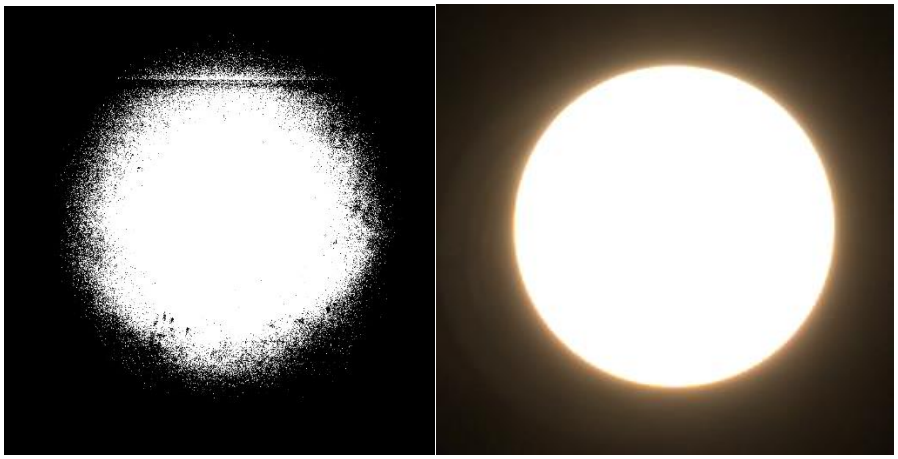


Figure 3.5 Left: light pattern as observed on a flat white surface, binary thresholded. The radius of the pattern grew in a uniformly circular pattern as the threshold level was adjusted, indicating a proper point light falloff. The horizontal artifact is a result of a rolling camera shutter. Right: Image of the spherical diffuser when the flash is activated.

Images were taken with a Canon EOS 80D camera in raw .CR2 mode captured at 24.2 megapixels. The camera was connected to a PocketWizard MultiMax II remote trigger with custom firmware developed in conjunction with the PocketWizard support team. One single trigger of the camera prompted the camera to shoot five images in quick succession. The attached MultiMax II was simultaneously triggered to cycle through 5 radio channels corresponding to 5 remote flashes listening in radio receivers in synchrony with the shutter. The initial camera trigger was operated with a pair of transceivers (PocketWizard Plus IIIs) to avoid touching the camera and perturbing the image framing at any point in the process. Capturing a 5-shot sequence in this manner took around half a second with exposure time set to 1/10 seconds. F-stop was set to 3.5 (the lowest setting) with ISO set to 100 at the minimum 18mm focal length. All of these settings were designed to maximize light capture by the camera sensor and by extension the signal-to-noise ratio of the image. The shutter interval was programmed to completely contain the ignition interval of each remote flash while allowing for the scene to reset to pitch blackness in between each shot to preserve single-point-light assumptions. Images were

converted to 16-bit TIFF files using *dcraw*. This preserved linearity of pixel intensity across the dynamic range of the image instead of an inaccurate logarithmic compression in 8-bit JPEG. An example sequence captured with a 2m tall mannequin is shown in Fig. 3.6.

Figure 3.6. Five photo sequence with automatically triggered remote flashes placed at differing heights and horizontal translations.

3.2 Results

Orthographic photometric stereo produced shapes that were artificially flattened and bent away from the center of projection. (Fig. 3.7) However, a high level of detail was observed in the surface geometry. Photometric stereo produces surfaces with a one-to-one correspondence between pixels and mesh vertices. The output mesh had the same vertex density as the pixel density of the image; small details such as stretch marks were visible in the mesh geometry in the absence of textures. (Fig. 3.8)

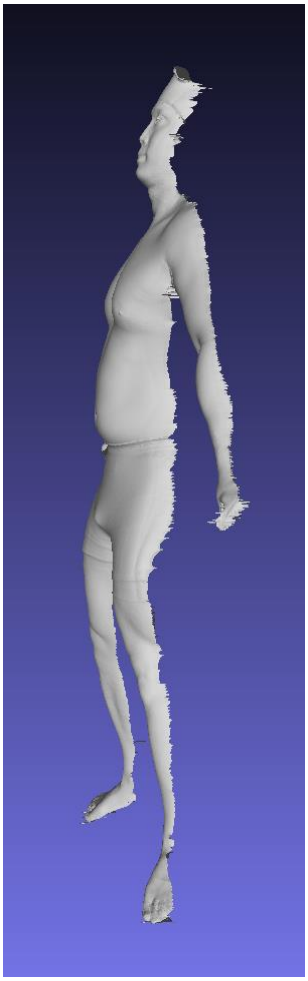


Figure 3.7 Surface reconstruction of test subject under orthographic (linear) projection assumptions. Note the exaggerated flattening of the feet and the appearance of bending away from the center of projection.

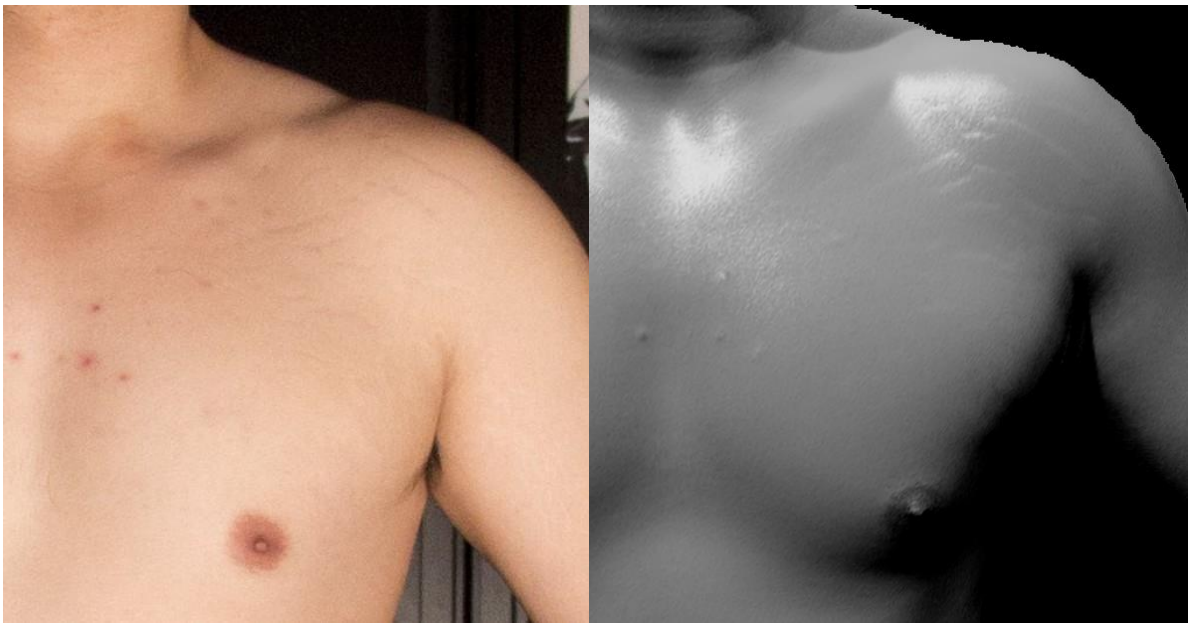


Figure 3.8. Close up of chest area on one of the photos in the photometric stereo sequence (left) compared to the reconstruction (right) with no textures present. Very small-scale surface deformations such as stretch marks were captured as geometry in the 3D reconstruction. The primary benefit of photometric stereo is the ability to capture pixel-resolution detail as surface geometry.

Perspective correction with a nonlinear optimization produced more realistic geometry than the orthographic solution. (Fig. 3.9) Note how the appearance of leaning back and the flattening of the feet is absent when perspective is accounted for. Loosening constraints such as allowing albedos and specular exponents to vary per pixel resulted in nonconvergent solutions due to an excessive number of parameters.

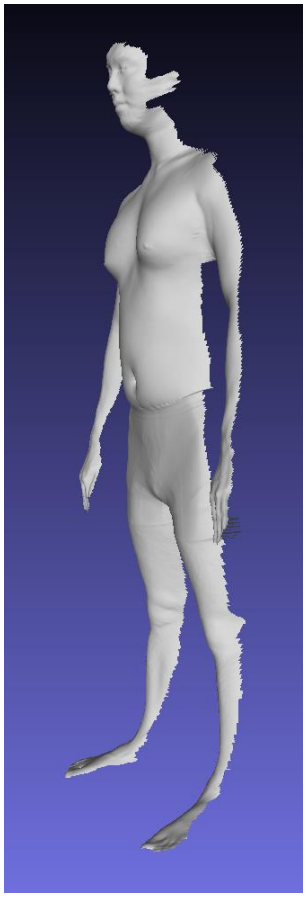


Figure 3.9. Reconstruction of a human subject under perspective projection and nonlinear optimization. Correcting for perspective projection allows for reconstruction of a more natural pose and correct feet geometry.

We tested reconstruction of a mannequin with sharp surface geometry. Synthetic test cases like this could be used to compare our method against a reference one, especially if a 3D CAD model from the manufacturer can be made available. (Fig. 3.10)

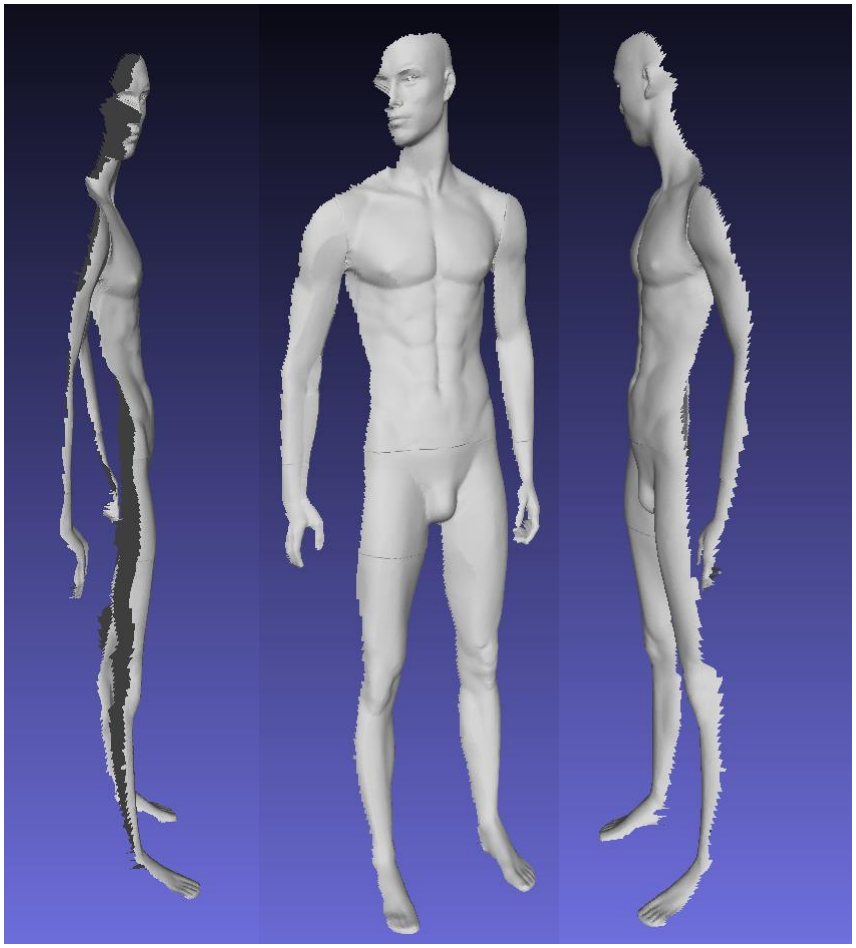


Figure 3.10. Multiple views of a reconstruction of a 2m tall mannequin with sharply detailed surface geometry. The mannequin was more specular than human skin due to its plastic material.

3.3 Discussion

In this project, we built a setup to perform pixel-resolution single viewpoint surface reconstruction of human body shape using calibrated photometric stereo. However, mounting problems with practicality and usability prevented us from continuing this project.

The accuracy of the 3D reconstruction under perspective projection was still susceptible to distortions. Errors accumulate in the optimization process from inaccurate reflectance functions (the Phong model is not energy preserving, for example), camera sensor noise, measurement errors in the scene and light positions, and a low dynamic range resulting from

dimming the flash power considerably with diffusers in order to achieve an isotropic point light emission pattern. We confirmed this distortion by reconstructing a rigid mannequin after 30 degrees rotation both clockwise and counterclockwise. The reconstructed surfaces do not perfectly align with each other or the center orientation reconstruction. (Fig. 3.11)

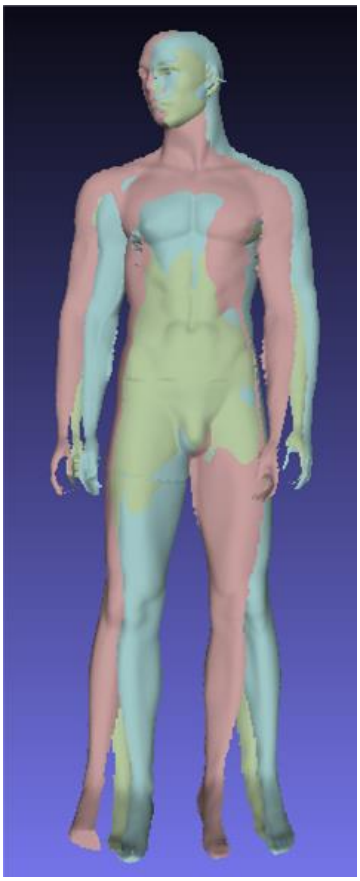


Figure 3.11. Reconstruction of 30-degree rotated mannequins (red: CCW, green: CW) compared to center facing reconstruction (blue). The surfaces are slightly distorted and cannot be made to perfectly register to each other.

The capture conditions required a large empty space with blackout conditions; at least 5 square meters with no light infiltration or reflective surfaces. Furthermore, the total cost of the equipment for a 5-light system with remote flash triggering was around $5000; this setup would have to be mirrored on the backside of the scanned subject in order to generate a full 360-degree

reconstruction in a single ~1 second take without reposing. These constraints made this solution impractical relative to commercial 3DO scanners such as Fit3D, which cost a comparable amount of money and only took up 1 square meter of space without necessitating blackout room lighting. Future work on learning from 3D human shape meshes were based on commercially built 3DO scanning systems built with structured light and time-of-flight mechanics.

4. Predicting 3D Body Shape and Body Composition from Conventional 2D Photography

For in-the-wild use cases, 2D photography with uncalibrated lighting was the most accessible and lowest cost data acquisition protocol for optical imaging of human bodies. We pivoted away from the high cost, highly calibrated photometric stereo setup in the previous chapter and investigated how much information about body composition could be determined from cheaply acquired data with a minimum amount of complexity; specifically, a single 2D photo paired with easily known personal measurements like height and weight.

In this chapter, we predicted 12 body composition variables from monocular 2D imaging by fitting a parameterized 3D shape model to the segmented frontal silhouette of a full body photo under perspective projection. Body composition regression functions were trained from the PCA parameter space of the 3D shape model; the parameters of the best fitting 3D shape were used to predict the most explanatory body composition metrics.

This chapter describes work originally published in Medical Physics:

[117] *Tian IY, Ng BK, Wong MC, Kennedy S, Hwaung P, Kelly NN, Liu YE, Garber AK, Curless B, Heymsfield SB, and Shepherd JA. Predicting 3D body shape and body composition from conventional 2D photography. Medical Physics 47 (12), 6232-6245. 2020*

4.1 Abstract

Purpose: Total and regional body composition are important indicators of health and mortality risk, but their measurement is usually restricted to controlled environments in clinical settings with expensive and specialized equipment. A method that approaches the accuracy of the current gold

standard method, dual-energy X-ray absorptiometry (DXA), while only requiring input from widely available consumer grade equipment, would enable the measurement of these important biometrics in the wild, enabling data collection at a scale that would have previously been prohibitive in time and expense. We describe an algorithm for predicting 3-dimensional body shape and composition from a single frontal 2-dimensional image acquired with a digital consumer camera.

Methods: Duplicate 3D optical scans, 2D optical images, and DXA whole body scans were available for 183 men and 233 women from the Shape Up! Adults Study. A principal component analysis vector basis was fit to 3D point clouds of a training subset of 152 men and 194 women. The relationship between this vector space and DXA-derived body composition was modeled with linear regression. The principal component 3D shape was then fitted to match a silhouette extracted from a 2D photograph of a novel body. Body composition was predicted from the resulting 3D shape match using the linear mapping between the principal component parameters and the DXA metrics. Accuracy of body composition estimates from the silhouette method was evaluated against a simple model using height and weight as a baseline, and against DXA measurements as ground truth. Test-retest precision of the silhouette method was evaluated using the duplicate 2D optical images and compared against precision of the duplicate DXA scans. Paired *t*-tests were performed to detect significant differences between the sets.

Results: Results were reported on a held-out set. Body composition prediction achieved R^2s of 0.81 and 0.74 for percent fat prediction of males and females, respectively, on a held-out test set consisting of 31 males and 39 females. Precision estimates for fat mass were 2.31% and 2.06% for males and females, respectively, compared to 1.26% and 0.68% for DXA scans. The *t*-tests

revealed no statistically significant differences between the silhouette method measurements and DXA measurements, or between retests.

Conclusion: Total and regional body composition measures can be estimated from a single frontal photograph of a human body. Body composition prediction using consumer level photography can enable early screening and monitoring of possible physiological indicators of metabolic disease in regions where medical imagery or clinical assessment is inaccessible.

Keywords*:* Body Composition, Dual-energy X-ray absorptiometry, Principal Components Analysis, Silhouette, Obesity, Nutritional Assessment.

Abbreviations*:* 3D, three-dimensional; BMI, body mass index; CAESAR, Civilian American and European Surface Anthropometry Resource Project; DXA, dual-energy X-ray absorptiometry; DSLR, digital single-lens reflex camera; FFM, Fat free mass; ICP, Iterative Closest Point algorithm; PCA, principal component analysis; RMSE, root-mean-square error.

4.2 Introduction

Predicting body composition has many useful clinical and research applications. Obesity is considered a primary risk factor for the development of type 2 diabetes, cardiovascular disease, and multiple forms of cancer. [144, 26, 17] Regional composition of selected body regions has been shown to be even more specific for prediction of the aforementioned health risks than whole body measures such as total body fat. Anthropometric surrogate measures of these regional tissue compartments such as waist circumference (WC), waist to hip ratio (WHR), surface markers of visceral adipose tissue (VAT) and related depots, have been shown to be stronger indicators of metabolic disease and mortality risk than total body fat [93, 59]. Mid-upper-arm circumference

(MUAC) is recognized by the World Health Organization as a marker of nutritional status, particularly in populations at risk for malnutrition [77]. Appendicular lean mass index is a marker for limb strength and can be used to diagnose muscle wasting disorders such as sarcopenia [52]. A criterion method for body composition assessment is Dual-Energy X-ray absorptiometry (DXA), an imaging technique that is currently considered the gold standard for measurement of total and regional body composition in clinical trials and research studies because of its precision and accuracy [69]. However, DXA is only available in specialized clinics and its use of ionizing radiation limits its frequent repetitive use.

The importance of body composition monitoring coupled with its high cost and low accessibility suggest a need for methods that can easily be used without access to a controlled clinical environment with cost prohibitive equipment and expertise to monitor the status of and changes in total and regional body composition compartments. Ideally, this technology would be affordable to middle- and low-income individuals, who are the populations most likely to be adversely affected by high costs and low access due to the increased risk of metabolic disease among lower socioeconomic brackets, and accessible through hardware that is widely distributed and commonly available outside of specialized clinics. Such a method would allow for measurement of body composition “in the wild” and would enable the outsourcing of body composition tracking from the professional clinic to the domestic household. This large-scale broadening of accessibility to clinically important body metrics can enable participation in self-monitoring and population health data analysis at previously infeasible scales. Commercial candidate solutions exist that are minimally invasive and relatively inexpensive by clinical standards. These include bioimpedance scales in both the bathroom scale format and in the tetrapolar configuration (BF-680W and MC-980U, Tanita Corporation, Arlington Heights, IL,

USA). Although tetrapolar scales are more accurate and can provide more regional composition information, they cost between $12,000 and $20,000 and are generally only purchased by commercial gyms. Another candidate technology is air-displacement plethysmography (ADP) such as the BodPod (Cosmed, Rome, Italy). This device has been shown to be similarly accurate as DXA but does not provide regional measures and is laboratory based. 3D optical scanners have recently been shown to be able to accurately measure body circumferences and estimate body composition in both adults and children [83, 135]. However, they too are not available for home use and can be expensive for individuals.

We propose a method for estimating fat and lean masses from a single front-facing 2D RGB photo taken from a consumer camera. Digital home photography is now easier and more accessible than ever with the mass popularity of mobile devices in the last decade. Cameras, whether standalone or integrated into a phone, are general purpose-devices that are not purchased solely for the purpose of body composition evaluation. The hardware is already widely accessible to people even in the lowest income brackets, requiring no additional cost to obtain composition metrics: 95% of Americans making less than $30,000 a year own some kind of cell phone, and 71% own some kind of smart phone [147]. Such a method could remove the barrier to preventative care and diagnostic evaluations that tend to disproportionately impact communities underserved by the medical profession by outsourcing the data collection method to household devices that are readily available.

The objective of this study was to show that DXA body composition measurements could be reliably estimated using a photograph of a human body. We first created a model to estimate DXA body composition from 3D optical scans. We then synthesized a 3D body shape that best matched the binary silhouette of the human body in a 2D image taken in front of a green background and

predicted the expected body composition from the parameters of the fitted 3D shape. The model for predicting DXA body composition from a 3D optical scan was thus extended to support a 2D optical image. We described the accuracy and precision of the 3D and 2D composition estimation models relative to DXA in a population of healthy adults.

4.2 Materials and Methods

We performed a prospectively acquired cross-sectional study on adults with a wide variety of age, Body Mass Index (BMI), and ethnicities for both sexes. All participants received duplicate whole body DXA scans, 3D optical scans, and 2D color photos. Advanced statistical methods were used to relate 2D and 3D body shapes to DXA body composition. The accuracy of the optical methods to DXA as well as their test-retest precision are described and reported below.

4.2.1 Study Population and Procedures

Participants were recruited in the Honolulu, HI area at the University of Hawaii at Manoa, in the San Francisco, CA area at the University of California, San Francisco, and in the Baton Rouge, LA area at Pennington Biomedical Research Center as part of the Shape Up! Adults Study (NIH R01 DK109008). Recruitment was stratified by age (18-40, 40-60, > 60 years), ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, Asian, and Native Hawaiian or Pacific Islander (NHOPI)), gender, and BMI (< 18, 18-25, 25-30, > 30 kg/m^2). Participants wore skintight underwear consisting of grey or black bike shorts and either a grey or black untextured and unstructured sports bra (women) or were shirtless (men). For optical scans, participants hid their hair in a swim cap. Following the Shape Up protocol, each participant

underwent duplicate whole-body DXA and 3D Optical (3DO) scans, blood tests for diabetes and lipid biomarkers, as well as handgrip and thigh strength tests. Handgrip strength was measured as the average of three squeezes on a handgrip dynamometer (JAMAR 5030J1, Sammons Preston Rolyan, Nottinghamshire, UK) on each hand. Leg strength was measured as isokinetic and isometric knee extension and flexion on a HUMAC NORM (Computer Sports Medicine Inc., Stoughton, MA, USA) or Biodex Systems (Biodex Medical System Inc., Shirley, NY, USA) dynamometer. Participants were excluded if they could not stand without aid for two minutes or lie flat for ten minutes without movement, had metal objects in their body, or previously had major body-shape-altering procedures (e.g., liposuction, amputations, etc.). Female participants were also excluded if pregnant or breast feeding. Written informed consent was obtained from each participant upon arrival and all procedures were approved by the Pennington Biomedical Research Center Institutional Review Board (IRB# 2016-053-PBRC), the UH Office Of Research Compliance (CHS# 2017-01018), and the Human Research Protection Program Institutional Review Board at the University of California, San Francisco (IRB# 15-18066). The study is publicly listed on ClinicalTrials.gov as ID NCT03637855.

4.2.2 DXA Scanning

As part of the data acquisition procedure for Shape Up, we captured two whole-body DXA scans, with body repositioning between scans, on either a Hologic Horizon/A system (UCSF) or a Discovery/A system (PBRC and UHCC) (Hologic Inc., Marlborough, MA, USA) for each participant. Participants were positioned and scanned according to each manufacturer's guidelines. All DXA scans were analyzed at UHCC by a single certified technologist using Hologic Apex version 5.6 with the National Health and Nutrition Examination Survey (NHANES) Body Composition Analysis calibration option disabled. DXA systems quality

control was performed by monitoring the weekly values of the Hologic Whole Body Phantom. Cross calibration was checked between sites using a whole-body phantom scanned at each site. No cross-calibration adjustments were needed [83]. Body composition measurements from DXA included total and regional (trunk, arms, legs) measures of total fat mass and fat free (lean) mass (FFM). Percent fat (% fat) is represented as fat mass divided by total mass.

4.2.3 3D Optical Scanning

For each participant, we also captured two 3DO whole-body surface scans on a Fit3D ProScanner (Fit3D, Inc., Redwood City, CA, USA). Subjects were repositioned between scans. Participants followed a manufacturer specified positioning protocol. The ProScanner captures 3D shape by rotating a stationary subject 360 degrees in front of one or more light-coding depth sensors. Scanning takes approximately 40 seconds to complete. The Iterative Closest Point (ICP) algorithm is used to align unorganized point clouds captured by the sensor as the subject rotates. The final body-shape-approximating point cloud is converted to a triangle mesh with approximately 350,000 vertices and 700,000 faces. All 3DO scan data were transferred from the measurement sites and stored securely at UHCC prior to statistical analysis.

4.3.3 2D Optical Scanning

Each participant was photographed twice in front of a green screen using a digital single-lens reflex (DSLR) camera and repositioned between the two photos. Participants stood in a neutral A-pose facing the camera with feet placed at fixed, marked locations on the floor 11 inches apart. This pose was chosen to best mimic the 3D optical pose. Each subject held a positioning bar that fixed the position of their arms such that their hands were 34.75 inches apart with straight elbows. Photos were de-identified by superimposing a black oval on the face

without obscuring the outline of the head. Images were captured in RAW format and converted into 16-bit linear TIFF files using an open-source software routine *dcraw*. This 16-bit conversion was vital as a standard 8-bit compression violates the assumption of a linear relationship between pixel value and captured light intensity.

4.3.4 Constructing 3D-to-composition model

Our training procedure is described below; separate models were created for each gender:

1. Prepare inputs: ground truth 3D scans, DXA-derived body composition measures, 2D photographs.
2. Construct 3D shape space using Principal Component Analysis (PCA) from mesh templates fitted to ground truth 3D optical scans [5].
3. Determine the best fit of a projection of the 3D model to the silhouette extracted from the 2D image.
4. Derive the body composition estimates from the PCA weight coefficients of the best fit 3D shape.

4.3.5 Applying 3D model to 2D images

The study procedure is then as follows for any new subject with input comprised of their height, weight, an RGB photo of the subject against a green screen, and the camera parameters:

1. Automatically detect 2D joint locations and segment subject from background. Manually correct any errors in the segmentation.

2. Initialize 3D shape with input height, weight. Initialize rigid transformation to align initial shape to detected joints on image. Fit 3D PCA shape to silhouette minimizing energy function E (described below).
3. Map optimized 3D PCA coefficients to body composition using the mapping learned in the training phase.

4.3.6 Training Procedure

Our pipeline mapped a 2D image to a 3D statistical shape, and then mapped the parameters of that shape to body composition statistics. The 3D statistical shape was represented by a PCA basis consisting of d column vectors of size n = 180,003. This PCA basis was constructed from eigen decomposition of a zero-mean-centered set of N body meshes represented as 1D column vectors of length 180,003, representing 60,001 3D points in XYZ interleaved format. Meshes were created by deforming a watertight template to fit ground truth 3D optical scans of each subject in the manner described by Allen *et al.* [5] (Fig. 4.1). Template fitting was required to maintain topological consistency and to give consistent positioning of vertex locations across subjects.

We can then describe any new body shape parameterized by this PCA basis as:

$$\mathbf{s}_{\text{PCA}} = \boldsymbol{\mu} + \sum w_i \, \mathbf{a}_i = \boldsymbol{\mu} + \mathbf{A}\mathbf{w} \qquad (1)$$

Where $\boldsymbol{\mu}$ is the mean of all training meshes, $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_d]$ is the PCA basis matrix, and $\mathbf{w} = [w_1 \ldots w_d]^{\mathrm{T}}$ is a length d vector of PCA coefficients that parameterize a given shape as an offset from the mean. The first 80 vectors of the PCA matrix sorted by descending eigenvalue represented

just over 99% of the shape variance in the training meshes for both males and females. In Ng *et al.* [83], we used the first 15 vectors which only explained 95% of the shape variance. However, as more data became available, we found that 95% representation resulted in overly smoothed shape reconstructions that insufficiently captured details such as fatty skin folds. We defined dimensionality d as 80 for the rest of this work. We also recorded the corresponding standard deviations σ_i of each principal component defined as the square root of the explained variance. The standard deviations are useful for regularizing the space of anatomically plausible human body shapes, as we will explain later.

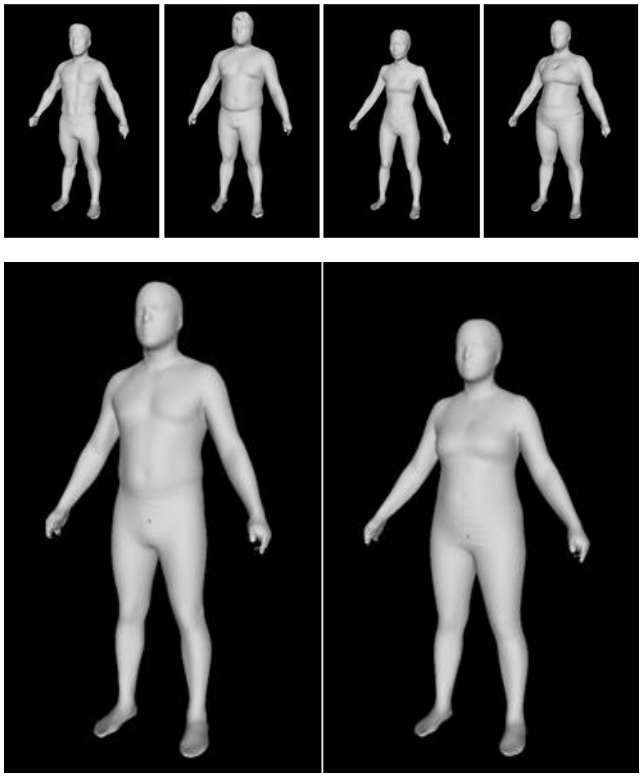


Figure 4.1. Top: Examples of template fitted 3D scans used to construct PCA space. Bottom: The mean male and female shape **μ** that are the starting points for all shape deformations.

A key contribution of this work is the ability to map between a 3D shape and its associated body composition metrics. Ng *et al.* defined a stepwise regression method mapping the first 15 PCA components to composition. We performed a simpler mapping using least squares and demonstrated that even such a naive method is quite effective despite using over five times the number of parameters.

For N training participants with M target features, we defined feature matrix $\mathbf{F}$ as:

$$\mathbf{F} = \begin{bmatrix} f_{1,1} & \cdots & f_{1,j} & \cdots & f_{1,N} \\ \vdots & & \vdots & & \vdots \\ f_{i,1} & \cdots & f_{i,j} & \cdots & f_{i,N} \\ \vdots & & \vdots & & \vdots \\ f_{M,1} & \cdots & f_{M,j} & \cdots & f_{M,N} \end{bmatrix} \quad (2)$$

Where the j^{th} column in $\mathbf{F}$ represents the feature vector (for example, [height, weight, % fat]$^{\text{T}}$) for subject j.

For the same N training participants, we define PCA weight matrix $\mathbf{W}$ as:

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,j} & \cdots & w_{1,N} \\ \vdots & & \vdots & & \vdots \\ w_{i,1} & \cdots & w_{i,j} & \cdots & w_{i,N} \\ \vdots & & \vdots & & \vdots \\ w_{d,1} & \cdots & w_{d,j} & \cdots & w_{d,N} \end{bmatrix} \quad (3)$$

Where the jth column in $\mathbf{W}$ is the PCA basis projection of the body shape mesh of subject j in d reduced dimensions.

We defined augmented matrix $\widetilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W} \\ \overline{1} \end{bmatrix}$, and the following linear relationship:

$$\mathbf{F} = \mathbf{M}_{\mathbf{w}\to\mathbf{f}}\widetilde{\mathbf{W}} \tag{4}$$

The augmented row of ones is necessary to allow for a non-zero intercept for the linear relationship. Matrix $\mathbf{M}_{\mathbf{w}\to\mathbf{f}}$ now represents a linear transformation between a PCA coefficient vector $\mathbf{w}$ and the predicted features $\mathbf{f}$. We can solve for the least-squares optimal solution for $\mathbf{M}_{\mathbf{w}\to\mathbf{f}}$ using the pseudoinverse $\widetilde{\mathbf{W}}^{+}$:

$$\mathbf{M}_{\mathbf{w}\to\mathbf{f}} = \mathbf{F}\widetilde{\mathbf{W}}^{+} \tag{5}$$

Conversely, we define augmented matrix $\tilde{\mathbf{F}} = \begin{bmatrix}\mathbf{F}\\ \hline 1\end{bmatrix}$ and:

$$\mathbf{W} = \mathbf{M}_{\mathbf{f}\to\mathbf{w}}\tilde{\mathbf{F}} \tag{6}$$

$\mathbf{M}_{\mathbf{f}\to\mathbf{w}}$ maps a vector of feature priors to a predicted shape $\mathbf{w}$. This is useful for initializing our shape parameter vector, e.g., given easily measured features like height and weight, to increase the convergence speed and accuracy of our optimization as we describe in the next section. We solve for the least squares optimal matrix using the pseudoinverse again as above:

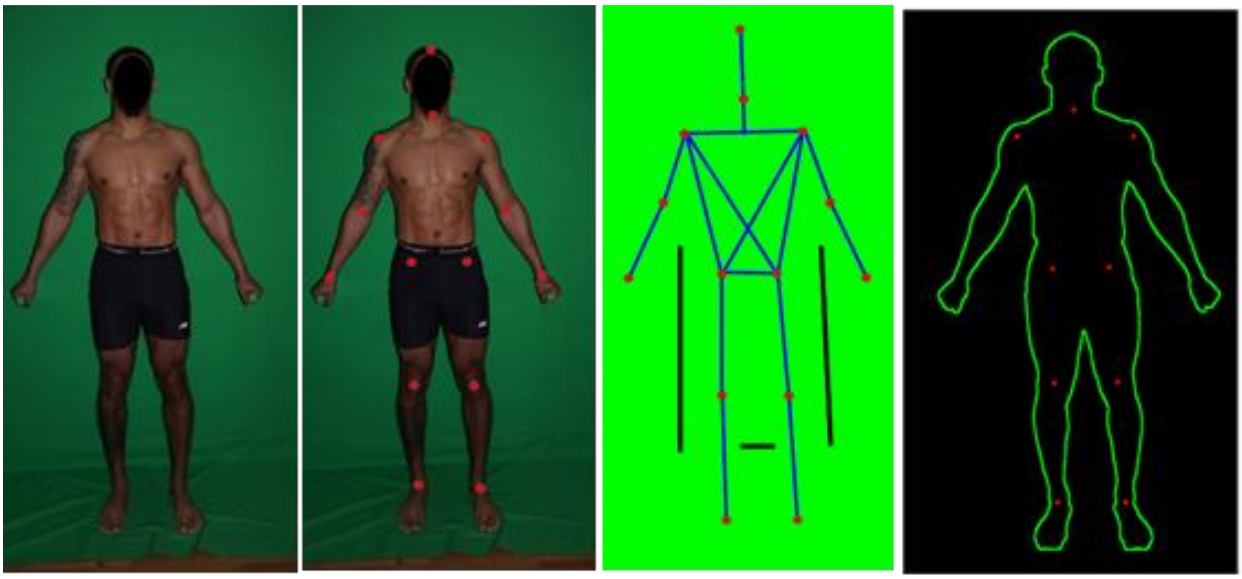
$$\mathbf{M}_{\mathbf{f}\to\mathbf{w}} = \mathbf{W}\tilde{\mathbf{F}}^{+} \tag{7}$$

4.3.7 Testing Procedure

The input to our algorithm was an RGB front-facing photo of a subject in a neutral pose in front of a green background, height of the subject in meters, weight of the subject in

kilograms, camera intrinsic parameters comprised of focal length and sensor dimensions, and an estimate of the distance between the camera and the subject.

As a pre-process, we extracted the approximate joint locations and the detailed silhouette of the subject. Given the input photo (Fig. 4.2a), we performed CNN-based automatic joint detection on the RGB image (Fig. 4.2b) using DeepCut [89]. The joints were used to initialize a skeleton foreground label (Fig. 4.2c) for automatic segmentation using GrabCut [101]. It is important to get as close to pixel accuracy as possible for the silhouette of the subject; therefore, it is sometimes necessary to manually patch holes or erase background in the automatic result. We used this mask to extract the silhouette pixels $\{B_j\}$, defined as the set of all foreground pixels that neighbor a background pixel (Fig. 4.2d). In addition, corresponding 3D joint locations were picked manually on the average template mesh once, and the vertex indices were saved for all further joint location references on the 3D mesh.



a b c d

Figure 4.2. Example of preprocessing an input image. a: the input RGB image. b: CNN detected joints. c: skeleton foreground seed label (blue) created by connecting detected joints. Background initialized as black marked lines. Initializations are done automatically. d: extracted silhouette (green) and joints used for fitting (red).

Because each subject did not stand in precisely the same location relative to the camera, it was necessary to allow for a rigid transformation, $\mathbf{T}$, of the PCA space to maximize the alignment with the detected silhouette both before and during the fitting procedure. Our goal was to solve for the 3D body shape $\mathbf{s}_{\mathrm{PCA}}(\mathbf{w})$ and camera transform $\mathbf{T}$ that best fits the subject seen in the 2D image. To achieve this fitting, we defined an objective comprised of multiple energy terms to be minimized together.

The first term $E_{\mathrm{sil}}(\mathbf{w}, \mathbf{T})$ minimized the distance between the silhouette of the perspective projection of the 3D PCA shape and the silhouette of the 2D input image:

$$E_{\mathrm{sil}}(\mathbf{w}, \mathbf{T}) = \sum_j \tau_j \mathrm{dist}^2 \left(B_j, \pi\big(\mathbf{T} \cdot \mathbf{s}_{\mathrm{PCA}}(\mathbf{w})\big)\right) \quad (8)$$

where dist() measures the distance between image silhouette point B_j and the nearest compatible silhouette point of the PCA mesh $\mathbf{s}_{\mathrm{PCA}}(\mathbf{w})$ transformed by $\mathbf{T}$ under camera projection π. Distances are weighted by τ_j depending on body part as described below. $E_{\mathrm{sil}}(\mathbf{w}, \mathbf{T})$ is the sum of pairwise 2D distances between the image silhouette points $\{B_j\}$ and matched PCA silhouette vertices defined as $\pi\big(\mathbf{T} \cdot \mathbf{s}_{\mathrm{PCA}}(\mathbf{w})\big)$. For every point on the image silhouette B_j, its nearest compatible PCA silhouette vertex was defined as the nearest transformed and projected neighbor that is a PCA silhouette vertex and shares a similar orientation.

A PCA silhouette vertex is a vertex whose normal is nearly orthogonal to the viewing ray, defined by the condition

$\mathbf{e}_i \cdot \mathbf{n}_i < 0.05$ for vertex normal $\mathbf{n}_i$ and viewing direction $\mathbf{e}_i$ taken from the camera center of projection to the current vertex, both transformed by rigid transformation $\mathbf{T}$. We matched each image silhouette pixel to a PCA vertex by performing a nearest neighbor search across the set of

candidate PCA silhouette vertices. The search was performed after the 3D PCA vertices were transformed by $\mathbf{T}$ and projected under perspective projection π to the same image coordinates as the image silhouette. We tracked the surface orientation of both the PCA boundary points and the image silhouette points. We rejected matches that did not have similar surface orientations to prevent incorrect registrations between different body surfaces due to poor alignment or initialization. Since deforming the PCA shape during fitting changes the candidate silhouette vertex coordinates, we repeat this registration in each iteration of the algorithm for intermediate shapes.

Additionally, limb misalignments were inevitable in our model as the 3D model our PCA space was trained on has no pose parameters. When participants were 3D scanned for the training set, everyone stood on the same footprints and grasped the same stationary handlebars, but differences in body proportions caused slight variations in limb angles and posture. The only way to attempt to match a discrepancy in limb alignment was to deform the entire body shape in the objective function. This deformation creates undesirable penalties in optimization energy when pose is slightly mismatched. Misaligned hands or feet contribute to large amounts of error in the energy function even if the rest of the body largely aligns. We introduced a term τ_j to give greater weight to the torso and hip silhouette points (6.0) relative to the limbs (1.0). We segmented the 3D average template mesh $\boldsymbol{\mu}$ in advance to identify points on the torso and hips.

The second term $E_{\text{joints}}(\mathbf{w}, \mathbf{T})$ is the sum of squared distances between the CNN-detected joints and the transformed and projected joint vertices on the 3D PCA model.

$$E_{\text{joints}}(\mathbf{w}, \mathbf{T}) = \sum_k \text{dist}^2 \left(J_k, \pi\left(\mathbf{T} \cdot \mathbf{J}_k^{\text{PCA}}(\mathbf{w})\right)\right) \quad (9)$$

where J_k is the *k*th detected joint on 2D image and $\mathbf{J}_k^{\mathrm{PCA}}(\mathbf{w})$ is the *k*th joint vertex on the 3D PCA mesh.

Joint vertices were picked once on the average template shape $\boldsymbol{\mu}$. Because topological consistency was guaranteed when the average shape was deformed to some new shape $\mathbf{w}$, the labeled joints had the same joint indices and were in approximately the same anatomical location. We used 10 joints representing shoulders, hips, knees, and ankles, plus a vertex for the crown of the head and a vertex for the base of the neck defined as the midpoint of the clavicles. This term provided a loose constraint on anatomical consistency for the fitting and favors a shape that has similar limb proportions under camera projection. Note that the detected elbows and wrists were not used in this term; arm position was highly variable and would have introduced noise to the fit.

The next two terms E_{height} and E_{mass} are regularizers based on the known prior height and mass of the subject to improve the anatomical accuracy of the shape fit:

$$E_{\mathrm{mass}}(\mathbf{w}) = (\mathbf{M}_{\mathbf{w}\rightarrow\mathbf{f}}\mathbf{w}[i_m] - m_0)^2 \quad (10)$$

$$E_{\mathrm{height}}(\mathbf{w}) = (||\mathbf{v}_{\mathrm{crown}}(\mathbf{w}) - \mathbf{v}_{\mathrm{heel}}(\mathbf{w})|| - h_0)^2 \quad (11)$$

$E_{\mathrm{mass}}(\mathbf{w})$ is the squared difference between the input known body mass m_0 and the predicted body weight using mapping matrix $\mathbf{M}_{\mathbf{w}\rightarrow\mathbf{f}}$ and the PCA shape vector of the estimated $\mathbf{w}$. $\mathbf{M}_{\mathbf{w}\rightarrow\mathbf{f}}$ in general produces a vector of body features: i_m gives the index of the total body mass feature in this vector. The predicted height was calculated simply as the squared 3D distance between a vertex at the crown of the head and a vertex at the base of the heel of the PCA model. The

position of these vertices are functions of $\mathbf{w}$. $E_{\text{height}}(\mathbf{w})$ is defined as the squared difference between this predicted height and the input height h_0.

The last term $E_\sigma(\mathbf{w})$ penalizes for large magnitudes of PCA shape vector $\mathbf{w}$, biasing the solution towards the mean. It is a weighted L2 regularization:

$$E_\sigma(\mathbf{w}) = \sum_{i=1}^{d} \left(\frac{w_i}{\sigma_i}\right)^2 \tag{12}$$

where w_i is the ith element of vector $\mathbf{w}$ and σ_i is the standard deviation of ith PCA vector. This regularizer prevents overfitting to the silhouette at the expense of producing unrealistic and unlikely body shapes. Shapes that are multiple standard deviations away from the mean (defined as $w_i = 0$ for all i) receive a larger penalty than shapes that deformed minimally from the origin (the mean).

Table 4.1. Hyperparameter optimal values.

Parameter	Description	Value
τ	Silhouette match weight	6.0 torso, 1.0 else
d	# of PCA components	80
α	Joint alignment weight	3.0
β	Height (m) alignment weight	5.0
γ	Mass (kg) alignment weight	1.0
λ	PCA std. dev. weight	0.001
ε	Convergence condition	0.3

We can now define the full energy function E as:

$$E(\mathbf{w}, \mathbf{T}) = E_{\text{sil}}(\mathbf{w}, \mathbf{T}) + \alpha E_{\text{joints}}(\mathbf{w}, \mathbf{T}) + \beta E_{\text{height}}(\mathbf{w}) + \gamma E_{\text{mass}}(\mathbf{w}) + \lambda E_{\sigma}(\mathbf{w}) \quad (13)$$

where α, β, γ, and λ are hyperparameters that determine the relative influence of each term in the energy function.

Due to the mesh projection step and the association of nearest compatible points, this is a non-linear objective. We iteratively optimized for $\mathbf{w}$ and $\mathbf{T}$ by minimizing $E(\mathbf{w}, \mathbf{T})$ using the Ceres [3] implementation of the Levenberg-Marquardt algorithm until the change in parameters $\mathbf{w}$ from the previous iteration was less than some cutoff ε. This difference was defined as the root sum of squared difference between the two vectors. Hyperparameters for (13) are listed in Table 4.1.

Using mapping matrix $\mathbf{M}_{\mathbf{f}\rightarrow\mathbf{w}}$, with $\mathbf{f}$ containing height and weight, we initialized shape parameters $\mathbf{w}$ as the PCA shape $\mathbf{w}_0 = \mathbf{M}_{\mathbf{f}\rightarrow\mathbf{w}}\mathbf{f}_0$ where $\mathbf{f}_0$ = [height, weight, 1]^T. This step initialized the PCA coefficients to an average person with the given height and weight, which increases the initial alignment with the target silhouette.

We initialized rigid transformation $\mathbf{T}$ by solving for the minimization of $E_{\text{joints}}(\mathbf{w}_0, \mathbf{T})$ with $\mathbf{w}_0$ fixed. A summary of our optimization loop is given in Algorithm 1. A visualization of the shape terms E_{sil} and E_{joints} is shown in Fig. 4.3.

Algorithm 1: 3D PCA to 2D Silhouette Alignment

Initialize $\mathbf{w}$ as $\mathbf{w}_0 = \mathbf{M}_{\mathbf{f}\rightarrow\mathbf{w}}\mathbf{f_0}$, where $\mathbf{f}_0$ is [height, weight, 1]$^{\mathrm{T}}$
Initialize $\mathbf{T}$ as min $E_{\text{joints}}(\mathbf{w}_0, \mathbf{T})$
while $||\Delta \mathbf{w}|| < \varepsilon$ **do**
    Recompute normals of intermediate shape $\mathbf{s}_{\text{PCA}}(\mathbf{w})$;
    for each image silhouette point B_j,
        find closest compatible point from $\pi(\mathbf{T}(\mathbf{s}_{\text{PCA}}(\mathbf{w})))$;
    Minimize $E(\mathbf{w}, \mathbf{T})$;
    Update $\mathbf{w}$ and $\mathbf{T}$ with converged result from this alignment
end
$\tilde{\mathbf{w}} = [\mathbf{w}, 1]^{\mathrm{T}}$
Compute body features $\mathbf{f} = \mathbf{M}_{\mathbf{w}\rightarrow\mathbf{f}}\tilde{\mathbf{w}}$;
return f

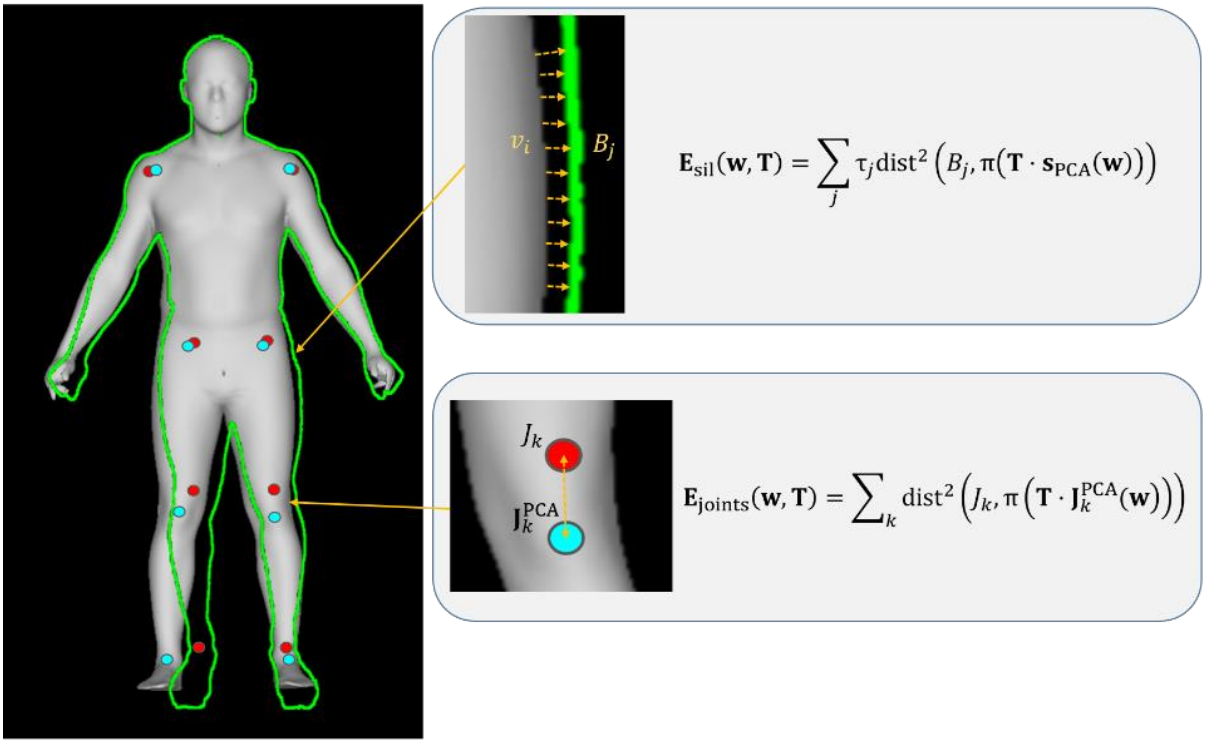


Figure 4.3 Visualization of the initial projected shape $\mathbf{w}_0$ overlaid onto the target silhouette (green). This projected 3D shape is fit by minimizing the closest pairwise distances between a boundary vertex and its closest silhouette point (top box) and by minimizing distances between detected joints on the silhouette (red) and the projected mesh joint vertices (blue) (bottom box).

4.3.8 Statistical Evaluation

We tested our method on a randomly selected held-out test set of 31 males and 39 females. Hyperparameters for reported results were chosen as indicated in Table 4.1 based on performance on a single male subject. Test set participants were not included in the PCA space construction, nor were they included in computing the mapping from PCA to body features. We performed 5-fold cross validation on this construction to verify the consistency of the PCA to composition regression. This was done by making $k = 5$ random folds of all subjects and creating 5 PCA spaces using each combination of k-1 folds. For each PCA space, we performed linear

regression between its fold members and their associated body statistics and reported validation results on the held-out fold representing 20% of total subjects. The experimental fold that we reported in the results section was a separate random fold and was not any of the above folds. Cross validation was necessary to demonstrate that our results are repeatable on arbitrary principal component spaces provided there is sufficient representation of body shapes and not just on a particularly favorable training – test split selected for this experiment.

We reported root-mean-square-error (RMSE) and the coefficient of determination (R^2) of our regression results from our predicted shapes using DXA measurements as the ground truth. We compared our predictions to a few different diagnostic scenarios to demonstrate the predictive quality of our silhouette fitting method. The lower bound scenario was demonstrated by predicting all body composition metrics on a simple linear regression from the known input scalars, height and weight, without any body geometry fitting. The upper bound scenario was demonstrated by taking the ground truth 3D scans of the test set and projecting them into principal component space by performing the inverse operation of (1); that is, subtracting out the mean shape and multiplying by the transpose of the PCA matrix. This produced a PCA coordinate vector that represented the projection of the 3D scan onto the principal component basis to give a prediction using the best possible geometric fit. We also reported the RMSE and R^2 of our 5-fold cross validation, using the sum total of prediction to ground-truth pairs across all 5 folds to compute these metrics. This demonstrated the robustness of the method against overfitting.

To ensure that our method is robust to natural variability in body pose and positioning we performed a test-retest precision evaluation on the experimental fold. Specifically, we evaluated a second set of images of the same test participants and compared predicted measurements

against those from the first set of images. Participants were repositioned between the two images, and thus stood in slightly different poses and positions. Precision of the 2D estimates was compared to the precision estimates from duplicate DXA scans. Coefficient of variation (%CV) results, defined as Glüer *et al*. [39] as the ratio of the standard deviation of repeat measurements to the mean of repeat measurements averaged across all test subjects, are shown in Table 4.2 and an example 3D to 2D fit in Fig. 4.4.

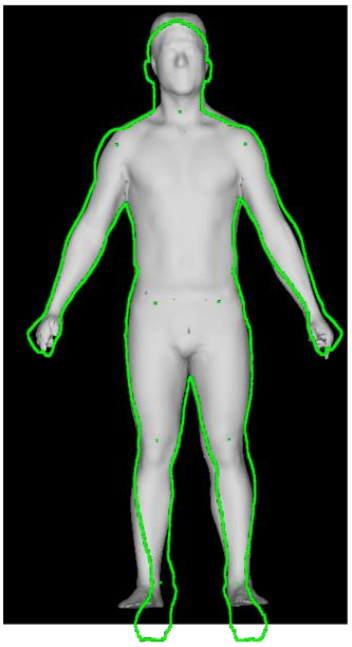


Figure 4.4. An example of a final aligned shape projected onto the target silhouette.

We performed a paired *t*-test on the test-retest trials for our method, the test-retest scans of DXA, and on the difference between our method and the DXA measurements. Since there were 12 different body composition measurements evaluated, a Bonferroni-corrected critical *p*-value of 0.05 / 12 = 0.004 was considered significant.

4.3 Results

Repeatability comparison to the DXA gold standard of measuring % fat is shown in Table 4.2 and represented as the coefficient of variation (CV). RMSE and R^2 values between the test and retest trials are also shown. %CV and RMSE values for our method were around 2-3 times larger than those from DXA. R^2 are all greater than 0.90 and are comparable to the DXA equivalents with the exception of female visceral fat and leg fat, at $R^2 = 0.60$ and 0.85 respectively. While reduced precision in limb compartment estimates may be explained by the lack of consistent pose alignment between photos of the same subject and the inability of our shape model to account for pose differences independent of body shape, the visceral fat imprecision suggests that particular measurement is not well modeled in females by our method.

Table 4.2. Test-retest precision of test set data compared to DXA.

	This Work						**DXA**					
	Male (n=31)			Female (n=39)			Male (n=31)			Female (n=39)		
	%CV	**R^2**	**RMSE**	%CV	**R^2**	**RMSE**	%CV	**R^2**	**RMSE**	%CV	**R^2**	**RMSE**
FMI [kg/m²]	2.40	**0.99**	**0.161**	2.19	**0.99**	**0.210**	1.27	**1.0**	**0.084**	0.68	**1.0**	**0.064**
FFMI [kg/m²]	0.78	**0.99**	**0.168**	1.22	**0.98**	**0.215**	0.37	**1.0**	**0.078**	0.44	**1.0**	**0.076**
Fat Mass [kg]	2.31	**0.99**	**0.469**	2.06	**0.99**	**0.512**	1.26	**1.0**	**0.252**	0.68	**1.0**	**0.168**
FFM [kg]	0.72	**0.99**	**0.469**	1.12	**0.98**	**0.512**	0.37	**1.0**	**0.232**	0.44	**1.0**	**0.199**
Percent Fat [%]	--	**0.97**	**0.502**	--	**0.94**	**0.671**	--	**1.0**	**0.242**	--	**0.99**	**0.243**
Visceral Fat [kg]	2.87	**0.98**	**0.016**	15.21	**0.60**	**0.065**	4.58	**0.96**	**0.023**	5.75	**0.96**	**0.022**
Trunk Fat Mass [kg]	2.67	**0.99**	**0.313**	2.76	**0.98**	**0.323**	2.21	**0.99**	**0.222**	1.73	**0.99**	**0.197**
Trunk FFM [kg]	0.62	**1.0**	**0.201**	1.68	**0.96**	**0.386**	0.93	**0.99**	**0.280**	0.817	**0.99**	**0.183**
Arms Fat Mass [kg]	3.64	**1.0**	**0.089**	3.96	**1.0**	**0.136**	2.49	**0.99**	**0.030**	2.12	**0.99**	**0.033**
Arms FFM [kg]	2.12	**0.96**	**0.195**	3.29	**0.90**	**0.173**	1.23	**0.99**	**0.052**	1.36	**0.98**	**0.032**
Legs Fat Mass [kg]	2.93	**0.98**	**0.193**	6.43	**0.85**	**0.636**	1.30	**1.0**	**0.042**	1.09	**1.0**	**0.050**
Legs FFM [kg]	0.92	**0.99**	**0.196**	1.25	**0.98**	**0.187**	0.93	**0.99**	**0.096**	0.80	**0.99**	**0.059**

The R^2 and RMSE values of every predicted body composition metric are shown in Table 4.3 and Table 4.4. In Table 4.3 we compared our results to 1) the 5-fold cross validation performance of each feature representing an estimate of the expected performance of the regression method on scans with known shape and PCA vectors, 2) the prediction produced only

by a linear regression of the known BMI of the subject, 3) the prediction produced only by a linear regression of the known initialization variables [height, weight] to each of the desired features, and 4) the prediction using the projection of the 3D scan of each subject to PCA basis space. The 5-fold cross validation comparison was necessary to demonstrate that our held-out test set was fairly representative of the predictive capabilities of the PCA method sampled across multiple training – test splits, rather than being an overperforming outlier set picked for the purposes of this publication. Comparison to linear regression using only BMI demonstrates the predictive power of this method relative to a common scalar analogue for % fat. Comparison to linear regression with the variables [height, weight] may seem redundant, but it is necessary to demonstrate that the silhouette fitting method adds predictive accuracy to the baseline input information of height and weight and represents a lower bound for performance. As this method is intended to be accessible to a nonprofessional audience, height and weight were chosen to be the initializer variables rather than BMI. We show that in every predicted variable, the silhouette fitting method improves upon the lower bound predictions that would have been available from using the initialization variables alone for both BMI and height + weight. Females were more accurately predicted by the initialization variables alone, showing 20% decreases in RMSE from the initialization result to the shape fitted result in fat and lean mass, as opposed to males which exhibited almost a 40% decrease.

Table 4.3. Prediction accuracy results on test data as measured by the coefficient of determination (R^2) and root-mean-square-error (RMSE). The five models from left to right represent ablation studies designed to determine the ceiling (Model 4) and floor (Models 2 and 3) of the presented method (Model 5). Model 1 represents the 5-fold cross validation results on all available data without a holdout set, comparable to previous work in Ng et al. For % fat, we included two methods of prediction, % fat = predicted Fat Mass / scale weight, and the linear regression of % fat from PCA vectors from (4) below in parentheses.

		Model 1: Combined 5-fold cross validation on all available scans		**Model 2: BMI regression only on test set**		**Model 3: Height & Weight regression on test set**		**Model 4: PC to body metrics regression using projected PCs from test set scans**		**Model 5: PC to body metrics regression using predicted body shape from image**	
Output Variable	**Gender**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Fat Mass [kg]	Male	**0.90**	0.90	**0.74**	5.88	**0.75**	5.78	**0.96**	**2.27**	**0.90**	3.63
	Female	**0.94**	0.94	**0.86**	3.43	**0.91**	2.83	**0.94**	**2.35**	**0.94**	2.29
FFM [kg]	Male	**0.93**	0.93	**0.31**	9.26	**0.73**	5.78	**0.94**	**2.78**	**0.89**	3.63
	Female	**0.91**	0.91	**0.57**	5.20	**0.87**	2.83	**0.89**	**2.70**	**0.92**	2.29
% Fat	Male	**0.68 (0.76)**	0.68 (0.76)	**0.41 (0.46)**	5.72 5.44	**0.41 (0.50)**	5.71 (5.27)	**0.90 (0.90)**	**2.36 (2.45)**	**0.725 (0.806)**	3.90 (3.27)
	Female	**0.77 (0.76)**	0.77 (0.76)	**0.50 (0.54)**	4.56 (4.43)	**0.65 (0.56)**	3.86 (4.31)	**0.75 (0.74)**	**3.06 (3.38)**	**0.74 (0.631)**	3.29 (3.94)
FMI [kg/m²]	Male	**0.89**	0.89	**0.75**	1.86	**0.75**	1.87	**0.96**	**0.73**	**0.90**	1.19
	Female	**0.94**	0.94	**0.88**	1.23	**0.91**	1.05	**0.93**	**0.91**	**0.94**	0.85
FFMI [kg/m²]	Male	**0.91**	0.91	**-0.42**	3.26	**0.53**	1.87	**0.90**	**0.90**	**0.81**	1.19
	Female	**0.89**	0.89	**0.43**	2.07	**0.85**	1.05	**0.87**	**1.02**	**0.91**	0.85
Visceral Fat Mass [kg]	Male	**0.67**	0.67	**0.12**	0.23	**0.18**	0.23	**0.75**	**0.13**	**0.66**	0.15
	Female	**0.76**	0.76	**0.25**	1.89	**0.27**	1.86	**0.71**	**0.12**	**0.36**	0.17
Trunk Fat Mass [kg]	Male	**0.92**	0.92	**0.72**	3.35	**0.74**	3.25	**0.95**	**1.40**	**0.92**	1.76
	Female	**0.94**	0.94	**0.84**	1.98	**0.89**	1.64	**0.92**	**1.37**	**0.91**	1.48
Trunk FFM [kg]	Male	**0.90**	0.90	**0.39**	4.26	**0.80**	2.45	**0.87**	**1.97**	**0.87**	1.97
	Female	**0.88**	0.88	**0.46**	2.91	**0.83**	1.64	**0.84**	**1.63**	**0.87**	1.43
Arms Fat Mass [kg]	Male	**0.80**	0.80	**0.70**	0.74	**0.74**	0.70	**0.89**	**0.45**	**0.81**	0.59
	Female	**0.88**	0.88	**0.76**	0.69	**0.81**	0.61	**0.83**	**0.58**	**0.84**	0.57
Arms FFM [kg]	Male	**0.88**	0.88	**-0.02**	1.80	**0.31**	1.48	**0.87**	**0.65**	**0.71**	0.96
	Female	**0.80**	0.80	**0.47**	0.76	**0.64**	0.63	**0.71**	**0.57**	**0.66**	0.61
Legs Fat Mass [kg]	Male	**0.78**	0.78	**0.63**	2.53	**0.62**	2.56	**0.91**	**1.26**	**0.75**	2.07
	Female	**0.91**	0.91	**0.66**	2.02	**0.67**	2.00	**0.90**	**1.13**	**0.85**	1.32
Legs FFM [kg]	Male	**0.90**	0.90	**0.26**	3.48	**0.67**	2.34	**0.87**	**1.44**	**0.84**	1.61
	Female	**0.88**	0.88	**0.59**	2.05	**0.80**	1.45	**0.87**	**1.19**	**0.85**	1.26

The prediction using the projected PCA coordinates of the 3D scan represented a rough upper bound of the prediction capability of the method. It is the approximate best-case scenario

of the regression function assuming shape prediction was perfect. This allowed us to evaluate how effective the shape fitting was at improving composition prediction independent of the noise inherent in the regression functions. However, this was not an exact upper bound because subjects were not photographed and scanned in the exact same motionless position. This introduced some variance to the shape caused by slight differences in limb pose and posture, which our shape model is currently not capable of separating from body shape. Some metrics in females, such as lean mass, showed higher R^2 and lower RMSE in our test prediction from 2D data than from the best-case 3D shape projection as a result.

Fat mass and fat free mass (FFM) estimates for females showed an RMSE of almost 40% lower than those for males. For trunk fat mass and fat free mass, females were 16% and 27% lower, respectively. Percent fat (% fat) was calculated in two ways: first by dividing the predicted fat mass by the known input body mass, and then by directly predicting percent fat as a feature in the linear regression described by (4). The first method achieved 15% lower RMSE on females, which is consistent with their lower fat mass error. However, linear regression of the percent fat variable produced the opposite effect, with males having 15% lower RMSE than females. We treat the first method as the standard method in future references to percent fat to be consistent with previous work. Every limb compartment fat and fat free mass estimate had lower RMSE for females, there was an accepted amount of limb misalignment for both genders due to pose variations in the dataset. Visceral fat was the only measurement for which the model for males notably outperformed the model for females (R^2 of 0.66 and 0.36, respectively).

Table 4.4 compares our results, which starts from a 2D input (camera photo) to Ng *et al.*, which starts from a 3D scan. We show that our method is comparable to this related method that also used PCA to predict body composition variables despite an additional step that requires

predicting the 3D body shape from the silhouette, rather than having the ground truth 3D shape as input. RMSE in our method was 7% higher in fat and lean mass for males, but 23% lower in females.

Table 4.4. Comparison of our work with Ng *et al*. which only reports % fat as predicted Fat Mass / scale weight.

		Ng et al, 3D PC Only; Stepwise Regression 5-fold CV		This work; prediction on 2D image, reported on test set scans	
Output Variable	**Gender**	**R^2**	**RMSE**	**R^2**	**RMSE**
Fat Mass [kg]	Male	**0.88**	3.38	**0.90**	3.63
	Female	**0.93**	2.96	**0.94**	2.29
FFM [kg]	Male	**0.93**	3.38	**0.89**	3.63
	Female	**0.90**	2.95	**0.92**	2.29
% Fat	Male	**0.65**	3.83	**0.725 (0.806)**	3.90 (3.27)
	Female	**0.70**	4.10	**0.74 (0.631)**	3.29 (3.94)
FMI [kg/m^2]	Male	**0.87**	1.11	**0.90**	1.19
	Female	**0.93**	1.13	**0.94**	0.85
FFMI [kg/m^2]	Male	**0.90**	1.11	**0.81**	1.19
	Female	**0.88**	1.12	**0.91**	0.85
Visceral Fat Mass [kg]	Male	**0.67**	0.16	**0.66**	0.15
	Female	**0.75**	0.14	**0.36**	0.17
Trunk Fat Mass [kg]	Male	**0.91**	1.68	**0.92**	1.76
	Female	**0.94**	1.43	**0.91**	1.48
Trunk FFM [kg]	Male	**0.90**	1.94	**0.87**	1.97
	Female	**0.87**	1.72	**0.87**	1.43
Arms Fat Mass [kg]	Male	**0.84**	0.26	**0.81**	0.59
	Female	**0.70**	0.58	**0.84**	0.57
Arms FFM [kg]	Male	**0.76**	0.52	**0.71**	0.96
	Female	**0.67**	0.33	**0.66**	0.61
Legs Fat Mass [kg]	Male	**0.71**	0.87	**0.75**	2.07
	Female	**0.83**	0.86	**0.85**	1.32
Legs FFM [kg]	Male	**0.89**	0.76	**0.84**	1.61
	Female	**0.83**	0.71	**0.85**	1.26

Table 4.5 shows *p*-values for a paired *t*-test performed on three pairs of body composition measurement sets: DXA retrials, test-retest of our method, and our method against DXA. T1 vs DXA1 tested the accuracy of our method (T1) against the accepted ground truth (DXA1).

Although a few tests produced *p*-values below a single-test critical value of 0.05, none were below the Bonferroni corrected critical *p*-value of 0.004. Importantly, total body fat and lean mass along with percent fat all greatly exceeded the individual significance level of 0.05. Thus, the mean differences between retrials and between our method and the DXA measured composition variables were not statistically significantly different from zero.

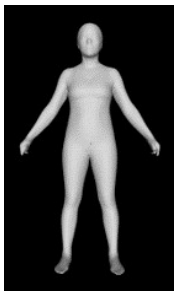
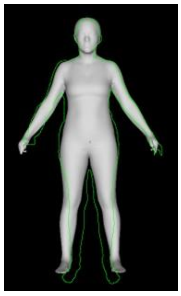
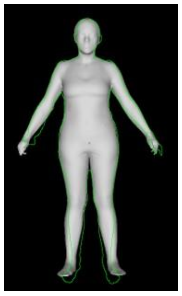
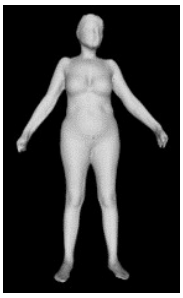
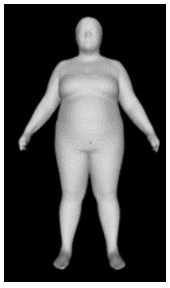
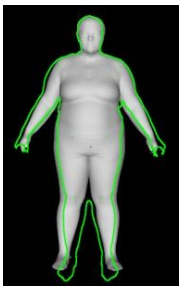
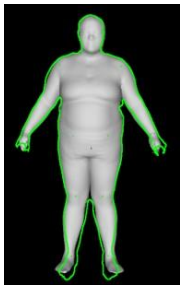
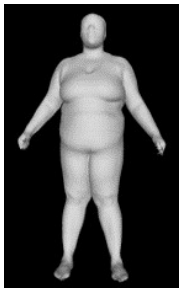
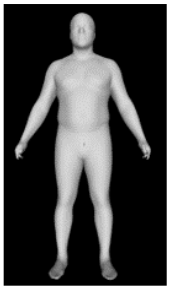
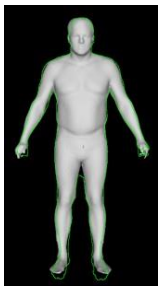
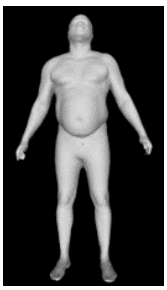
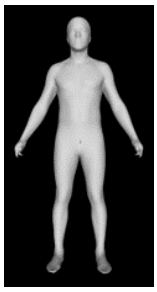
Table 4.5. *p*-values for paired *t*-tests. $p < 0.004$ was used to test for statistically significant differences. DXA1 and DXA2 are the two DXA measurements, T1 and T2 are the two trials of our method on separate sets of photographs of the same individuals. T1 was the test set that we treated as the result trial in Table 4.3, 4.4, and 4.6.

Output Variable	Gender	DXA1 vs DXA2	T1 vs T2	T1 vs DXA1
Fat Mass [kg]	Male	0.21	0.50	0.35
	Female	0.58	0.30	0.45
FFM [kg]	Male	0.30	0.50	0.35
	Female	0.68	0.30	0.45
% Fat	Male	0.17	0.71	0.46
	Female	0.44	0.78	0.50
FMI [kg/m^2]	Male	0.18	0.40	0.35
	Female	0.52	0.28	0.53
FFMI [kg/m^2]	Male	0.30	0.40	0.35
	Female	0.60	0.28	0.53
Visceral Fat Mass [kg]	Male	0.10	0.51	0.20
	Female	0.13	0.13	0.03
Trunk Fat Mass [kg]	Male	0.74	0.56	0.76
	Female	0.24	0.11	0.62
Trunk FFM [kg]	Male	0.76	0.82	0.01
	Female	0.10	0.19	0.03
Arms Fat Mass [kg]	Male	0.30	0.59	0.02
	Female	0.23	0.66	0.50
Arms FFM [kg]	Male	0.86	0.54	0.07
	Female	0.21	0.22	0.87
Legs Fat Mass [kg]	Male	0.06	0.49	0.18
	Female	0.43	0.10	0.35
Legs FFM [kg]	Male	0.02	0.38	0.92
	Female	0.48	0.22	0.31

We show some examples of our method on individual subjects from the test set in Table 4.6. From left to right, we show the input 2D photo, the initial shape as predicted by input height and weight, the extracted silhouette from the 2D photo aligned with the initial shape, the optimal

converged shape aligned with the same silhouette, and the 3D scan. The 3D scan cannot be regarded as explicitly ground truth because subjects were not scanned in the exact same pose or location as the 2D photo, but it shows the level of detail that can be expected of an actual optical scanner compared to our prediction method. On individual examples, percent fat prediction accuracy ranged from <1% to as high as 6%. Because our method was not able to factor in depth cues such as the shading of the torso region, indicating either a convex abdomen or a lean figure with defined musculature, many of the higher error examples tended to have proportions that were not well predicted by the silhouette alone. Subjects that had average waist breadth but were deep in the sagittal plane tended to be underpredicted in fat mass and percent fat, while subjects that were wide shouldered and muscular while being somewhat lean tended to be overpredicted.

Table 4.6. Visualized results. Results viewed under camera projection π. Columns in order show: a) The camera image input b) the seed shape defined by the known height and weight c) the seed shape optimized for the rigid transformation to align best to the joint positions d) the final optimized shape deformation and transformation e) the ground truth scan. Note that participants are not scanned in the exact same position they were photographed in. f) Predicted and ground truth % fat values from the direct regression method, picked for consistency.

a) Input	b) HW	c) Init	d) Final	e) Scan	f) Values
					p: 33.36% gt: 32.6%
					p: 37.36% gt: 41.36%
					p: 21.68% gt: 28.02%
					p: 16.27% gt: 11.85%

4.4 Discussion

In the current study we demonstrated that composition of a human body can be inferred from a 2D silhouette taken from an RGB image given known height and weight. Previous publications have presented work in both computer vision and medical research that parallel parts of our project, but to the best of our knowledge, no other publication has gone from a single 2D image to body composition estimates using 3D shape prediction as an intermediate. Guan *et al.* [43] presented an early method of mapping a 3D human shape space to a single monocular RGB image. This method has the advantage of modeling pose variation and shading, which ours does not, but there is no subsequent mapping to clinical metrics. Bogo *et al.* [11] used a more advanced posable shape model, the skinned multi-person linear model (SMPL), to estimate a 3D shape from arbitrary poses, but the actual 2D to 3D mapping was based solely on joint projections without silhouette fitting, resulting in very coarse fits. Using Shape Up! 3D optical depth scans, we had previously derived a PCA model of body shape and related those PCA vectors to criterion body composition measures from DXA. Here we extend that work using only the 2D photograph, the camera focal length, and the subject's height and weight to predict the PCA parameterized body shape in cases where 3D depth scans are not available. We estimated the composition of these predicted body shapes using linear regression from PCA parameters to criterion measures derived from DXA. Affuso *et al.* [2] presented a method that uses both front and side images to generate features for a support vector regression that achieved an R^2 of 0.78 for percent fat across all adults in 3-fold cross validation. Our method achieved R^2 of 0.73 and 0.74 on randomly held-out sets of males and females respectively using only a single frontal image, with 5-fold cross validation results showing 0.68 and 0.77 respectively. Unlike this work, we separated our experiment by males and females and did not include children. Farina *et al.*

[29] presented a method that predicts fat mass from a single side-profile photograph. We believe our method is more robust due to the larger sample size (152 males, 194 females compared to 54 males, 63 females) and verification on a separate held-out set. The R^2 values greater than 0.95 in Farina *et al.* appear to be reported on the training set, leaving the generalizability of this method uncertain. Furthermore, the methods are not reproducible because they depend on an undisclosed, proprietary body segmentation algorithm as part of their training procedure. More recently, Lu *et al.* [70] predicted body fat directly from a 3D body mesh with machine learning methods. This method was trained on a limited sample of 50 adult males and makes the prediction on a 3D scan with a minimum RMSE on percent fat of 3.17. This result was reported using the leave-one-out method, where training was performed on n-1 samples and testing done on just one. Our method achieved comparable RMSE of 3.9 and 3.3 on males and females respectively, using one consistent model on a randomly selected held-out test set and only requiring a 2D photo, height, and weight as input.

Although effective, our method could be improved by going beyond silhouettes and including shading information in the input images. Guan *et al*. [43] demonstrated a method that optimizes geometry to explain the observed shading over the surface of the subject with a single light source. Although the shading model was not based on human skin reflectance models, it was shown to improve the fit to the silhouette and pose of images that feature human participants in differing poses. Including a shading term in our optimization could produce more accurate 3D reconstructions, as we currently only use the silhouette pixels and ignore the interior pixel information. While Guan *et al*. only used the shading term to enhance the geometric similarity between predicted shapes and ground truth geometry, this additional detail may enhance the accuracy of our body composition prediction.

Our shape models in this work were not constructed to explicitly handle pose-dependent shape variation. A posable model with joint angle parameters would allow pose to be optimized separately from "intrinsic" body shape, as in Guan *et al*. and *Bogo et al*. Although our pose space is constrained to only frontal images of participants standing on footprints with handlebars, the amount of variation between people of different sizes fixing their extremities to static points in space is substantial enough to affect the PCA formulation. Differences in the lean, leg spread, and arm spread were misconstrued as fundamental body shape variations by our PCA model. This pose variation causes fitting issues when differences in leg position cannot be isolated from height or girth, or conversely when limbs cannot be matched without compromising the accuracy of the torso alignment. Building our PCA model on top of a posable model such as SMPL will allow us to isolate pose from shape and theoretically produce better reconstructions and results.

In the absence of a posable model that can account for variations in arm and leg angles, we created a demo of a smartphone app that facilitates the collection of 2D image data in the wild for non-professionals. Our app projected a stick figure to the camera screen of the phone, indicating to the photographer how the subject should be aligned in frame to best fit the expected pose of the PCA space. Silhouette accuracy is extremely important and requires near pixel accurate segmentation of the human body, ideally clothed with no more than a skintight bathing suit equivalent. While this is easy to accomplish with standard methods against a green screen background, reliable automatic segmentation against arbitrary real-world backgrounds such as the one shown in requires more advanced computer vision methods that are beyond the scope of this work. Fig. 4.5 shows a screenshot from a smartphone interface illustrating an in-the-wild use case for this algorithm taken from an app developed in collaboration with Wolox, an Argentinian software contractor. On-device computation was not resolved at the time of this writing.

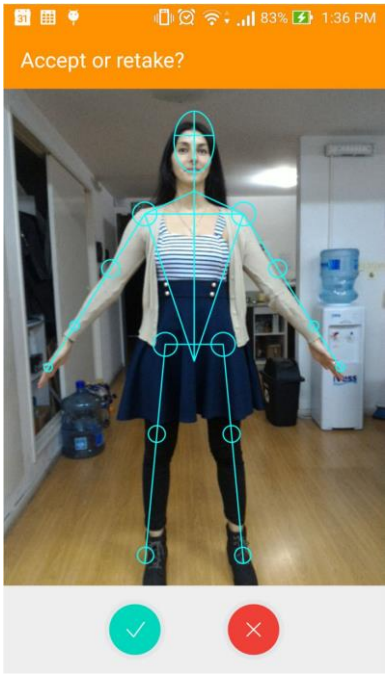


Fig. 4.5. Smartphone app screenshot indicating pose alignment landmarks.

Our mapping function **M** was assumed to be linear and derived from a simple least-squares regression. It is possible that a more ideal function can be more complex, such as a polynomial kernel or a neural network function, an area for future work. Our initial experiments using fully connected networks were unsuccessful as the predictions were very quickly overfitted.

As with all machine learning based methods, our predictive power is strongly based on the quality and variety of training data. Additional training data should add to the robustness and consistency of the model.

Finally, hyperparameters from Table 4.1 were tuned by trial and error on a single randomly chosen individual. Ideally, we would tune our hyperparameters on a third, held-out set that is not part of either the training or test set to tune our hyperparameters on (the validation

set). Due to the low subject count, we did not further fragment our subject set to robustly optimize the many hyperparameters.

4.5 Conclusion

Frontal body silhouette provides substantial information on the body composition of a subject in the absence of other views or additional imaging information such as depth. This method requires minimal data inputs and can be employed in a much wider scope of practice than traditional medical imaging methods. Given the clinical significance of both total and regional body adiposity for predicting metabolic disease and mortality risk, our method may be an impactful first step in propagating low-cost early screenings that can be performed outside of medical clinics by non-professionals for patients that may not warrant or cannot afford a clinical evaluation and gold-standard medical imaging. Future implementations of this project can deploy this algorithm to mobile devices, making it an attractive low-cost approximation of advanced imaging in more remote areas with lower rates of medical access.

4.6 Appendix

4.6.1 Selfie with shading

Fitting a projected body shape to only the silhouette of a body discards the majority of the information contained in a 2D photo. The shading of the body surface inside the silhouette boundary provides visual cues about the convexity or concavity of the body geometry that cannot be captured in a single frontal view. We experimented with sampling the interior pixel intensities

and pairing it to a shading error function like that specified in the photometric stereo experiment of Section 3.2 to optimize for a body shape that jointly fits the silhouette outline and the observed shading on the image.

The shading error function was a Phong [88] model with the normal vector at a vertex expressed as a function of the PCA parameters **w**. The normal vector of a triangle can be expressed as the cross product of two of its edges, each of which is determined by the positions of its endpoints. The position of the *i*th point parameterized by **w** is the *i*th row of the shape vector [**Aw** + **μ**]. Thus, the normal at the *i*th vertex $\mathbf{n}_i$ can be written as the area-weighted average of the normals of its neighboring faces, each of which is a cross product parameterized by **w**. (Fig. 4.6)

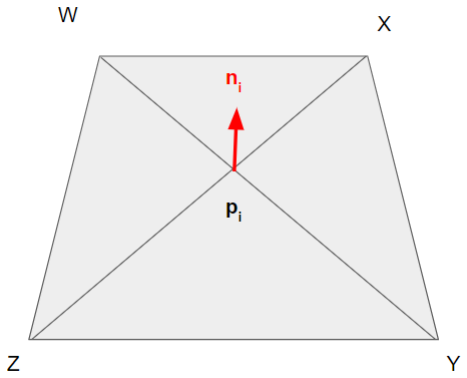


Figure. 4.6 The normal n_i at p_i is the weighted average of the normals of its neighboring triangles; the normal of each triangle is a cross product of its vertices i.e. (p_i, Z) × (p_i, Y). The position of these vertices can be parameterized in terms of the PCA parameters *w*.

This problem was not well posed as deforming the 3D shape model changed which pixel selected vertices projected to, violating the paired pixel-vertex assumptions of the shading cost function. A more correct formulation is fixing vertex-pixel pairs at the start of the optimization and only allowing each vertex to translate along the viewing ray. The parameter counts in the

second formulation scaled linearly with the number of pixels in the image, densely sampled, rather than remaining constant with the number of PCA parameters; furthermore, each vertex was dependent on the position of its neighbors and required joint optimization rather than parallelized independent solutions. This problem was too large to optimize, and this experiment direction was abandoned in favor of deep network solutions that targeted body composition as outcomes without explicitly estimating shading.

4.6.2 End-to-end deep network prediction from silhouettes

Rather than explicitly optimizing the difference between rendered 3D body shape and observed pixel intensity with shading cost functions, we experimented with using a deep convolutional neural net (CNN) to directly regress the target body composition variables from images end-to-end. RGB images containing skin and underwear textures created an overfitted model with no generalization to test data; this is possibly due to the low number of images (a few hundred) relative to the typical dataset size used to train 2D image CNNs (2-3 orders of magnitude higher). Instead, we opted to take binary masks of the human body 2D image as input and use a side view in the sagittal plane to substitute for depth ambiguities inherent to monocular solutions. Images were automatically converted to binary segmentation masks using the GHUM3D model from the MediaPipe model (https://github.com/google/mediapipe/blob/master/docs/solutions/pose.md) before training or testing to eliminate variance caused by texture or lighting.

Front and side binary masks were fed into two parallel subnetworks with independent weights but identical architectures. The two terminal feature vector outputs of each parallel subnetwork were concatenated together along with known height and weight measurements of the photographed subject and fed to a fully connected regression layer to finally estimate target

body composition variables. Additional layers and nonlinearities did not improve the results. Fig. 4.7 shows a diagram of our deep convolutional regression network.

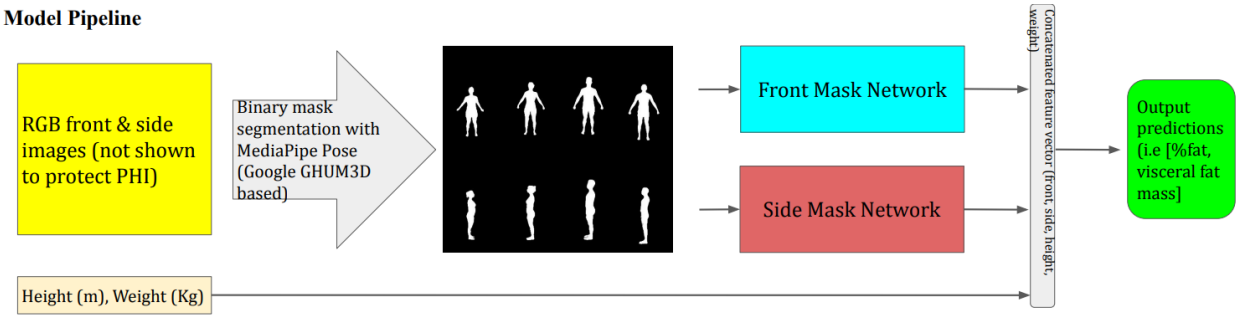


Figure 4.7. Diagram of network architecture. The front mask and side mask networks were identical pretrained DenseNet-121 networks initialized on ImageNet features. The final layer was a fully connected regression layer.

We split the dataset into 643 pairs of training images: 173 validation, and 106 test. The subnetworks were initialized to DenseNet-121 architectures with pretrained ImageNet classification weights. The original DenseNet terminal classification layer was removed and replaced with a three-node continuous regression target calculating a L2 mean-squared-error (MSE) loss. Networks were trained for 512 epochs. Initial learning rate was set to 1e-4. We used the Adam optimized. Training images were augmented by applying the same random flip, rotation, and 224x224 pixel crop to each image pair.

Regression results are shown in Table 4.7. Although visceral fat in females performed better than the results presented in Chapter 4.3, male visceral fat prediction failed completely. Total fat mass and percent fat were comparable to but not strictly better than previous results.

Table 4.7. R^2 and RMSE for percent fat, visceral fat, and fat mass as predicted by a DenseNet-121 initialized two-view binary silhouette to body composition network.

Percent Fat	Visceral Fat	Fat Mass
M: 0.65 (4.25) F: 0.65 (3.96)	M: 0.01 (0.29) F: 0.63 (0.17)	M: 0.87 (3.90) F: 0.92 (3.53)

Additional training data beyond a few hundred may improve these results; we investigated using the binary silhouette model of Klarqvist et al. [56] as the initial pretrained weights instead of ImageNet features. The Klarqvist model was trained to regress visceral fat from segmentations of MRI images as opposed to classification of RGB images using over 40,000 images from the UK Biobank dataset. We did not continue this experiment due to IRB restrictions on obtaining the UK Biobank data and lack of public access to the Klarqvist model. We built on the concept of seeding a deep network with tens of thousands of adjacently similar data to mitigate low target training data availability to train a 3D mesh autoencoder network in Chapter 8.

5. A Device-Agnostic Shape Model for Automated Body Composition Estimates from 3D Optical Scans

Body composition estimates derived from 3D shape models in the last chapter and in its contemporary work were data limited by the number of templated 3D body meshes available with paired reference DXA measurements. Raw 3DO scans could not be directly assimilated into an ongoing PCA model as they contained unordered vertex data with differing connectivity and quantity both between and within the same scanning device. 3D scans had to be fit with a watertight, manifold human body template mesh to ensure topological consistency between all members of the dataset. The fitting process was a nonrigid surface-to-surface deformation between the template mesh and the target scan, guided by manually placed anatomical landmark correspondence assigned by a trained technician. This pipeline was prohibitively slow for large datasets; an automated method that did not require a human labor in the loop proportional to the size of the dataset could substantially improve existing models by more than doubling the number of available templated meshes. Furthermore, existing work restricted both shape templating and body composition analysis to scans collected on a single scanner. We proposed implementing an automated pipeline that used PCA-parameterized 3D shape deformation in lieu of manual landmark placement to register templated meshes to scans. Initializing a 3D PCA parameterized template shape based on known height and weight measurements of a scanned subject and deforming it in the principal component domain allowed us to reliably generate an anatomically consistent model that roughly aligned to the raw optical scan in dimensions and proportions. A markerless closest-surface nonrigid deformation completed the registration of the template mesh to the raw surface data. We showed this pipeline was capable of generalizing to

multiple scanning devices, accounting for slight variations in pose and geometry between them with high accuracy relative to prior work.

This chapter describes work originally published in Medical Physics:

[118] *Tian, IY, Wong, MC, Kennedy, S, et al. A device-agnostic shape model for automated body composition estimates from 3D optical scans. Medical Physics 49 (10), 49: 6395– 6409. 2022.*

5.1 Abstract

Background: Many predictors of morbidity caused by metabolic disease are associated with body shape. 3D optical (3DO) scanning captures body shape and has been shown to accurately and precisely predict body composition variables associated with mortality risk. 3DO is safer, less expensive, and more accessible than criterion body composition assessment methods such as dual-energy X-ray absorptiometry (DXA). However, 3DO scanning has not been standardized across manufacturers for pose, mesh resolution, and post processing methods.

Purpose: We introduce a scanner-agnostic algorithm that automatically fits a topologically consistent human mesh to 3DO scanned point clouds and predicts clinically important body metrics using a standardized body shape model. Our models transform raw scans captured by any 3DO scanner into fixed topology meshes with anatomical consistency, standardizing the outputs of 3DO scans across manufacturers and allowing for the use of common prediction models across scanning devices.

Methods: A fixed-topology body mesh template was automatically registered to 848 training scans from three different 3DO systems. Participants were between 18 and 89 years old with BMI ranging from 14 to 52 kg/m^2. Scans were registered by first performing a coarse nearest neighbor alignment between the template and the input scan with an anatomically constrained PCA domain

deformation using a device and gender specific bootstrap basis trained on 70 seed scans each. The template mesh was then optimized to fit the target with a smooth per-vertex surface-to-surface deformation. A combined unified PCA model was created from the superset of all automatically fit training scans including all three devices. Body composition predictions to DXA measurements were learned from the training mesh PCA coefficients using linear regression. Using this final unified model, we tested the accuracy of our body composition models on a withheld sample of 562 scans by fitting a PCA parameterized template mesh to each raw scan and predicting the expected body composition metrics from the principal components using the learned regression model.

Results: We achieved coefficients of determination (R^2) above 0.8 on all nine fat and lean predictions except female visceral fat (0.77). R^2 was as high as 0.94 (total fat and lean, trunk fat) and all root-mean-squared errors (RMSE) were below 3.0 kg. All predicted body composition variables were not significantly different from reference DXA measurements except for visceral fat and female trunk fat. Repeatability precision as measured by the coefficient of variation (%CV) was around 2-3x worse than DXA precision, with visceral fat %CV below 2x DXA %CV and female total fat mass at 5x.

Conclusions: Our method provides an accurate, automated, and scanner agnostic framework for standardizing 3DO scans and a low cost, radiation-free alternative to criterion radiology imaging for body composition analysis. We published a web-app version of this work at https://shapeup.shepherdresearchlab.org/3do-bodycomp-analyzer/ that accepts mesh file uploads and returns templated meshes with body composition predictions for demo purposes.

Keywords:

Body composition, 3D Scanning, Obesity, Regional composition, Dual X-ray absorptiometry, Principal component analysis, Linear regression, Fixed topology mesh

5.2 Introduction

Metabolic syndrome is strongly correlated with the leading causes of death in the US and the world. [51] Many clinical studies have shown the importance of regional body composition as a predictor for metabolic disease risk and increased mortality even when controlling for total body variables such as weight and body mass index (BMI). Goodpaster *et al*. [41] showed that males in a normal BMI range with high visceral fat mass were twice as likely to have metabolic syndrome that can lead to higher risk of heart attack and stroke. Wilson *et al*. [133] showed that high trunk-to-leg volume ratios predict greater diabetes risk, with the highest quintile of the population 6.8 times as likely to become diabetic. Zhang *et al*. [144] demonstrated a correlation between abdominal obesity and death from cancer over a 16-year longitudinal study, with a 63% increase in mortality risk in the highest quintile. Although the criterion methods for measuring body composition and metabolic risk factors reside mainly in radiology facilities, many mortality predicting variables have visually observable external effects on body shape. [81]

3D optical (3DO) scanners capture external body shape and are relatively inexpensive, noninvasive tools for gathering data that can predict or prevent metabolic disease. [6] However, algorithms that estimate body composition and metabolic risk factors from 3D scans are often proprietary and unique to a particular scanning system, such as the independent black boxed methods of Fit3D, Styku, and Size Stream. Even within data from a single system, the order, placement, and number of vertices in the file vary from scan to scan, making unified cross-compatible prediction algorithms based on surface geometry difficult to develop and verify. These

limitations undermine the accessibility of 3DO scanning as a universal clinical tool due to possible inconsistent interpretations depending on the scanning system deployed.

Previous works have attempted to bridge the divide between inconsistent system data and consistent translation to medically informative statistics. Ng *et al*. [83] fit a 3D human mesh template to raw scans from a single scanning system using the mesh deformation methods of Allen *et al*. [5] This work translated unorganized raw scans into topologically and anatomically consistent 60,001 vertex meshes (60k), which abstracted away the specific output format of the scanning device and allowed statistical methods such as principal component analysis (PCA) to be applied.

While this method in theory was not dependent on the input system, it was only benchmarked on a single system and did not scale well to large datasets due to the manual effort involved. Point correspondences between the standard template mesh and the raw scan required manual annotation on each raw scan to initialize the mesh correspondences and deformation. PCA models built on one scanning system did not generalize to scans from novel systems due to strict pose constraints inherent in the single scanning device. A scalable, device agnostic solution needs to allow for fast automatic mesh template fitting across multiple devices while showing high accuracy on body composition prediction in held-out test data.

The primary objective of this study was to develop an algorithm for automatically standardizing 3DO scans from multiple total body scanning systems captured in standard anatomical poses and predicting body composition metrics from the resulting standardized body meshes using a unified scanner agnostic model. While Ng *et al*. [83] directly applied the per-vertex deformation algorithm detailed in Allen *et al*. [5] with the assistance of labelled correspondences between template and raw scan, we first used a global deformation constrained by a bootstrap

principal component basis created with a subset of manually guided template fits (see Supplementary Methods) to enforce anatomical correspondence. We refined the surface-to-surface alignment with the same per-vertex deformation but with zero manually annotated markers as part of the optimization as anatomical and topological consistency were constrained by the principal component domain. We extended the manual fitting method of Ng *et al.* to automatically fit a standard template to 848 training scans from Systems 1-3.

We performed a cross-sectional study on a diverse sample of convenience that received metabolic status measures and 3DO scans on 4 different technologies, including one unseen technology that was not used in model training. A secondary aim of this study was to quantify the test-retest precision of body composition estimates using our automatic mesh templating method against DXA.

5.3 Methods

5.3.1 Study Design

We constructed a parameterized 3D body shape statistical model to both perform the 3D geometric surface registration and the subsequent body composition prediction from the standardized mesh template. Our method was trained and tested on adult participants from the Shape Up! Adults Study (NIH R01 DK109008) and were recruited in the Honolulu, HI area at the University of Hawaii at Manoa (UH), in the San Francisco, CA area at the University of California, San Francisco (UCSF), and in the Baton Rouge, LA area at Pennington Biomedical Research Center (PBRC) as described in Tian *et al.* [117] Recruitment was stratified by age (18–40, 40–60,

>60 yr), ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, Asian, and Native Hawaiian or Pacific Islander (NHOPI)), gender and BMI (<20, 20–25, 25–30, >30 kg/m^2).

Participants were excluded if they could not stand unassisted for two minutes or lie supine for ten minutes without movement, had metal objects in their body, or had major body-shape-altering procedures (e.g., liposuction, amputations, etc.). Female participants were excluded if pregnant or lactating. Written informed consent was obtained from each participant upon arrival and all procedures were approved by the Pennington Biomedical Research Center Institutional Review Board (IRB# 2016-053-PBRC), the UH Office of Research Compliance (CHS# 2017-01018), and the Human Research Protection Program Institutional Review Board at the University of California, San Francisco (IRB# 15-18066). The study is publicly listed on ClinicalTrials.gov as ID NCT03637855.

Ground truth total and compartmental body composition measurements were defined by DXA. We acquired duplicate whole-body scans of each participant on either a Hologic Horizon/A system (UCSF) or a Discovery/A system (PBRC and UHCC) (Hologic Inc., Marlborough, MA, USA). Participants were positioned and scanned according to each manufacturer's guidelines. All scans were analyzed at UHCC by a single certified technologist using Hologic Apex version 5.6 with the National Health and Nutrition Examination Survey (NHANES) Body Composition Analysis calibration option disabled. DXA systems quality control was performed by monitoring the weekly values of the Hologic Whole Body Phantom. Cross-calibration was checked between sites using a whole-body phantom scanned at each site. No cross-calibration adjustments were needed.[7]

Each participant was scanned in one or more 3DO systems pending availability at each recruiting location. We used four different 3DO system manufacturers across all sites: System 1

(Fit3D Proscanner 4.x, Fit3D Inc, Redwood City, CA, USA), System 2 (Styku S100 4.1, Styku LLC, Los Angeles, CA, USA), System 3 (Size Stream SS20, Size Stream, Cary, NC, USA) and System 4 (Naked Body Scanner, Naked Labs, Redwood City, CA, USA). Scans from different systems differed slightly in pose, although all were upright with straight elbows and knees in a neutral A-pose, with elbows and knees held in maximum extension and arms and legs abducted slightly away from the midline of the body, and differed significantly in vertex count, spanning three orders of magnitude from 4,000 to 400,000 points. Participants stood with arms and legs held straight and slightly away from the midline of the body, but the exact angles varied with the position of the handrails, foot marker placement (if present at all) and height of the subject. System 2 had more extreme arm position variance as there were no fixed handrails. In the case of System 4, the restricted field of view of the optical sensor created much more variance in the pose of participants, who had to stoop or bend their limbs to fit within the constraints of the system. Participants were scanned twice to gather a test-retest precision evaluation data set.

The number of raw scans included in this study was as follows: for males, System 1: 241, System 2: 216, System 3: 116, and System 4: 59; for females, System 1: 294, System 2: 276, System 3: 135, and System 4: 73. Some participants were represented by more than one system. These systems spanned a wide range of output resolution, with some System 3 scans having as few as 4,000 vertices and the largest System 1 scans having over 400,000. Scanner properties are summarized in Table 5.1. We standardized all scans to System 1 reference coordinates shown in Fig. 5.1. We trained our PCA model on a subset of the first three systems and treated System 4 as a completely unseen validation method with extreme pose variation. For each training system, the participants for each gender were split into three sets: a 70-member bootstrap set, and the remainder split 50/50 into train and test.

Table 5.1. File properties and scan times of 3DO scanners in study.

	Vertices	**File Size**	**Scan Time**
Fit3D Proscanner (System 1)	300-500k	40-60 MB	40 sec
Styku S100 (System 2)	30-60k	4-7 MB	20 sec
Size Stream SS20 (System 3)	4 – 25k	0.5 – 3 MB	3 sec
Naked Body Scanner (System 4)	100k	12 MB	5 sec

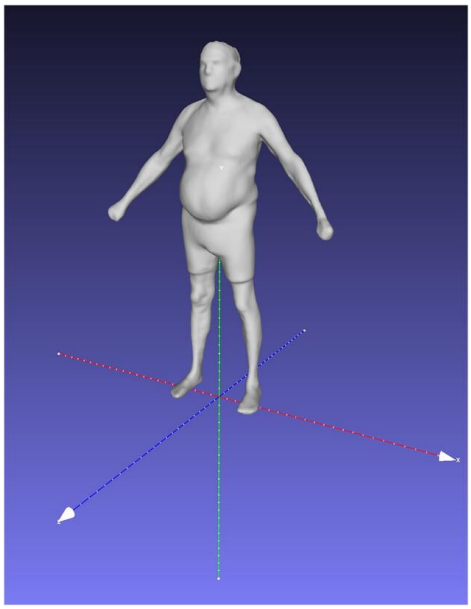


Figure 5.1. Diagram of rotation/translation relative to the origin and standard basis. Scale is metric. This reference frame was the factory default for System 1, but in principle any common transformation could be used if all scans aligned at the feet

5.3.2 Resolving Data Inconsistencies using Standardized Mesh Templates

A key advancement on previous work predicting body composition from unorganized 3DO scans was substituting the manual annotation of anatomical landmarks for a coarse initial deformation in principal component domain.

To fit our 60k standardized mesh template to the scan point cloud, we iteratively minimized closest-point Euclidean distances between vertices in our template and the input scan in two phases. First, we constructed six PCA bases using manually fit [83] template meshes from the bootstrap set to initialize the automatic fitting algorithm. There were three scanning systems, and PCA spaces were separated by gender. Each bootstrap set consisted of 70 fitted meshes. We performed a global mesh deformation of our template body mesh in the dimension-reduced bootstrap principal component domain corresponding to the mesh's gender and scanning system to constrain surface smoothness and anatomical consistency as described in Tian *et al.* [117] and achieve a coarse shape fit. Closest point pairs were determined with a nearest-neighbor algorithm and pairwise distances were minimized with a linear least-squares solver. This process was repeated iteratively until the difference between iterations fell below the convergence tolerance hyperparameter. Second, we refined the surface alignment between the deformed template and the scan mesh by optimizing for per-vertex 3D rigid transformations to minimize surface-to-surface distance as shown in Allen *et al.* to produce the final fit. This step was analogous to the mesh fitting in Ng *et al.* but was fully automatic with no annotated anatomical landmarks, as deformation in the principal component domain constrained the coarse fit to be topologically and anatomically consistent with a human body shape. We first fit the 420 scans from all bootstrap sets with the initial manually bootstrapped PCA models. Mathematical details are provided in the Supplementary Materials.

5.3.3 Verifying Anatomical and Topological Consistency of Automatic Template Fits

We compared our automatic markerless body shape fits to the manually fit equivalents from Ng *et al.* to check for topological and anatomical consistency between the two shape fitting methods. The 60k mesh template had 74 anatomical markers placed at anthropometric positions originally defined by the CAESAR study [97] by a trained medical professional. The original landmarks corresponded to skeletal features determined by palpation on a live subject. Our manually annotated landmarks were placed in a 3D model viewer by a trained technician to correspond to the physically palpated ones.

We can recover the 3D position of these landmarks in any shape fit with the 60k template by querying the coordinates of the neighboring vertices that define the placement of each landmark. We compared the landmarks recovered from automatic template fits to the equivalent manually clicked landmarks, using the latter as the gold standard, and compared their precision with manually annotated equivalents to the benchmarked precision of repeat manual point placement on the same scan.

5.3.4 Expansion of Body Shape Model Using Markerless Automated Fitting

We repeated building the initial six separate bootstrap PCA models but with the resulting automatically fit mesh templates and compared their body composition prediction accuracy against the manually fit baselines to ensure there was no loss of accuracy due to dropping manual annotation from the pipeline. Each model was incompatible with scans from a different system. This incompatibility was due to the strict hand and feet endpoints imposed by each system, which introduced minor pose variations but caused meshes scanned with one system to be

unrepresentable in the PCA domain of another. Fig 5.2 shows an example of a System 2 scan fit using a System 1 PCA male model.

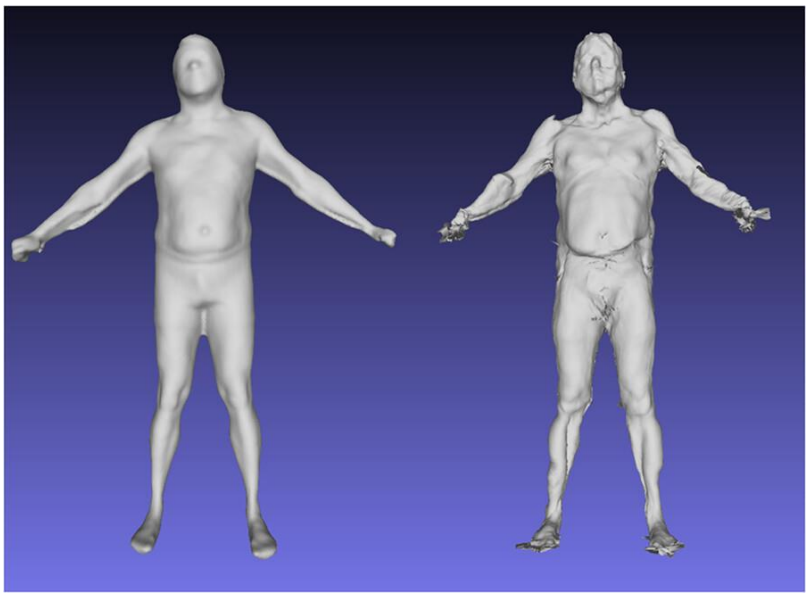


Figure 5.2. Example of Failure of Single Device Model. Raw System 2 scan on the left, and fit attempt with System 1 only model on the right. The pose difference between the two systems (System 2 has no fixed hand position) makes models trained on a single system nongeneralizable to other systems. This demonstrates the need for a unified model that can interpolate the pose differences between different systems.

Our goal was to train a single PCA model that could automatically fit templates to raw scans from any system and predict accurate body composition from the standardized fit. We used the second automatically fit bootstrap model for each gender and system combination to fit all the raw scans from their respective training sets. These coarse PCA fits were then refined with surface-to-surface alignment as previously described. The refined fits from all systems were grouped into one training set per gender.

We constructed the final unified PCA model by merging the refined fits of all training data for all scanning systems into one training set per gender. The combined training data included the

training set and the bootstrap set, which was a specific subset of the training data. This system-agnostic model included 391 males and 457 females. The unified model was used to fit raw scans from the held-out test set and predict their body composition.

5.3.5 Predicting Body Composition Using Scanner Agnostic Shape Space

We solved for a linear mapping between PCA basis coordinates and body composition by performing linear regression between the principal components of the automatic template fits to training scans and their corresponding criterion body composition measurements taken with DXA as described in Tian *et al*. Linear models for predicting body composition as a function of principal component basis weights were solved with least squares regression between the principal components and the ground truth DXA variables. For a regression of the form $\mathbf{Mx} = \mathbf{y}$, we augmented the vector $\mathbf{x}$ with the value 1 to allow for non-zero intercepts.

We reported all prediction accuracies on test meshes that were not used in PCA model construction or regression training. Statistical significance for body composition prediction was computed with a paired two-sided *t*-test. Body composition prediction from 60k mesh templates were paired with the DXA derived gold standard measurement. The test was successful if the null hypothesis could not be rejected, i.e., the difference between DXA and our prediction was not significantly different from 0. The Bonferroni correction for the *p*-value was 0.004. n= 182 and 248 total test set meshes for males and females, respectively, with an additional separate 59 and 73 validation meshes from System 4. Mathematical details are written in the Supplementary Materials.

5.3.6 Algorithm Workflow Summary

Training Procedure:

START: 70 male and 70 female bootstrap scans from each of Systems 1 – 3. (420 total)

1. Manually assign anatomical correspondences between 60k template and bootstrap scans. Perform marker guided surface to surface deformation to make 6 bootstrap template sets.
2. For each device AND gender combination (6 total System # + M/F combinations), make an independent 70-member bootstrap basis with PCA.
3. For each bootstrap scan, deform the 60k template to satisfy nearest neighbor alignment in its gender & scanner specific PCA bootstrap domain. See Supplementary Materials.
4. Starting with coarse alignments from step 3, smoothly deform each mesh to register to target scan with per-vertex surface-to-surface alignment (Allen *et al.* [5])
5. Repeat step 2 with the results from step 4. Verify marker alignment with manual placements.
6. Repeat step 3-4 using the PCA model from step 5 to fit all training scans.
7. Merge all refined fits from step 6 into a single superset per gender consisting of fitted training meshes from all three scanning devices. Perform PCA on this fixed topology set to get the unified PCA basis. (1 male, 1 female)
8. Project all refined fits from step 4 onto the PCA basis from step 7 to get PC basis coordinates for each training mesh. Learn per gender linear regressions from PC coordinates to DXA measurements. (Tian *et al.* [117])

Testing Procedure:

For any NEW test scan (any device):

1. Fit a 60k mesh template using PCA initialization as described in training step 3 & 4, using the PCA superset model from training step 7 as the coarse deformation prior.
2. Project the refined test mesh fit onto the unified PCA basis as in training step 8. Use regression matrix learned in training step 8 to estimate DXA measurements.

5.4 Results

Participant characteristics for the training and test sets are shown in Table 5.2. There were 1278 scans from three scanning devices randomly divided into training and testing. We reserved 70 manually fitted scans of each gender from the training set of each of the three systems to bootstrap the shape model. There was no significant difference between the means of training and test data.

Table 5.2. Population statistics for training and test scans. Plus-minus values are standard deviation. p-value of 0.004 was considered significant after Bonferroni correction. All metrics were not significantly different between test and train except for female height, denoted by the asterisk, which is not a predicted measurement. Body composition measurements were taken from DXA.

	Male					
	Train (N = 391)			Test (N = 241)		
	Mean ± SD	Min	Max	Mean ± SD	Min	Max
Age (Years)	44.92 ± 16.06	18	79	44.07 ± 16.15	18	79
Height (m)	1.76 ± 0.08	1.51	2.02	1.75 ± 0.07	1.55	1.91
Mass (kg)	88.42 ± 21.34	40.74	172.4	84.25 ± 18.91	40.77	135.58
BMI	28.44 ± 6.21	16.52	52.55	27.44 ± 5.55	16.96	45.82
Percent Fat	22.66 ± 6.86	9.03	47.69	22.07 ± 6.71	9.03	38.58
Lean Mass (kg)	67.48 ± 12.99	33.95	107.82	64.87 ± 11.64	33.95	93.1
Fat Mass (kg)	20.94 ± 10.67	5.01	66.48	19.39 ± 9.47	5.13	45.94
Visceral Fat (kg)	0.49 ± 0.27	0.16	1.64	0.5 ± 0.32	0.16	1.64
Leg Lean (kg)	10.97 ± 2.28	5.43	18.95	10.48 ± 1.99	5.43	14.74
Arm Lean (kg)	4.46 ± 1.05	2.05	8.33	4.26 ± 0.98	2.05	7.38
Trunk Lean (kg)	32.36 ± 6.4	15.95	51.16	31.24 ± 5,82	15.95	48.02
Trunk Fat (kg)	10.5 ± 6.15	1.76	34.12	9.72 ± 5.71	1.97	26.25
Leg Fat (kg)	3.41 ± 1.71	0.89	11.86	3.12 ± 1.44	0.85	9.02
Arm Fat (kg)	1.25 ± 0.68	0.29	4.34	1.15 ± 0.61	0.29	3.72
FMI	6.74 ± 3.37	1.68	20.78	6.32 ± 3.01	1.91	15.49
FFMI	21.72 ± 3.53	14.18	35.65	21.16 ± 3.23	14.18	30.72

	Female					
	Train (N = 457)			Test (N = 321)		
	Mean ± SD	Min	Max	Mean ± SD	Min	Max
Age (Years)	46.24 ± 16.13	18	89	47.53 ± 16.71	18	89
Height (m)	1.62 ± 0.07	1.44	1.80	1.61 ± 0.07 *	1.44	1.76
Mass (kg)	72.43 ± 20.93	35.44	153.05	69.39 ± 19.60	35.44	153.05
BMI	27.48 ± 7.65	14.16	51.86	26.81 ± 7.05	14.16	51.86
Percent Fat	34.06 ± 7.88	12.63	53.3	33.78 ± 7.44	17.18	53.3
Lean Mass (kg)	46.49 ± 9.48	28.56	80.37	44.91 ± 9.42	28.56	80.37
Fat Mass (kg)	25.85 ± 12.72	6.3	72.68	24.39 ± 11.38	6.88	72.68
Visceral Fat (kg)	0.45 ± 0.31	0.06	1.37	0.43 ± 0.3	0.05	1.22
Leg Lean (kg)	7.48 ± 1.73	4.25	14.03	7.21 ± 1.78	4.42	13.19
Arm Lean (kg)	2.42 ± 0.57	1.31	4.42	2.34 ± 0.58	1.44	4.63
Trunk Lean (kg)	23.13 ± 4.91	13.71	41.59	22.26 ± 4.72	13.71	41.14
Trunk Fat (kg)	11.94 ± 6.75	2.37	35.6	11.27 ± 6.21	2.48	35.6
Leg Fat (kg)	4.81 ± 2.23	1.23	12.92	4.54 ± 1.98	1.23	12.34
Arm Fat (kg)	1.67 ± 0.99	0.28	6.08	1.55 ± 0.84	0.4	6.08
FMI	9.82 ± 4.79	2.01	26.57	9.44 ± 4.24	2.75	24.68
FFMI	17.6 ± 3.28	11.43	29	17.33 ± 3.16	10.93	28.88

5.4.1 Landmark Consistency between Automatic and Manual Fits

To compare the topological consistency of our automatic fitting method to the manually guided fitting method of Ng *et al.*, [83] we compared the pairwise Euclidean distances between the 74 anatomical landmarks specified in the CAESAR dataset on both sets of template fits. The landmarks on the base template mesh were readily mapped to any automatically fit mesh as a function of neighboring vertices. Between topologically consistent template fits to the same raw scan, vertices and therefore landmarks should be analogous. For the 420 scans in the bootstrap set that had manually placed markers for reference, mean error was 11 mm with a standard deviation

of 12 mm. By comparison, the precision of a trained analyst manually assigning markers on the same mesh for a set of 15 males and 15 females was 8 mm with a standard deviation of 7 mm.

A subset of test set meshes had manual fit equivalents, although they were not used in training. To check the anatomical correspondences for held-out test meshes, we compared the markers of 95 male and 135 female test set meshes fit with our automatic template fitting across all three scanning systems to their manually placed counterparts. Test set meshes were fit with the final unified PCA space, which was composed of automatically template fit meshes from the training sets of all three systems. The mean distance between the recovered markers and the manually placed equivalents was 21 mm, with a standard deviation of 23 mm. Fig. 5.3 shows a visual comparison of manually placed markers to the markers recovered from the auto-templated fit.

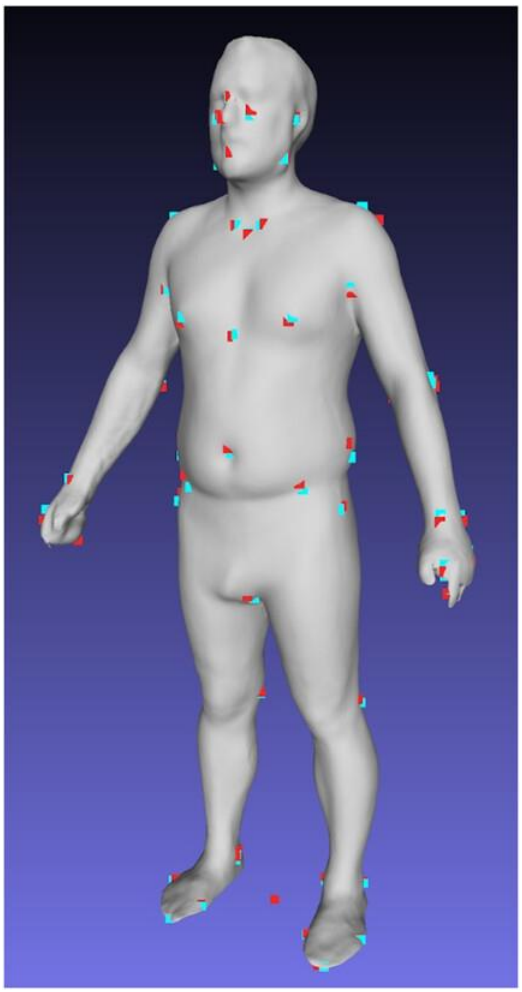


Figure 5.3. Topological consistency of auto template mesh. Auto fit (red) and manually placed (blue) anthropometric landmarks on a fitted test set mesh. The mean difference was 21 mm for 74 landmarks across 230 meshes. The red dot between the middle of the feet is the origin of the coordinate system.

Automatic mesh templating and standardization exhibits greater point placement error compared to repeat manual annotations, although surface-to-surface registration was visually equivalent. We show below that the slightly varied vertex distribution on a common surface did not decrease prediction accuracy of body composition prediction from body shape.

5.4.2 Accuracy of Markerless, Unified Shape Models against Manual Baseline

Fig 5.4 shows automatic template fitting using the unified superset model for coarse shape alignment on the heaviest female and tallest male in the test set. A visualization of our fitting method on raw scans of a single participant scanned on four different systems is shown in Fig 5.5.

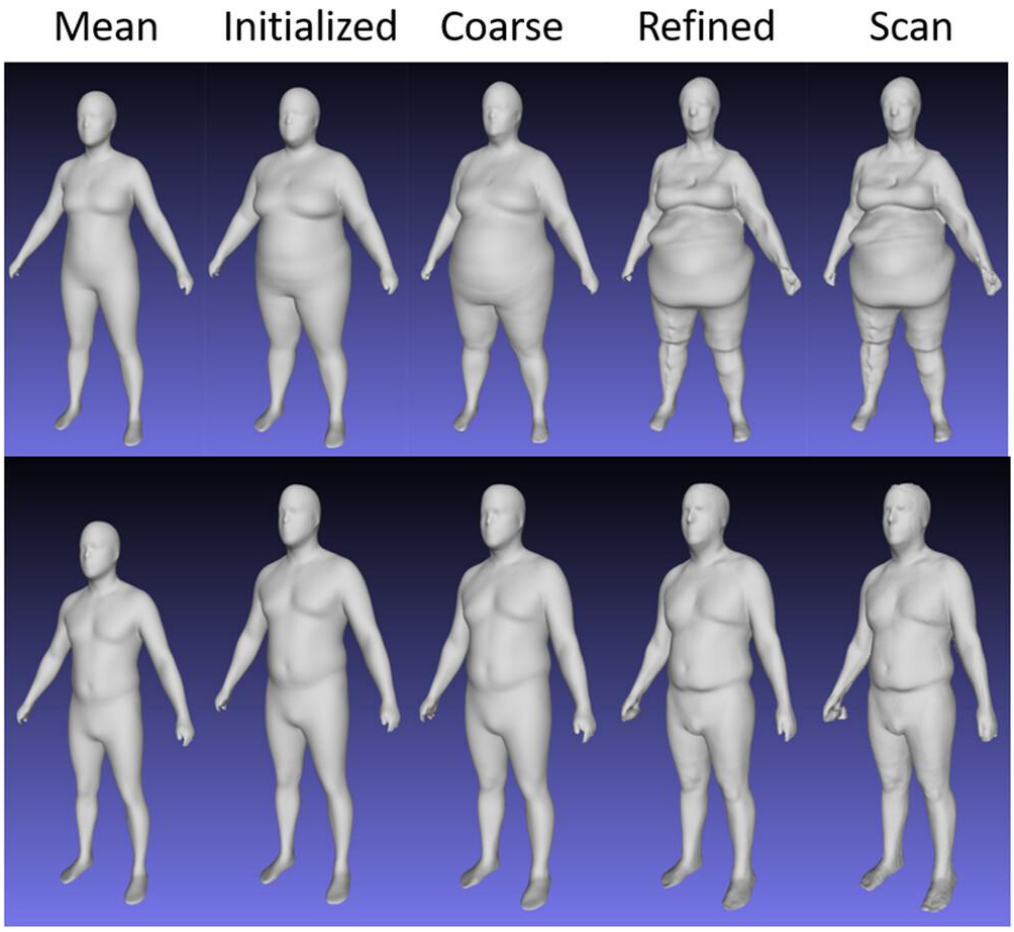


Figure 5.4. Shape fitting progression. Top: From left to right: The mean shape, the initialization shape, the coarse fit, the final fit, and the raw input from System 1. This example represents one of the worst-case scenarios as this individual was the heaviest female in our test set at 163.4 kg and is one of the farthest shapes from the mean. Bottom: Same progression as above but using the tallest male in the test set, at 190.8 cm. Note the large size increase after initializing the body shape with the known height and weight of the participant

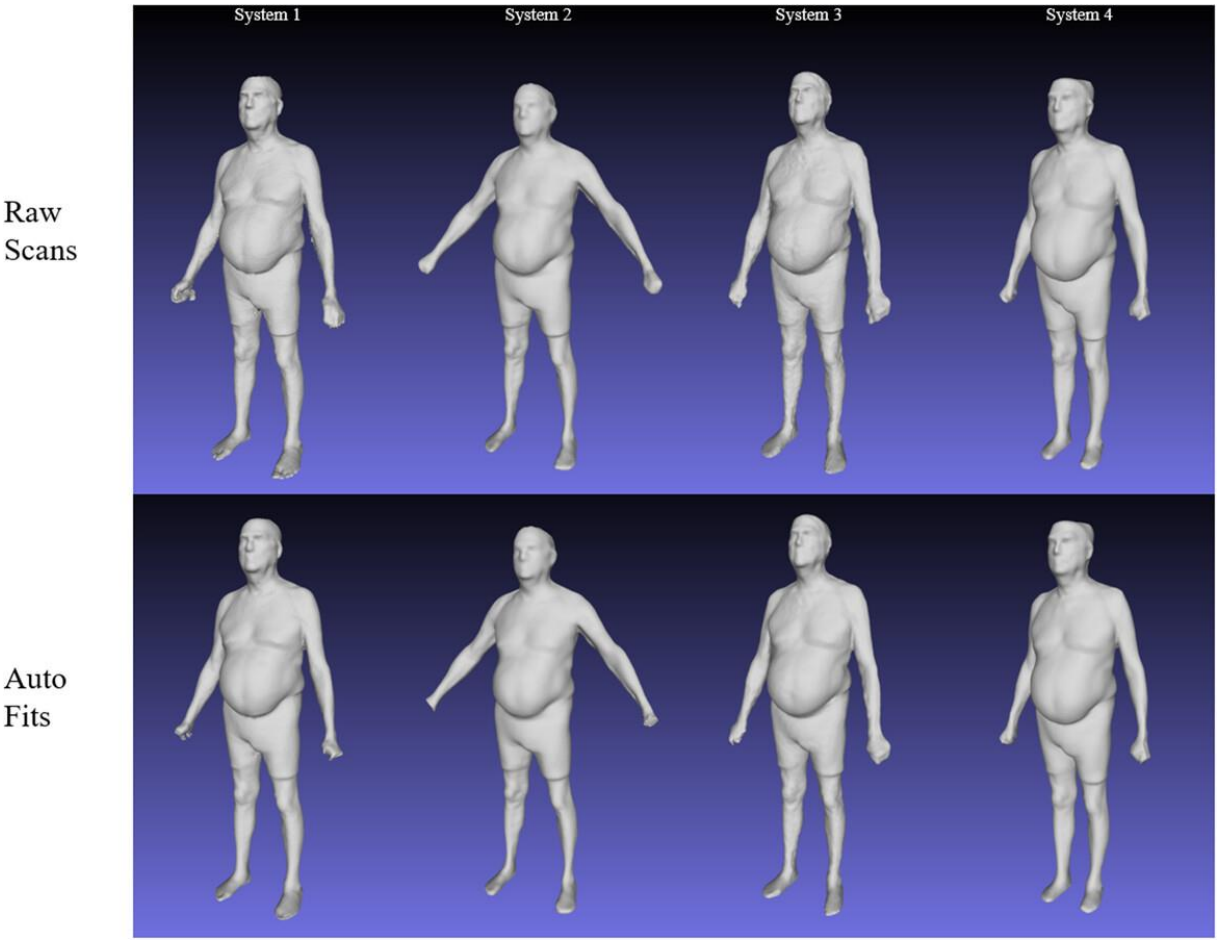


Figure 5.5. Examples of auto template fits on a single participant. One participant was represented in the test set for all four scanning systems. Raw scan input on the top and our automatic mesh fit on the bottom. Although visually similar, the raw scan had as high as 400 000 (System 1) unorganized vertices and as low as 25 000 (System 3). We fit a 60 k anatomically consistent mesh to each scan using d = 391 principal components

We recreated the same six manual bootstrap PCA models using automatically templated meshes to isolate the effect of automatic template fitting on body composition prediction in the absence of additional training data. Fig 5.6 shows the difference in R^2 prediction on test data between automatically and manually fitted PCA models containing the same 70 bootstrap meshes for training. Most metrics gain predictive accuracy for both genders, with small increases in error of <0.1 for some measurements such as arm fat in females. R^2 values for the auto-templated models

increased by an average of 0.054 and 0.023 for males and females, respectively, averaged across all three systems and 12 composition metrics. Although the training meshes were the same, there was an overall increase in predictive performance. This boost can be attributed to the smoothing operation followed by markerless nearest-surface smooth alignment. The resulting fits from our method had fewer artifacts that were introduced as a byproduct of the manually initialized deformation, where the initial template mesh was very misaligned with the target scan and necessitated very large nonlinear transformations.

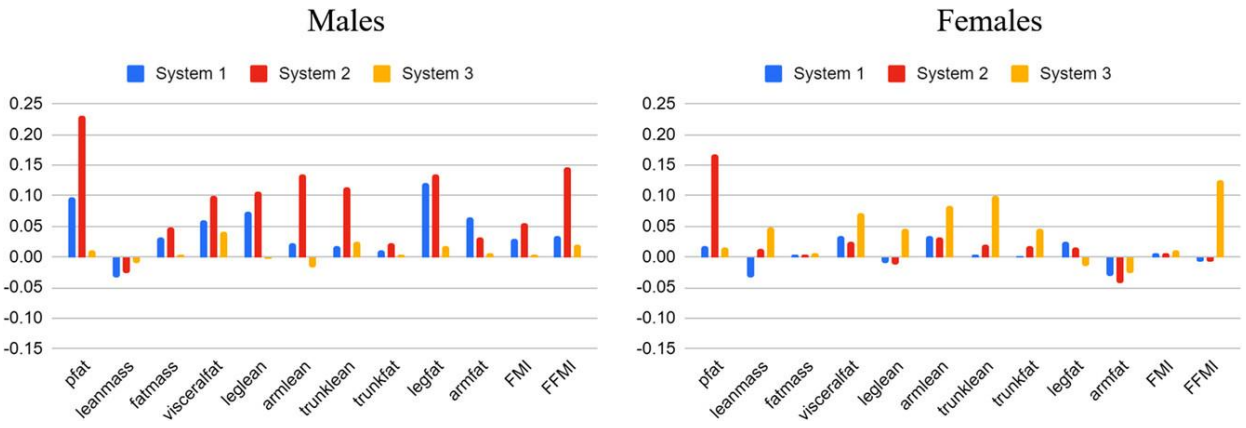


Figure 5.6. R^2 difference in body composition predictions between automatic and manual fit bootstrap models. R^2 differences between predictions from automatically and manually fit meshes are shown on held-out test set. Training set members were the same 70 scans for each system and gender combination for both automatic and manual fitting. The only variable changed was the method of template fitting to the raw scans. R^2 increased by 0.05 and 0.02 on average across males and females, respectively, indicating substituting markerless automatic fitting for manual fitting did not cause enough topological inconsistency to impact the resulting body composition regressions.

We present the unified superset model consisting of the union of all automatically fit template meshes from the training sets of all scanning systems per gender as the benchmark model in this work. Fig 5.7 shows the body composition prediction R^2 difference on system specific test scans between predictions learned from the unified superset PCA model and the 70 member

manually fit bootstrap of each scanning system. Our unified superset had many more training examples than the bootstrap model (391 for males and 457 for females) but had increased noise introduced by pose variance. Our results show that the unified superset model produced good geometric fits to withheld test scans and improved on the body composition predictions made by the initial manual model despite the inclusion of training scans that exhibited differing pose.

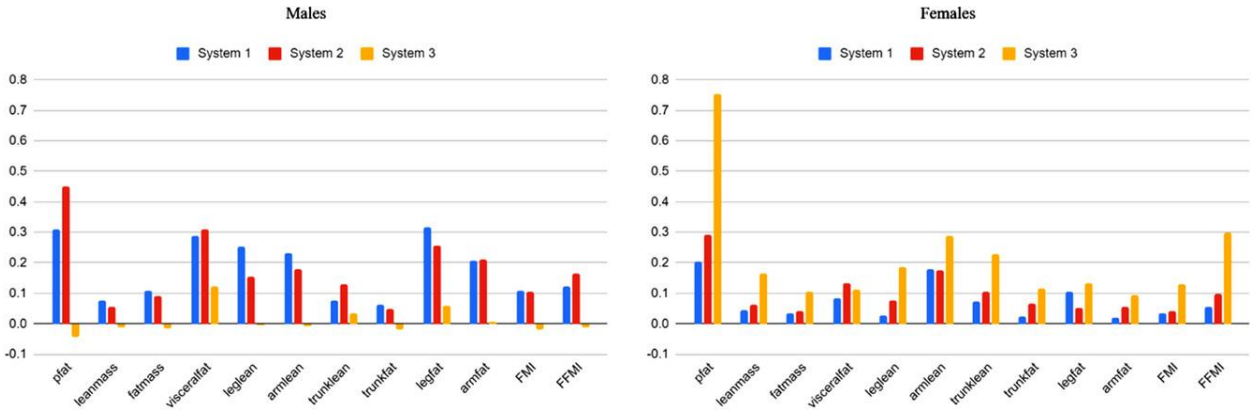


Figure 5.7. R^2 difference in body composition predictions between final and initial model. R^2 differences are shown on held-out test set, which was the same for both the final and initial models. Initial model was composed of 70 manually guided template fits per system and gender. Each system and gender combination was a separate principal component analysis (PCA) model, resulting in six models. Models could only be tested on the system they were trained on. Final models were unified across three scanning systems but split per gender. Final male and female models had 391 and 457 training scans, respectively, and all were automatically template fit with our method. The same model could be applied to test scans from all systems. Overall, prediction accuracy increased on test data despite using no manual fits in our final model and injecting noise in the body shape model due to including pose variance from multiple scanning systems in our unified model. Per-system composition prediction metrics are shown in Table S1

5.4.3 Body Composition Prediction Accuracy and Precision on Test Data from Multiple Systems

We automatically fit standardized mesh templates to test set scans from all three systems and used the projected PCA basis coordinates to predict body compositions from our regression matrices. Table 5.3 shows the R^2, RMSE, and *p*-values of the predictions on test scans using the

final unified model and all available PCA components (d = 391 for males and 457 for females) compared to DXA as reference. Percent fat was calculated as fat mass / scale weight, fat mass index (FMI) was calculated as fat mass / height^2, and fat free mass index (FFMI) was calculated as lean mass / height^2. Lean mass was calculated as weight - fat mass. All masses are reported in kilograms (kg). Scans were sourced from three input systems and varied between 4,000 and 400,000 vertices. All test scans were held out of PCA model training and composition regression. p-values were calculated with a paired t-test against respective DXA measurements to determine the presence of bias in our method. For 12 simultaneous measures, a $p < 0.05 / 12 = 0.004$ was considered significant, meaning the mean difference between our predictions and DXA was statistically significantly different from 0. Our bias check passed for every measure except visceral fat in both genders, and in trunk lean in females. Our results were comparable to the PC-only model of Ng *et al.*, but were reported on completely held out test data, tested on scans from two additional scanning systems, and required no manual intervention to initialize the fit. A breakdown of the results by input device is displayed in Table S1.

Table 5.3 Body Composition Prediction Metrics on Test Set. R^2, RMSE, and paired p-value with DXA measurements for all test set fits using all available principal component dimensions (391 and 457 for males and females). One PCA model was used to fit body shape and predict composition per gender, while input scans came from three independent manufacturers. Only visceral fat for both genders and female trunk lean were significantly different from DXA measurements in a paired t-test. R^2 and RMSE values were comparable to the manually guided methods of (2) but were reported on held out data rather than scans that were in the PCA domain.

	Male, d = 391, n=182		**Female, d = 457, n=248**	
	R^2 **(RMSE)**	***p*-value**	R^2 **(RMSE)**	***p*-value**
Percent Fat	0.77 **(3.24)**	0.32	0.68 **(4.22)**	0.27
Lean Mass (kg)	0.95 **(2.57)**	0.23	0.91 **(2.84)**	0.64
Fat Mass (kg)	0.93 **(2.57)**	0.23	0.94 **(2.84)**	0.64
Visc. Fat (kg)	0.86 **(0.12)**	**0.0 ***	0.77 **(0.14)**	**0.0 ***
Leg Lean (kg)	0.89 **(0.68)**	0.04	0.91 **(0.55)**	0.01
Arm Lean (kg)	0.81 **(0.42)**	0.05	0.84 **(0.23)**	0.35
Trunk Lean (kg)	0.91 **(1.73)**	0.29	0.88 **(1.64)**	**0.002 ***
Trunk Fat (kg)	0.94 **(1.45)**	0.84	0.93 **(1.67)**	0.33
Leg Fat (kg)	0.85 **(0.56)**	0.04	0.92 **(0.56)**	0.58
Arm Fat (kg)	0.88 **(0.22)**	0.38	0.89 **(0.28)**	0.83
FMI	0.92 **(0.85)**	0.19	0.93 **(1.12)**	0.62
FFMI	0.93 (**0.85)**	0.19	0.88 **(1.12)**	0.62

Table 5.4 shows our test-retest precision measured with the coefficient of variation (%CV) Glüer *et al*. [39] Same-day duplicate scans were taken of the subjects in the test set, and the fitting and predictions were repeated using the final unified model. Coefficient of variation (%CV) was defined as the ratio of the standard deviation of repeat measurements to the mean of repeat measurements averaged across all test subjects. The closer this value was to zero, the more precise our predictions for repeat scans of the same participant. Not every participant had a duplicate scan,

so the test-retest pairs were less than the total test set size. We compared test-retest precision against the duplicate DXA measurements for each participant to determine precision of our method for predicting the composition of fitted scans relative to the gold standard method.

Table 5.4. Test-retest precision of test set scans. Coefficient of variation (%CV) and RMSE on test set compared to repeat DXA scans of same participants. %CV was comparable to (*4*) and around 2 to 4 times the magnitude of corresponding DXA precision metrics.

	This Work		**DXA**	
	Male n=146 %CV **(RMSE)**	Female n=208 %CV **(RMSE)**	Male n=143 %CV **(RMSE)**	Female n=205 %CV (**RMSE)**
Percent Fat	**(1.90)**	**(2.91)**	**(0.50)**	**(0.48)**
Lean Mass (kg)	1.22 **(1.55)**	2.09 **(1.87)**	0.38 **(0.49)**	0.45 **(0.40)**
Fat Mass (kg)	4.04 **(1.55)**	3.92 **(1.87)**	1.24 **(0.47)**	0.77 **(0.37)**
Visc. Fat (kg)	8.83 **(0.08)**	12.25 **(0.09)**	4.79 **(0.05)**	6.3 **(0.05)**
Leg Lean (kg)	2.23 **(0.46)**	2.19 **(0.32)**	0.86 **(0.18)**	0.82 **(0.12)**
Arm Lean (kg)	2.8 **(0.23)**	3.64 **(0.17)**	1.05 **(0.09)**	1.51 **(0.07)**
Trunk Lean (kg)	1.86 **(1.14)**	1.99 **(0.90)**	0.75 **(0.46)**	0.78 **(0.35)**
Trunk Fat (kg)	4.53 **(0.86)**	4.78 **(1.04)**	2.06 **(0.39)**	1.6 **(0.35)**
Leg Fat (kg)	5.03 **(0.32)**	3.89 **(0.35)**	1.88 **(0.11)**	1.08 **(0.10)**
Arm Fat (kg)	6.06 **(0.14)**	5.77 **(0.18)**	2.32 **(0.05)**	2.3 **(0.07)**
FMI	3.98 **(0.50)**	3.94 **(0.73)**	1.26 **(0.16)**	0.76 **(0.14)**
FFMI	1.20 **(0.50)**	2.11 **(0.73)**	0.38 **(0.16)**	0.44 **(0.15)**

5.4.4 Template Fitting and Body Composition Prediction on Novel System (System 4) Input

To test the generalizability of our fitting and prediction method to an unseen scanning technology, we performed automatic template fitting and body composition prediction using the

unified system-agnostic model on 59 males and 73 females scanned with System 4. This device had no representation in the training set and presented an additional challenge of having the most non-conforming pose of any of the optical systems. The very limited field of view of this system necessitated many participants to hold their arms very close to their body often in bent positions, and taller individuals had to stoop either by bending at the back, waist, or knees to fit into the scanning volume. These meshes were not well conforming to the A-pose constraints specified in the meshes of the training set, in which the abduction of the arms and legs varied but all subjects stood upright with fully extended knees and elbows. R^2 and RMSE are reported in Table 5.5 and test-retest precision is reported in Table 5.6 Although the difference between our prediction and DXA was statistically significant from zero for more compositional measures relative to the devices that were included in the training data, the total body lean mass, fat mass, and percent fat predictions were not significantly different from the gold standard despite the unfavorable poses in this validation set.

Table 5.5. Body Composition Prediction on Novel System (System 4) Input. R^2, RMSE, and *p*-value of System 4 test set. No scans from this device were included in any training data, and participants scanned with this system often had bent limbs or hunched postures due to a narrower field of view for the sensor. Total body lean mass, fat mass, and percent fat were still not significantly different from DXA measurements.

	Male, d = 391, n=59		**Female, d = 457, n=73**	
	R^2 **(RMSE)**	***p*-value**	R^2 **(RMSE)**	***p*-value**
Percent Fat	0.85 **(2.93)**	0.95	0.72 **(4.13)**	0.46
Lean Mass (kg)	0.96 **(2.41)**	0.56	0.93 **(2.56)**	0.22
Fat Mass (kg)	0.94 **(2.41)**	0.56	0.96 **(2.56)**	0.22
Visc. Fat (kg)	0.63 **(0.17)**	**0 ***	0.85 **(0.13)**	0.015
Leg Lean (kg)	0.69 **(1.09)**	**0 ***	0.77 **(0.85)**	**0 ***
Arm Lean (kg)	0.15 **(0.83)**	**0 ***	0.82 **(0.24)**	0.006
Trunk Lean (kg)	0.72 **(3.06)**	**0 ***	0.85 **(2.01)**	**0 ***
Trunk Fat (kg)	0.94 **(1.43)**	0.28	0.96 **(1.44)**	**0 ***
Leg Fat (kg)	0.88 **(0.52)**	0.09	0.86 **(0.75)**	**0 ***
Arm Fat (kg)	0.83 **(0.25)**	**0 ***	0.93 **(0.26)**	0.66
FMI	0.94 **(0.82)**	0.64	0.96 **(0.98)**	0.26
FFMI	0.93 **(0.82)**	0.64	0.91 **(0.98)**	0.26

Table 5.6. Test-retest precision on novel system (System 4) compared to DXA. %CV and RMSE for System 4. These results represent the precision of our method on completely novel input from scanning systems not represented in the training data. The precision error is worse than our results on a held-out test set of systems 1 through 3, ranging 3 to 5 times that of DXA.

	This Work		**DXA**	
	Male n=41 %CV **(RMSE)**	Female n=58 %CV **(RMSE)**	Male n=41 %CV **(RMSE)**	Female n=58 %CV **(RMSE)**
Percent Fat	**(2.59)**	**(3.73)**	**(0.55)**	**(0.55)**
Lean Mass (kg)	1.40 **(1.86)**	2.46 **(2.26)**	0.41 **(0.54)**	0.49 **(0.46)**
Fat Mass (kg)	4.69 **(1.86)**	4.34 **(2.26)**	1.43 **(0.55)**	0.83 **(0.43)**
Visc. Fat (kg)	13.57 **(0.11)**	14.23 **(0.12)**	3.80 **(0.04)**	6.65 **(0.06)**
Leg Lean (kg)	1.91 **(0.38)**	3.87 **(0.53)**	0.86 **(0.18)**	0.9 **(0.13)**
Arm Lean (kg)	3.21 **(0.23)**	5.65 **(0.27)**	1.25 **(0.11)**	1.38 **(0.07)**
Trunk Lean (kg)	2.55 **(1.51)**	2.99 **(1.35)**	0.74 **(0.48)**	0.96 **(0.45)**
Trunk Fat (kg)	6.16 **(1.20)**	4.79 **(1.12)**	2.21 **(0.44)**	1.73 **(0.43)**
Leg Fat (kg)	4.75 **(0.31)**	5.09 **(0.52)**	1.80 **(0.11)**	0.92 **(0.09)**
Arm Fat (kg)	7.13 **(0.18)**	5.10 **(0.17)**	3.12 **(0.07)**	2.58 **(0.09)**
FMI	4.78 **(0.62)**	4.33 **(0.86)**	1.44 **(0.18)**	0.82 **(0.16)**
FFMI	1.43 **(0.62)**	2.45 **(0.86)**	0.40 **(0.18)**	0.48 **(0.17)**

Our method worked well for most scans on this validation system but did not align properly to some scans with larger pose differences, such as excessively bent elbows. These scans were excluded from our results as they violated the parameters of the model. Our method can be used to fit a standardized templated to scans from any system that captures a human in an A-pose, which is any pose with fully extended elbows and knees with arms and legs abducted around 30 degrees

from the midline. Some variation in limb abduction at the shoulders and hips is accounted for by the model, but large body part rotations create nonlinearities not representable by PCA.

5.5 Discussion

In this work we developed a scanning system agnostic algorithm for standardizing unorganized 3DO body scans with vertex counts spanning three orders of magnitude and slight variations in pose. We constructed a PCA model using training data from three different systems and validated our shape fitting and body composition prediction accuracy on held out scans from all three systems and an additional fourth unseen system. Our resulting prediction models trained on automatically fit training set scans predicted total and regional body fat with R^2 and RMSE comparable to the principal component only model from Ng *et al.* [83] Our model was more accurate in many cases even on the validation system. For example, R^2 for total fat mass for System 4 was 0.94 and 0.96 for males and females, respectively, compared to 0.88 and 0.93 in Ng *et al*. Our results are even stronger when compared to previous work as our fits and predictions were performed on test scans that were not included in the PCA training data. In Ng *et al.*, 5-fold cross validation was performed to train the linear model mapping principal components to body composition, but the PCA domain included all the available data and thus contained no withheld data for blinded validation. Furthermore, our scans were sourced from four different systems and generalized to a system that was not represented by training data with no manual initialization. We demonstrated that our automatic template fitting algorithm can generalize to inputs from novel scanning systems exhibiting small variations in pose. Inclusion of these system-agnostic template fits into a new expanded PCA domain is likely to further increase the predictive accuracy of our regression models.

Our retest precision error on the test set was two to four times that of the criterion DXA scans. This was a slightly less precise result than Ng *et al.* but was reported on a greater number of subjects (as opposed to just 119 of each gender) scanned on three different systems (as opposed to one). Furthermore, all test scans were held out of the PCA construction while Ng *et al.* made no such distinction. Although our precision is trailing that of criterion radiology, this can be mitigated by averaging predictions from repeat scans. The least significant change (LSC) [108] between the average of multiple measurements sampled at baseline and follow-up is defined as:

$$LSC = Z\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (14)$$

for a Z defined by the two-sided 95% confidence interval z-score and precision error σ, with n_1 measurements at baseline and n_2 measurements at follow-up. A difference greater than this value means there is 95% confidence a true change in body composition has occurred, and a lower value implies greater resolution of change over time. Z is constant at 1.96 and σ is inherent to the measurement method, but we can set $n_1 = n_1 = 9$ scans at baseline and follow-up to drop the LSC by a factor of 3. This would account for the difference in precision error between our method and a single DXA scan on most metrics. As 3DO scanners take a minute or less to complete and can be repeated without threat of radiation injury, collecting nine scans is not unreasonably burdensome relative to the gold standard radiology. As we have shown that the accuracy of our method increases with larger training sets, the precision of our method could potentially also be increased with additional data.

Our automated markerless fitting and prediction method generalized well to scans that were adjacent to an A-pose. We tested incorporating additional training scans in a T-pose stance, with arms held out parallel to the ground and legs straight with feet together, to determine how robust

our method was to scans with more extreme pose variation. Although we were able to achieve good geometric fits to test scans in both the A and T pose, the predictive accuracy of the regression models decreased drastically. Such huge differences in the limb positions were not correlated with body composition but had major effects on the PCA shape coordinates of the mesh. We decided against recommending inclusion of this degree of extreme pose variability in the training set for the results of this paper as the variance introduced by the extreme pose difference created excessive amounts of noise in the prediction regression. An example of a T-pose fit using our model is shown in Fig 5.8.

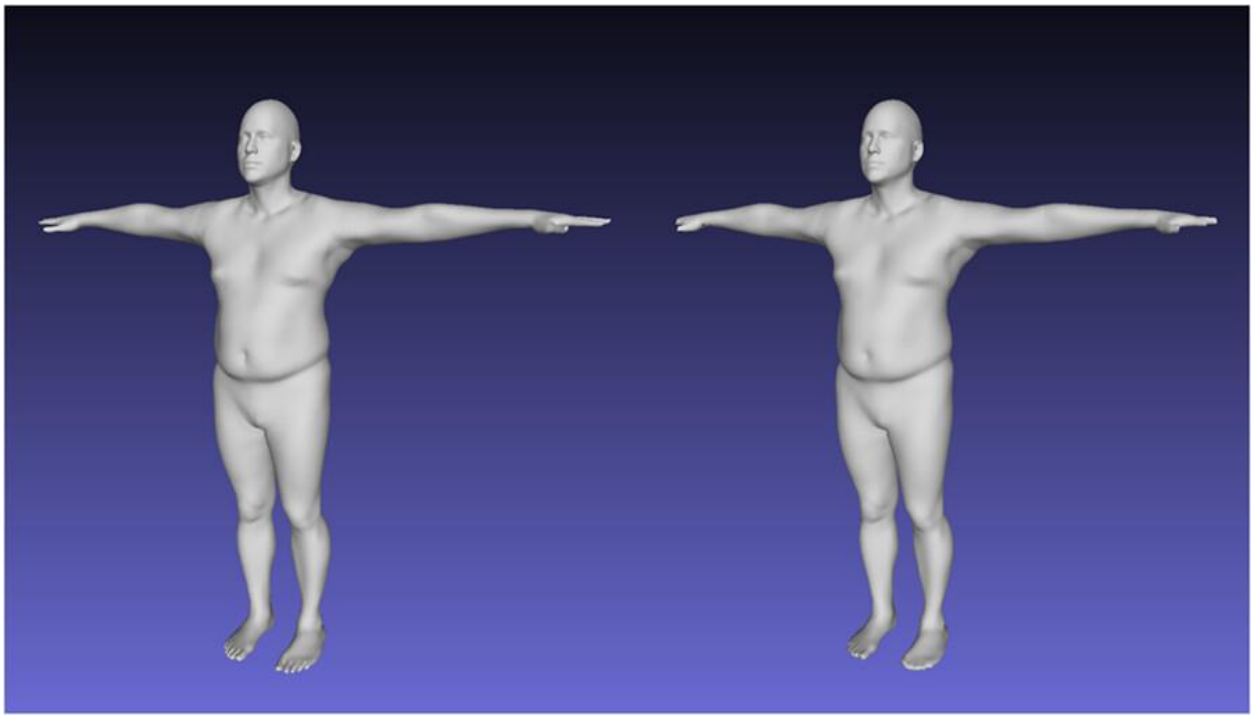


Figure 5.8. Robustness of fitting algorithm to extreme pose variation. T-pose example, input scan on the left with 110k vertices and our 60k fit on the right.

We demonstrated that our method was able to fit to a range of A-poses if there was representation of the target pose in the training data. However, these pose differences were

modeled as linear interpolations of the limbs between example meshes in the training data exhibiting pose variation rather than rotational transformations at joints. A body mesh that separately models pose and shape using a jointed skeleton and skinning function may provide more accurate fits and predictions for scans that are captured in poses that are further from anatomical neutral, i.e., the hands-up pose in an airport system. Such a model may also produce increased prediction accuracy for body composition even on A-pose scans, as the vertex deformations caused by small differences in limb position are not discriminated from differences in anatomical body shape and create substantial noise in learning the relationship between shape and composition. Although prior work such as SMPL [67] exists in the field of human pose and shape estimation using skinned models, these methods were not created with clinical application in mind. SMPL (Skinned Multi Person Linear model) learned a linear body shape PCA from the CAESAR database and decoupled pose from the model by standardizing all fitted templates to a T-pose with parallel arms and vertical legs. This was made possible with an animated model containing a skeleton with vertices mapped to a hierarchical tree (known as a skinned model). However, they did not collect medical measurements along with shape and pose data for their model construction. Subsequent work mainly sought to recreate visually plausible reconstructions of human shapes from "in the wild" photos [11] and lack correlated clinical variables. Future work may merge the flexibility of a skinned posable model with our clinical data.

While our work is not the first to propose templated encoding spaces for human body shapes [90], our implementation was the first to include associated body composition data with 3D optical scanning, and also the first to demonstrate a robust pipeline that works with limited pose variation across different scanning devices. Due to the logistical challenges of collecting multi-person data, CAESAR [97] remains the largest fixed pose, multi-shape dataset almost 20 years

after its initial publication. Many recent works on human body encoding focus on limited person, multi-pose shape models [146] and do not address the medical implications of body shape. Future work could use our method to combine the scans of CAESAR and Shape Up! into a singular multi-person shape model and learn the correlations to body composition on the subset of training data with associated DXA measurements using more robust shape descriptors learned on the entire large dataset.

Certain fine details in 3D scans are not useful in predicting body composition, such as the position and shape of the hands and fingers or head shape deformation due to large hair volume. Further work could replace these regions with smooth surfaces, effectively eliminating the variance in shape caused by these uncorrelated details. Such a reduction in noise could potentially increase the accuracy of our prediction models.

Our algorithms predict analogs for mortality risk due to metabolic disease from 3D body shape. The relationship between scalar metrics such as compartmental body fat masses and disease have been previously studied in the clinical setting using radiology. Our method allows the use of 3D optical scanning as an alternative to traditional medical imaging for predicting these measurements. It is possible that there are direct relationships between 3D body shape and disease risk latent in the templated data that can be recovered from the parameters of the principal component domain or through deep-learning methods on the mesh coordinates directly. The standardization of 3D scans allows for the application of these methods, but we cannot yet form such conclusions without a longitudinal dataset that directly associates body shape with mortality from metabolic disease. Our algorithm may be able to recover higher dimensional relationships between body shape and mortality risk in the future, bypassing scalar analogs such as prediction of visceral fat mass, when such data becomes available.

Our prediction model was determined with a least-squares linear regression between the principal component coordinates of our training data and their associated DXA composition metrics. Our method worked well in Tian *et al*. [117] and was the most conservative model for relating standardized 60k mesh templates to body composition. A better performing prediction model with non-linear terms, such as one produced by a graph convolution network in Bouritsas *et al*. [15], may offer better predictions from either the dimension reduced PC coordinates or the 60k mesh vertices themselves. As these models required thousands of training examples, we relied on linear models that could learn good correlations with only a few hundred scans. Creating larger standardized body mesh databases using our method with future data collection may facilitate work in this direction.

5.6 Conclusions

At the conclusion of this study, we automatically fit a common 60k vertex body template to 1410 raw scans from four different scanning technologies with an automated PCA initialized two-stage deformation process. We built a device agnostic unified PCA model from 848 training scans and learned a regression from projected PCA basis coordinates to DXA body composition measurements. We used the same fitting procedure with this device agnostic model to automatically template 562 held-out test scans and derive body composition predictions from PCA coordinates. Our model predicted body composition metrics with accuracy comparable to or better than previous models built with manually targeted mesh fitting, achieving 0.8 or better R^2 for all fat and lean mass predictions on test scans except female visceral fat, with all RMSEs below 3.0 kg. Processing data at this scale by hand would be prohibitive in time and cost. Furthermore, our results were reported on held-out test data, which further strengthens our results relative to

previous work published on training accuracy only. Our work seeks to make 3D optical scanning an accurate, automated, and device agnostic tool for body shape modeling and composition analysis with many potentials for clinical and diagnostic application, such as serial monitoring of changes in body weight and risk for chronic disease.

5.7 Appendix

5.7.1 Mathematical Details of Shape Fitting

In Ng *et al.* [83], fitting a consistent 60k template to an arbitrary scan was contingent on manually annotating anatomical landmarks on the scan. The template, which also has corresponding markers at specific vertices, was deformed to minimize the combined data, smoothness, and marker difference error using the method described in Allen et al. [5] Given an existing PCA shape model, in this case constructed using the fitted meshes in the bootstrap set, we can substitute the manual marker annotation step with an automated registration algorithm that deforms the starting mesh in the principal component domain rather than the 3D coordinate domain. For a 60k mesh in *x, y, z* coordinates, the flattened 3D vector would be $\mathbb{R}^{180,003}$.

Deforming the body mesh in principal component domain reduces the dimensionality greatly (for a 70-member bootstrap set, the maximum dimensionality is $\mathbb{R}^{70}$) and constrains the allowable deformations to searches in the range of plausible human body shapes. Tian et al. [117] demonstrated that such a constrained deformation was sufficient to match a principal component mesh model to a silhouette extracted from a 2D photograph, subject to additional perspective projection and rigid transformation operations. We used the same method described in Tian *et al.* in Chapter 4.3 to fit a principal component model to a raw scan.

A body shape can be reconstructed in PCA domain with

$$\mathbf{s} = \mathbf{A}\mathbf{w} + \boldsymbol{\mu} \quad (15)$$

Where $\mathbf{w}$ is a d length PCA parameter vector and $\boldsymbol{\mu}$ is the mean shape of the training set.

The most critical condition to satisfy for automatic mesh registration is alignment and anatomically analogous vertex correspondences between template and target. Manual marker annotation was necessary in Allen *et al.* and Ng *et al.* because the template mesh was often strongly misaligned with the target scan due to size differences. Using very accessible scalar priors such as height and weight to initialize the size of the template mesh resolves most of the alignment issues caused by differences in body size. This enables vertex correspondences to be matched with simple nearest neighbor search.

We produced a coarse fit to the target scan by iteratively registering each vertex on the current template mesh shape to its closest point in the target scan. Nearest neighbor search was performed with FLANN [79]. Given a target scan $\mathbf{x}$ and the initial shape, $\mathbf{s}_0 = \mathbf{A}\mathbf{w}_0 + \boldsymbol{\mu}$, we solve:

For each vertex s_{0i}

$$y_i = \text{FLANN}(s_{0i}, \mathbf{x}) \quad (16)$$

where y_i is the point in the scan $\mathbf{x}$ that is closest to initial shape vertex s_{0i}.

We minimized the difference between the template and target surfaces by solving the following L2-regularized regression:

$$E(\mathbf{w}) = \left|\left|\mathbf{A}\mathbf{w} - \mathbf{b}\right|\right| + \left|\left|\boldsymbol{\Lambda}(\mathbf{w} - \mathbf{w}_0)\right|\right|_2^2 \quad (17)$$

where

$$\mathbf{b} = \mathbf{y} - \boldsymbol{\mu} \quad (18)$$

and each y_i in $\mathbf{y}$ is the nearest neighbor on the target scan corresponding to the *i*th point of the template scan. $\mathbf{y}$ and $\boldsymbol{\mu}$ are flattened vectors of size n = 180,003. $\boldsymbol{\Lambda}$ is the Tikhonov matrix, a size d diagonal matrix with the value $\frac{\sqrt{\lambda}}{\sigma_i}$ on the diagonal where σ_i is the standard deviation of the *i*th principal component of $\mathbf{A}$ and λ is a hyperparameter denoting regularization weight. This regularization term constrains acceptable fits to the shape of the expected height and weight by penalizing the difference $\mathbf{w} - \mathbf{w}_0$ rather than the mean shape (which is the zero vector) in the same manner as Tian *et al*.

The result is an L2-regularized least-squares problem with a unique linear solution at the derivative:

$$\mathbf{w}^* = (\mathbf{A}^T\mathbf{A} + \boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}(\mathbf{A}^T\mathbf{b} + \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\mathbf{w}_0) \quad (19)$$

Nearest neighbor matching and surface distance minimization were performed iteratively until the sum squared difference between the principal component vectors of the current and previous iterations fell below a convergence parameter.

$\mathbf{w}^* = \mathbf{w}_0$
Repeat
    $\mathbf{w} = \mathbf{w}^*$
    $\mathbf{s} = \mathbf{A}\mathbf{w} + \boldsymbol{\mu}$
    for each vertex s_i
        $y_i = \text{FLANN}(s_i, \mathbf{x})$
    $\mathbf{w}^* = \min_{\mathbf{w}} \mathbf{E}(\mathbf{w},\mathbf{y})$
until $\|\mathbf{w}^* - \mathbf{w}\| < \varepsilon$

where convergence parameter ε = 0.1 and shape regularization parameter λ = 0.01

To produce the final mesh fit, we first smooth the converged shape $\mathbf{s}^*$ with the HC-Laplacian operation presented by Vollmer *et al.* [123] to remove noise and non-smooth surface irregularities that can arise from the principal component basis deformation. The result is a smooth

body shape that generally aligns to the raw scan in body proportions and limb pose. At this point, the optimized template mesh $\mathbf{s}^*$ and target scan $\mathbf{x}$ is in close enough alignment such that registering $\mathbf{s}^*$ to $\mathbf{x}$ with a minimum surface distance heuristic without annotated anatomical markers is sufficient. We can then use the mesh fitting function from Allen *et al.* with zero weight on the marker matching term to perform a refined fit between the final smoothed coarse fit and the raw target scan to get the final surface $\mathbf{s}_{final}$.

Fig. S1 shows the differences in R^2 between successive iterations of our algorithm. This process was charted to demonstrate that our automatic template mesh fitting method produced comparable or better body composition predictions when compared to the manually initialized method and that adding variance in the model by unifying scans from all systems into one shape space did not substantially penalize prediction performance.

5.7.2 Body Composition Prediction from Templated Fits

In this work, we followed the linear regression model of Tian *et al.* to train a mapping matrix $\mathbf{M}$ between PCA components $\mathbf{w}$ and DXA features $\mathbf{f}$.

$$F = \mathbf{M}_f\mathbf{W} \tag{20}$$

For a refined templated mesh fit $\mathbf{s}_{final}$, we find its projection onto the principal component basis $\mathbf{A}$ by multiplying it with the transpose of the PCA transformation matrix:

$$\mathbf{w}_{final} = \mathbf{A}^T\mathbf{s}_{final} \tag{21}$$

The predicted features of this final refined mesh are:

$$\mathbf{f}_{final} = \mathbf{M}_f\mathbf{w}_{final} \tag{22}$$

In this work, we computed the following feature vector: [percent fat (fat mass / weight), lean mass, fat mass, visceral fat, leg lean mass, arm lean mass, trunk lean mass, trunk fat mass, leg fat mass, arm fat mass, fat mass index (FMI), fat free mass index (FFMI)]

5.7.3 Iterative Improvement of Shape Model

We tested the predictive accuracy of our intermediate shape models to isolate the effects of automatic markerless fitting and inclusion of additional training data. Fig. S1 shows the change in R^2 for body composition prediction on each intermediate iteration of our algorithm.

Iteration 0 was initialized with 70 manually guided templated meshes created with the procedures detailed in [83] specific to each pair of scanning systems (1, 2, and 3) and gender (Male/Female). The scans in this initial set comprised 6 bootstrap sets totaling 420 scans. We performed PCA on each device and gender split independently, producing six models.

In iteration 1, we recreated the same six principal component models from iteration 0 using automatically templated meshes automatically fit with the PCA coarse deformation from iteration 0. This was to test the effect of automatic template fitting on body composition prediction in the absence of additional training data. Most metrics gain predictive accuracy for both genders, with small increases in error of <0.1 for some measurements such as arm fat in females. R^2 values for the auto-templated models increased by an average of 0.054 and 0.023 for males and females, respectively, averaged across all three systems and 12 composition metrics. Although the training set membership was the same, there was an overall increase in predictive performance. This boost can be attributed to the smoothing operation followed by markerless nearest-surface smooth alignment. The resulting fits from our method had fewer artifacts that were introduced as a

byproduct of the manually initialized deformation, where the initial template mesh was very misaligned with the target scan and necessitated very large nonlinear transformations.

In iteration 2, the PCA shape models from iteration 1 were used to fit additional raw scans from the training set. These newly automatically templated training scans were combined with the templated meshes from iteration 1 and new PCA models were reconstructed. This step was necessary because as per the problem definition, it cannot be assumed that a previously unseen input scan has a corresponding manually curated 60k mesh template to use for composition prediction. This step demonstrates that a scan with no manually annotated counterpart can be templated and used as additional training data, thereby allowing our method to generate more training data quickly and automatically in the future without loss of accuracy. Increasing the training set members for each device increased the performance of composition prediction on the test set while holding the parameter count at 70 principal components.

In iteration 3, we consolidated our six system specific PCA models from iteration 2 into two, one for each gender, by merging the automatically templated training and bootstrap meshes from all systems into a single training set. Models built on solely one system could not be applied to other systems due to unmodeled pose variations between devices. Merging training sets from all systems introduced pose variance into the training shape domain, which injected some amount of noise in body composition prediction but allowed us to fit and predict any scan that fell within the range of the modeled pose variation with a single model. Although predictions were not uniformly better across all devices and measures, the important advance in this iteration is the application of a single unified model to all test set scans from all modalities. The unified superset model gained at most 1.0kg of RMSE in fat mass for males and 1% body fat RMSE for males, which we deemed an acceptable tradeoff for introducing limited robustness to pose difference and

agnosticism to scanning technology. We held the number of parameters used for composition prediction at 70 to show a fair comparison with the previous two iterations.

In our final iteration we included all available PCs for mapping the relationship between body shape and composition. The PCA training set members were the same as iteration 3. For males, $d = 391$ and for females $d = 457$. As these predictions were all made on held out test scans, despite the high number of components there does not appear to be overfitting.

Males, Iterative PCA Progression

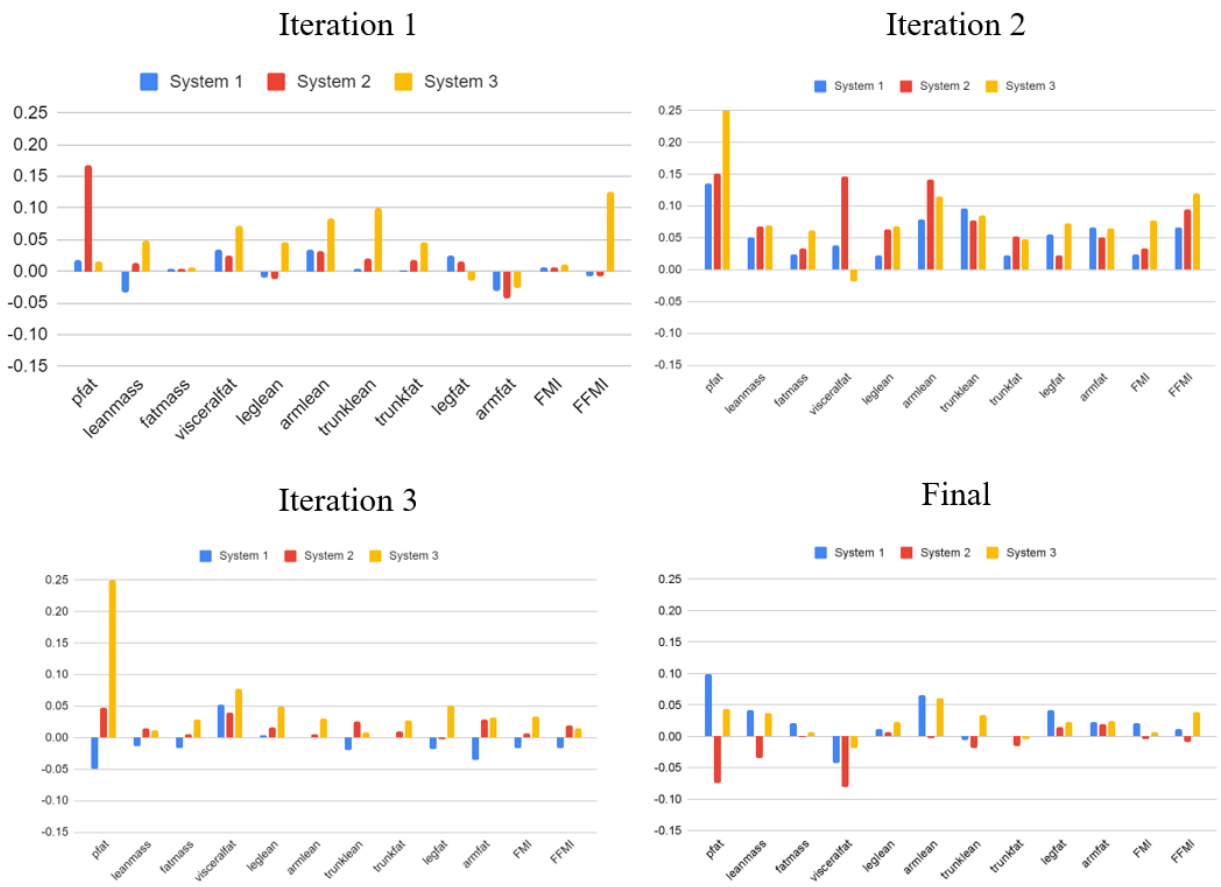


Figure S1. RMSE differences in body composition prediction of test set scans between iterations. Iteration 1: predictions of PCA model trained with auto templated bootstrap set compared to predictions using PCA model trained with manually guided fitted meshes, also only with bootstrap set. 70 meshes and PCs per gender and device. Automatic fitting produced predictions that were mostly better than the manual equivalents. Iteration 2: R^2 change when PCA model from iteration 1 was extended with training set meshes fit using the iteration 1 model. This shows the validity of using the fitting algorithm to generate more training data for expanded models with previously unseen scans that have no manually fitted counterpart. Predictions used 70 PC parameters to show a fair comparison to previous iteration, which only had 70 parameters total. Iteration 3: R^2 change from iteration 2 when a single PCA model per gender was created by merging the training data from all three systems into one superset. No new training data was added. 70 PCs were used for prediction. Final: R^2 change from iteration 3 when all principal components were used. Max PCs were equal to the total number of training scans in the PCA model. For males, d=391 and for females, d=457. Because all fits and composition prediction benchmarks were performed on a held-out test set spanning all three systems, this precludes the risk of overfitting and justifies the use of all principal components in body composition prediction.

Table S1. Per-system body composition accuracy using unified PCA model. R^2 and RMSE values for composition predictions of final refined test set fits, one principal component model per gender.

	Male, d = 391						**Female, d = 457**					
	Sys. 1, n=86		**Sys. 2, n=73**		**Sys. 3, n=23**		**Sys. 1, n=112**		**Sys. 2, n=103**		**Sys. 3, n=33**	
	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**
% fat	0.84	2.56	0.70	3.64	0.73	4.08	0.74	3.73	0.57	4.94	0.80	3.28
Lean Mass (kg)	0.96	2.14	0.94	2.78	0.95	3.24	0.93	2.36	0.88	3.48	0.95	1.97
Fat Mass (kg)	0.94	2.14	0.91	2.78	0.93	3.24	0.947	2.359	0.922	3.48	0.97	1.97
Visc. Fat (kg)	0.87	0.10	0.85	0.14	0.86	0.12	0.79	0.13	0.72	0.16	0.84	0.13
Leg Lean (kg)	0.89	0.59	0.86	0.78	0.92	0.65	0.89	0.57	0.92	0.54	0.90	0.49
Arm Lean (kg)	0.87	0.33	0.70	0.52	0.89	0.39	0.84	0.22	0.82	0.26	0.89	0.17
Trunk Lean (kg)	0.91	1.60	0.90	1.84	0.93	1.85	0.86	1.59	0.88	1.80	0.94	1.17
Trunk Fat (kg)	0.94	1.19	0.93	1.55	0.93	1.90	0.95	1.31	0.91	2.09	0.96	1.17
Leg Fat (kg)	0.87	0.50	0.80	0.62	0.90	0.55	0.92	0.53	0.92	0.61	0.94	0.47
Arm Fat (kg)	0.88	0.21	0.87	0.19	0.85	0.28	0.85	0.27	0.89	0.31	0.95	0.19
FMI	0.94	0.70	0.90	0.94	0.92	1.06	0.95	0.92	0.91	1.38	0.97	0.77
FFMI	0.94	0.70	0.92	0.94	0.92	1.06	0.91	0.92	0.84	1.38	0.92	0.77

6. Automated body composition estimation from device-agnostic 3D optical scans in pediatric populations

Using the automated mesh processing method of the previous chapter, we fit templated meshes to all scans of the Shape Up! Kids dataset available at the time of writing and built PCA shape models and regression models with methods parallel to Chapter 5. In this work we also investigated several alternative regression methods, such as AdaBoost, gradient boost, decision trees, and random forest algorithms. We found that Least Angle Regression (LARS), a continuous generalization of stepwise regression in which additional parameters are added in weighted increments rather than with binary selection, produced slightly better prediction results than least squares with all parameters included. Works prior to this chapter used either stepwise regression or full OLS regression to predict body composition from PCA parameters. This work suggests LARS may serve as an intermediate method analogous to an L2 regularization of the regression model rather than an L1 model that benefits from preferring smaller magnitude parameters without losing signal from higher order parameters due to statistical elimination.

This work was the first PCA shape model constructed on a pediatric population and was necessary to bring the Shape Up! Kids results up to date with the most current Shape Up! Adults analyses. Growth and physical maturation representing pre-pubescent through late-adolescent developmental stages was uniquely represented in this dataset when compared to adult data. We demonstrated in this work that visual indicators of Tanner staging were represented in the PCA shape model as a function of age even though such classifications were not formally reported in the dataset collection nor targeted during training.

This chapter was based on work published in Clinical Nutrition:

[119] *Tian IY, Wong MC, Nguyen WM, Kennedy S, McCarthy C, Kelly NN, et al. Automated body composition estimation from device-agnostic 3D optical scans in pediatric populations. Clinical Nutrition 2023;42:1619–30. doi:10.1016/j.clnu.2023.07.012.*

6.1 Abstract

Background: Excess adiposity in children is strongly correlated with obesity-related metabolic disease in adulthood, including diabetes, cardiovascular disease, and 13 types of cancer. Despite the many long-term health risks of childhood obesity, body mass index (BMI) Z-score is typically the only adiposity marker used in pediatric studies and clinical applications. The effects of regional adiposity are not captured in a single scalar measurement, and their effects on short- and long-term metabolic health are largely unknown. However, clinicians and researchers rarely deploy gold-standard methods for measuring compartmental fat such as magnetic resonance imaging (MRI) and dual X-ray absorptiometry (DXA) on children and adolescents due to cost or radiation concerns. Three-dimensional optical (3DO) scans are relatively inexpensive to obtain and use non-invasive and radiation-free imaging techniques to capture the external surface geometry of a patient's body. This 3D shape contains cues about the body composition that can be learned from a structured correlation between 3D body shape parameters and reference DXA scans obtained on a sample population.

Study Aim: This study seeks to introduce a radiation-free, automated 3D optical imaging solution for monitoring body shape and composition in children aged 5-17.

Methods: We introduce an automated, linear learning method to predict total and regional body composition of children aged 5-17 from 3DO scans. We collected 145 male and 206 female 3DO scans on children between the ages of 5 and 17 with three scanners from independent manufacturers. We used an automated shape templating method first introduced on an adult

population to fit a topologically consistent 60,000 vertex (60k) mesh to 3DO scans of arbitrary scanning source and mesh topology. We constructed a parameterized body shape space using principal component analysis (PCA) and estimated a regression matrix between the shape parameters and their associated DXA measurements. We automatically fit scans of 30 male and 38 female participants from a held-out test set and predicted 12 body composition measurements.

Results: The coefficient of determination (R^2) between 3DO predicted body composition and DXA measurements was at least 0.85 for all measurements with the exception of visceral fat on 3D scan predictions. Precision error was 1-4 times larger than that of DXA. No predicted variable was significantly different from DXA measurement except for male trunk lean mass.

Conclusion: Optical imaging can quickly, safely, and inexpensively estimate regional body composition in children aged 5-17. Frequent repeat measurements can be taken to chart changes in body adiposity over time without risk of radiation overexposure.

6.2 Introduction

Obesity in childhood and adolescents strongly predicts lifelong obesity. More than half of obese children develop into obese adolescents, and around 70% of obese adolescents remain obese past age 30 [110]. Over the last 30 years, obesity has more than doubled in children under age 12, quadrupled in adolescents, and is linked to cardiovascular disease, insulin resistance, and 13 different types of cancer [112, 32, 104, 94, 93]. Body mass index (BMI) and its z-score (BMI-Z) are often used as analogues for adiposity but fail to capture both total and regional compositional differences between fat and lean mass, both of which are correlated with cardiometabolic risk. [111]

Despite the connections between pediatric obesity and lifelong metabolic disease risk, study of body composition analysis in pediatrics is limited. Gold standard imaging techniques to measure total and regional adiposity such as dual X-ray absorptiometry (DXA) requires exposure to ionizing radiation, which are potentially more harmful in children and adolescents and is used sparingly. MRI, although radiation free, is expensive to capture and unsustainable in both time and cost for frequent repeat monitoring. Children with behavioral conditions may be ineligible for both DXA and MRI due to compliance limitations at this age bracket [16]. This is particularly problematic in pediatric research as people experience the largest rates of change in body mass and composition between the ages of 5 and 18 [129].

Intermediate alternatives to radiological imaging for adiposity measurement, such as bioelectric impedance, offer more information than BMI alone but have limited resolution in regional adiposity measurement. These methods extrapolate adiposity from scalar signals instead of directly imaging the body compartments contributing to total and regional adiposity [1, 128]. A non-radiological method that can accurately estimate the metrics measured by reference methods such as DXA without exposure to ionizing radiation could dramatically increase the survey frequency for metabolically at-risk children and adolescents during their critical developmental period of maximum compositional fluctuations.

3D optical (3DO) scanning is a radiation-free method of capturing external body shape using structured light (SL) or time of flight (ToF) imaging that is faster, safer, and relatively inexpensive compared to radiology. [120] 3DO scanning produces a 3D image of the surface geometry of a human body represented as a point cloud connected by edges in a graph structure. These 3D surfaces represented as unions of discrete points and their 3D spatial graph connections are defined as 3D meshes. Previous work on an adult cohort demonstrated that 3DO

captured body shape is a strong signal for many of the body composition metrics measured by DXA. [81, 83] Additional studies on children aged 5-17 indicated similar correlations between 3D shape and body composition for a pediatric population [135]. A 3DO model for predicting body composition from scans for children would allow for safe and repeatable measurements of total and regional adiposity. As was the case in adults, this step was missing due to a lack of paired 3D to DXA data and the lack of standardized formats and interpretations for unstructured 3D mesh data. [118]

The purpose of our investigation was to develop a unified, input device-agnostic 3DO body composition prediction model for a pediatric population spanning the age range of 5-17. Our method standardizes randomly ordered point clouds captured by 3DO scanners into a common topology and builds a principal component parameterized shape space from the standardized mesh set. We derived predictions to DXA metrics from this shape space and tested its predictive capacity on held out 3DO scans. This method could be used for body composition monitoring in juvenile and adolescent patients without exposing them to potentially harmful radiation doses while also increasing the survey frequency for data acquisition, enabling longitudinal monitoring over shorter intervals. We hypothesized that our method would generate body composition predictions with accuracy and precision comparable to DXA, thus allowing for greater freedom and resolution in longitudinal studies of body composition for children and adolescents.

6.3 Methods

This work builds upon the methods of [118] and is a cross-sectional study of a highly stratified sample of children that estimates body composition from unstructured 3DO scans and

benchmarks accuracy against whole body DXA measures. We present an automated, device agnostic pipeline for converting raw, unorganized 3D optical scans into watertight and topologically consistent 60,001 vertex (60k) fitted mesh templates. From a set of templated meshes produced by our method, we construct a parameterized 3D shape space using PCA and demonstrate its capacity to predict body composition measures with accuracies measured against DXA. The subset of DXA variables predicted in this study are fat mass, fat-free mass, percent fat, visceral fat, arm fat, leg fat, trunk fat, arm lean, leg lean, trunk lean, fat mass index, and fat-free mass index. In addition, we inverted the prediction equations to visualize how increments in certain body variables affect the 3D shape of a scanned subject, potentially providing clinicians a visualization tool for estimating longitudinal differences in a pediatric population whose bodies change much more rapidly in short time scales than an adult cohort.

We used the methods from [118] to standardize 145 male and 206 female scans from three different 3DO scanners to construct statistical models of body shape using principal component analysis (PCA). We computed linear regressions between the shape parameters of the PCA model and paired DXA composition measurements for all participants in the training set. We tested the predictive accuracy of our model on 132 held out 3DO scans (52 males). We used our shape model to extrapolate the body shape and composition of a younger child to an older age to visualize both the effects of intervention and aging versus no intervention.

6.3.1 Experimental Cohort

Our method was trained and tested on children and adolescents from the Shape Up! Kids Study (NIH R01 DK111698) and were recruited in the Honolulu, HI area at the University of

Hawaii Cancer Center (UH), in the San Francisco, CA area at the University of California, San Francisco (UCSF), and in the Baton Rouge, LA area at Pennington Biomedical Research Center (PBRC). This work represents an intermediate analysis of an in-progress data collection project.

This was a cross-sectional study stratified by age (5-17yr), ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, Asian, and Native Hawaiian or Pacific Islander (NHOPI)), sex, and BMI z-score. [135] Recruitment quotas were designed to represent a continuum of body types spanning underweight, healthy, and overweight ranges as determined by BMI z-score. BMI z-score cutoff boundaries were -2, -1, 0, 1, and 2. Tanner staging was self-reported by participants and not used as a stratification variable. Age in years was used as the surrogate variable. This cohort was deliberately designed to continuously span the most comprehensive range possible of body shapes and compositions of children aged 5-17. A diverse dataset with densely sampled examples across the spectrum of expected body compositions and ages enables the derivation of more accurate and body shape and composition models that generalize better to unseen data. The scope of inclusion of our study allows for our model to be compatible with any child who falls within the ranges of our dataset specified in Table 6.1.

Participants were screened for study eligibility via phone interview and excluded if they could not stand unassisted for two minutes or lie motionless for ten minutes, had metal implants, or had major body-shape-altering procedures (e.g., liposuction, amputations, etc.). No pregnant or lactating individuals were included in the study. Written informed consent was obtained from each participant upon arrival and all procedures were approved by the Pennington Biomedical Research Center (PBRC; IRB study #2017-10, Federalwide Assurance #00006218); University of California, San Francisco (UCSF; IRB #16-20197); and University of Hawaii Office of Research

Compliance (UH ORC; Center of Health Sciences #24282). The study is publicly listed on ClinicalTrials.gov as ID NCT03706612.

DXA Scans. Reference total and compartmental body composition measurements were defined by DXA. We acquired single whole-body scans of each participant on either a Hologic Horizon/A system (UCSF) or a Discovery/A system (PBRC and UHCC) (Hologic Inc., Marlborough, MA, USA). Participants were positioned and scanned according to the respective manufacturer's guidelines. All scans were analyzed by a single certified technologist at UHCC using Hologic Apex version 5.6 with the National Health and Nutrition Examination Survey (NHANES) Body Composition Analysis calibration option disabled [82]. DXA systems quality control was performed by monitoring the weekly values of the Hologic Whole Body Phantom. Cross-calibration was checked between sites using a whole-body phantom scanned at each site, and calibrations were performed to compensate for systemic bias in all DXA measurements.

3D Optical Scans. Inputs to our prediction model for training and testing were collected via 3D optical scanning. Scanners output unordered and unstructured point clouds representing the 3D surface geometry of the scanned participant. Participants wore form-fitting tights, a swim cap, and sports bra if female. Each participant was scanned in one or more 3DO systems pending availability at each recruiting location. We used three different 3DO system manufacturers across all sites: System 1 (Fit3D Proscanner 4.x, Fit3D Inc, Redwood City, CA, USA), System 2 (Styku S100 4.1, Styku LLC, Los Angeles, CA, USA), and System 3 (Size Stream SS20, Size Stream, Cary, NC, USA). Scans from different systems differed slightly in pose, although all were upright with straight elbows and knees in a neutral A-pose, and differed significantly in vertex count, spanning three orders of magnitude from 4,000 to 400,000 points. Scans were

scaled to metric units and rotated and translated to align with System 1 orientation and positioning. This rigid transformation aligned the feet of all scans were aligned with the ZX plane as the ground plane and the origin as the center point midway between the arches, as shown in Fig 6.1. Height normalization was not performed on a per-subject basis as scanners were calibrated to agree on metric length scales. No pose normalization was performed on the scans. Minor discrepancies in scale or pose on an individually varying basis after these preprocessing steps were treated as data sampling noise consistent with future in-the-wild applications that our model was expected to learn to be agnostic towards. No other variables from the 3D scanner other than the raw 3D position information of the body surface were used in our model. Each participant was scanned on all three devices. This was done to train the shape and regression models to behave agnostically with respect to scanning machinery specific variations. Pose and scan quality varied slightly based on scanning technology. Sampling each participant on multiple scanners allows us to generate augmentation data on a limited cohort and train the model to account for scanner induced noise in the input data [118]. Participants were scanned twice on each system to gather a test-retest precision evaluation data set. As these scanning devices were often designed for adult dimensions and were not calibrated for the stature of small children, some scans were excluded from the study due to scanning defects. Scans were excluded using quality control procedures looking for a priori quality issues including scan artifacts, excessive spatial noise, and incomplete scans [135]. No differences were noted in the study outcome variables between scans included and excluded due to participant demographics. An example of the differences present in different device captures of the same individual is shown in Fig 6.1.

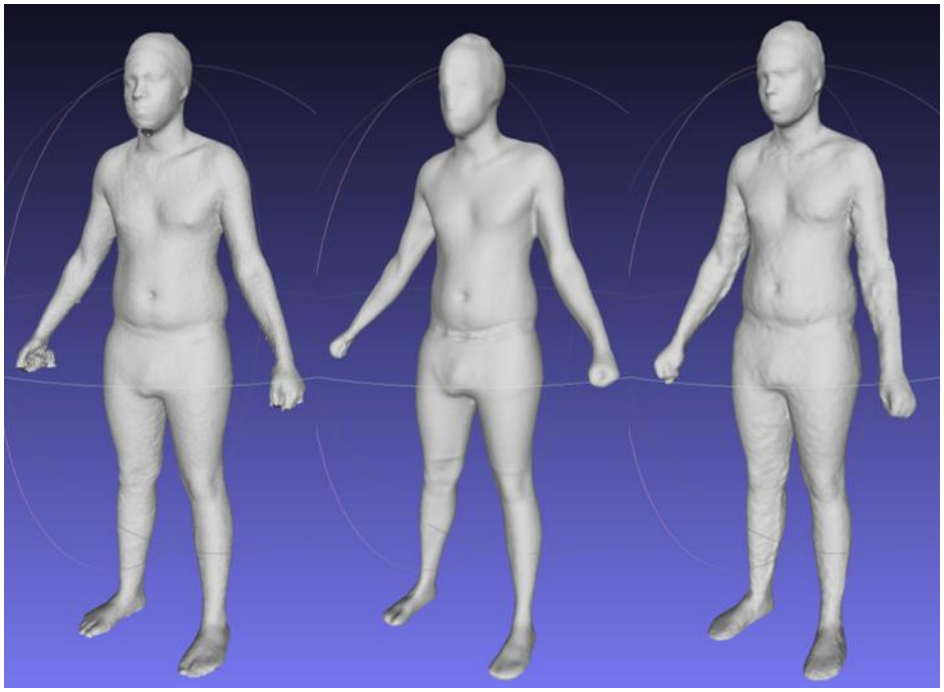
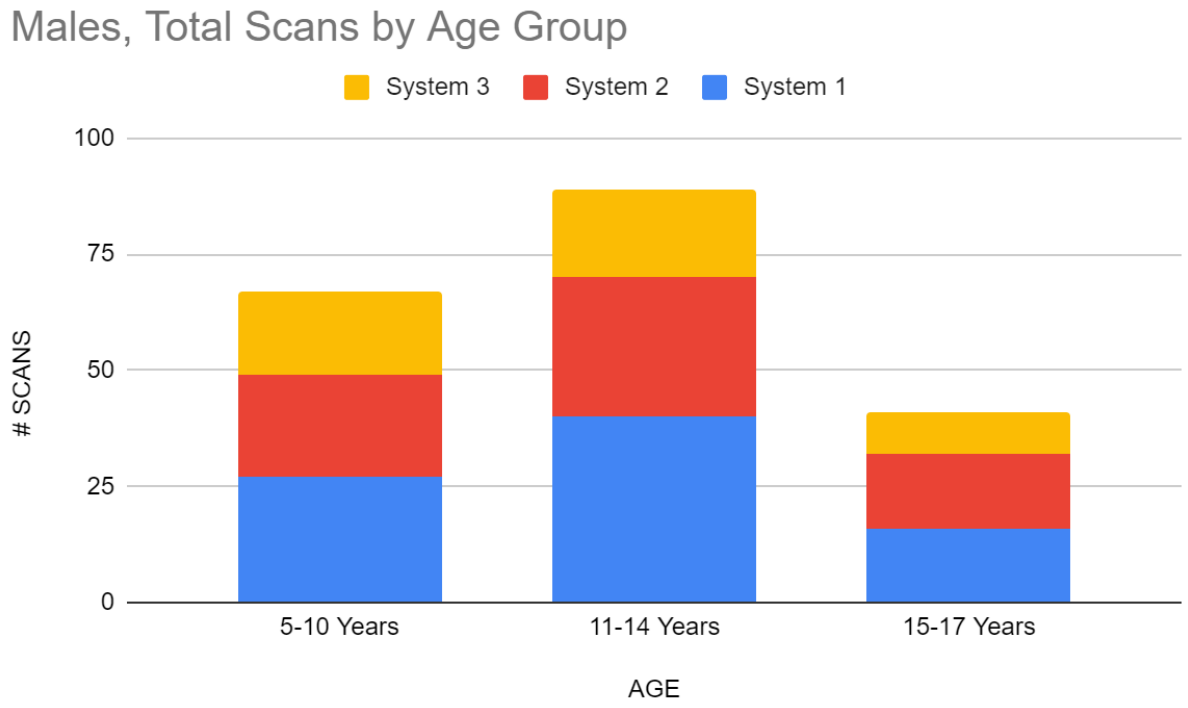


Figure 6.1. Scans of a 14-year-old male on Systems 1, 2, and 3 respectively.

There were 145 and 206 meshes in the training set for males and females, respectively. For Systems 1, 2, and 3, respectively, there were 57, 53, and 35 males scans and 88, 68, and 50 female scans in the training set. There were 52 and 80 meshes in the test set, with 26, 15, 11 males and 35, 25, 11 females for Systems 1, 2, and 3 respectively. Fig 6.2 shows the age stratification of our dataset. Using a standardized significance level of 0.05, a desired power level of 0.8, and a standardized effect size of 0.2, 0.5, and 0.8 for small, medium, and large effect sizes respectively as suggested by Cohen's d, the number of observations required for a well-powered study are 394, 64, and 26. 0.56 is the minimum effect size our study is powered to detect in the male model given the smaller 52-member cohort while 0.45 is the minimum for females with 80 test scans. Our study is well-powered to detect a large effect size in males and a medium effect size in females.

Male and female models were trained and tested separately. All participants in the dataset were represented by between one and three scans. Data acquisition protocol specified two scans on each of three scanning devices in order to generate augmentation data and retest precision data. However, some scans were dropped from the dataset due to poor quality or incompleteness [135, 118]. Because participants were scanned with more than one scanning system, there were only 83 and 119 unique participants in the training set for males and females respectively, and 30 and 38 unique participants in the test set. The test set was a randomly selected set of 20% of the total unique participants.



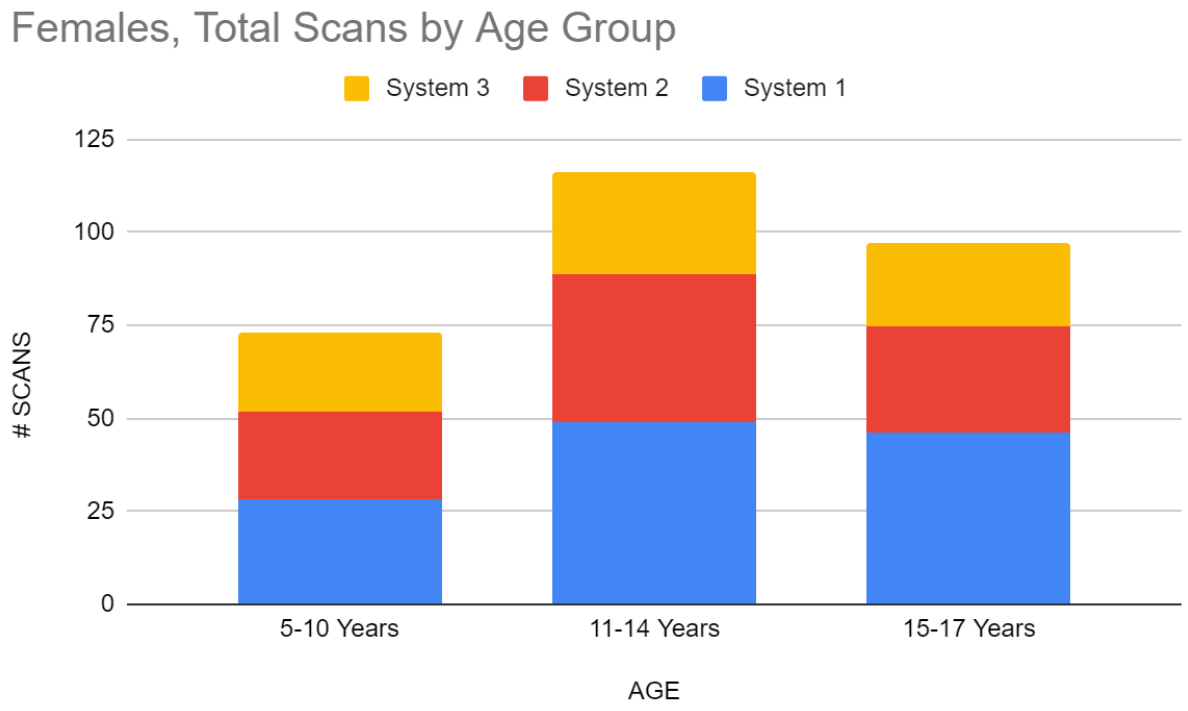


Figure 6.2. Scan count for each scanning device in each age category.

6.3.2 3D Scan Templating and Shape Model Construction

To estimate the body composition of novel unorganized 3D scans, we constructed a unified PCA body shape model spanning the minor pose variance observed among our three scanning devices using an automatic templating procedure as specified in Tian et al. [118]. Our initial bootstrap sets consisted of 107, 28, and 25 fitted female meshes from System 1, 2, and 3, respectively, and 75, 12, and 9 fitted male meshes. The bootstrap set meshes were template-fitted with manually annotated landmarks as specified in Allen et al. [5] using a topologically consistent 60k mesh and combined into a single initial bootstrap PCA shape model.

The bootstrap PCA model was used to initialize an automated fitting procedure to produce a final unified autofit PCA model consisting of 60k autofits of all training scan meshes. We created a random 80/20 training/test set split sorted on participant IDs instead of individual mesh files, resulting in 39 female and 30 male IDs in the test set. Since we split training and test by participant ID and not by mesh ID, we ensured no participant had one system's scan in training while another separate mesh was in test. This resulted in a test set mesh count consisting of 80 female and 52 male raw scans as some participants were scanned in up to three different scanning systems. We deformed a 60k template by optimizing its principal component domain parameters in the bootstrap model to coarsely fit all raw training scans from all devices, totaling 88, 68, and 50 female scans, and 57, 53, and 35 male scans from Systems 1, 2, and 3 respectively. This was done by optimizing the L2 regression objective:

$$E(\mathbf{w}) = \left|\left|\mathbf{A}\mathbf{w} - \mathbf{b}\right|\right|_2 + \left|\left|\mathbf{\Lambda}(\mathbf{w} - \mathbf{w}_0)\right|\right|_2 \qquad (23)$$

where:

$$\mathbf{b} = \mathbf{y} - \boldsymbol{\mu} \quad (24)$$

for PCA transformation matrix **A**, PCA mean vector $\boldsymbol{\mu}$, regularization matrix **Λ**, target surface **y**, and initialization shape vector $\mathbf{w}_0$. We solve for w by performing iterative least squares optimization. At each step, we populate the target surface vector y with the closest points in the 3DO scan to the vertices of the current fitted surface estimate represented by **Aw + b**, and then solve for w via linear least squares. Since this method assumes rough anatomical alignment between the template and the target meshes, an initialization vector comprised of [height, weight, age] was used to seed the template mesh shape and start the vertex pairing from $\mathbf{w}_0$, the expected shape of an individual with those defined dimensions, instead of $\boldsymbol{\mu}$, the population mean. **Λ** was a diagonal matrix with $\frac{\sqrt{\lambda}}{\sigma_i}$ at each entry, where λ was a regularization strength hyperparameter (set to 0.001) and σ_i was the standard deviation of the *i*th PCA component. PCA parameters **w** were optimized and new pairings **y** were computed iteratively until norm of the difference between $\mathbf{w}^k$ and $\mathbf{w}^{k-1}$ was less than a convergence hyperparameter (0.1) for iteration *k*.

This coarse PCA deformation resulted in an intermediate mesh shape that was close to the target raw scan and was anatomically constrained by the standard deviations of each PCA basis parameter instead of manually assigned vertex targets. This allowed us to automate the fitting procedure and scale our method to much larger datasets. A surface-to-surface alignment from [5] was used to create a final refined fit that brought the coarse-fit 60k template into close alignment with the raw scan. This method was described in greater mathematical detail in [118] for adult participants.

We constructed a new unified autofit PCA space from the automatically templated training meshes consisting of 206 females and 145 males from all three scanning systems. This

PCA model is the final model used for test set fitting and body composition prediction. We repeated the automated fitting procedure using the unified autofit PCA model as specified in [118] on the raw test set scans. A diagram of this multi-step process is shown in Fig 6.3.

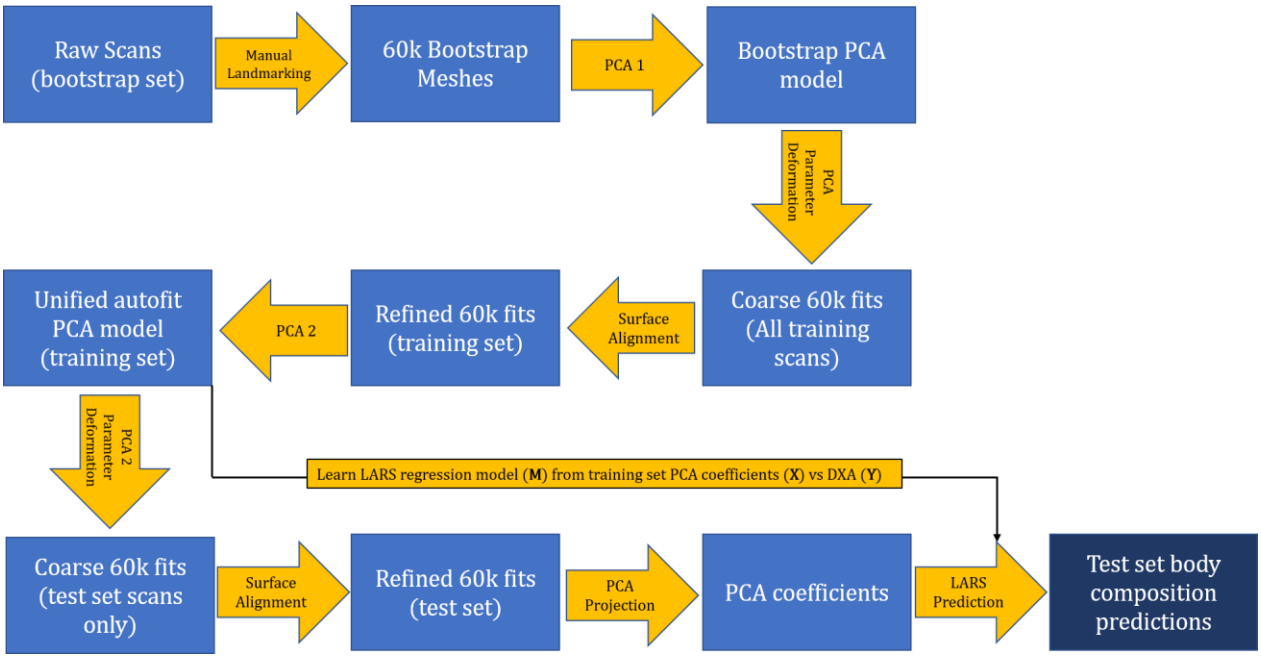


Figure 6.3. Diagram showing workflow to produce body composition predictions for unseen test scans. Operations in each row corresponds to bootstrap, training, and test data respectively.

Statistical Considerations. In this study, our goal was to generate estimates of body composition from 3DO scans. We learned body composition metrics from the projected PCA coordinates of our final refined training mesh fits in a manner similar to [118]. For all templated 60k meshes in the training set, we found the projected PCA coordinate vector **w** by multiplying its flattened 3D coordinate vector **s** with the transpose of the PCA basis, $\mathbf{A}^T$, which is also its inverse.

$$\mathbf{w} = \mathbf{A}^T\mathbf{s} \qquad (25)$$

We learned a regression matrix **M** that best maps the projected PCA basis coefficients of all training meshes **W** to the DXA body composition measurement vector of all training meshes **Y**.

$$\mathbf{Y} = \mathbf{MW} \quad (26)$$

Previous work learned the regression from shape PCA to body composition on the first *k* components that explained 95% of the shape variance in the training data [83] or all available PCA components using ordinary least squares (OLS). [118] In this work, we used a Least Angle Regression (LARS) [27] procedure to select the weights of our prediction model. We chose *k*=61 and *k*=40 for females and males respectively, as those were the number of components that explained 99.9% of the shape variance in the training set for each gender. Height, weight, age, and BMI were appended as additional regression features. We used this LARS model trained on the unified autofit training set to predict the body composition of the held-out test set scans after automatic template fitting and refined surface alignment. For comparison to LARS, we estimated body composition using four additional kinds of ensemble learning methods when mapping optimized PC coordinates to body composition: Decision tree [80], random forest [95], AdaBoost [33], and Gradient Boost [115].

We compared the predicted body composition values against DXA reference with the coefficient of determination (R^2), root-mean-squared error (RMSE), and Student's t-test.

A *p*-value of less than the Bonferroni corrected significance threshold of 0.004 for 12 measurements is undesirable as it indicates a significant difference between the means of DXA and our predictions from 3D templated scans.

To visualize the effect of changing scalar metrics on body shape, we solved the inverse regression problem that maps changes in a set of metrics of interest $\Delta\mathbf{Y}$ to a change in principal component coefficients $\Delta\mathbf{W}$ with matrix $\mathbf{X}$ analogous to section 4.3 of [5]:

$$\Delta\mathbf{W} = \mathbf{X}\Delta\mathbf{Y} \quad (27)$$

We apply the shape offset $\Delta\mathbf{W}$ to the scanned shape as discussed in [118] and [5] to simulate a change in a body composition variable, or for growing children, an advancement in age.

6.4 Results

DXA reference values for the study population are shown in Table 6.1. Training and test set properties were not significantly different for any measured variable (p-value > 0.004 after Bonferonni correction). A visualization of the distribution of scans by device for each age group in the training set is shown in Fig. 6.2. Age categories were delimited based on recruitment stratification in Shape Up! Kids.

Table 6.1. DXA reference measurement statistics for unique individuals in the training and test sets. Plus-minus values are standard deviation. p-values represent significance values for the t-test between the means of training and test sets.

	Male						
	Train (N = 83)			Test (N = 30)			*p*-value
	Mean ± SD	Min	Max	Mean ± SD	Min	Max	
Age (Years)	11.7 ± 3.0	5	17	11.8 ± 2.8	7	17	0.9
Height (m)	1.5 ± 0.2	1.1	1.9	1.5 ± 0.2	1.3	1.8	1.0
Mass (kg)	55.2 ± 24.5	20.7	145.6	53.3 ± 21.5	23.0	107.2	0.7
BMI	22.1 ± 6.6	14.4	53.0	21.8 ± 5.8	14.6	39.5	0.8
BMI-z	0.8 ± 1.2	-2.3	3.2	0.7 ± 1.2	-2.1	3.0	0.8
% Fat	24.1 ± 9.6	8.5	46.7	23.7 ± 9.2	9.0	50.1	0.8
Lean Mass (kg)	41.2 ± 16.3	16.7	86.9	40.2 ± 15.7	17.0	80.1	0.8
Fat Mass (kg)	14.0 ± 11.2	3.9	68.0	13.1 ± 8.7	4.3	35.9	0.7
Visceral Fat (kg)	0.2 ± 0.1	0.02	0.9	0.2 ± 0.1	0.1	0.4	1.0
Leg Lean (kg)	7.0 ± 3.0	2.4	15.8	6.8 ± 2.9	2.4	14.1	0.8
Arm Lean (kg)	2.3 ± 1.1	0.8	5.6	2.3 ± 1.1	0.8	5.0	1.0
Trunk Lean (kg)	19.1 ± 7.7	7.3	39.7	18.3 ± 7.5	7.7	37.4	0.6
Trunk Fat (kg)	5.4 ± 5.5	1.0	34.9	4.9 ± 3.9	1.2	14.7	0.6
Leg Fat (kg)	3.0 ± 2.1	0.8	11.9	2.9 ± 1.8	1.0	8.0	0.8
Arm Fat (kg)	0.9 ± 0.8	0.2	4.1	0.8 ± 0.8	0.2	2.4	0.6
FMI	5.8 ± 3.9	1.6	21.9	5.5 ± 3.7	2.0	19.7	0.8
FFMI	16.6 ± 4.1	11.6	36.8	16.2 ± 3.1	10.8	24.1	0.6

	Female						
	Train (N = 119)			Test (N = 38)			*p*-value
	Mean ± SD	Min	Max	Mean ± SD	Min	Max	
Age (Years)	12.2 ± 3.2	5	17	12.9 ± 3.2	6	17	0.2
Height (m)	1.5 ± 0.1	1.1	1.8	1.5 ± 0.1	1.2	1.9	0.5
Mass (kg)	53.5 ± 20.8	18.4	140.4	59.6 ± 22.0	27.1	124.1	0.1
BMI	22.7 ± 6.4	13.1	52.2	25.0 ± 7.3	14.1	46.9	0.1
BMI-z	0.8 ± 1.2	-2.4	2.9	1.0 ± 1.2	-1.5	2.6	0.3
% Fat	31.2 ± 7.3	13.9	47.9	33.1 ± 8.0	17.7	47.2	0.2
Lean Mass (kg)	35.8 ± 11.0	13.3	74.4	38.8 ± 11.7	17.2	68.2	0.2
Fat Mass (kg)	17.7 ± 10.8	4.0	66.0	20.8 ± 11.4	6.2	55.9	0.1
Visceral Fat (kg)	0.2 ± 0.2	0.01	0.8	0.2 ± 0.2	0.03	0.7	0.5
Leg Lean (kg)	6.0 ± 2.1	1.7	12.3	6.5 ± 2.1	2.4	11.2	0.2
Arm Lean (kg)	1.9 ± 0.6	0.7	4.1	2.0 ± 0.7	0.9	3.9	0.1
Trunk Lean (kg)	16.9 ± 5.5	6.2	37.5	18.4 ± 5.8	8.0	34.3	0.2
Trunk Fat (kg)	7.2 ± 5.4	1.1	33.4	8.9 ± 6.2	1.8	31.5	0.1
Leg Fat (kg)	3.7 ± 2.0	0.9	11.0	4.2 ± 1.9	1.5	7.9	0.2
Arm Fat (kg)	1.1 ± 0.8	0.2	4.8	1.3 ± 0.8	0.3	4.0	0.1
FMI	7.4 ± 3.8	2.2	24.4	8.7 ± 4.3	2.8	21	0.1
FFMI	15.2 ± 2.9	10.1	27.5	16.2 ± 3.3	10.7	25.7	0.1

An example of an automated 60k templated fit from the test set fit using our method is shown in Fig. 6.4.

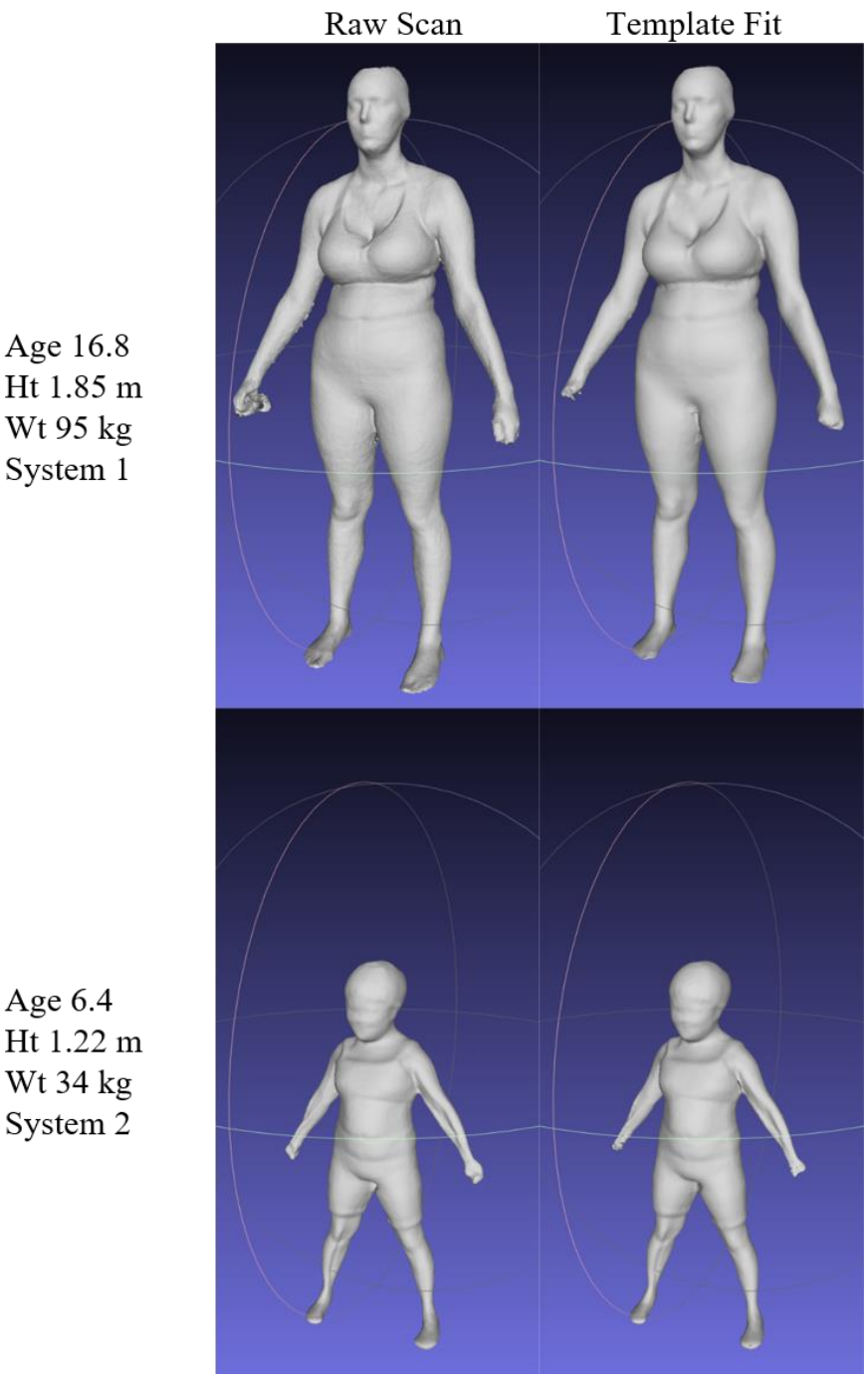


Figure 6.4. Automated template fitting results on two females from the test set at opposite ends of the physical maturity spectrum. Top subject was a very tall 16-year-old and had adult like proportions, scanned on System 1 with 404k vertices in the raw file. Bottom subject was a very small 6-year-old scanned on System 2 with 51k vertices.

Body composition prediction accuracy and *t*-tests against DXA from automated 3D optical scan fits for the LARS method are shown in Table 6.2. The values are mostly higher than the analogous results on adults in Tian et al. with the exception of male visceral fat. No predictions were significantly different from DXA measurements except for male trunk lean mass. The Demographics Only column shows the baseline regression accuracy using only the supplied initialization variables [height, weight, age] in the absence of any PCA shape information. Improvement upon these values justifies our methods for increasing the composition prediction accuracy beyond regressions from just the initialization scalars. All predicted metrics were better in our 3D template fit predictions than the baseline predicted by height, weight, and age with no shape fitting. Even in metrics that started at very high R^2 values such as lean mass and trunk lean, the RMSE values improved by as much as 20-40% after 3D fitting. A breakdown of prediction values by scanning device is shown in Table S1 of the Supplemental Data to detect biases in any single scanning system. Only visceral fat in both sexes was significantly different from DXA and underperformed relative to the combined cohort. No other biases were detected. None of the other ensemble learning methods (decision tree, random forest, AdaBoost, and Gradient Boost) performed as well as LARS method (data not shown). LARS regression was in practice only different from an OLS solution on the full component vector by around 0.02 R^2 on test data but was a more conservative model as it used less than a third of the total variables to learn the mapping.

Table 6.2. Body composition prediction accuracy against DXA reference measurements on test data for 3D automated template fits. Shapes were fit using the full PCA basis (145 components for males, 206 components for females) but body composition was predicted on k components selected using LARS, where k = 40 for males and 61 for females. % fat was calculated as fat mass divided by known weight. p-values indicate difference from DXA measurement, where the Bonferroni corrected threshold for significance for 12 measurements is <0.004.

	Male, d=145, n=52			**Female, d=206, n=80**		
	R^2 (RMSE)	***p*-value (comparison of means)**	**Demographics Only R^2** (RMSE)	**R^2** (RMSE)	***p*-value (comparison of means)**	**Demographics Only R^2** (RMSE)
% fat	**0.88** (3.55)	0.09	**0.70** (5.14)	**0.85** (3.20)	0.52	**0.68** (4.47)
Lean Mass (kg)	**0.99** (1.74)	0.12	**0.96** (3.30)	**0.98** (1.68)	0.44	**0.97** (2.17)
Fat Mass (kg)	**0.96** (1.74)	0.12	**0.86** (3.30)	**0.98** (1.68)	0.44	**0.97** (2.17)
Visc. Fat (kg)	**0.65** (0.06)	0.004	**0.45** (0.07)	**0.78** (0.08)	0.004	**0.64** (0.1)
Leg Lean (kg)	**0.98** (0.4)	0.81	**0.96** (0.61)	**0.96** (0.45)	0.23	**0.92** (0.6)
Arm Lean (kg)	**0.96** (0.21)	0.81	**0.87** (0.39)	**0.90** (0.21)	0.020	**0.88** (0.24)
Trunk Lean (kg)	**0.98** (0.93)	**0.002**	**0.96** (1.52)	**0.98** (0.92)	0.14	**0.96** (1.14)
Trunk Fat (kg)	**0.91** (1.22)	0.93	**0.81** (1.78)	**0.96** (1.24)	0.24	**0.93** (1.64)
Leg Fat (kg)	**0.95** (0.43)	0.71	**0.83** (0.76)	**0.93** (0.5)	0.008	**0.89** (0.64)
Arm Fat (kg)	**0.93** (0.16)	0.07	**0.87** (0.22)	**0.95** (0.19)	0.22	**0.91** (0.25)
FMI	**0.97** (0.74)	0.08	**0.90** (1.19)	**0.97** (0.76)	0.41	**0.96** (0.9)
FFMI	**0.94** (0.74)	0.08	**0.86** (1.19)	**0.95** (0.76)	0.41	**0.93** (0.9)

Table 6.3 shows test-retest precision of the 3DO estimates as measured by the coefficient of variation (%CV) and defined in Glüer *et al*. [39]. Duplicate DXA scans were not acquired to be radiation dose conserving. Shepherd *et al*. [109] showed that the %CV of DXA in children aged 6-16 on bone mineral density and content was age dependent but typically less than 2%. Although we lose some precision with our optical method, we note that we gain the ability to

collect multiple data points in rapid succession or over a longitudinal study without risk of radiation exposure.

Table 6.3. Test-retest precision for 3D auto templated composition predictions. All components were used for shape fitting. k components were selected using LARS for prediction regression, where k = 40 and 61 for males and females respectively.

	This Work			
	Male n=43		Female n=65	
	%CV	**RMSE**	%CV	**RMSE**
% fat	-	1.88	-	1.77
Lean Mass (kg)	1.84	0.75	2.23	0.88
Fat Mass (kg)	6.09	0.75	4.01	0.88
Visc. Fat (kg)	10.1	0.02	10.75	0.03
Leg Lean (kg)	2.53	0.17	2.06	0.14
Arm Lean (kg)	3.52	0.08	4.67	0.09
Trunk Lean (kg)	1.77	0.34	1.96	0.37
Trunk Fat (kg)	10.14	0.47	4.44	0.42
Leg Fat (kg)	6.79	0.19	5.07	0.23
Arm Fat (kg)	8.71	0.06	3.19	0.04
FMI	6.99	0.35	4.28	0.39
FFMI	2.18	0.35	2.37	0.39

We computed the delta-weight vector corresponding to age changes and added this vector to a templated scan multiplied by a constant to represent change in years as in section 4.3 of [5]. This process can generalize to any number of variables, but in our visualization, we solved for a hyperplane over the age / BMI domain and kept BMI fixed while incrementing age in order to visualize the effect of aging for an individual of those particular dimensions. Fig. 6.5 shows examples of extrapolating young children to 18 years of age while keeping predicted BMI constant. A high and a low BMI example was tested for a male and a female participant.

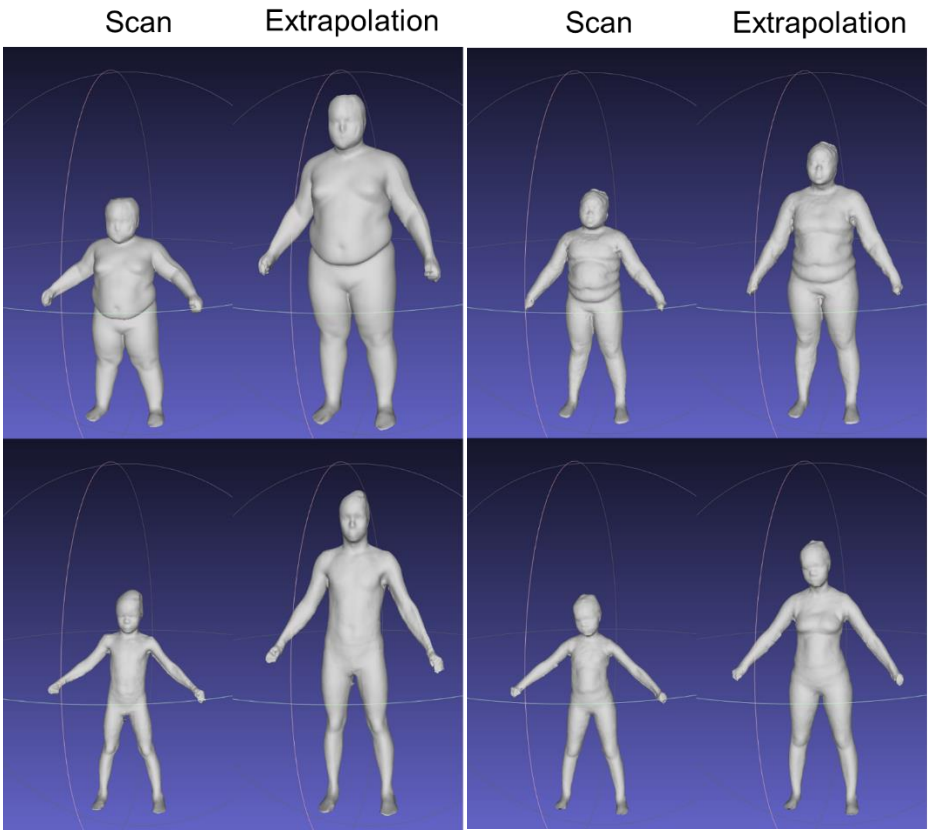


Figure 6.5. (A) Two males scanned at 7 (top) and 8 (bottom) years of age with BMIs of 40 and 16, extrapolated to age 18. Starting DXA measured percent fats were 50% and 21% respectively. Extrapolated percent fats were 35% and 18%. (B) Two females scanned at 9 (top) and 7 (bottom) years of age with BMIs of 29 and 17, extrapolated to age 18. Starting DXA measured percent fats were 47% and 37%. Extrapolated percent fats were 39% and 33% respectively.

These extrapolated images capture the effects of puberty in its age-progression deformation, such as changes in limb proportions and development of secondary sex characteristics. Since Tanner staging was self-reported and thus not used as a stratification variable in this study, age served as the prior for pubertal maturity in our model. Fig. 6.6 shows expected body shapes generated using our PCA model for a female measuring 1.6m tall and 50kg at ages 9, 13, and 17. This regularization allowed our model to disentangle pubertal

maturity from individuals of similar stature, a distinction that was not necessary in prior work on adult cohorts.

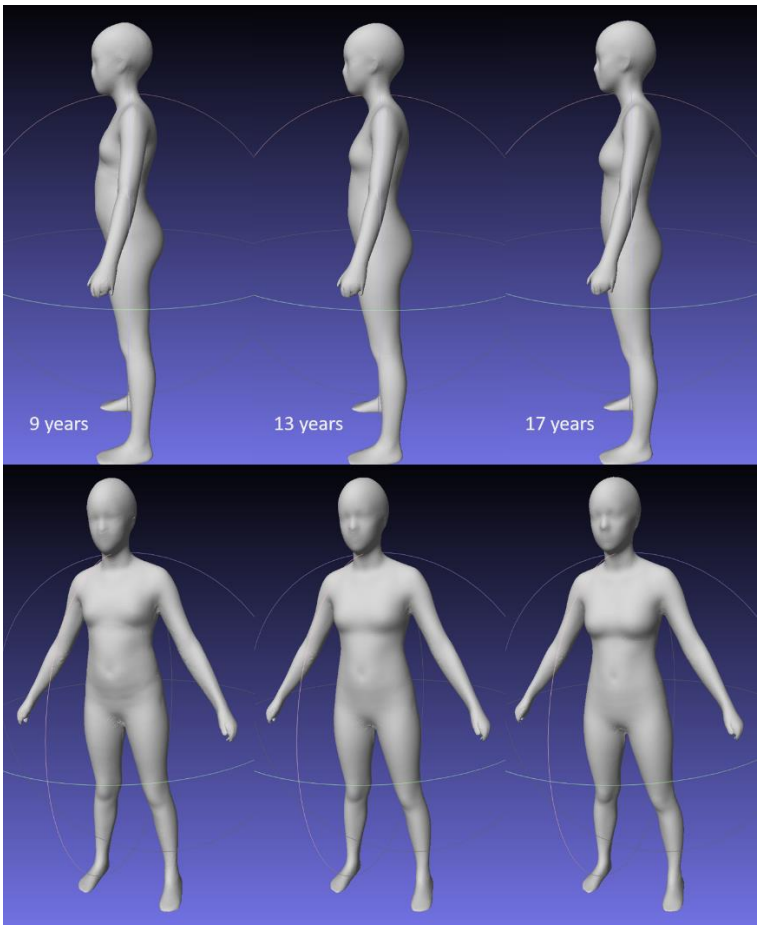


Figure 6.6. Generated average shapes for a 1.6m, 50kg female at ages 9, 13, and 17. Tanner staging was self-reported and not used in this study; we use age as a surrogate in this study to act as the prior for pubertal development in our model.

Fig. 6.7 shows Bland-Altman plots with associated best-fit lines for fat mass, lean mass, and visceral fat for female and male test set predictions respectively. Equations for the best-fit lines on each plot and the *p*-values for the coefficients are shown in the top right corners. This study was designed to develop a model that could perform generalized device, age, and shape

agnostic body shape and composition predictions to any 3DO input that fell within the bounds of our training parameters and not to distinguish differences between scanners or groups. Prediction models learned continuous interpolations of densely sampled body shape variation within a representative population and thus required inclusion of a diverse range of training and testing examples in the experimental cohort. The plots on Fig. 6.7 show that outliers are uniformly distributed. Our method achieves consistently accurate predictions across the range of body compositions in our test data without significant bias.

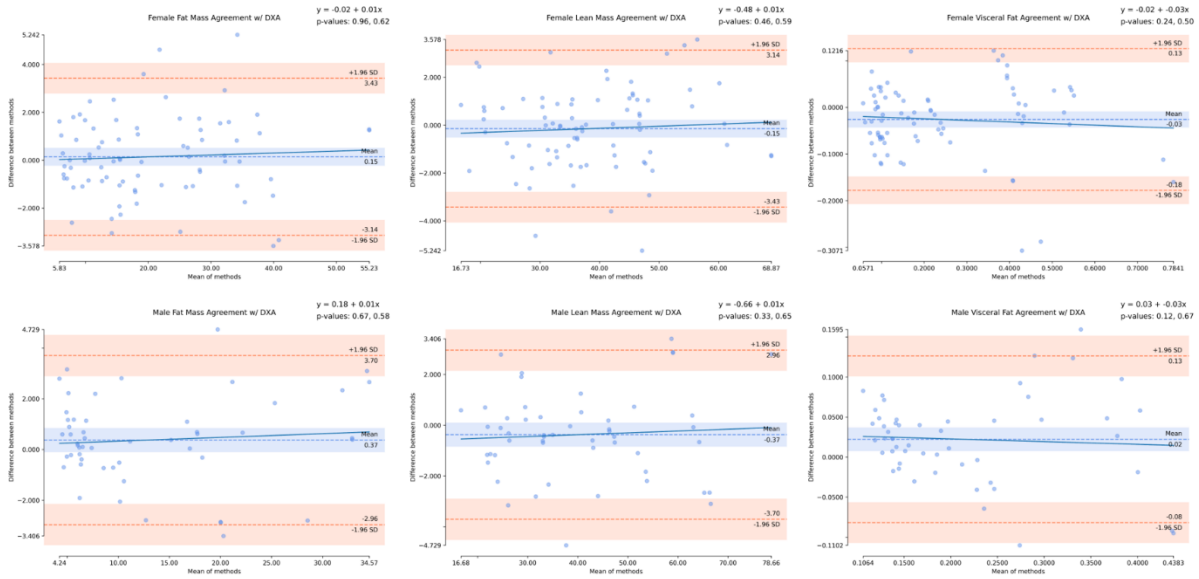


Figure 6.7. Bland-Altman plots for fat mass, lean mass, and visceral fat for female and male test set predictions with best-fit line equations. Best fit line equation and p-value for the coefficients are shown in the top right corners. No significant biases were detected.

6.5 Discussion

We presented the most complete 3DO-based solution for total and regional body composition analysis in children and adolescents at the time of writing. Wells et al. [130] studied

a sample of 1484 children between the ages of 5-11 with 3DO scanning but only tested the accuracy of collecting scalar anthropometric features from scans without consideration of body composition. Santos et al. [103] used the features collected by this method to construct a PCA space similar to our work, but only analyzed the correlation between individual principal components with anthropometric and body composition measurements and did not actually predict the body composition metrics of test data. Our method takes raw, unformatted 3DO scans as input and predicts 12 body composition variables end-to-end with no intermediate manual processing required. Our automated system enabled fast processing of raw 3DO scans that yielded 3-5 times more templated meshes for some scanning systems than were previously available through the manual annotation pipeline. Our method worked on young children and adolescents despite the greater range of body size variation among this age group relative to an adult-only cohort.

Our study is currently underpowered for small effect sizes in females and medium effect sizes in males. With a target recruitment of 720 total individuals for Shape Up! Kids (NCT03706612) and an approximate 50/50 sex distribution, our study is not powered to detect a small effect size of d=0.2 between DXA and 3DO prediction. The power level of our study will not affect the primary metric of accuracy used for determining agreement between two measurements, R^2 and RMSE, or the precision of repeat trials (%CV). A study underpowered for small effect sizes may prevent us from detecting a statistically significant, systemic bias between 3DO prediction and DXA as measured by the p-value in Table 6.2.

The wide variation of body shapes between young children and more mature teens presents additional challenges for evaluating body composition in a pediatric population. For example, although a BMI of 17 in Fig. 5B is not alarming, an associated 37% body fat increases

cardiovascular disease risk by 2-4 times in adults [72]. Our model predicted this value to fall to 33% at age 18. Intervention could be recommended based on current body fat percentage but not current BMI. Additional work is required to accurately model how individual bodies change shape and composition over time as influenced by puberty and maturity.

We tried four kinds of ensemble learning methods when mapping optimized PC coordinates to body composition: Decision tree, random forest, AdaBoost, and Gradient Boost. We found that these alternatives to LARS did not perform as well with this data type. None of them were close to achieving an 0.85 R^2 for fat mass on the test data. Such methods may be more useful when larger datasets are available, as fragmenting an age 5-17 cohort into many subsets creates heavily biased, weak estimators that do not perform well on test data that may be drastically different from the training data.

We tried alternatives to PCA for constructing a body shape latent space, namely supervised dimensionality reduction using linear discriminant analysis (LDA) [140] with percent fat and visceral fat mass as discriminant variables, and nonparameterized prediction methods such as Gaussian Process Regression (GPR) [106]. PCA had advantages over both methods. LDA created projections that cluster data points around their assigned labels, which did not translate well to continuous regression. GPR performed better than parameterized regressions in our previous adult study [118] but the test dataset was too small to adequately compare PCA to GPR.

Children have a lower safety tolerance for X-ray radiation than adults [140], and this imposes a single scan limit with a long minimum time interval between scans. Although the precision of our 3DO method for body composition estimation is 2-5 times lower that of DXA

bone density precision on kids, this limitation can be mitigated due to the ease of repeating of 3DO scanning. As we showed in Tian et al. [118], the least significant change (LSC) [108] is:

$$LSC = 2.77 * RMSE_{precision} \quad (28)$$

This value drops with the square root of the sample number [14]. The LSC for fat mass is around 2kg for children. Four samples at a given time point is enough to drop the LSC below 1kg, which is comparable to the LSC of DXA on adults from Tian et al [118]. As each scan takes around a minute, is radiation-free, and can be processed automatically, this increase in procedure time is an acceptable trade off to improving the monitoring precision. The speed of the scan capture is especially valuable for very young children, as they have greater difficulty staying still for the requisite amount of time needed for DXA or MRI (10-30 minutes).

2 scans on each of 3 scanning systems were collected as part of the data protocol to train our model to be robust and agnostic to pose variation and scanner specific geometry characteristics. However, some scans were dropped due to hardware or software errors on the side of the scanner manufacturer. Future data captures may opt to collect 3 or 4 scans on each system with the intent of both lowering the LSC and to build in redundancy in the data capture such that a minimum floor of 6 scans, two on each device, can be guaranteed for every participant in the dataset.

Pose variation is more extreme in younger children than in adults or postpubescent teens. Fixed handrail and feet placements in some scanners will naturally cause more arm and leg abduction for smaller bodies and could introduce additional variations in the shape model that do not correlate well to body compositions. We experimented with a regression from all 60k

vertices in the test set to body composition instead of from PCA coefficients, eliminating compression loss from the PCA encoding as a source of error. These results were similar (within 0.02 R2 in either direction) to PCA coefficient regression, suggesting the variance in the shapes themselves contribute to most of the inaccuracy. Pose normalization [134] may focus the variance of the 3D shapes on physiological change and decrease the impact of pose on the shape model, leading to more precise predictions of body composition.

This study has several limitations left to be addressed by future work. Shape differences between individuals of different ages are much more drastic in a 5-17 aged cohort than in an adult study and can vary substantially between individuals of the same age depending on rate of growth and maturity. It is necessary to have more data relative to an adult study to fully capture the continuum of growth and variation in the preteen to teen age range. The currently available participant cohort is about half the size of the adult study population in. Additional data availability may increase the accuracy of the predictions by fully representing the variance between age ranges and scanners, enabling the use of more flexible nonlinear models such as GPR. Furthermore, because the bodies of juveniles and adolescents can change dramatically in a period of time that may be shorter than a typical clinical follow-up, longitudinal studies in this age range are required to isolate the relationship between body composition and body shape controlled for natural growth and maturation.

6.6 Conclusion

3DO imaging technologies can estimate body composition in children and adolescents with accuracy well correlated to and not significantly different from reference DXA

measurements and provides estimates on many variables not accounted for by commonly used body composition measurement proxies, such as BMI. Precision error of our 3DO method is within 2-5 times that of DXA, a tradeoff that can be compensated for with repeat measurements. We applied a method for automatically fitting a topologically consistent 60k template mesh to arbitrary 3DO scans and predicting body composition metrics from the standardized meshes. Our previous work using this technique was only trained and tested on an adult population (age 18+) and was not assumed a priori to work on children and adolescents. Although the precision error as measured by %CV was worse than DXA bone mineral density benchmarks, the rapid (<1 minute) scan time and the lack of ionizing radiation allows for multiple samples to be collected either in quick succession to mitigate prediction noise or over a monitoring period to plot longitudinal change. Future work can focus on further reducing the barrier to collecting body composition data by allowing for predictions on 2D image inputs such as in [117].

Our method is safe and fast, with end-to-end automation between 3DO inputs and predicted outputs and can be deployed in non-clinical settings where radiological imaging would be prohibitively expensive. 3DO imaging is a promising tool for monitoring childhood obesity during a crucial developmental period that may have long-term implications on future health and metabolic risk.

6.7 Appendix

Table S1. Results from Table 6.2 broken down by input system to check for bias towards or against a single scanning device. Visceral fat in System 1 underperformed for both genders, but all other metrics were consistent with combined metrics.

	Males System 1 n=26		**Males System 2 n=15**		**Males System 3 n=11**	
	R^2 (RMSE)	***p*-value**	**R^2** (RMSE)	***p*-value**	**R^2** (RMSE)	***p*-value**
% fat	**0.89** (3.1)	0.05	**0.89** (3.64)	0.62	**0.84** (4.34)	0.16
Lean Mass (kg)	**0.99** (1.68)	0.12	**0.98** (1.96)	0.94	**0.98** (1.54)	0.31
Fat Mass (kg)	**0.96** (1.68)	0.12	**0.96** (1.96)	0.94	**0.97** (1.54)	0.31
Visc. Fat (kg)	**0.54** (0.07)	0.03	**0.77** (0.05)	0.14	**0.67** (0.05)	0.36
Leg Lean (kg)	**0.98** (0.41)	0.65	**0.98** (0.37)	0.65	**0.97** (0.39)	0.46
Arm Lean (kg)	**0.97** (0.19)	0.59	**0.94** (0.26)	0.48	**0.97** (0.14)	0.25
Trunk Lean (kg)	**0.99** (0.89)	0.03	**0.98** (1.04)	0.40	**0.98** (0.85)	0.01
Trunk Fat (kg)	**0.94** (1.01)	0.45	**0.89** (1.52)	0.78	**0.89** (1.23)	0.74
Leg Fat (kg)	**0.94** (0.46)	0.70	**0.95** (0.45)	0.66	**0.97** (0.29)	0.31
Arm Fat (kg)	**0.93** (0.16)	0.23	**0.93** (0.17)	0.51	**0.95** (0.14)	0.23
FMI	**0.97** (0.69)	0.07	**0.97** (0.82)	0.89	**0.98** (0.73)	0.19
FFMI	**0.94** (0.69)	0.07	**0.92** (0.82)	0.89	**0.92** (0.73)	0.19

	Females System 1 n=35		Females System 2 n=25		Females System 3 n=11	
	R^2 (RMSE)	*p*-value	R^2 (RMSE)	*p*-value	R^2 (RMSE)	*p*-value
% fat	**0.84** (3.06)	0.73	**0.86** (3.22)	0.42	**0.85** (3.41)	0.09
Lean Mass (kg)	**0.97** (1.82)	1.0	**0.99** (1.34)	0.72	**0.97** (1.8)	0.08
Fat Mass (kg)	**0.98** (1.82)	1.0	**0.99** (1.34)	0.72	**0.97** (1.8)	0.08
Visc. Fat (kg)	**0.65** (0.1)	0.04	**0.88** (0.06)	0.01	**0.85** (0.07)	0.79
Leg Lean (kg)	**0.95** (0.47)	0.35	**0.97** (0.4)	0.78	**0.95** (0.45)	0.44
Arm Lean (kg)	**0.87** (0.24)	0.025	**0.93** (0.18)	0.23	**0.93** (0.16)	0.97
Trunk Lean (kg)	**0.97** (0.95)	0.71	**0.99** (0.77)	0.64	**0.96** (1.04)	0.004
Trunk Fat (kg)	**0.95** (1.39)	0.90	**0.97** (1.13)	0.38	**0.96** (1.07)	0.16
Leg Fat (kg)	**0.9** (0.58)	0.008	**0.95** (0.43)	0.10	**0.96** (0.42)	0.74
Arm Fat (kg)	**0.95** (0.18)	0.46	**0.96** (0.17)	0.81	**0.91** (0.22)	0.29
FMI	**0.97** (0.77)	0.90	**0.98** (0.68)	0.66	**0.96** (0.85)	0.08
FFMI	**0.95** (0.77)	0.90	**0.96** (0.68)	0.66	**0.92** (0.85)	0.08

7. Deep 3D Autoencoder Model Accuracy on Detailed 3D Human Full Body Shape: A Systematic Review

Work presented in the prior chapters and in the preceding literature exclusively based their models on linear methods. In the context of body composition analysis from optical imager, PCA was the only shape model used for body shape modeling. A natural evolution of ongoing work was to investigate the viability of nonlinear methods for shape modeling and regression training. Since PCA was a linear autoencoder algorithm, we searched the computer science and machine learning literature for all deep 3D autoencoder methods that dealt with modeling human body shape to gain a broad understanding of the architectures used in this application when agnostic to clinical applications. An architecture from this review that covers the weaknesses of PCA, such as lack of local spatiality in the feature space and a restriction to linear subspaces, may be adopted for the purpose of body composition analysis.

7.1 Abstract

Deep 3D autoencoders have seen a recent explosion of development with the rising accessibility of deep networks for 3D data. Deep representation of total 3D human shape is a subfield of 3D autoencoders that presents unique challenges due to the nonrigid nature of the target data. Specific application of deep 3D learning to human shape has many potential applications including ergonomics and medical care. However, comparative modeling accuracies are not well understood due to a lack of uniformity in the datasets used and in the performance metrics reported. This opaqueness makes it difficult to draw conclusions about the competitive performance of different network architectures or direct future efforts towards the most promising methods. In this systematic review, we surveyed the state of the art in deep 3D

autoencoders for human body shape with unsupervised machine learning between 2003 and 2022. We identified 25 works representing multiple categories of shape modeling methods and directly compared quantitative results across works where available normalized for the degree of dimensionality reduction using a newly defined metric, the reconstruction-compression quotient (RCQ).

7.2 Introduction

Three-dimensional autoencoders (3DAE) for total human body embedding pose a unique set of challenges when compared to the general body of work on learning deep representations of rigid 3D shapes [54]. Human bodies deform non-rigidly both between individuals and within the same individual due to differences in skeletal proportions, body mass and composition, posing, and motion. Learning accurate models of these body variations among a diversified population can advance the accuracy and realism of human shape representation in many applications that depend on high fidelity virtual humans, such as virtual reality [138], telecommunications [98, 42], gaming [145], fitness tracking [120], garment simulation [127], and healthcare [61].

One of the most widely cited methods for generative human body shape modeling is SMPL [67], which is built on a linear principal component analysis (PCA) derived shape parameterization paired with a skinned kinematic skeleton for reposing. PCA is a linear autoencoder that simplifies body shape variation to linear deformations and is insensitive to local features due to its lack of spatial kernels. These limitations introduce inherent inaccuracies in high-fidelity body shape embedding that may be overcome with deep nonlinear autoencoders with local feature encoding. However, the competitive performance and accuracy of different 3DAEs on total human body shape modeling is not well understood due to a lack of a standardized benchmark

such as ShapeNet [18]. A comprehensive survey of the state of the art in this specific field of 3D shape learning is required to identify the best-performing methods, the most important variables that improve model performance, and the most fruitful directions to explore to further advance the field.

In this work, we perform a systematic review of the computer science and machine learning literature to determine which architectures, datasets, and performance metrics represent the current state of the art in deep representation of total human body shape. We quantitatively compared the reconstruction accuracies of various 3DAE human body shape methods and qualitatively analyzed and discussed the viability of various methods based on end-use scenarios. Based on our experience in medical and clinical applications of 3D human body shape, we offer commentary on limitations and deficiencies of current work and provide recommendations on future work to develop models that maximize end-user usability and impact on various applications.

To our knowledge, no systematic review has been completed on deep 3D autoencoder reconstruction of total human body shapes. Recently, Muhammad et al. [78] conducted a review of 3D human pose and shape recovery. This review was a broad survey of 3D human shape and pose reconstruction from 2D image or video inputs, focused on a narrow time period between 2018 - 2021 including only three conferences: ECCV, CVPR and ICCV. The authors surveyed in-the-wild pose and shape recovery methods including monocular reconstruction from 2D images and did not focus on evaluating the comparative quantitative performances of deep 3D autoencoders where the reference 3D shape is known a priori. In this work we will perform a systematic review focusing on a narrow inclusion criterion that targets reconstruction of unclothed 3D human shape with ground truth evaluation and millimeter-scale resolution while

broadly surveying the 20-year period between 2003 and 2022 spanning a majority of journals and conferences in the computer science literature. We will only include works that introduce new learned models, methods, or datasets that can advance the state-of-the-art in high-fidelity total body encoding and reconstruction, as opposed to works that apply an existing model such as SMPL to an estimation task without fundamentally altering the learned model.

Cheng et al. [20] reviewed parametric modeling of 3D human body shape in 2018 but only included six models with linear parameterizations built on PCA and did not discuss recent techniques involving deep network autoencoders on 3D meshes, many of which were published after their survey. Nonlinear autoencoding for 3D human body meshes is an open problem without a deterministic and optimal solution like linear PCA parameterization. As such, the encoder-decoder design can vary greatly and the only way to compare relative performance is to trace fragmented comparative results in each individual primary study across publications in multiple conferences and journals. We will systematically compare the design and performance of the various unsupervised learning studies on 3D human parametric models with a focus on deep, nonlinear autoencoder methods.

Our contributions in this work are listed as follows:

1. Standardized taxonomy: We identify and define the defining characteristics of 3DAEs for total human body shape and introduce a normalized error metric for comparing current and future work in this space.
2. Dataset comprehension: We describe the various 3D human shape datasets used in current work in 3D shape modeling for human bodies and discriminate use cases for which certain datasets may be more or less appropriate. Future work may refer to this

resource when designing experiments or design data capture to fill in gaps left by the current state of human body shape data.

3. Future directions: We propose standards and conventions for future work on 3D human shape modeling to adhere to in order to generate results comparable to prior work for easy comparison. We identify components of human shape encoders whose isolated effects on reconstruction performance are unresolved by the current literature. Future work can target specific unanswered questions about human shape encoding with standardized procedures, dataset benchmarks, and ablation studies of architecture.

7.3 Definition of Terminologies

Deep 3D shape models are data driven and computed from a set of training shapes composed of 3D total body surface scans of real participants. Bartol et al. [8] describes the common technologies of 3D body scanning used to capture real world body shape data in their survey. Scan databases can be clothed or unclothed (i.e. wearing skintight underwear that does not obstruct the base body shape) and can be pose constrained or pose varied in their captures. Datasets containing unclothed, pose constrained captures of a diverse cast of participants are the most valuable for building a parametric model of human body shape variation as all other sources of shape variance are normalized. However, **multi-identity** scan databases containing hundreds to thousands of unique individuals are rare and difficult to acquire relative to **multi-pose** databases containing just a few individuals performing many static or dynamic pose variations. In this work, we define a multi-identity dataset as one whose number of unique individuals is the same order of magnitude as the total scan count. A multi-pose dataset may contain more than one individual captured in up to thousands of poses, but relative to its total data volume the variation in identity-driven shape is

small and likely introduces overfitting to the particular individuals represented without additional data collection. As such, we will include studies trained on both kinds of data in this review but make a distinction when only multi-pose, limited identity data was used for model training and validation. Fig. 7.1 depicts examples of multi-identity and multi-pose data.

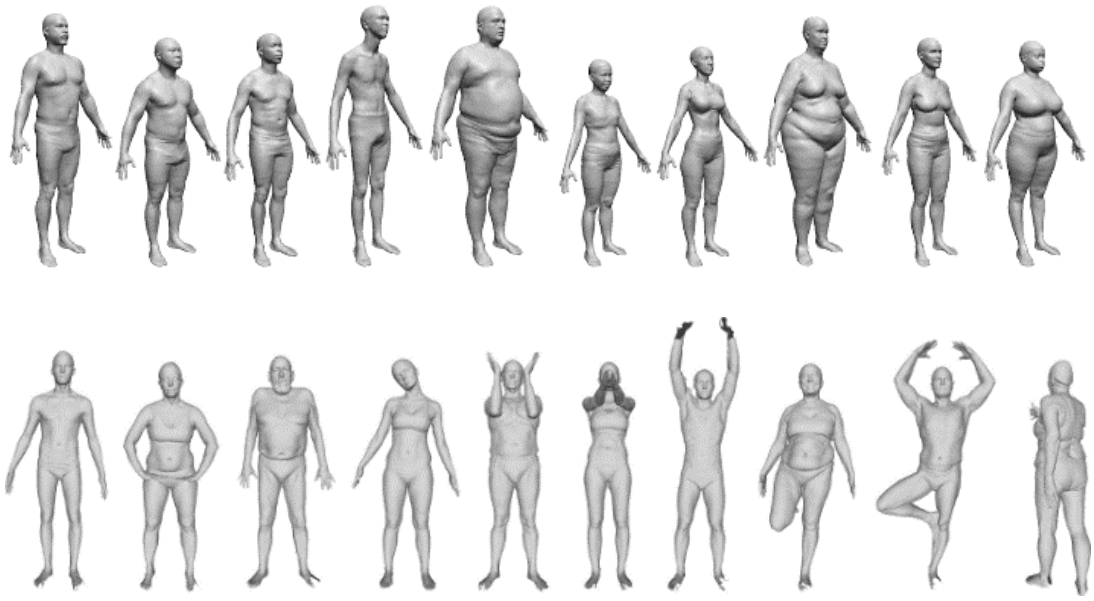


Figure. 7.1. Multi-identity (processed CAESAR [97], top) vs multi-pose (FAUST [12], bottom) data commonly used in 3D parametric shape model training. Note that the multi-pose dataset still contains more than one unique individual.

Body datasets are distinct from shape models as they represent the original real-world captures of human subjects irrespective of downstream parameterization and processing. All 3D shape models can be traced back to a source dataset unless they introduce a novel data capture as part of their study. Table 7.1 lists the datasets used by the works analyzed in this review.

Table 7.1. 3D human body datasets used to train models surveyed in this review.

Name	# of scans	Demographics	# of individuals	Multi-pose?	Multi-identity?
DFAUST [13]	40,000	5 men, 5 women	10	Yes	No
FAUST [12]	300	5 men, 5 women	10	Yes	No
Dyna [91]	70,000	3 men, 3 women ranging from low to high BMI	6	Yes	No
MANO [99]	27,156	5 men, 5 women	10	Yes	No
AMASS [73]	No new scans. Reparameterization of 15 existing datasets	N/A	344 unique subjects	Yes	No
SCAPE [6]	70	1 man	1	Yes	No
Hasler et al. [47]	554	59 men, 55 women	114	Yes	Yes
CAESAR [97]	4,309	Stratified sampling of North American population (Age, Weight, Height, Sex)	4,309	No	Yes
Shape Up! Adults [83]	1,278	292 men, 356 women stratified by age, race, BMI captured on 3 different scanners	648	No	Yes
Shape Up! Kids [135]	483	113 boys, 157 girls stratified by age, race, BMI captured on 3 different scanners	270	No	Yes
GHS3D [141]	60,000+	Unknown	Unknown	Yes	Unknown
SURREAL [122]	N/A, synthetic dataset generated from SMPL parameters	Same as SMPL	Synthesized	Synthesized	Synthesized

3DAEs fall under different shape model categories depending on their implementation architecture, including fully connected, spatially convolutional, flat convolutional, spectral, and implicit surface functions. Some methods use a variational autoencoder (VAE) encoding and decoding structure. The impact of these design differences on 3DAE accuracy for human body shape is not well understood and will be a variable of interest in this review A summary of 3DAE categories is shown in Table 7.2.

Table 7.2. Summary of shape model classifications used to designate autoencoder architecture type in this review.

Shape Model	Summary
Fully Connected	3D mesh is flattened into a 1D vector. Every node in the next layer is a linear combination of every single node (vertex component) in the original layer. No spatial locality is preserved.
Irregular Convolution	Convolutional operation is applied to mesh data without regard to spatial relationships, i.e. 1D window on a flattened vertex vector. Unlike image convolution, the window neighborhood encodes no locality as the flattened vector is ordered arbitrarily.
Graph Convolution (Spectral)	A nonlinear analog to PCA, these methods decompose a mesh into representative components via eigendecomposition of the graph Laplacian and perform convolutional operations on the dimension-reduced components.
Implicit Surface	Shape is represented by a surface boundary rather than a collection of vertices and edges. Boundary is demarcated by a union of mathematical functions such as occupancy or signed distance function (SDF) plus a high frequency residual modeled by a neural network and a skinning function.
Graph Convolution (Spatial)	A convolution kernel is defined and applied to a local neighborhood based on some heuristic, such as spiral convolution or precomputed pooling neighborhoods. Usually, topological consistency is required to ensure the heuristic holds for all meshes in a dataset. This method is most analogous to 2D image convolution.

The generalized methods and algorithms for the 3DAE implementations listed in Table 7.2 are extended from analogous methods in 2D learning and were surveyed by Wu et al. [139] and Li et al. [64]. We classified the surveyed works in Table 7.3 by their shape modeling method.

7.3.1 Standardized Reconstruction Error Evaluation Across Different Works

3DAEs perform dimensionality reduction when encoding shape variation into a reduced

latent space. Methods need to optimize the tradeoff between information loss and encoding size. A transformation that is simply identity has no information loss but performs no data compression. Similarly, a transformation that reduces the dimensionality to 1 (i.e. taking an average of all inputs) may lose too much shape information and produce a parameter space that is not a good representation of the original 3D data. Thus, the primary query this review will be resolving is the normalized comparative reconstruction accuracies of different parametric modeling methods using 3D human body shapes as input. To standardize the performances of different methods, we define a new normalized error metric, the reconstruction-compression quotient (RCQ), which we define here as:

$$\text{RCQ} = \text{Error} / (1 - \text{Compression Ratio}) = \frac{MAE}{1-(\frac{d_c}{d_i})} \quad (29)$$

where error is the per-vertex reconstruction error and ratio is the *compression ratio* defined as the reduced dimensionality size divided by the original parameter size. d_c and d_i are the dimensionalities of the compressed and initial parameter spaces, respectively. The denominator is a normalization factor that asymptotically approaches 1 as the compression ratio gets close to 0 and goes to infinity as the compression ratio approaches 1. Intuitively, a 3DAE with a compression ratio of 1 is an identity function that is lossless but also useless at encoding information. This normalization factor adjusts the reconstruction accuracies across different works so that networks with less compression in their latent embeddings do not artificially appear more accurate. Fig 7.2. illustrates the normalization factor between the domain of (0, 1).

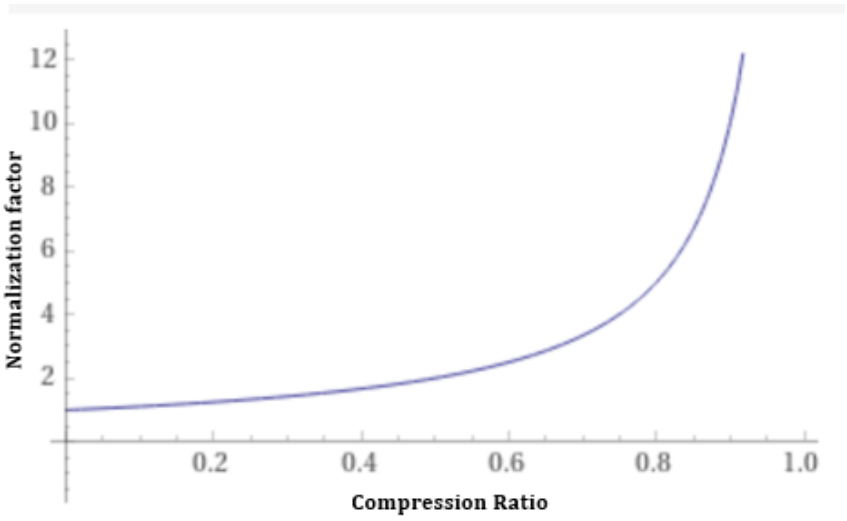


Figure 7.2. Normalization factor vs compression ratio for the RCQ.

Given the structures and definitions in the above section, our motivating research question for this review was the following:

What is the reconstruction accuracy of deep 3D auto encoders on human total body shape?

We answer this question quantitatively with a compilation of comparative reconstruction errors across multiple works normalized by the RCQ. However, the implications for applications of 3DAEs to human shape learning are not fully resolved by a ranking of recent methods. Many inconsistencies with the dataset selection and network architecture cloud the determination of the best method for 3D human shape representation. We comment on these confounding factors in our discussion based on our experience in applying 3D human shape learning to clinical variable estimations and present recommendations for future work in creating 3DAE human shape models that maximize impact and usability for their hypothetical end-use scenarios.

7.4 Systematic Review

We searched the comprehensive computer science literature from the 20-year period spanning 2003 - 2022, inclusive, to summarize, document, and compare results from the various works discussing autoencoders for 3D human bodies. We referenced three databases to perform this query: IEEE Xplore, ACM Digital Library, and dblp.org. Together these cover the vast majority of reputable and highly cited journals and conferences in the computer science literature likely to publish on 3D autoencoders. Detailed search parameters are listed in the Appendices.

7.4.1 Inclusion Criteria

A study is included in our review if it presents a template-based parametric model for representing 3D surface meshes of unclothed total human bodies. The work needs to be published in a peer-reviewed journal or conference between the years of 2003 and 2022. The work needs to train a new 3D parametric model with a deep autoencoder and apply it to a 3D human full body shape dataset.

Unpublished preprints are not considered in this review. 2D image autoencoders operating on pictures of people are not included, nor are works targeted for a specific body part such as the face or the hands. A priori parameterizations such as Lu et al. [70], where the feature vector was explicitly defined as axial circumferences at fixed intervals along the longitudinal axis of the body, are also not included. Works that apply an existing parametric model, such as SMPL, to fit new data such as image, video, or depth map inputs without retraining the parametric model itself are not included. Nonparametric 3D shape reconstruction, such as Furukawa and Ponce [34], is not considered in this review. Autoencoders that are only tested on rigid bodies and inanimate objects

such as vehicles [84] are theoretically applicable to human bodies but are not included in this review as they do not have comparable human body reconstruction accuracies to report.

7.4.2 Survey Results

We found and screened 1,023 titles from IEEE Xplore, 1,045 titles from ACM Digital Library, and 166 titles from dblp.org, of which 18, 2, and 4 works were selected for primary review respectively. Works were screened by title, abstract, and methods in order and rejected at the earliest stage if the content clearly classified the material as off-topic. Forward reference tracking was performed where a cited work's quantitative accuracy was directly compared against the included work. Only Deng et al. [24] was added during this process.

Table 7.3 lists the identified papers meeting the selection criteria along with a summary of important properties.

Table 7.3. List of works included in the systematic review. Shape model classification, training dataset classification, and a summary of the method are provided.

Title	Date	Shape Model	Multi-Identity?	Method Description
Variational Autoencoders for Deforming 3D Mesh Models [116]	Tan et al. 2018	Fully Connected	No	VAE using RIMD features as input
Neural 3D Morphable Models: Spiral Convolutional Networks for 3D Shape Representation	Bouritsas et al. 2019	Graph Convolution (Spatial)	No	Autoencoder uses anisotropic spiral convolution operator

Learning and Generation [15]				
PointAE: Point Auto-Encoder for 3D Statistical Shape and Texture Modelling [23]	Dai & Shao 2019	Irregular Convolution (1D)	**Yes**	1D convolutional, non-variational AE for flattened n x 3 point clouds
DSPP: Deep Shape and Pose Priors of Humans [49]	Hu, Shum, & Mucherino 2019	Fully Connected	**Yes**	GAN trained on latent space learns generative application of encoder output.
Local Deep Implicit Functions for 3D Shape [38]	Genova et al. 2020	Implicit Surface (Occupancy)	No	Implicit surface representation as union of 3D Gaussians + shape residual learned by deep network
Disentangled Human Body Embedding Based on Deep Hierarchical Neural Network [54]	Jiang, Zhang, & Cai 2020	Fully connected	**Yes**	VAE using ACAP [] feature input that separates shape and pose by learning jointly from paired posed and neutral stance data
GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models [141]	Xu et al. 2020	Fully Connected	**Yes**	VAE initialized to PCA coefficients that models shape, pose, hand, and facial expression simultaneously
Mesh Variational Autoencoders with Edge Contraction Pooling [143]	Yuan et al. 2020	Graph Convolution (Spectral)	No	Convolution on ACAP feature with single novel pooling layer
DEMEA: Deep Mesh Autoencoders for Non-rigidly Deforming Objects [121]	Tretschk et al. 2020	Graph Convolution (Spatial)	No	Spiral convolution AE that learns part deformations instead of vertex positions
Fully Convolutional Mesh Autoencoder using Efficient	Zhou et al. 2020	Graph Convolution (Spatial)	No	Recursive pooling layers are precomputed and fixed for a given topology

Spatially Varying Kernels [146]				
NASA: Neural Articulated Shape Approximation [24]	Deng et al. 2020	Implicit Surface (Occupancy)	No	Piecewise occupancy function representation of shape as a function of pose. Not designed for multi-shape usage.
LatentHuman: Shape-and-Pose Disentangled Latent Representation for Human Bodies [66]	Lombardi et al. 2021	Implicit Surface (SDF)	No	SDF per body part coupled with kinematic tree with SMPL parameters and fine pose dependent deformations
imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose [4]	Alldieck, Xu, & Sminchisescu 2021	Implicit Surface (SDF)	**Yes**	Piecewise implicit surface separately representing highly detailed face, hands, and bodies
LEAP: Learning Articulated Occupancy of People [76]	Mihajlovic et al. 2021	Implicit Surface (Occupancy)	No	Occupancy functions for posed, multi-shape data. Shape encoder is a PointNet [Qi] encoding of linear SMPL shapes
Learning Feature Aggregation for Deep 3D Morphable Models [19]	Chen & Kim 2021	Graph Convolution (Spatial)	No	Attention based pooling operator for existing spatial convolution operators that are learned from data rather than fixed in preprocessing
Learning Interpretable Representation for 3D Point Clouds [114]	Su, Lin, & Wang 2021	Unspecified	No	Shape and pose disentanglement from point cloud inputs. Does not address high resolution meshes.
Deep Learning-Based Automated Extraction of Anthropometric Measurements from a Single 3-D Scan [55]	Kaashki, Hu, & Munteanu 2021	Irregular Convolution	No	Convolutional AE for point clouds aggregating features in nearest neighbor windows at different scales. Does not compress data.

Multiscale Mesh Deformation Component Analysis with Attention-Based Autoencoders [142]	Yang et al. 2021	Graph convolution (Spatial)	No	1-ring ACAP feature convolution with additional chained attention-based AEs for fine tuning local details in individual body parts
Variational Autoencoders for Localized Mesh Deformation Component Analysis [116]	Tan et al. 2022	Graph Convolution (Spectral)	**Yes**	Spectral VAE on ACAP deformation feature input
Mesh Convolutional Autoencoder for Semi-Regular Meshes of Different Sizes [45]	Hahner & Garcke 2022	Graph Convolution (Spatial)	No	Remeshes inputs to create 6 neighbor connectivity at all vertices and uses hexagonal convolution operators. Remeshing loses considerable high frequency detail.
Deep Polynomial Neural Networks [22]	Chrysos et al. 2022	Nonlinear activation layer	No	Introduces a new neural network layer that expresses outputs as polynomial functions of the input rather than linear operations followed by nonlinear activations. Layers are inserted into existing architectures.
3D Shape Variational Autoencoder Latent Disentanglement via Mini-Batch Feature Swapping for Bodies and Faces [31]	Foti et al. 2022	Graph Convolution (Spatial)	No	Spiral convolution same as Gong et al. Trained on synthetic data generated from linear models (SMPL based) to maximize local shape control in generative applications.
Representation learning of 3D meshes using an Autoencoder in the spectral domain [62]	Lemeunier et al. 2022	Graph Convolution (Spectral)	No	1-D convolutional AE that encodes and decodes in the spectral domain of the connectivity matrix. Spectral coefficients are defined as the eigenvalues of the graph Laplacian multiplied by the spatial coordinates.
Dual octree graph networks for learning adaptive	Wang, Liu, & Tong 2022	Implicit Surface (Occupancy)	No	Point cloud based graph convolution network for learning implicit surface function as a 3D volumetric field

volumetric shape representations [124]				

We identified two categories of works that are adjacent to our criteria but do not warrant systematic review and thus excluded from further consideration. The first category consists of all papers found that are based off of PCA. As PCA is a deterministic operation with a single optimal solution per dataset, the efficiency of PCA can be understood from a single representative work instead of reiterating its performance across multiple trials. We use the reconstruction accuracy of Loper et al. to represent PCA.

The other category consists of papers that present deep learning algorithms on 3D human shapes but are not autoencoders. Instead, these papers report results on pose classification or body part segmentation analogous to many works dealing with convolutional neural nets (CNNs) on 2D images. These are listed in **Supplementary Table 7.1** and are not included in our analysis, but rather are cited to clearly define the boundaries of this systematic review. This table is not comprehensive; rather, it only covers the works that turned up within our search parameters that are closely related to our primary analysis.

7.4.3 Reconstruction accuracy comparison

Where there are multiple latent vector sizes presented, the result with the lowest error (usually with the highest latent dimension size) was chosen for RCQ computation. Reconstruction accuracy results are shown in Table 7.4 in ascending error order if the units of measurement are clearly specified in the original work, allowing for a fair comparison of the RCQ across works. Mean errors in Table 7.4 are confirmed to be either mean absolute errors (MAE) between registered vertex pairs in the input and reconstruction or Chamfer L1 distance between a reconstruction and

a nonregistered point cloud in millimeters (mm). Not every work reported results on a common dataset or with the same metrics. The dataset tested by the authors is listed in the final column. In general, we expect multi-identity data (CAESAR and Shape Up!) to be more challenging than multi-pose, limited identity data such as DFAUST. These cases are identified by boldface.

Cases where authors cited the method of another paper and compare its accuracy against theirs are included in rows indicated by the format *<Method by AUTHOR A>* → *<Cited and retested by AUTHOR B>* in Table 7.4. We only included these examples in cases where a quantitative result was missing in the original work or was reported differently, either on a separate dataset or error metric. Two methods, COMA (Ranjan et al. [96]) and SpiralNet++ (Gong et al. [40]), are listed in this table as the original work only represented face shape instead of total body and thus did not meet inclusion criteria for this review, but are quantitatively evaluated on DFAUST in further studies.

Table 7.4. RCQ accuracy comparisons of surveyed works.

Paper	Mean Error (mm)	Compression Ratio	RCQ	Dataset Tested
Jiang, Zhang, & Cai	2.75	0.13%	2.75	FAUST + Dyna + MANO
Xu et al. GHUM	2.81	0.05%	**2.81**	**CAESAR + GHS3D**
Lombardi et al.	3.04 (Chamfer L1)	0.46%	3.05	DFAUST
Lombardi et al.	3.14 (Chamfer L1)	0.46%	3.16	AMASS MoVi
Loper et al. PCA 300 Parameters	3.2	1.45%	**3.25**	**CAESAR**
Xu et al GHUML	3.27	0.17%	**3.28**	**CAESAR + GHS3D**
Mihajlovic et al. → Lombardi et al.	3.33 (Chamfer L1)	3.02%	3.43	DFAUST
Mihajlovic et al. → Lombardi et al.	3.52 (Chamfer L1)	3.02%	3.63	AMASS MoVi
Jiang, Zhang, & Cai	4.67	0.13%	**4.68**	**CAESAR + SCAPE + Hasler**
Loper et al PCA 50 Parameters	4.8	0.24%	**4.81**	**CAESAR**
Chen & Kim	5	0.31%	5	DFAUST
Bouritsas et al.	5	0.39%	5.02	DFAUST
Zhou et al.	5.01	0.31%	5.03	DFAUST
Deng et al → Lombardi et al.	7.19	3.02%	7.41	DFAUST
Ranjan et al → Chen & Kim	9	0.31%	9.03	DFAUST
Deng et al → Lombardi et al.	8.85	3.02%	9.13	AMASS MoVi
Gong et al → Lemeunier et al.	10	0.31%	10.03	DFAUST
Lemeunier et al.	10.3	0.31%	10.3	DFAUST
Yang et al.	21.7	0.16%	21.7	DFAUST
Tretschk et al.	22.3	0.16%	22.3	DFAUST
Kaashki, Hu, & Munteanu	0.027	>100%	< 0	SURREAL

In many cases, quantitative reconstruction accuracies in reviewed works are either missing or reported with unspecified units of measurement. We noted these works in Supplementary Table 7.2 for completeness.

7.5 Discussion

In this review, we compared the reconstruction accuracies of 3D human body shape autoencoders from 24 works published in the most widely cited computer science literature searchable through IEEE, ACM, and dblp. From these 24 works, we identified reconstruction error results that were comparable across publications and listed them in Table 7.4. Errors were comparable if they were reported as L1 mean absolute errors (MAE) with clearly defined metric units. Works that did not report quantitative errors or reported incompatible metrics such as L2 errors or indeterminate units were summarized in Supplementary Table 7.2.

Reconstruction errors were normalized according to a novel standardized error definition, the RCQ. This factor is close to 1 for small compression ratios but penalizes the error sharply when the encoder achieves smaller reconstruction error by performing less dimensionality reduction (1.25 at 20% compression)

Jiang et al. achieved the lowest RCQ in this survey at 2.75 mm. The authors trained on an ensemble dataset of both multi-pose and multi-identity data with a fully connected variational autoencoder that separated pose parameters from body shape parameters.

Future works in the space of 3D human body shape encoding should always include quantitative test set reconstruction results with defined units and an L1 error normalized by the RCQ. Agreeing on a standardized quantitative error measurement allows all human shape autoencoders to be compared against prior work without ambiguity in unit of measurement or error definition. Furthermore, to preserve consistency and comparability of results, future work should agree on a common test dataset representing both multi-pose and multi-identity data, such as a subset of DFAUST combined with a subset of CAESAR. An L2 error such as those reported

in Supplementary Table 7.2 could be reported in lieu of or in conjunction with the L1 MAE. The L1 error was preferred in this review due to the lack of units of measurement associated with L2 errors in the surveyed work and the observation that many works optimized the L1 MAE loss function during training, possibly due to the sensitivity and potential instability of L2 mean-squared-error when there are outliers in a mesh with thousands of vertices.

The RCQ adjustment is designed to penalize AE architectures that report high reconstruction accuracy without performing dimensionality reduction. In practice, this had negligible impact on the raw accuracies as almost all included works reported latent spaces with compression ratios <1%. The exception was Kaashki et al., who only increased the number of variables from the initial input and had a compression ratio greater than 1, resulting in a negative RCQ.

Fig. 7.3 sorts entries of Table 7.4 by ascending order color coded by network architecture class. The two fully connected networks were among the best performing out of all surveyed works. This contrasts with the two spectral convolution networks, which did not achieve below 5mm of RCQ. This difference is interesting as both methods lack local feature sensitivity much like the linear, global variance decomposition of PCA yet sit on opposite ends of the reconstruction performance spectrum. Spectral convolution learns features from decompositions of the connectivity matrix of the whole mesh determined by its neighbor structure rather than on the world vertex positions such as in PCA. Like PCA, the features learned by this class of method tend to be biased towards global rather than local support. Fully connected networks do not explicitly model the spatial relationship of neighboring vertices with a local kernel and assume all nodes of a network layer can be affected by all the nodes of the previous layer. Analogous to 2D image networks, this class of method should suffer inferior performance on reconstructing local details

due to the increased parameter count of the network and the implication of global features with no restricted local support [58].

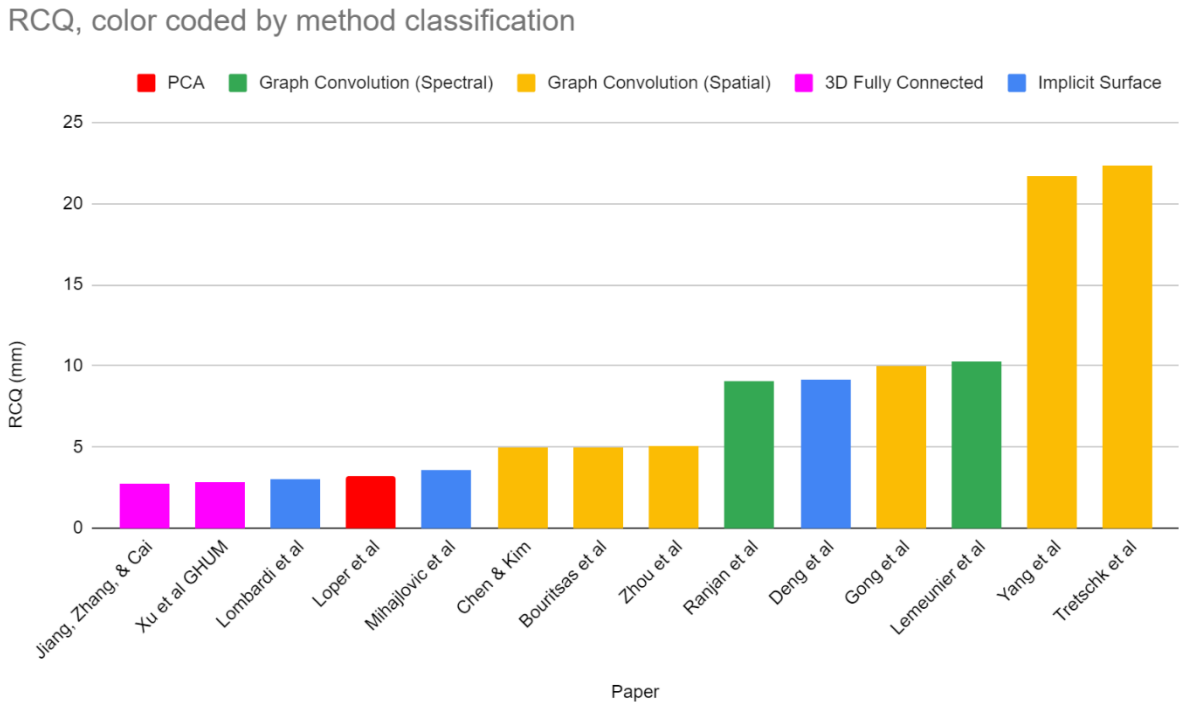


Figure 7.3. RCQ values for selected works in Table 7.4. The lowest error reported from each work was used in cases where there were multiple datasets trialed.

By contrast, spatial graph convolutional networks define a kernel operator on a local neighborhood akin to 2D image filters. Implicit surface methods as applied to human body parameterization are typically modeled per body part as in Lombardi et al. Thus, they have defined local support due to their piecewise implementation.

The graph convolutional methods surveyed in this review do not disentangle pose from body shape. This could result in decreased reconstruction performance in multi-pose dataset benchmarks as most of the shape variance between meshes is caused solely by part rotations explainable by a kinematic tree. All graph convolutional methods in this survey except Tian et al.

performed worse than Jiang et al., Xu et al., and the two implicit surface models trained on more than one individual, all of which modeled pose separately from the shape encoding space. Explicitly modeling the pose with a skinned and rigged skeleton regularizes the network to a defined animation parameterization and reduces the amount of remaining unexplained variance the autoencoder needs to represent [136]. Tian et al. achieved the lowest reconstruction error across all surveyed works by finetuning and testing a spatial graph convolutional model on neutral posed scans only. Future work should standardize pose by either removing pose variation from the input data or by modeling the skeleton parameters explicitly to achieve the lowest reconstruction errors.

The impact of a variational autoencoder (VAE) architecture on reconstruction accuracy is also difficult to isolate. Out of the surveyed works in Table 7.3, only five (Tan et al., Tan et al., Jiang et al., Foti et al., and Xu et al.) implemented a VAE, and two of these works achieved the lowest RCQ scores in the survey. A VAE architecture regularizes the latent encoding space and could prevent overfitting to training data, resulting in better generalization error on held out scans. Future work implementing VAE embeddings with graph convolutional networks may better resolve the relative contribution of the architectures. A more complete comparison of quantitative reconstruction accuracies including all the works in Supplementary Table 7.2 could help clarify the performance of VAE architectures relative to competing works. We leave it for future work to investigate the performance of these architectures using a common error metric and dataset.

Jiang et al. and Xu et al. differed from the rest of the surveyed work in that they were able to train and test on a more diverse dataset than competing works. These three networks were the only surveyed works in Table 7.4 that trained on a multi-identity dataset in addition to a multi-pose dataset. DFAUST, a multi-pose dataset consisting of 40,000 4D scans of 10 individuals, was the most represented dataset in this survey and was often the only dataset surveyed works used in

training and testing. The lack of a standardized dataset in 3D human body learning analogous to ImageNet or ShapeNet creates additional difficulty in comparing model quality across different works. DFAUST, by virtue of its high frequency of use, may be the de facto benchmark in current literature. However, DFAUST suffers from a lack of representation of diverse body shapes due to it only having 10 unique individuals in its capture. All works in this review that trained on a multi-identity dataset did so on a processed version of CAESAR published by Pishchulin et al. [90] This dataset lost much of the resolution of the original scan as they were bottlenecked through a linear PCA projection with a small number of parameters, as illustrated in Fig. 7.4. The network does not learn the original shape variation but rather a projection of that variation onto a linear subspace, which decreases some of the potential benefits of using a nonlinear encoding method such as a 3DAE. Future work on deep 3D shape encoding of human bodies should train and test on a standardized high-resolution dataset representing both identity dependent and pose-dependent deformations with no preprocessing that degrades the surface alignments of the data to the original scans.

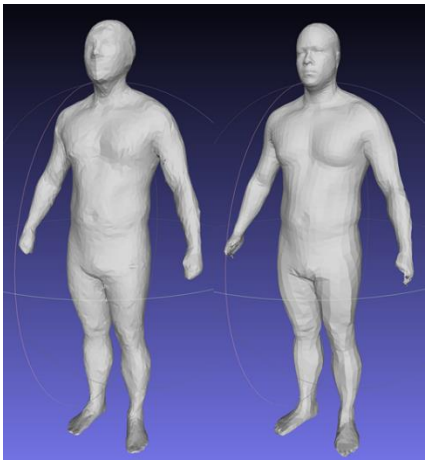


Figure 7.4. Mesh from Pishchulin et al. dataset (left) and a remesh of the same CAESAR scan subject used in Tian et al. Both meshes were standardized to a common topology containing 6890 vertices. All 3DAE works training on multi-identity data depend on meshes that have lost considerable detail from the ground truth.

Apart from Jiang et al., Xu et al., and Lombardi et al., the majority of 3DAE methods do not outperform the highest reported reconstruction accuracy of SMPL, a linear PCA method. When the compression ratio is reduced to 50 allowed parameters, SMPL is worse than all of the other multi-identity trained methods and is more comparable to but still slightly better than the spatial graph convolution methods such as Zhou et al. SMPL is most directly comparable to Jiang et al. and Xu et al. as it is a pose disentangled model trained jointly on multi-pose and multi-identity data (CAESAR), methods that it is clearly doing worse than at a comparable compression ratio. Tian et al. achieved a lower reconstruction accuracy than SMPL at an equivalent compression ratio and parameter count but reported worse accuracy at 49 latent variables than the comparable 50 parameter SMPL model. Linear PCA appears to be preferable at low parameter counts as it can guarantee an orthogonal division of its latent encoding space that maximally explains the observed data variance. However, this survey supports the assertion that deep, nonlinear 3DAE models can better model higher levels of detail at higher parameter counts than an equivalently sized PCA model.

7.6 Conclusion

Deep 3D autoencoders with local feature support trained on pose normalized or pose disentangled data representing a high-resolution sampling of a diverse cross-section of unique individuals outperform linear autoencoder methods such as PCA and all existing 3DAE literature as of the writing of this survey. Pose disentanglement is desirable for methods that target inputs with variable posing but may be extraneous when pose is normalized as part of data capture. A VAE architecture may further improve the performance of the 3DAE model by regularizing the

latent space so that it generalizes better to unseen inputs: however, the current implementation without the variational component outperforms the surveyed VAE methods. Future work in 3DAEs on total human body representation should leverage a common benchmark dataset with high resolution, multi-identity scans such as the one presented in this work in order to standardize comparisons between methods.

7.7 Appendix

IEEE Xplore query:

("Full Text & Metadata":autoencoder? OR "Full Text & Metadata":unsupervised OR "Full Text & Metadata":parametric OR "Full Text & Metadata":parameterization OR "Full Text & Metadata":"statistical shape model") AND (("Full Text & Metadata":"human bod" OR "Full Text & Metadata":"full bod*" OR "Full Text & Metadata":"total bod*" OR "Full Text & Metadata":"whole bod*") AND ("Full Text & Metadata":"3d mesh*" OR "Full Text & Metadata":"3-d mesh*" OR "Full Text & Metadata":"3-d human*" OR "Full Text & Metadata":"3d human*") AND NOT ("Document Title": photo OR "Document Title": monocular OR "Document Title": video OR "Document Title": image OR "Document Title": hand))*

ACM Digital Library query:

[[All: autoencoder?] OR [All: unsupervised] OR [All: parametric] OR [All: parameterization] OR [All: "statistical shape model"]] AND [[All: "human body"] OR [All: "human bodies"]] AND [[All: 3d] OR [All: 3-d]] AND [All: mesh] AND [E-Publication Date: (01/01/2003 TO 12/31/2022)]*

Dblp.org query:

“Mesh autoencoder”

Supplementary Table 7.1 Reviewed works with incompatible accuracy metrics.

Title	Authors
Subdivision-based Mesh Convolution Networks	Hu et al. 2022 [50]
Multi-chart generative surface modeling	Ben-Hamu et al. 2019 [9]
GRASS: generative recursive autoencoders for shape structures	Li et al. 2017 [63]
CurvaNet: Geometric Deep Learning based on Directional Curvature for 3D Shape Analysis	He et al. 2020 [48]
Convolutional neural networks on surfaces via seamless toric covers	Maron et al. 2017 [75]
MeshCNN: a network with an edge	Hanocka et al. 2019 [46]
HodgeNet: learning spectral geometry on triangle meshes	Smirnov & Solomon 2021 [113]
MeshWalker: deep mesh understanding by random walks	Lahav & Tal 2020 [60]
DiffusionNet: Discretization Agnostic Learning on Surfaces	Sharp, Crane, & Ovsjanikov 2022 [107]
Multi-directional geodesic neural networks via equivariant convolution	Poulenard & Ovsjankiov 2018 [92]
CNNs on surfaces using rotation-equivariant features	Wiersma, Eisemann, & Hildebrandt 2020 [132]
MeshMAE: Masked Autoencoders for 3D Mesh Data Analysis	Liang et al. 2022 [65]

Supplementary Table 7.2 Reviewed works with incompatible accuracy metrics.

Paper	Error Value	Measurement Error	Dataset
Hu, Shum, & Mucherino	N/A		
Genova et al	N/A		
Foti et al	N/A		
Wang, Liu, & Tong	N/A		
Alldieck, Xu, & Sminchisescu	0.36 x 10-4	Chamfer L2	**GHS3D + CAESAR**
Su, Lin, & Wang	0.79 x 10-4	Chamfer L2	DFAUST
Mihajlovic et al	2.27 x 10-4	Chamfer L2	DFAUST
Mihajlovic et al	2.80 x 10-4	Chamfer L2	MoVi
Tan et al	5.25	RMSE, unknown units	**CAESAR (Females)**
Tan et al	1.10 x 10-3	RMSE	Dyna
Hahner & Garcke	1.79 x 10-2	RMSE	FAUST
Yuan et al	2.49 x 10-2	RMSE	Dyna
Chrysos et al	11	Generalization, Unknown units	DFAUST
Deng et al	4 x 10-5	Chamfer L1, Unknown units	DFAUST
Dai & Shao	<79.7	Generalization, Unknown units	**CAESAR**

8. Using Deep Learning for Nonlinear Estimation of Body Composition from 3D Optical Scans

From the surveyed works in Chapter 7, we selected the 3D graph convolutional network of Zhou et al. [146] to retrain as a deep feature extractor for body composition regression. This was a novel first application of 3D deep networks to the task of body composition prediction from human shape. We investigated the effects of replacing linear regression used in prior chapters and works with a nonlinear regression model, specifically Gaussian process regression (GPR) as first seen in Wang et al. [125]

Zhou et al. was selected over the works with lower reconstruction error due to its reduced model complexity because of omitting pose parameterization from its architecture. The Shape Up! datasets are small and were pose controlled during capture. Our retrained model may be detrimentally affected by a larger architecture designed to estimate pose parameters simultaneously with identity determined body shape. The implementation of Zhou et al. was presented as a generalization of 2D convolutional neural networks to an irregular 3D graph. The authors claimed this design gave deep features local control over a spatial neighborhood akin to the saliency response of 2D convolutional filters. This was a desirable alternative to test in contrast to PCA, which by its mathematical definition lacks local control in its individual features.

We assembled a novel single pose, multi-identity ensemble dataset comprised of all the Shape Up! data from Chapters 5 and 6 as well as the North American CAESAR scans using the automatic mesh templating method presented in Chapter 5 and retrained the network of Zhou et al. starting from an initialization state pretrained on DFAUST. This two-step training procedure

allowed us to start from where Zhou et al. ended rather than initializing the network with random weights, thus allowing our limited dataset to achieve greater training accuracy. We systematically compared the performance of the novel nonlinear methods (deep 3D autoencoder feature extraction and GPR) relative to the baselines established by PCA and linear regression in prior work to establish evidence supporting a definitive conclusion on the best state-of-the-art pipeline for body composition prediction from 3DO image input of human body shape.

This chapter describes work originally submitted to Medical Physics.

8.1 Abstract

Background: Total and regional body composition are strongly correlated with metabolic syndrome, an array of biological and physical measures that contribute to many leading causes of death in the US and around the world. Quantitative body morphology measured using 3D optical scans is a direct result of the local distribution of fat and muscle. Prior work has relied on linear algorithms for dimensionality reduction of body morphology and for regression to criterion body composition measures when changes in body shape may be highly nonlinear.

Purpose: In this study, we asked if a deep 3D autoencoder model for body morphology parameterization and a nonlinear Gaussian process regression for body composition would provide more accurate predictions of body composition than previously studied linear methods.

Methods: We built a novel ensemble training dataset consisting of 4286 3D optical scans from four different collection sources: DFAUST, CAESAR, Shape Up! Adults, and Shape Up! Kids. The Shape Up! data contained paired reference DXA measurements for body composition. We standardized all scans to a topologically consistent 6890-vertex mesh template format. We trained a deep 3D graph convolutional autoencoder on the ensemble dataset using an additional

separate 462 CAESAR scans for evaluation and predicted body composition metrics using extracted deep features at multiple levels with nonlinear Gaussian process regression (GPR). We built a linear baseline model using principal component analysis (PCA) on the same training dataset and predicted 12 measures of body composition from its shape features with both linear regression and GPR. We measured the accuracy and precision of all our models on a common test set with 424 held-out meshes to evaluate the marginal contribution of nonlinear shape feature extraction and nonlinear regression relative to predictions derived from linear PCA shape features and linear regression.

Results: Nonlinear GPR produced up to 20% reduction in prediction error and up to 30% increase in precision over linear regression for both sexes in 12 tested body composition variables. Deep shape features produced 6-8% reduction in prediction error over linear PCA features for males. Our best performing nonlinear model predicting body composition from deep features outperformed prior work using linear methods on all tested body composition prediction metrics in both precision and accuracy. All coefficients of determination (R^2) for all predicted variables were above 0.86.

Conclusions: GPR is a more precise and accurate method for modeling body composition mappings from body shape features than linear regression. Deep 3D features learned by a graph convolutional autoencoder only improved male body composition accuracy. We summarized our findings by presenting a best performing GPR model trained from deep features that achieved lower RMSEs than any work published to date on 12 metrics of body composition with emphasis on the least accurate predictions reported in linear methods, percent fat and visceral fat.

8.2 Introduction

Total and regional body composition are correlated with many of the leading causes of death in the US and around the world. [133, 57, 7, 41] High visceral fat deposits doubled the risk for metabolic syndrome in males with otherwise normal BMIs [41] and increased mortality rates from cancer by up to 63% [144]. Metabolic syndrome, a condition indicated by an array of biological and physical vital measurements such as blood glucose and waist circumference, is strongly associated with many chronic conditions such as cancer, heart failure, and diabetes [51, 25, 133]. Management of these conditions is also heavily impacted by body composition, specifically low lean mass, which predicts poor treatment outcomes [35, 102] and up to 17-fold increased mortality [74, 100]. However, body composition assessment historically required more advanced instruments such as dual X-ray absorptiometry (DXA) or air displacement plethysmography (ADP) [30]. Unlike ADP, DXA can measure regional compartmental fat and lean deposits but exposes participants to potentially harmful ionizing radiation and is not recommended for frequently repeated measurements, particularly in at-risk groups such as young children and pregnant women. Radiation exposure must be limited to exigent circumstances even in healthy adults. An ideal alternative assessment system should achieve accuracy and precision on total and regional body composition measurement comparable to DXA without utilizing ionizing radiation. The system should be relatively inexpensive and also fast and simple to use, requiring no special training or certifications to operate and returning results in a minute or less.

Recent work showed that 3D optical (3DO) imaging can serve as an accurate and precise, low-cost, noninvasive surrogate to DXA imaging [83, 118, 120]. 3DO measures the surface geometry of the human body using light in the optical spectrum as opposed to the penetrating radiation of DXA. It does not require the injection of an isotope contrast like MRI and can scan

an entire adult body in under one minute. Scanning systems cost on the order of $10,000 and are programmed by the manufacturer to operate automatically without the need of certified technicians. The external 3D shape of a human body contains strong signals about its internal structure and composition that can be learned by machine learning algorithms. Recent advances in 3D scanning technology have made 3DO scanning of human bodies more accessible and widely distributed than ever [120, 131]. However, current work on learning body composition from 3DO scans relies on linear mathematical models, such as principal component analysis (PCA) and linear regression. These simplified linear assumptions impose potentially erroneous prior assumptions on the parameterization of both the 3D shape model and the mapping between shape parameters and body composition. Nonlinear methods that relax restrictive assumptions on the estimated functions may align better with the true relationship between shape and body composition and provide better prediction accuracy of target variables. An accurate model for estimating total and regional body composition metrics from 3DO scans could standardize body composition assessment by removing the costs and risks associated with clinical evaluation and irradiation. We propose to use deep, nonlinear methods to better estimate shape parameterization and body composition relative to prior work with linear baselines.

In this work, we trained a deep 3D graph convolutional autoencoder on a diverse sample of full-body 3DO scans. A subset of these scans was captured with paired DXA scans for ground truth body composition training targets. We trained a regression model from extracted 3D graph convolutional features in the deep autoencoder network to DXA body composition variables using a nonlinear Gaussian process regression (GPR). We thoroughly investigated the effect of differing depths of convolutional features and compared novel applications of nonlinear methods to linear methods of prior work through ablation studies measuring 3D reconstruction error and

body composition prediction accuracy. To our knowledge, this is the first application of a 3D autoencoder to body composition prediction specifically and to medical statistics learning in general. Our results indicate that nonlinear regression using a GPR with a squared dot product kernel provides greater accuracy and precision in body composition estimation on test data than previous linear methods. This was the first application of a 3D graph autoencoder to the task of body composition estimation for clinical evaluation.

8.3 Methods

This work performs a comprehensive comparative analysis between linear and nonlinear methods for shape modeling and body composition estimation of 3D human body shape. We assembled a composite dataset of 3DO human body shape from four independent sources and standardized them to a common topology. We trained a deep 3D autoencoder (3DAE) based on the work of Zhou et al. [146] on the composite dataset at multiple parameter sizes and evaluated its 3D reconstruction error on test data. We trained nonlinear GPR models from 3DAE deep features at multiple scales to predict DXA body composition measurements. We then trained a PCA model as a linear baseline using the same training data and parameter counts. We trained linear least squares regression and nonlinear GPR models from PCA features to body composition to create fully linear and linear-nonlinear hybrid models respectively.

8.3.1 Experimental Cohort

Four data sets were used for this study: CAESAR [97], Shape Up! Adults (SUA) (NIH R01 DK109008) [83], Shape Up! Kids (SUK) (NIH R01 DK111698) [135], and DFAUST [13]. Table 8.1 shows how this data was subdivided to train and test our models.

The CAESAR 3DO scan dataset represented 2400 American and Canadian adults aged 18-65 3D scanned in a neutral A-pose, with legs and arms held fully extended and abducted ~30 degrees from the midline of the body. Participants were mostly unclothed except for form-fitting gray underwear and a swim cap to standardize hair appearance. Roughly half of the recruited cohort were female. Sex, height, and weight were the only demographic variables used to construct the shape model; no other demographic collected was used in this study. Subjects were scanned on a custom-built Cyberwar WB4 3D scanner.

The SUA and SUK datasets (ClinicalTrials.gov ID NCT03706612 (SUK) and ID NCT03637855 (SUA)) were cross-sectional and stratified by age (SUK: 5-8, 9-12, 13-17 yr., SUA: 18–39, 40–59, ≥60 yr.), ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, Asian, and Native Hawaiian or other Pacific Islander (NHOPI)), sex, and BMI Z-score. Along with extensive demographics, quantitative measures included whole body DXA scans and 3DO scans. We acquired duplicate whole-body DXA scans of each participant on either a Hologic Horizon/A system (UCSF) or a Discovery/A system (PBRC and UHCC) (Hologic Inc., Marlborough, MA, USA). Participants were positioned and scanned according to guidelines specified by the respective system manufacturers. All scans were analyzed at UHCC by a single certified technologist using Hologic Apex version 5.6 with the National Health and Nutrition Examination Survey (NHANES) Body Composition Analysis calibration option disabled. DXA systems quality control was performed by monitoring the weekly values of the Hologic Whole Body Phantom. Two independent 3DO scans were acquired for each participant in up to three scanning devices: Fit3D Proscanner 4.x, Fit3D Inc, Redwood City, CA, USA, Styku S100 4.1, Styku LLC, Los Angeles, CA, USA, and Size Stream SS20, Size Stream, Cary, NC, USA. All 3DO scans were captured in a neutral A-pose that closely mirrored the pose of the CAESAR dataset.

The DFAUST dataset was a 4D capture of human pose. A templated mesh was registered to over 41,000 snapshots from continuous sequence of 3D point clouds (3dMD LLC, Atlanta, GA) of 10 unique individuals performing dynamic movements captured at 60 frames per second. 4264 meshes were reserved for testing and were not used in this study. No clinical data such as body composition was reported.

The DFAUST dataset was used to pretrain the 3DAE model. The pretraining step initialized the network weights to a state similar to the presentation in the original work of Zhou et al. [146] An ensemble dataset of CAESAR, SUA, and SUK was used as training data to finetune the model. During deep network training, a held-out data split, the evaluation set, was reserved to benchmark the performance of the model at the conclusion of each training epoch as determined by its geometric reconstruction loss. The model state at that epoch was saved as the current best iteration if the evaluation error is lower than the previous best; this guards against overfitting to training data without violating the integrity of the test set. 20% of the CAESAR data was reserved for the finetuning evaluation steps (FT eval). To create a standardized benchmark between the methods, the same training and test split in Tian et al. [118] was preserved to investigate the performance of the new deep autoencoder and nonlinear GPR on the same test set.

Table 8.4. 3D scan count and unique individual representation for the pretrain, finetune, evaluation, and test sets. The pretraining step used only DFAUST data and included its own evaluation split consisting of 20% of the DFAUST scans. The finetune step was an iterative training step performed using only CAESAR, SUA, and SUK data starting with the network weights determined by the pretraining step. The evaluation set was used to determine the best performing model state to save without causing overfitting to the training data. The test set was kept the same as prior work to create controlled comparisons.

Dataset	Total 3D scans / Total unique individuals			
	Pretraining	**Finetuning**	**FT Eval**	**Test**
DFAUST	36,956 / 10	0	0	0
CAESAR	0	2140 / 2014	462 / 462	0
Shape Up! Adults	0	1500 / 542	0	424 / 336
Shape Up! Kids	0	646 / 200	0	0
Total	-	4286 / 2882	-	-

To enforce topological consistency between all 3DO meshes, all scans in Table 8.1 were standardized to a constant topology containing 6890 vertices (referred to as the SMPL [67] topology) except for the DFAUST data, which is already in this format natively. Raw 3DO scans were first automatically registered to a 60,001-vertex template (60k) using the automatic 3D template fitting method of Tian et al. [118]. A 60k template was fit to a single reference A-pose mesh in SMPL topology from the DFAUST dataset. A sparse matrix **C** was solved to determine a linear mapping between the 60k topology and the SMPL topology of the reference DFAUST mesh as a least-squares solution. For each row *I* of **C**, all columns were zero except for the index *j* corresponding to the vertex on the 60k mesh that was the nearest neighbor of the *i*tch vertex in the SMPL mesh. All 60k template fits from CAESAR, SUA, and SUK were multiplied by **C** to

create rough mappings to SMPL topology, then rigidly deformed with the method described in Allen et al. [5] to maximize surface-to-surface alignment with the original 60k mesh. Meshes that contained geometric defects after templating, such as collapsed or missing body parts, were removed from the dataset and not counted in Table 8.1. A visual example of this process is shown in Fig. 8.1.

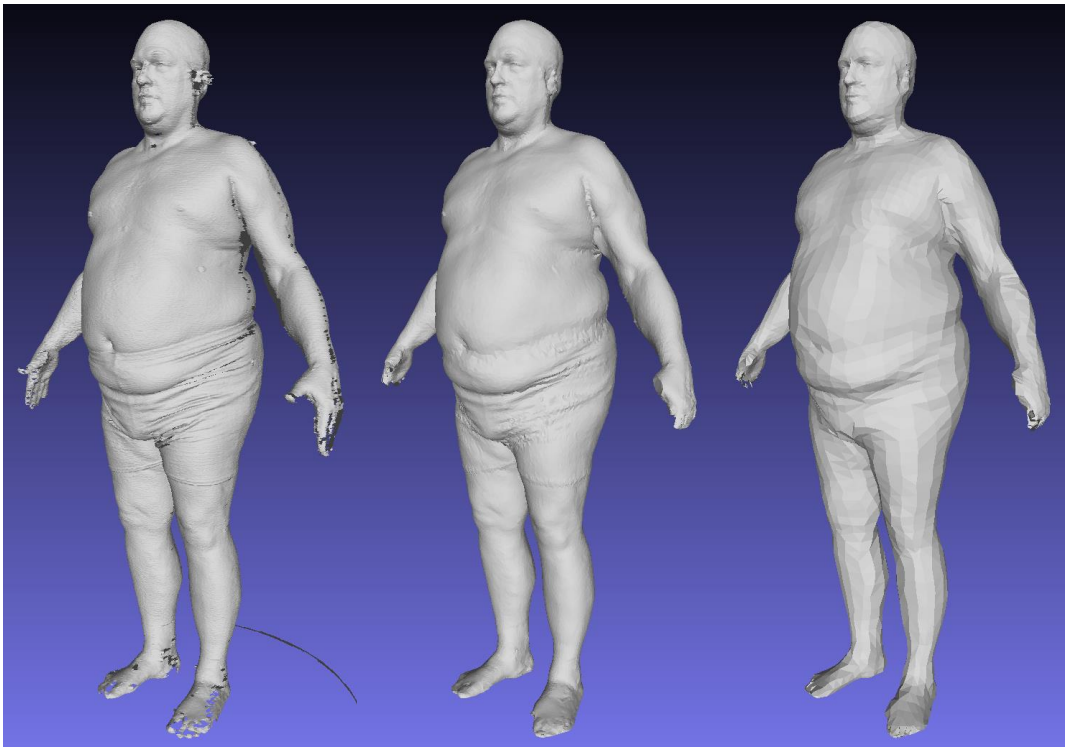


Figure 8.1. Left: a raw CAESAR scan with unordered vertices and incomplete surfaces. Center: automatic 60k template registration using the methods of Tian et al. Right: SMPL (6890) topology transformation of the 60k fit, created by multiplying the 60k mesh with a sparse nearest-neighbor matrix **C** and deforming the result to align with the 60k mesh.

The ensemble 3DO data selected for this study collectively covered many of the limitations inherent to each individual dataset. Clinically sampled, multi-identity datasets like SUA and SUK are custom-made for studies modeling human shape variation and body composition estimation

across a diverse, cross-sectional sample of the population. SUA and SUK contain 3D meshes and body composition reference values for adults and kids respectively but are low in participation due to the additional overhead and difficulty of collecting clinical data. CAESAR augments this dataset by doubling the number of single-pose 3DO scans available while almost quadrupling the number of unique individuals represented. However, since there is no clinical data associated with CAESAR, it can only be used to train the 3D shape model and not the mapping to body composition. Only CAESAR scans were held out for model evaluation during the finetuning step of 3DAE training to preserve as many SUA and SUK scans with paired DXA measurements as possible for body composition regression training and testing. The DFAUST data was different from the other three as it contained very few unique individuals (10, 5 male and 5 female) but the largest number of unique 3DO scans (~37,000). This dataset varied pose instead of individual identity to create shape diversity. Modeling shape variation due to posing is not an objective of this study and DFAUST was not used to jointly train the 3DAE shape model with the other datasets given its outsized data representation. However, DFAUST was very useful in initializing the 3DAE weights to a better-than-random starting state for training on the CAESAR-SUA-SUK ensemble dataset.

8.3.2 3D Deep Autoencoder with Graph Convolutional Network

Raw 3DO scans contain potentially hundreds of thousands of unorganized vertices and are not suitable inputs for regression algorithms without first processing then into standardized feature vectors for all dataset members. In this study, we use a 3D graph convolutional neural network adapted from Zhou et al. [146] to perform nonlinear dimensionality reduction and feature extraction on the 3D body shape space. This deep network is a 3D autoencoder that possesses many attributes inherent to deep convolutional neural networks (CNNs). A local kernel operator

paired with layered feature pooling and unspooling, equivalent to multilevel filters in 2D image CNNs, gives this method local feature sensitivity while enabling the representation of nonlinear relationships between the latent encoding and the decoded shape. For a 3D autoencoder, the inputs are the 3D mesh vertex coordinates represented as 6890x3 sized tensors. The loss is a geometric mean absolute error (MAE) loss minimizing the reconstruction error against the original input coordinates.

Unlike image convolution on a regular square grid, the 3D mesh graph is irregular with varying connectivity. This requires some architectural modifications in the network in order to apply convolutional operations to human body scans. Our chosen implementation of a 3DAE assumes topological consistency of all mesh inputs. This simplifies the network design by allowing us to determine the spatial down sampling and upsampling operations once per mesh template in a preprocessing step. We defined both the convolutional kernel radius and the spatial kernel stride as two to be consistent with Zhou et al. [146] A visualization of the precomputed graph down sampling for each layer of the autoencoder is shown in Fig 8.2. Additional implementation details are written in the Appendix and in Zhou et al.

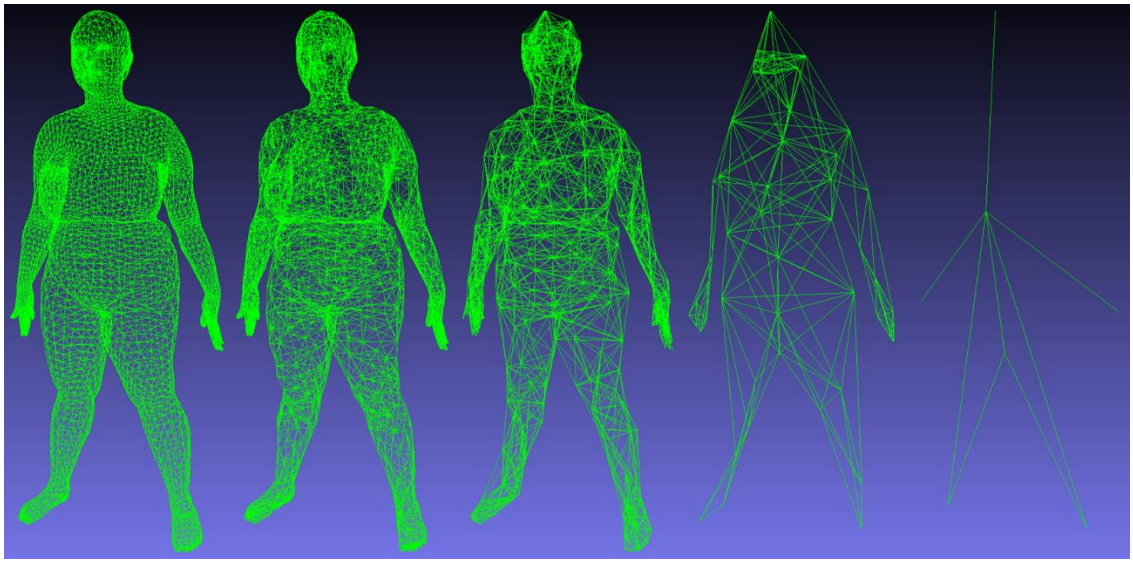


Figure 8.2. Successive convolutional downsampling of a human mesh with a radius and stride of 2. The images represent in order the input mesh (6890x3) and the intermediate graph topologies containing (# vertices x # channels): (6890x16), (1925x32), (400x64), (54x128), (7 x *f*). The final layer is the bottleneck layer, whose channel count *f* is variable in our experiments.

Data paucity was one of the primary reasons for using linear algorithms rather than a deep network in prior work as 3DO data paired with ground truth DXA measurements are absent in the literature outside of the limited SUA and SUK collections. We initialized our model with the DFAUST dataset as augmentation data to mitigate the lack of data availability. We pretrained our 3DAE on DFAUST for 200 epochs, then trained on a composite of SUA, SUK, and CAESAR for an additional 400 epochs excluding DFAUST. This two-tiered training approach was necessary due to the divergent nature of the datasets: SUA, SUK, and CAESAR captured thousands of unique individuals in a consistent pose, while DFAUST captured 10 individuals in thousands of poses. For identity-dependent body shape representation and body composition prediction, a model concentrated on pose-restricted scans of a maximum number of unique individuals is desired. We trained 3DAE models bottleneck channel depths of 7, 43, 90, and 601. With 7 nodes at the bottleneck layer, these channel depths defined total latent feature vector sizes

of 49, 301, and 630, and 4284 as shown in Fig 8.1. All models were trained with a kernel basis size of $M = 37$. The learning rate was set to 1e-4. The training batch size was set to 16. (See Zhou et al. for implementation details) All variations had the same architecture and training data aside from the bottleneck layer which determines the size of the latent space.

8.3.3 Learning a nonlinear transfer function with GPR

For body composition analysis, we performed nonlinear Gaussian process regression between the features extracted from the SUA dataset and their paired DXA measurements. The Shape Up! Data was unique among human body datasets because it captured dual X-ray absorptiometry (DXA) images paired with same day 3D surface scans. This gave every scan in the dataset criterion body composition measurements, allowing us to learn a mapping between 3D scan representation and body composition. GPR was previously used in Wang et al. [125] for visceral adipose tissue estimation. We trained a GPR model to learn a nonlinear mapping between the encoded latent vectors of the training data and the DXA measurements for 12 body composition measures. We chose a squared dot product kernel for GPR under the assumption that relationships between body shape features and body composition were nonlinear but monotonic in both the first and second derivatives. Our assumption of monotonicity between body shape and body composition is consistent with the observation that variables such as visceral fat and percent fat are positively correlated with anthropometric measurements such as waist circumference. Positive correlations between internal composition and exterior shape result from the predictable density of lean and fat body tissue which cause proportionate increases in body volume with increased mass [120]. We experimented with other kernels such as the radial basis function (RBF) and higher powers of dot product kernels and found they performed worse

than a squared dot product kernel. We present results for the squared dot product kernel in this paper. We trained GPR models for male and female participants separately.

GPR derivation details can be referenced in the Appendix and in greater detail in [106] and [44]. We implemented GPR using scikit-learn [87]. We concatenated [height, weight, age] of each participant to the feature vector input to GPR in accordance with the procedures established in previous work. To comprehensively search the feature representation space across multiple scales for the best predictive inputs, we performed GPR on all intermediate feature layers of the 3DAE shape model to predict body composition targets. For each of the four levels of deep features shown in Fig 8.2, we performed GPR to body composition targets. We reported the prediction accuracies for the bottleneck layer containing 4284 (7x612) features and the feature layer producing the lowest RMSE in GPR prediction.

8.3.4 Comprehensive performance analysis vs. linear methods via model permutations

PCA can be considered the most common linear approach to dimensionality reduction for 3D human body shape due to the widespread adoption of methods such as SMPL [67] and was the shape modeling method used in prior work on 3DO body composition estimation. PCA is a deterministic linear operation with a globally optimal solution and produces feature vectors over a space of orthogonal components, making it well-behaved even for datasets containing just tens to hundreds of scans [118, 5].

To test the comparative performance of new nonlinear models against linear baselines established in previous work, we trained a PCA model using the same 4286 finetune training meshes of the 3DAE. DFAUST was not used for training PCA shape models. Although the test

set membership of this study is the same as that of Tian et al (2022) [118], the training set of this work is greatly expanded. Recreating PCA models with the same data selection allowed us to isolate the effects of deep 3DAE shape encoding and nonlinear GPR prediction relative to a baseline of PCA and ordinary least squares (OLS) on the exact same data. Due to the predefined downsampling of the 6890-vertex mesh topology shown in Figure 8.2, the bottleneck layer of the 3DAE must be a multiple of 7. We set the maximum bottleneck (latent code) size of our 3DAE to 4284 as it was the closest multiple of 7 to 4286, the maximum number of possible PCA components (also corresponding to the number of meshes in the training set)**.** While the 3DAE bottleneck layer dimension could be increased to an arbitrarily high number, resulting in lower reconstruction error, there would not be an equivalent linear PCA model to function as a comparative baseline.

To further test our hypothesis regarding the better prediction accuracy of nonlinear GPR relative to OLS, we tested training a GPR on the same PCA basis from Tian et al. (2022) [118], labeled PCA-GPR M391/F457 which indicates 391 and 457 components for males and females respectively.

8.3.5 Statistical Analysis

We measured the geometric reconstruction accuracy of both linear and nonlinear shape models at different model sizes to assess the marginal contribution of nonlinear autoencoders to shape modeling accuracy relative to a linear PCA method used in prior work. We then comprehensively iterated through different model permutations consisting of feature extraction with linear and nonlinear shape models followed by linear and nonlinear regression to DXA body composition measurements to determine the marginal contributions of GPR and 3DAE to

the precision and accuracy of body composition prediction relative to the baselines established by their linear counterparts.

We compared 3D geometric reconstruction error between the PCA baseline and 3DAE shape models at four model sizes as the per-vertex mean absolute error (MAE). We compared the accuracy of body composition regression for the [height, weight, age] feature vector only GPR baseline, the linear baseline (PCA-OLS), hybrid (PCA-GPR), and fully nonlinear (3DAE-GPR) pipelines as measured by the root-mean-squared-error (RMSE) to reference DXA measurements and plotted the normalized RMSE of each predicted metric relative to the PCA-OLS fully linear baseline. The coefficient of determination (R^2) for agreement to DXA was summarized for the 3DAE-GPR model.

We predicted body composition on test-retest data pairs to compare the precision of the above pipeline permutations to prior work and to DXA scanners using the coefficient of variation (%CV) of visceral fat and the repeat RMSE of percent fat. We compared the precision of 3DAE-GPR at two model sizes against the PCA-GPR and PCA-OLS permutations trained and tested on the exact same training data and test-retest pairs. For 3DAE-GPR, we trained the GPR component on the bottleneck layer (301 and 4284 features for the small and large model sizes respectively) to show the marginal precision differences of each method permutation holding all variables but one constant between trials.

We compared the accuracy of our best model performance to a comprehensive list of prior work using percent fat and visceral fat prediction as the benchmark and showed that we can

achieve state of the art results on 3DO body composition prediction using deep convolutional feature extraction and nonlinear regression.

We conducted ablation studies to evaluate the effects of skipping model training steps or withholding training data on geometric reconstruction accuracy and body composition prediction accuracy to confirm that all training procedures and all subsets of the data described in the method positively contributed to the accuracy of our results. We recorded the reconstruction accuracy of the 3DAE using a random initialization without the DFAUST pretrained initialization. We then tested the inverse data withholding condition by training a 3DAE on DFAUST only with no further finetuning with our multi-identity ensemble dataset on SUA test data.

Initially, all scans from SUA, SUK, and CAESAR were merged into a single ensemble dataset under the hypothesis that more data of higher diversity improves autoencoder shape embedding. We withheld CAESAR, SUK and both CAESAR and SUK data from the training sets to test if dissimilarities between data subsets may have reduced reconstruction or regression performance. We tested the geometric reconstruction accuracy and the body composition prediction accuracy for each exclusion scenario using the same procedures described in the

statistical analysis section. Ablation trials were tested on a *d*=301 sized bottleneck network due to its much lower training time.

8.4 Results

The Shape Up! Adults study population has been previously described [118] and summarized in Table 8.2. Only this subset of the data was used for body composition regression training and testing.

Table 8.5. Plus-minus values are standard deviation. p-value of 0.004 was considered significant after Bonferroni correction. All metrics were not significantly different between test and train except for female height, denoted by the asterisk, which is not a predicted measurement. Body composition measurements were taken from DXA.

	Male					
	Train (N = 391)			Test (N = 181)		
	Mean ± SD	Min	Max	Mean ± SD	Min	Max
Age (Years)	44.92 ± 16.06	18	79	44.07 ± 16.15	18	79
Height (m)	1.76 ± 0.08	1.51	2.02	1.75 ± 0.07	1.55	1.91
Mass (kg)	88.42 ± 21.34	40.74	172.4	84.25 ± 18.91	40.77	135.58
BMI	28.44 ± 6.21	16.52	52.55	27.44 ± 5.55	16.96	45.82
Percent Fat	22.66 ± 6.86	9.03	47.69	22.07 ± 6.71	9.03	38.58
Lean Mass (kg)	67.48 ± 12.99	33.95	107.82	64.87 ± 11.64	33.95	93.1
Fat Mass (kg)	20.94 ± 10.67	5.01	66.48	19.39 ± 9.47	5.13	45.94
Visceral Fat (kg)	0.49 ± 0.27	0.16	1.64	0.5 ± 0.32	0.16	1.64
Leg Lean (kg)	10.97 ± 2.28	5.43	18.95	10.48 ± 1.99	5.43	14.74
Arm Lean (kg)	4.46 ± 1.05	2.05	8.33	4.26 ± 0.98	2.05	7.38
Trunk Lean (kg)	32.36 ± 6.4	15.95	51.16	31.24 ± 5,82	15.95	48.02
Trunk Fat (kg)	10.5 ± 6.15	1.76	34.12	9.72 ± 5.71	1.97	26.25
Leg Fat (kg)	3.41 ± 1.71	0.89	11.86	3.12 ± 1.44	0.85	9.02
Arm Fat (kg)	1.25 ± 0.68	0.29	4.34	1.15 ± 0.61	0.29	3.72
FMI	6.74 ± 3.37	1.68	20.78	6.32 ± 3.01	1.91	15.49
FFMI	21.72 ± 3.53	14.18	35.65	21.16 ± 3.23	14.18	30.72

	Female					
	Train (N = 457)			Test (N = 239)		
	Mean ± SD	Min	Max	Mean ± SD	Min	Max
Age (Years)	46.24 ± 16.13	18	89	47.53 ± 16.71	18	89
Height (m)	1.62 ± 0.07	1.44	1.80	1.61 ± 0.07 *	1.44	1.76
Mass (kg)	72.43 ± 20.93	35.44	153.05	69.39 ± 19.60	35.44	153.05
BMI	27.48 ± 7.65	14.16	51.86	26.81 ± 7.05	14.16	51.86
Percent Fat	34.06 ± 7.88	12.63	53.3	33.78 ± 7.44	17.18	53.3
Lean Mass (kg)	46.49 ± 9.48	28.56	80.37	44.91 ± 9.42	28.56	80.37
Fat Mass (kg)	25.85 ± 12.72	6.3	72.68	24.39 ± 11.38	6.88	72.68
Visceral Fat (kg)	0.45 ± 0.31	0.06	1.37	0.43 ± 0.3	0.05	1.22
Leg Lean (kg)	7.48 ± 1.73	4.25	14.03	7.21 ± 1.78	4.42	13.19
Arm Lean (kg)	2.42 ± 0.57	1.31	4.42	2.34 ± 0.58	1.44	4.63
Trunk Lean (kg)	23.13 ± 4.91	13.71	41.59	22.26 ± 4.72	13.71	41.14
Trunk Fat (kg)	11.94 ± 6.75	2.37	35.6	11.27 ± 6.21	2.48	35.6
Leg Fat (kg)	4.81 ± 2.23	1.23	12.92	4.54 ± 1.98	1.23	12.34
Arm Fat (kg)	1.67 ± 0.99	0.28	6.08	1.55 ± 0.84	0.4	6.08
FMI	9.82 ± 4.79	2.01	26.57	9.44 ± 4.24	2.75	24.68
FFMI	17.6 ± 3.28	11.43	29	17.33 ± 3.16	10.93	28.88

Geometric reconstruction error for both 3DAE and PCA shape models of four increasing sizes are shown in Table 8.3. The dimensionality *d* represents either the number of PCA coefficients used to parameterize shape or the number of latent variables in the bottleneck layer of the 3DAE connecting the encoder module to the symmetric decoder module. Reconstruction error was calculated as the geometric mean absolute error (MAE) between original and reconstructed vertex 3D positions. As expected, models reducing to a larger minimum parameter count were able to reconstruct the test data with lower error. Both linear and deep methods are comparable in terms of geometric reconstruction accuracy for the first three model sizes. PCA

achieved lower geometric reconstruction error at the highest parameter count while 3DAE reconstruction error appears to level off at above 2mm MAE.

Table 8.6. Per-vertex reconstruction error measured as mean absolute error (MAE) between input and decoded meshes for held-out test meshes from Shape Up! Adults (n=424). The dimensionality (*d*) represents the number of latent layer variables in the autoencoder bottleneck or the number of principal components used for a linear shape space reconstruction.

	d=49	*d*=301	*d*=630	*d*=4284
3DAE MAE (mm)	5.16	2.57	2.18	2.02
PCA MAE (mm)	5.26	2.41	2.23	1.0

Fig. 8.3 depicts the body composition prediction results from *d*=4284 sized models using different combinations of shape feature extraction and body composition regression methods. Excluding the baseline column, subsequent column represents a change of exactly one pipeline parameter holding all others constant; i.e. OLS to GPR, PCA to 3DAE, 4824 total parameters to 400x64 parameters. All model permutations were trained and tested on the exact same data. The smaller *d*=301 3DAE had worse accuracy on all predicted variables than the larger model on males and was comparable on female; however, we found that no level of feature extraction with a 3DAE improved upon using the 3D mesh coordinates directly as input features for females in either model. The *d*=301 sized model was omitted from the charts for clarity and brevity.

The baseline prediction accuracy of GPR is the RMSE resulting from a regression using only known features [height, weight, age] without any conditioning on the 3DO shape features as was previously done in [118] and [117]. This RMSE was higher than any subsequent regression model using any kind of shape features as input and shows the validity of using 3DO shape as

features for body composition prediction even when nonlinear regression algorithms are employed.

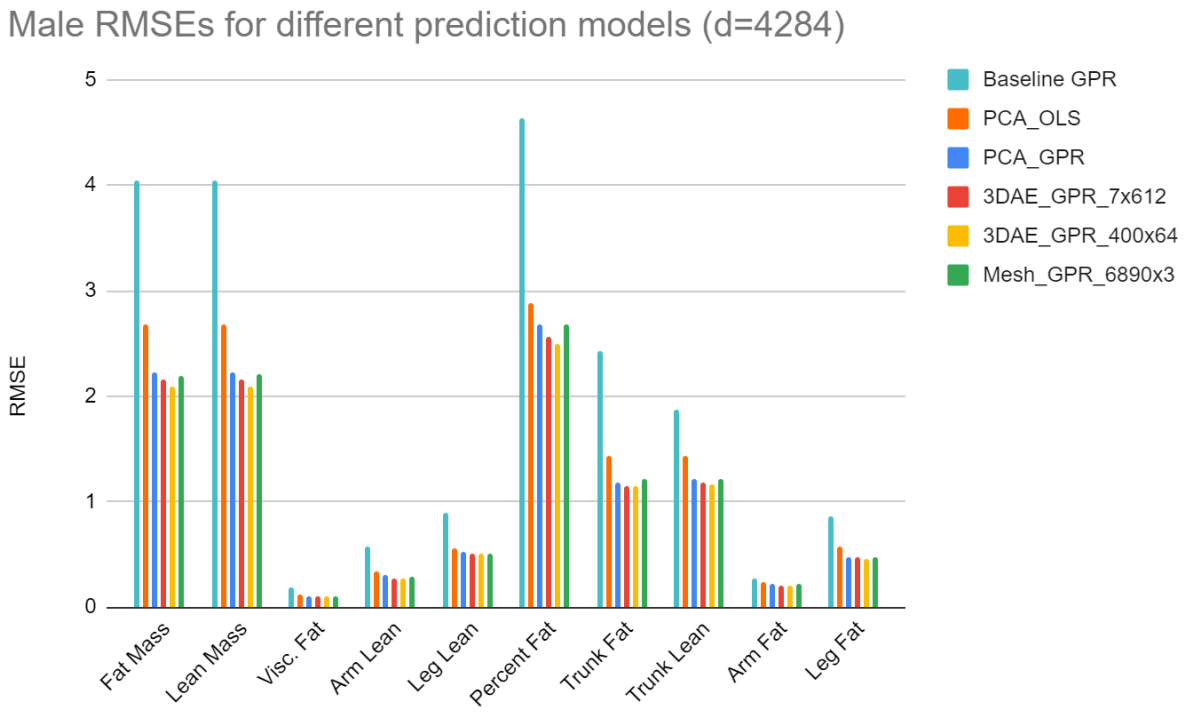
PCA-OLS is the linear model baseline meant to represent the methods of prior work. This is a linear regression model for body composition prediction taking PCA features as inputs. The PCA model had 4284 parameters to stay exactly consistent with the dimensionality reduction factor of the 3DAE and represented the best-case scenario for linear prediction of body composition from 3DO in this dataset. Prediction RMSEs for the PCA-OLS pipeline were identical to OLS regression using the mesh vertex positions as inputs, indicating that at ~1mm of reconstruction error the shape features of PCA with d=4284 is also near lossless relative to the original 3DO scan for body composition prediction.

PCA-GPR represents a hybrid pipeline predicting body composition with nonlinear GPR from linear PCA features. Nonlinear regression from linear shape features achieved lower RMSE on every predicted metric except arm lean on females, which was equal, and leg lean on females, which was higher by 0.02 kg (or 5%). This indicated that GPR was a more accurate regression method than OLS when all other factors were held constant for the majority of body composition targets. Just as in the OLS case, we tested the prediction accuracy of GPR using the 3D mesh vertices as input and noted that the RMSEs were comparable to the 4284 PCA feature input.

3DAE-GPR represents the fully nonlinear pipeline where body composition is predicted from GPR using 3DAE deep features as inputs. The results for the bottleneck layer (7x612) and the lowest error layer are shown. The third layer was the most accurate feature layer for males and the first layer was the best for females. In males, 3DAE feature extraction lowered RMSE relative to the previous model permutations, but in females no level of feature extraction

outperformed GPR on the raw mesh coordinates. This result suggests the features extracted from female meshes were less informative relative to male features or were highly correlated regardless of the method and model size. GPR always improves accuracy relative to OLS, but 3DAE feature extraction only improved accuracy for males. This is supported by our observation in Table 8.7 that 3DAE-GPR for males trained on the bottleneck layer with 4284 features achieved lower RMSEs than PCA-GPR with the exact same parameter count, while the same was not true for females. Coefficients of determination (R^2) for the best 3DAE-GPR models were greater than or equal to 0.86 for all predicted variables.

OLS regression to body composition with 3DAE features resulted in very low correlations with DXA due to the nonlinearity of the deep features and was not reported. We also tested concatenating all feature layers into a single multi-scale feature vector for GPR regression but did not achieve lower errors in doing so.



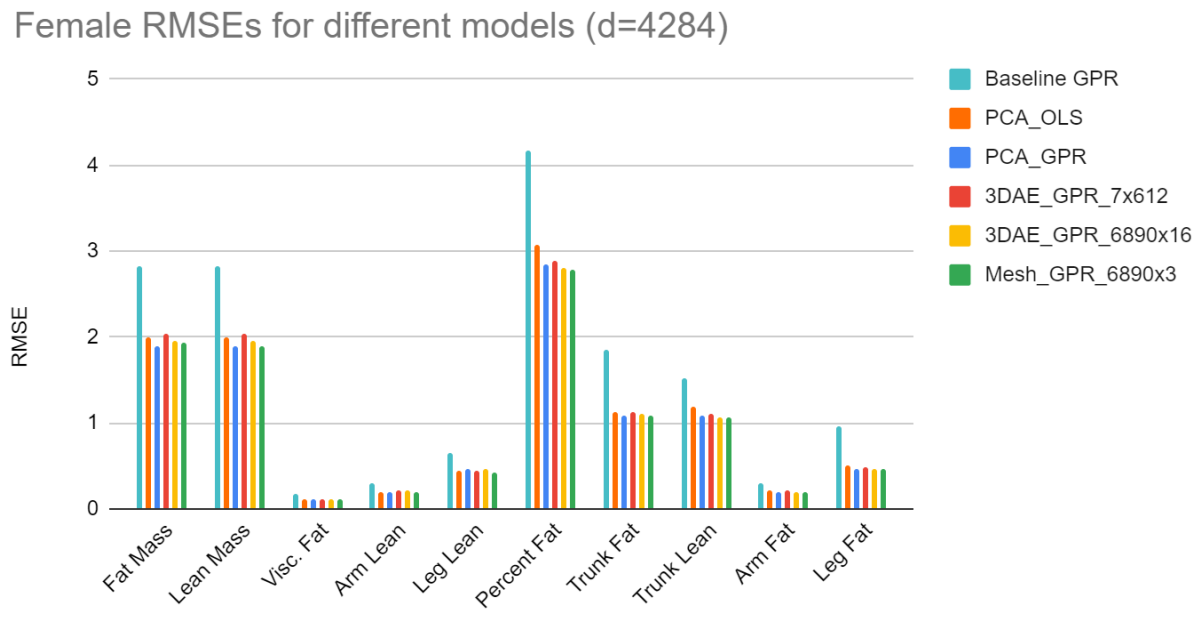
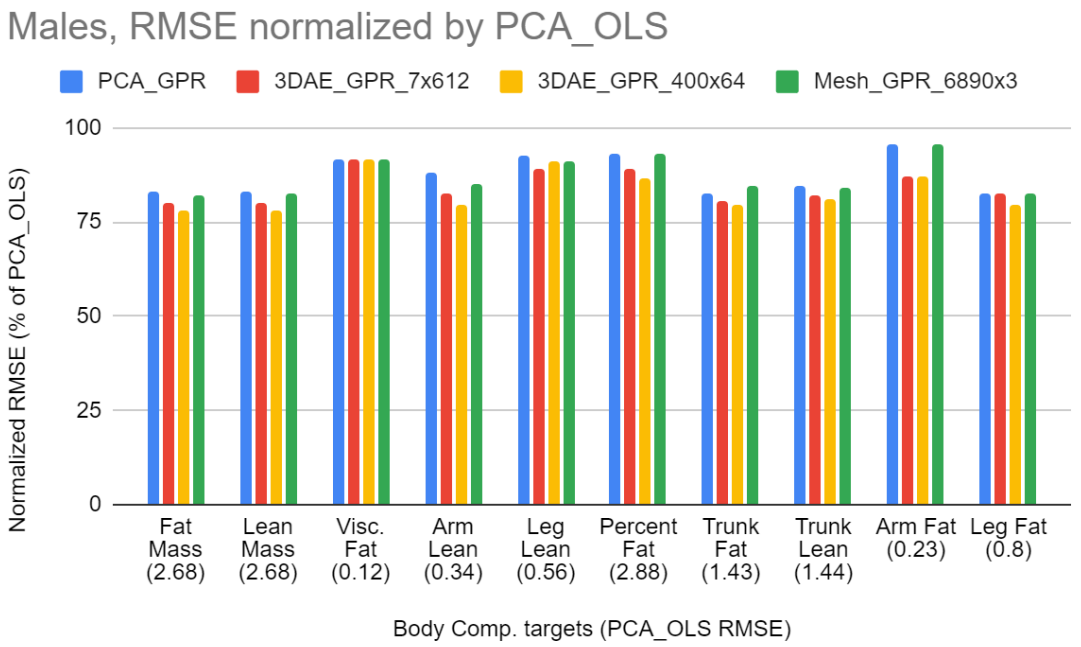


Figure 8.3. RMSE prediction errors for different model permutations. Every column represents exactly one modification to the pipeline parameters from the previous column (i.e. OLS to GPR) excluding the baseline column. RMSE is shown in kg except for PFAT, which is expressed as a percentage. Baseline is GPR prediction from known priors [height, weight, age] only. PCA_OLS is linear regression from linear PCA features. This result was identical to OLS from the 3DO vertex coordinates. PCA_GPR is nonlinear GPR prediction from linear PCA features. Mesh_GPR is GPR prediction from 3DO vertex coordinates with no feature extraction. 3DAE_GPR_X is the most accurate GPR model trained from any feature layer of the 3DAE network with the feature dimension indicated as X.

Table 8.7. Body composition prediction RMSEs comparison between 3DAE-GPR and PCA-GPR holding all variables constant except for shape feature type. 3DAE was trained with a d=4284 sized bottleneck. The bottleneck layer (with a size 4284 feature vector) was used to train the GPR to body composition. PCA-GPR was trained with 4284 PCA components, the same feature vector size as 3DAE-GPR. Males exhibited lower RMSE when 3DAE was used as the shape feature extractor, while females exhibited comparable performance with mixed results. 3DAE features are informative for males but not necessarily so for females.

	RMSEs			
	Male PCA-GPR (d=4284)	**Male 3DAE-GPR (7x612)**	***Female PCA-GPR (d=4284)***	***Female 3DAE-GPR (7x612)***
Fat Mass (kg)	2.22	2.15	*1.9*	*2.03*
Lean Mass (kg)	2.22	2.15	*1.9*	*2.03*
Visc. Fat (kg)	0.11	0.11	*0.11*	*0.11*
Arm Lean (kg)	0.3	0.28	*0.2*	*0.22*
Leg Lean (kg)	0.52	0.5	*0.46*	*0.45*
Percent Fat (%)	2.68	2.56	*2.85*	*2.89*
Trunk Fat (kg)	1.18	1.15	*1.09*	*1.13*
Trunk Lean (kg)	1.22	1.18	*1.08*	*1.1*
Arm Fat (kg)	0.22	0.2	*0.2*	*0.21*
Leg Fat (kg)	0.48	0.48	*0.47*	*0.49*

We rescaled the charts in Fig 8.3 by normalizing each column by the RMSE of the fully linear PCA_OLS model and plotted the values in Fig. 8.4 for visual clarity. In males, all subsequent model permutations had RMSEs less than that of PCA_OLS (indicated by a normalized RMSE of 100%). For females, only leg lean exceeded 100% of PCA_OLS RMSE when moving from OLS to GPR. However, the normalized RMSE for leg lean fell to 97.7% of PCA_OLS when GPR prediction takes the mesh vertices as input. Since PCA_OLS achieved the same RMSEs as Mesh_OLS, this result supports the conclusion that GPR improved upon OLS when the feature input was held constant as the 3DO mesh vertices. All *p*-values between the predicted body composition values in Figure 4 and their paired reference DXA values as calculated by a two-sided *t*-test were greater than the Bonferroni corrected significance level of 0.004 for 12 simultaneous measurements. Our *t*-test indicates no significant difference between any of our prediction methods and the measurement of the reference method (DXA), which is the desired result for an analysis testing for agreement between two methods and suggests no significant bias in our model.



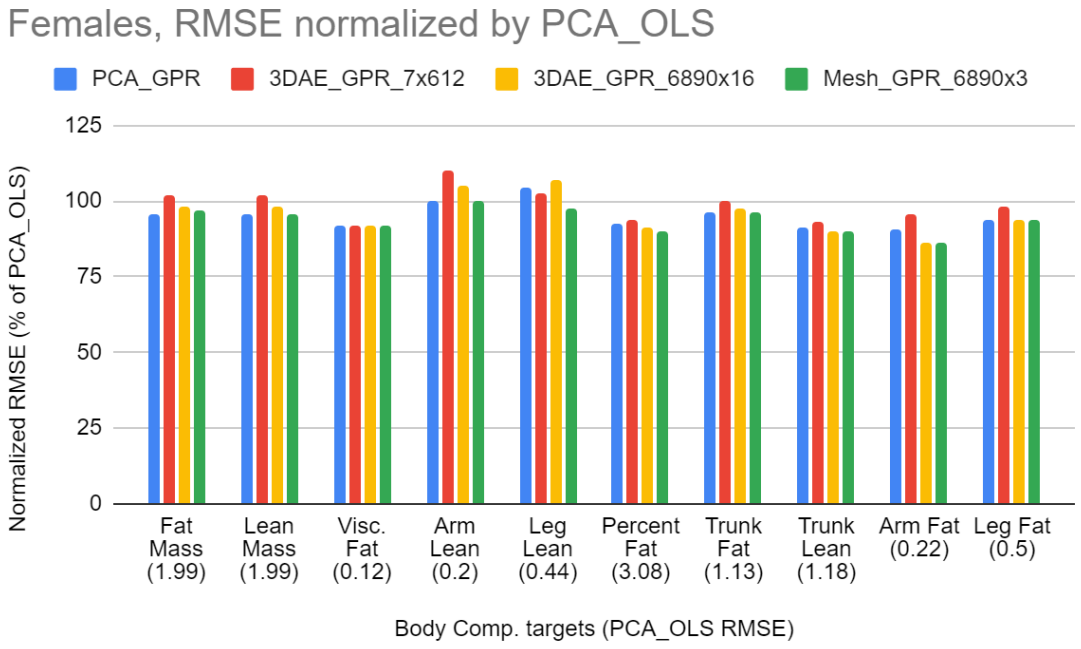


Figure 8.4. Normalized RMSE values from Figure 3 as a percentage of the linear baseline (PCA_OLS), shown in parentheses below every body composition metric.

Test-retest precision of percent fat and visceral fat estimation is shown in Table 8.5. For visceral fat mass precision, Coefficient of Variation (%CV) was calculated according to the definition in Glüer et al. [39] between two scans of the same individual in the test set taken on the same day. RMSE between trial 1 and trial 2 was shown for percent fat precision as the %CV of a percentage measurement is not used in convention.

GPR was more precise on retests than OLS as shown by the PCA-OLS versus PCA-GPR trials, where the latter resulted in up to a 30% decrease in precision error. 3DAE decreases precision error relative to PCA, as illustrated by the 3DAE-GPR versus PCA-GPR trials. 3DAE coupled with nonlinear GPR had the highest precision. Compared to DXA, percent fat precision error for 3DAE-GPR was roughly twice as high. However, visceral fat precision was lower than DXA, indicating the 3DAE model paired with GPR prediction is much more reliable on retest accuracy than prior work. The larger d=4284 3DAE model was comparable in precision to the smaller model. 3DAE was more precise than PCA, and GPR was more precise than OLS.

Table 8.8. Test-retest precision between repeat 3DO scan pairs of each participant in the SUA test set on the same scan device. 3DAE models were benchmarked with GPR trained from their bottleneck layers to standardize the comparison between model sizes. 3DAE was more precise than PCA, and GPR was more precise than OLS.

Visceral fat % CV	3DAE-GPR 301	3DAE-GPR 4284	PCA-GPR 4284	PCA-OLS 4284	Tian et al. 2022 [118]	DXA (Criterion)
Males (n = 143)	4.4%	5.0%	5.2%	7.2%	8.8%	4.8%
Females (n=199)	5.5%	5.5%	6.2%	8.6%	12.3%	6.3%

Percent fat precision RMSE	3DAE-GPR 301	3DAE-GPR 4284	PCA-GPR 4284	PCA-OLS 4284	Tian et al. 2022 [118]	DXA (Criterion)
Males (n = 143)	1.0%	0.9%	1.1%	1.6%	1.9%	0.5%
Females (n=199)	1.2%	1.2%	1.4%	2.0%	2.9%	0.5%

Comparisons of our 3DAE-GPR models against prior work are shown in Table 8.6 using the d=4284 latent size model and Gaussian process regression with the best performing feature layer (third for males, first for females). Percent fat (PFAT) and visceral fat mass (VFAT) were selected as the comparative target variable as they tended to have the lowest accuracy in previous methods based on linear models.

GPR using the exact same PCA features of Tian et al. (PCA-GPR M391/F457) lowered the RMSE from OLS for percent fat in both sexes and in female visceral fat but increased it slightly in male visceral fat by 8%. The models presented in [118] were built on a dataset containing an order of magnitude fewer members than what we presented in this work. These

results suggest linear pipelines may be more competitive with nonlinear methods when training data quantity is more limited. We included RMSEs for our best performing, maximum size PCA model (PCA-GPR 4284) to demonstrate increased parameter count does not cause overfitting and degrade accuracy relative to a sparser model.

Our fully nonlinear model, 3DAE-GPR, produced the lowest error on visceral fat and percent fat estimation compared to all prior work on 3DO body composition estimation (bolded). We note that compared to Tian et al (2022), our best 3DAE-GPR model achieved lower RMSE on all the 12 body composition metrics measured in Figure 8.4. However, predicted metrics other than percent fat and visceral fat already showed high correlation with DXA in prior work using linear methods. Thus, we concentrate the comparison to the metabolically significant and previously underperforming predictions of percent fat and visceral fat in Table 8.9. The test set used in this work was held the same as [118]. However, the training dataset was greatly expanded in size and scope relative to prior works.

Table 8.9. Root-mean-squared errors (RMSE) for predicted percent fat (PFAT) and visceral fat (VFAT) of all current 3D-optical body composition prediction literature on Shape Up! Adults compared to the 3DAE-GPR prediction of the d=301 and d=4284 models using the most accurate feature layer identified in Figure 3. Best performing values are bolded.

Paper	*N* test meshes	PFAT RMSE (%)	VFAT RMSE (kg)
Ng et al (2019) Anthro only	M: 177 F: 230	M: 4.03 F: 3.99	M: 0.15 F: 0.14
Ng et al (2019) [83]	M: 177 F: 230	M: 3.55 F: 3.88	M: 0.14 F: 0.13
Tian et al (2020) [117]	M: 31 F: 39	M: 3.90 F: 3.29	M: 0.15 F: 0.17
Wong et al (2021) [136]	M: 159 F: 202	M: 2.73 F: 3.46	M: 0.13 F: 0.13
Tian et al (2022) [118]	M: 182 F: 248	M: 3.24 F: 4.22	M: 0.12 F: 0.14
PCA-GPR M391/F457 [118]	M: 182 F: 248	M: 2.79 F: 3.09	M: 0.13 F: 0.12
PCA-GPR 4284	M: 181 F: 239	M: 2.68 F: 2.85	**M: 0.11** **F: 0.11**
3DAE-GPR 4284	M: 181 F: 239	**M: 2.50** **F: 2.81**	**M: 0.11** **F: 0.11**

8.4.1 Ablation studies

Fig. 8.5 shows a mesh reconstruction using a *d*=4284 model trained with 400 epochs on the finetuning ensemble training data from 1) a random initialization state and 2) from a pretrained initialization trained with 200 epochs using DFAUST data only. The model trained

from a random initialization achieved 22.7mm MAE on the test data, more than 10x the error of the error show in Table 8.3.

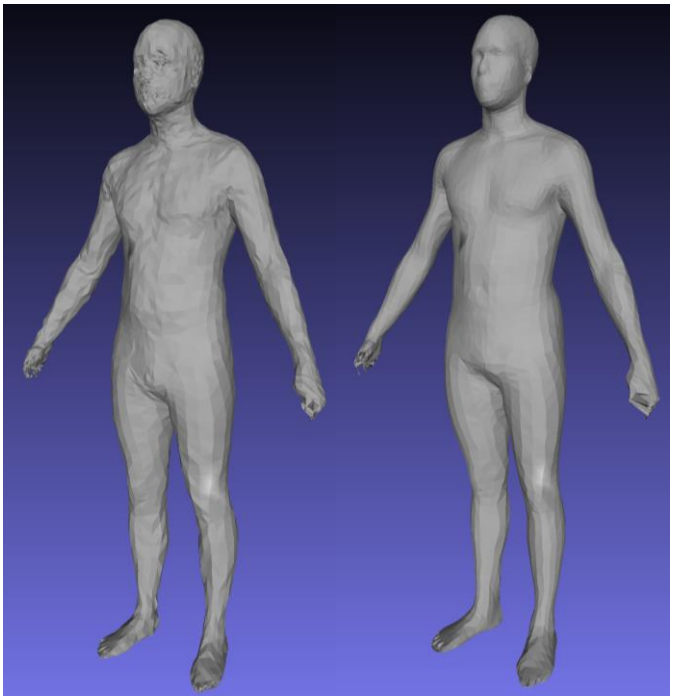


Figure 8.5. Pretraining on 40,000 DFAUST meshes improves reconstruction fidelity in single pose, multi-identity data. Left: Evaluation set mesh reconstruction after training 400 epochs from random initialization. Test set reconstruction error was 22.7mm. Right: Same mesh reconstruction using a model pretrained first on 200 epochs with DFAUST only followed by 400 epochs of finetuning. Test set reconstruction error was 2.0mm.

Fig. 8.6 shows visualizations of geometric reconstruction error as heat maps for a male and female subject before and after fine-tuning the d=4284 3DAE model with high resolution Shape Up! and CAESAR data. Without fine-tuning, the 3DAE model was equivalent to the work presented in Zhou et al. [146] trained exclusively on DFAUST data and achieved a 3D reconstruction error of 8.7 mm, as opposed to the 2.0 mm shown in Table 8.3 for the finetuned model. This model generalized very poorly to unseen scans of many unique individuals such as in Shape Up!, as DFAUST only contained 10 unique individuals captured in thousands of

different poses. A 3DAE model for clinical machine learning applications needs to generalize well to any individual scanned from the general population in a neutral pose. Our fine-tuned model represented the non-rigid, identity-dependent deformations of unique individuals much more accurately.

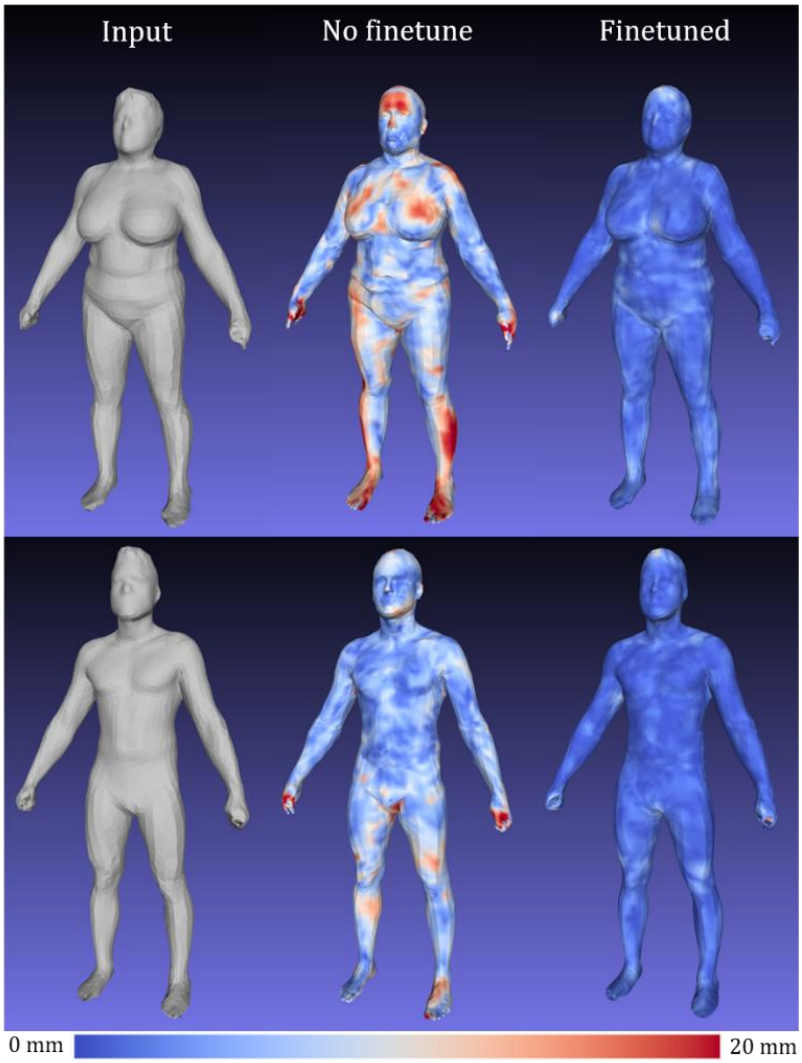


Figure 8.6. Finetune training on 4286-member multi-identity ensemble dataset reduces reconstruction error on test set meshes relative to 3DAE models trained on DFAUST only. Left: input 3DO test mesh. Center: reconstruction error heatmap using model state following 200 epochs of pretraining on DFAUST only. Test set MAE was 8.7mm. Right: heatmap on same test mesh after an additional 400 epochs of finetuning on multi-identity, single pose ensemble dataset. Error heatmap is in the range of [0,20] mm.

Withholding non-SUA data from 3DAE training did not improve 3D reconstruction error on SUA test meshes. In all three ablation trials, reconstruction accuracy on the same SUA test set (2.58mm, 2.71mm, 2.76mm for removing only SUK, only CAESAR, and both, respectively) was worse than the model trained on all data combined (reconstruction accuracy of 2.57) in Table 8.3. This ablation result validates the inclusion of a diverse dataset across multiple collection protocols.

GPR models trained on the 3DAE model excluding SUK data showed 1% difference in RMSE in both directions. Not all target variables were uniformly lower in error when including versus excluding SUK; we conclude that the variation could be attributed to noise and does not justify excluding SUK. Excluding CAESAR data from the training scans had similarly negligible effects on male prediction accuracy but increased female RMSEs without exception by up to 4%. Excluding both SUK and CAESAR produced results like the case where only CAESAR was excluded, but with even higher errors in females. Overall, we found that including all available 3D mesh data when training shape and regression models produced the lowest errors for 3D reconstruction and body composition prediction.

8.5 Discussion

This study showed two primary findings. First, GPR generally improves body composition prediction accuracy and precision relative to linear regression regardless of whether the feature domain is linear (PCA) or nonlinear (3DAE). By comprehensively exploring model permutations and changing exactly one variable at each iteration, we were able to demonstrate that nonlinear regression with GPR performed better than OLS for every body composition

target in both males and females. Second, 3DAE feature extraction improved body composition prediction for males. However, deep features were not more informative on females. A combination of an expanded and diversified training set with deep feature extraction and nonlinear GPR prediction allowed us to build an end-to-end model that outperformed all prior works on body composition prediction accuracy. The improved accuracy and precision demonstrated here builds on our prior work [83, 118] and further establishes 3DO as a reliable and accessible clinical tool for the assessment of body composition. Such a tool has promising clinical implications for the management of chronic disease, where body composition has known associations with morbidity and mortality.

To our knowledge this was the first application of deep nonlinear autoencoder networks to 3DO body composition estimation. We explicitly used spatial graph convolutional network for our autoencoder approach. Other autoencoder approaches, such as implicit surface encoding [4], non-convolutional variational architectures [54], and spectral graph convolution [62] have been used to study body shape, but without associations to clinical outcomes. Future work may find that these other approaches have advantages to the work presented here.

We hypothesized that lower geometric reconstruction error in a shape autoencoder model was correlated with higher body composition prediction accuracy during regression from the extracted features. However, although a linear PCA model of size d=4284 produced the lowest shape reconstruction error, it did not always outperform 3DAE on body composition prediction accuracy or precision. Geometric reconstruction accuracy may be affected by shape deformations irrelevant to body composition variation such as pose or face detail. PCA may be outperforming 3DAE in shape reconstruction at high parameter counts due to 3DAE potentially learning redundant and correlated features when the model size is large. Unlike 3DAE, PCA is guaranteed

to learn uncorrelated, orthogonal features due to its mathematical construction. Future work should investigate the relationship between geometric reconstruction accuracy and body composition prediction accuracy by controlling for uncorrelated geometric variations and using different autoencoder architectures.

PCA's high performance on shape reconstruction validate the methods of past work built on PCA models with linear regression [83, 118, 135] despite the restrictive linear assumptions of the algorithm. Prior work restricted the shape model and body composition prediction features to a sparse subset of the total number of PCA components to avoid overfitting during regression. Our extensive testing with different model sizes and permutations does not support the assumption that large parameter counts overfit on test data. PCA-GPR model using 4284 components achieved lower RMSEs on percent fat and visceral fat on both males and females relative to models trained with 391 or 457 features. Restricting shape and regression models to the first *n* components that describe 95% or 99% of the shape variance may not be justified in future work as it potentially handicaps prediction accuracies unnecessarily.

Our chosen autoencoder architecture does not explicitly disentangle pose deformation from identity-dependent deformations, such as in Wong et al. [136] or Jiang et al. [54]. Factoring out slight pose variation across our dataset may improve the reconstruction accuracy and regression accuracy of our method. Increased sensitivity to small changes in body shape may allow our method to improve monitoring body composition change over time in the same individual [134]. As our dataset was transformed to adopt the mesh topology of SMPL [67] while preserving the geometric surface detail of the original 3DO scans, it may be straightforward in future work to "unpose" every mesh in our 4286-member ensemble dataset to

a neutral T-pose like that of Wong et al. using the skinning weights and joint position definitions of the SMPL template. The accuracy of this procedure will depend on how anatomically consistent the template correspondences are between our dataset and the meshes in Loper et al.

Future work can explore the impact of different nonlinear regression methods on body composition prediction accuracy. GPR regularized the shape of the regression function to its kernel. A less restricted but more flexible regression method such as a multi-layer perceptron (MLP) may achieve better prediction error with proper regularization; however, deeper models run the risk of overfitting to the training data especially if the latent parameter count is large. We observed that GPR using a radial basis function kernel (RBF) already exhibits symptoms of overfitting, showing zero error on the training data but greatly increased error on the test data.

Like prior works using linear models, this work separately trains a shape feature extractor followed by a body composition regression from features. Future work may integrate the current two-step pipeline into a single end-to-end deep network. This can be achieved by connecting the 3DAE encoder layers directly to a neural network regression model that targets body composition as its output instead of self-reconstruction. This end-to-end model can be initialized to state reported in this work by importing our trained 3DAE weights into the encoder of the network and the weights of a separately trained regression model into the neural network regression layers. GPR may be recreated as a neural network in PyTorch with GPyTorch [37]. This implementation could allow the combined network to achieve greater accuracy by starting from the state presented in this work and further optimizing all features to target body composition prediction with no intermediate objectives.

Our 3D deep shape model was trained on the largest collection of high quality, multi-identity 3DO scan data currently available, spanning CAESAR, SUA, SUK, and DFAUST over

more than a 20-year period. However, with 2,900 unique individuals, this dataset is still orders of magnitude smaller than those used in analogous networks for deep 2D image learning. This restriction on dataset size and population variation sampling density can create gaps in our model that exhibit low reconstruction and prediction accuracy where training data was not available or under sampled, especially at the extremes of body shapes as shown in Fig. 8.7. Ongoing data collection of high resolution 3DO scans will improve the reconstruction accuracy of our method. Improvement of the nonlinear regression model will require additional paired DXA scans.

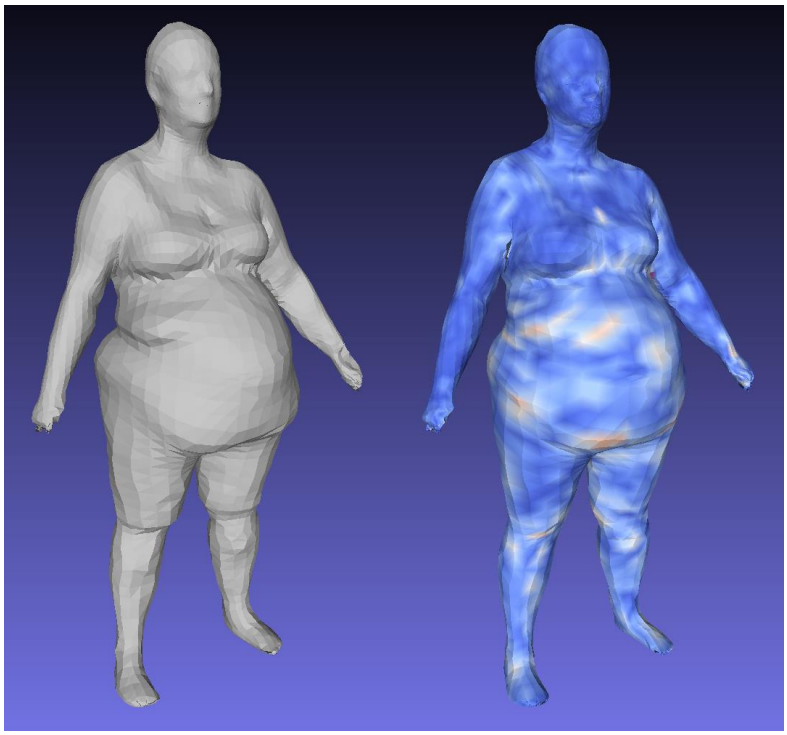


Figure 8.7. Left: an input test mesh with high BMI; right, the heat map of the reconstruction showing where the greatest errors occurred. The numerous folds in the abdomen of this individual (47.3% body fat) were unique to only a few scans in the dataset and were possibly insufficiently modeled in the latent encoding.

The CAESAR data used in this work was built from the original 3D scan data collected by Robinette et al. [97] and not on a derived reconstruction such as the MPI CAESAR dataset produced by Pishchulin et al. [90]. The MPI dataset is a template standardization of CAESAR using a limited number of PCA features for reconstruction. The resulting shapes are linearly projected compressions of the original 3D scan data and do not preserve original high-resolution detail well, as shown in Fig. 8.8. Including these shapes into our training database would bias our deep nonlinear model towards an approximation of a linear solution. Our remeshing of the CAESAR scans into the SMPL format includes a nonrigid surface-to-surface deformation that produced templated meshes consistent with the original scan geometry and did not constrain the training data to projections onto a linear subspace. Future work studying human shape variation as a function of varying identity rather than pose may build upon our results by utilizing the

higher fidelity standardized templates of our dataset rather than training on potentially degraded 3D shape geometry.

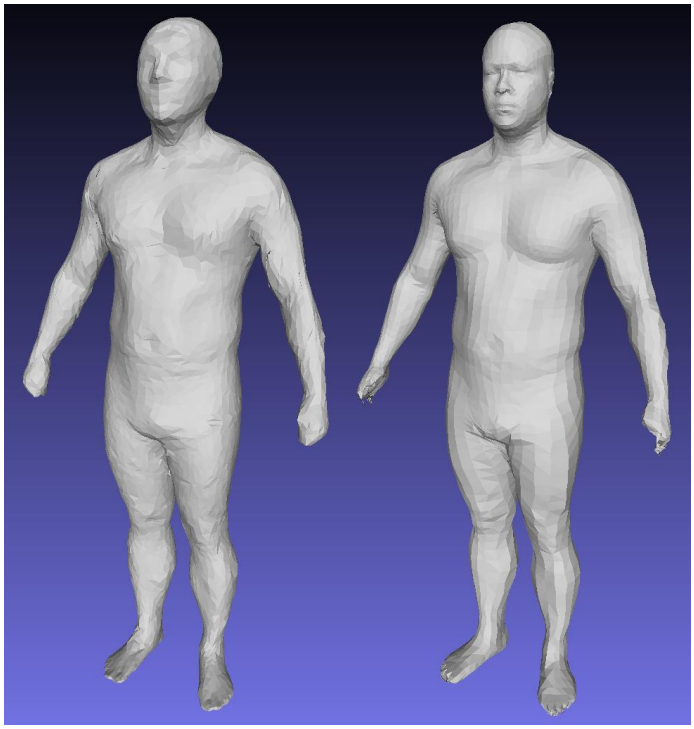


Figure 8.8. MPI CAESAR mesh compared to our remeshing of the same individual. The MPI shape was projected onto a linear PCA basis and lost considerable identifying detail.

8.6 Conclusion

A comprehensive comparison of nonlinear methods for shape feature extraction and body composition regression against previous linear algorithms showed that nonlinear GPR improved body composition prediction accuracy and precision relative to linear regression for 12 metrics of body composition in both males and females. Nonlinear GPR produced up to 20% reduction in prediction error and up to 30% increase in precision over linear regression for both sexes in 12 tested body composition variables. Feature extraction with a deep 3D autoencoder provided marginal improvements to prediction accuracy for males but did not supersede the performance

of GPR on raw mesh vertex positions for females. Deep shape features produced 6-8% reduction in prediction error over linear PCA features for males. Our best performing nonlinear pipeline using 3DAE-GPR outperformed prior works on body composition prediction accuracy for all metrics. Precision error of our method is within 1-2x that of DXA, the gold standard for compartmental body composition measurement. The agreement with DXA as measured by R^2 values were greater than or equal to 0.86 for all predicted metrics. These findings improve the clinical utility of 3DO as an accurate and accessible tool for the assessment of body composition. Future work in this space should explore the effects of disentangling pose variations from 3DO scans along with different deep network architectures such as variational latent encodings, end-to-end deep models, and predicting longitudinal change over time in individual subjects.

8.7 Appendix

8.7.1 Graph Convolutional Network Implementation

For each layer of the network, each vertex on its downsampled graph is a weighted average of all vertices within a 2-ring neighborhood of the center vertex in the previous layer. Center vertices and their neighborhoods are determined during preprocessing via an adjacency matrix; see Zhou et al. for details. Every vertex in each downsampled graph level contains a feature vector of arbitrary channel depth determined by the network architecture akin to a 2D convolution. The downsampling mapping determines which local neighborhood of features get pooled to a single node at each layer of the network akin to a convolutional filter window.

In a 2D convolutional network, the output feature $\mathbf{y}_i$ for all input features $\mathbf{x}_{i,j}$ within a size j kernel at position i is calculated as

$$\mathbf{y}_i = \sum_{\mathbf{x}_{i,j}} \mathbf{W}_j^{\mathrm{T}} \mathbf{x}_{i,j} + \mathbf{b} \quad (30)$$

for an I channel input feature $\mathbf{x}_{i,j}$, an $I \times O$ weight matrix $\mathbf{W}_j$, and some learned bias $\mathbf{b}$. This formulation works when meshes are preprocessed to have the same connectivity at all vertices [45] but is invalid on an irregular graph as there is not a one-to-one correspondence between the number of neighboring graph vertices and kernel indices. Zhou et al. defines a generalization that handles irregular convolutional kernel size by redefining the kernel as a set of M basis vectors with no fixed spatial correspondence. For each index $\mathbf{x}_{i,j}$ within a 2-ring spatial neighborhood, the corresponding weight matrix does not correspond to a single kernel index but rather is redefined as a linear combination of all basis vectors $\mathbf{B}_k$:

$$\mathbf{W}_{i,j} = \sum_{k=1}^{M} \alpha_{i,j,k} \mathbf{B}_k \quad (31)$$

where M is 37 in our experiments. Thus, every vertex within a variably sized spatial convolutional kernel is a function of a shared set of M basis vectors analogous to a convolutional filter window in 2D. See Zhou et al. for derivations.

8.7.2 Gaussian Process Regression Summary

A Gaussian process $f(x)$ is characterized by a multivariate Gaussian probability distribution $f(x) \sim N(\mu, \sigma^2 \mathbf{K})$ for a feature vector $\mathbf{x}$, where μ is the mean, σ^2 is the variance, and $\mathbf{K}$ is the symmetric covariance matrix where each entry $\mathbf{K}_{i,j} = K_{\boldsymbol{\theta}}(x_i, x_j)$ for some kernel function $K_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ computed at all paired combinations of training data $\mathbf{x}_i, \mathbf{x}_j$. The kernel function $K_{\boldsymbol{\theta}}$ is a measure of distance between two sample points x_i, x_j in kernel space and implies that inputs close together in kernel space should have similar predicted targets y_i, y_j. For our experiments, $K_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1(\mathbf{x}_i \cdot \mathbf{x}_j + \theta_2)^2 + \theta_3$ for kernel parameters $\boldsymbol{\theta} =$

$\{\theta_1, \theta_2, \theta_3\}$.The parameters of the Gaussian process $\mu, \sigma^2, \boldsymbol{\Theta}$ are optimized via maximum likelihood estimation (MLE) of the probability density function (PDF) of the multivariate Gaussian for *n* observations:

$$L(\mathbf{y}|\mu, \sigma^2, \boldsymbol{\Theta}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n|\mathbf{K}|}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{1}\mu)^{\mathrm{T}}\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{1}\mu)\right) \quad (32)$$

where **1** is an *n* length vector of 1's. The parameters $\mu, \sigma^2, \boldsymbol{\Theta}$ are optimized by setting the derivative of the log-likelihood $\ln(L)$ to 0 and solving with standard methods:

$$\ln(L) = \frac{-1}{2\sigma^2}(\boldsymbol{y} - \mathbf{1}\mu)^{\mathrm{T}}\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{1}\mu)$$

$$\hat{\mu}, \hat{\sigma}^2, \widehat{\boldsymbol{\Theta}} = \arg\max \ln(L) \quad (33)$$

Sampling from the optimized posterior distribution at $\mathbf{x}^*$ to calculate a prediction for regression is equivalent to determining the conditional distribution $P(f^*|\boldsymbol{y})$ where f^* is the Gaussian distribution describing the PDF of the target variable at input feature vector $\mathbf{x}^*$. The predicted value is conventionally interpreted as the mean of the Gaussian f^*, which can be analytically derived as:

$$\mu^* = \hat{\mu} + \mathbf{k}^{*T}\mathbf{K}_{\widehat{\boldsymbol{\Theta}}}^{-\mathbf{1}}(\boldsymbol{y} - \mathbf{1}\hat{\mu}) \quad (34)$$

Where $\mathbf{k}^*$ is a vector containing all pairs $K_{\boldsymbol{\Theta}}(\mathbf{x}^*, \mathbf{x}_i)$ for all *n* training observations $\mathbf{x}_i$, and $\hat{\mu}$, $\mathbf{K}_{\widehat{\boldsymbol{\Theta}}}^{-\mathbf{1}}$ are the same optimized parameters as Eq. 33.

8.7.3 Initial investigations with pose-standardized training and test meshes

As stated in Chapter 8.5, prior work established [136] that artificially normalizing the pose of all 3DO mesh scans post-hoc using skinned kinematic deformations improved the accuracy of body composition estimation by roughly 10% on visceral fat and up to 30% in total fat mass. Intuitively, body composition does not vary with changing pose in the same individual, making even minor pose inconsistencies in the 3D shape data detrimental to the signal-to-noise ratio of the input. Wong et al. [136] outsourced pose normalization to Meshcapade GmbH., which used a proprietary algorithm based on the SMPL model to rig, skin, and repose raw 3DO scans into a standardized pose, in this case a T-pose with arms in full extension parallel to the floor and legs in full extension perpendicular to the floor. We tested an in-house solution to approximate the reposing procedure of Meshcapade using skin weights defined in SMPL-X [86] and tested the effects of pose normalization on our 3DAE body composition estimation procedure. This follow-up experiment enables rapid low-cost pose standardization of scanner-agnostic 3DO meshes.

Since all meshes described in Table 8.4 were transformed to the SMPL standardized topology containing 6890 vertices, we transferred the skeleton, joints, and per-vertex skinning weights of the SMPL template [67, 86] directly to every mesh in our ensemble dataset. SMPL skeleton and joint positions are defined as a function of their surrounding surface vertices; therefore, every mesh in our dataset implicitly defines a kinematic skeleton with predetermined skin weights optimized on a dynamic multi-pose dataset in Loper et al. [67] For every mesh in Table 8.4, we determined the initial joint coordinates as a function of the mesh vertices. We determined angle of the upper arm and leg segments (corresponding to the humerus and femur) of the kinematic skeleton defined by the initial joint coordinates relative to the spine of the

skeleton (defined as the segment between SMPL joint indices 0 and 9) and solved for the relative rotations $\boldsymbol{\theta}$ required at the shoulder (SMPL joints 16 and 17)) and hip joints (SMPL joints 1 and 2) to position the arms and legs at a consistently defined angle relative to the spine segment. We set the arms at 30 degrees abduction and the legs at 15 degrees abduction for every pose normalization and rotated about the Z-axis. A parallel and perpendicular T-pose did not work well for this application as large deformations from the initial pose created more artifacts and self-intersections at the shoulders and armpits.

Every vertex on the source mesh was transformed according to the skin weight $W_{k,i}$ corresponding to the weighted influence of the rotation of joint k on the ith mesh vertex using pose parameters $\boldsymbol{\theta}$. Part rotations were hierarchically structured as defined by the SMPL kinematic skeleton tree structure in section 3 of Loper et al. [67] This kinematic skeleton transformation was implemented with linear blend skinning (LBS) as part of the SMPL-X distribution. [86] This transformation normalizes every 3DO mesh to a consistent pose defined by the anatomy of the individual using the spine segment, allowing for robustness to deviations in posture or rigid global transformations. We omitted the dynamic pose-dependent shape deformation defined in SMPL in this initial trial. Examples of transformed meshes are shown in Fig. 8.9.

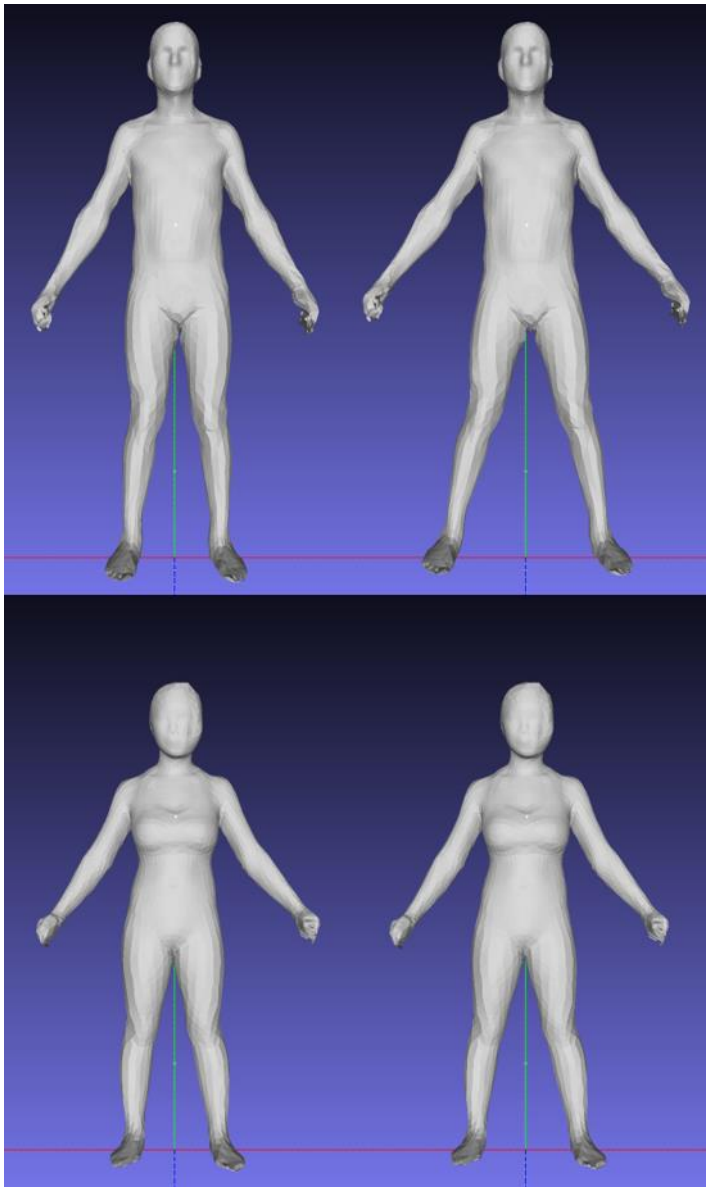


Figure 8.9. Examples of reposing on two individuals of different heights and sexes. The left column shows the original scanned pose; the right column shows the scan reposed with arms abducted 30 degrees relative to the SMPL kinematic skeleton spine and legs abducted 15 degrees relative to the spine.

We repeated the procedures detailed in Chapter 8.3 on the reposed mesh data and retrained a 3DAE network starting from the same pretrained DFAUST initialization using the same training data members and training parameters with the only difference being SMPL skinning constrained pose normalization on all input data. We trained GPR regressions to DXA

body composition using the latent encodings of the reposed training meshes and tested the performance of body composition prediction accuracy on the same test set as Chapter 8.4 after applying pose normalization to all test meshes.

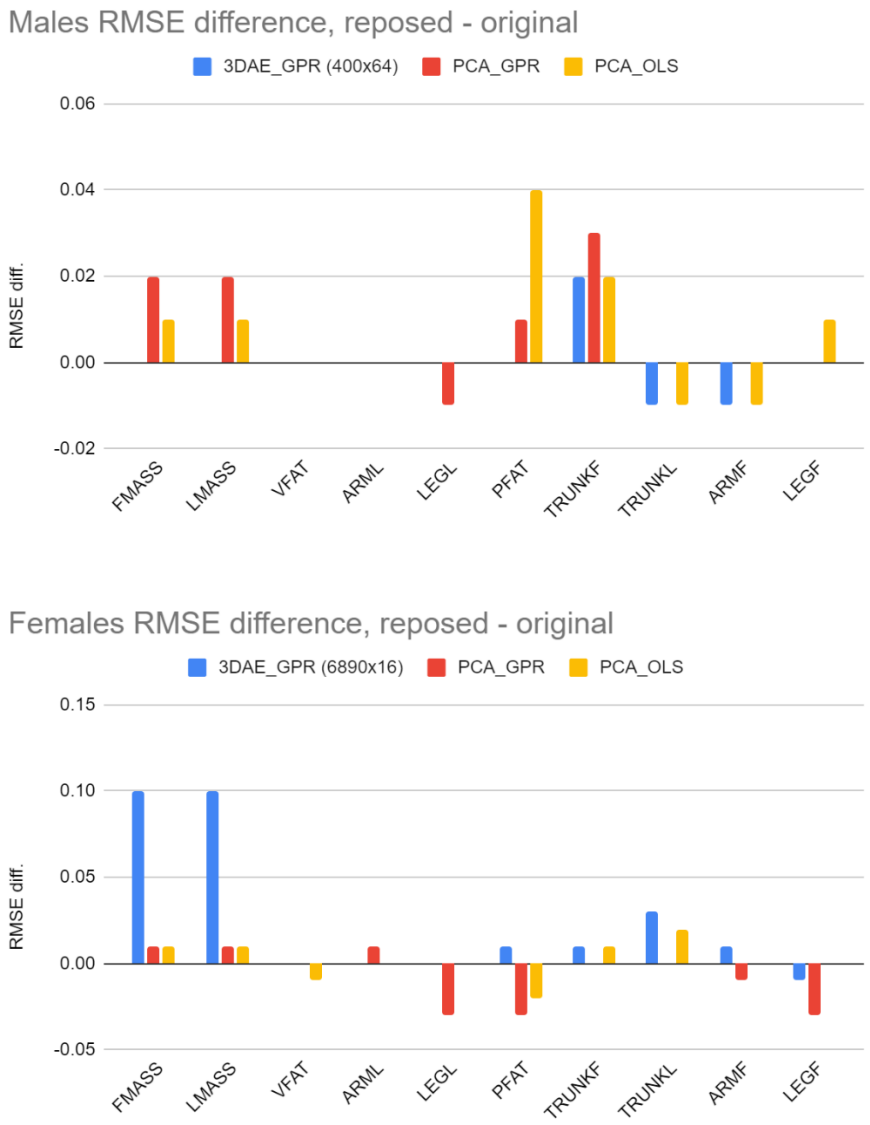


Figure 8.10. Difference in RMSE between body composition predictions trained and tested on a reposed dataset (30-degree arm abduction, 15-degree leg abduction) vs the original predictions on the raw scan data observed in Fig. 8.3. Reposing meshes to a standardized arm and leg angle did not create consistent reductions in body composition prediction RMSE across all metrics.

Fig. 8.10 shows the RMSE difference between predictions made with shape models and regressions trained on reposed (pose normalized) data vs. the original predictions from Chapter 8.4 made with the original scanned poses. Three comparative trials were conducted; one with the 3DAE-GPR model from Chapter 8.4 exhibiting the lowest RMSE on the original data, one with a linear PCA model paired with nonlinear GPR, and a fully linear PCA-OLS model comparable to the method used in Wong et al. [136] In all three trials for both males and females, the only variable changed between the reposed and the original prediction model was the pose normalization (or lack thereof) of all training and test data.

There was consistent difference between predictions on reposed and original mesh data, contradicting the findings in Wong et al. Wong et al. conducted experiments on a smaller set (n=540) of meshes scanned only on the Fit3D ProScanner, which fixed the endpoints of the hands and feet to handlebars and foot placement markers and did not include any augmentation data for pose variability in the same individual. This experiment suggests that given a dataset that augments for pose diversity within the same individual, unsupervised learning algorithms can learn to ignore the noise resulting from inconsistent posing when predicting body composition from 3DO scans, rendering pose normalization operations potentially unecessary.

9. Conclusion

In this thesis, I presented a comprehensive total and regional body composition analysis progression from optical image inputs. I presented multiple body composition prediction models for twelve target metrics that are used for evaluation of obesity and metabolic syndrome and demonstrated state-of-the-art precision and accuracy relative to reference DXA measurements. My work covers 2D photo and 3D scan input, adult, and pediatric cohorts, and linear vs. nonlinear algorithms for both shape modeling and body composition regression. The work in this thesis merges modern conventions and practices of training large machine learning and computer vision models with medical research standards to present clinically relevant results validated with statistical and computational rigor. This thesis described the following contributions:

1. A functional photometric stereo implementation that could reconstruct finely detailed geometry of human bodies in a scene of ~3 square meters. Although this project was not pursued past initial proof of concept, future analyses may be able to merge the high resolution of photometric stereo with the undistorted accuracy of time-of-flight and structured light scanners used in the subsequent works to generate more detailed 3D captures of human body shape.
2. A method for predicting 12 metrics of total and regional body composition from a single frontal 2D photo given simple additional parameters of height, weight, distance, and camera focal length. This method was designed to present the lowest cost, most accessible solution for in-the-wild measurement of body composition and at the time of publication was the only such method that operated on a monocular input and had

validation statistics on held-out test data. Prediction accuracies were comparable to prior work operating on the same dataset that had access to 3D PCA information as regression variables.

3. A method for automatically standardizing 3D optical scans from arbitrary scanning devices into manifold template meshes with topological consistency between all data set members. This method was a solution to the practical problem of manual data processing bottlenecks in the recruitment protocol. This work presented a unified body composition model derived from a composite dataset consisting of meshes scanned from three devices that was robust to slight variations in pose that were naturally occurring between scanners, which previous works had not addressed. The methods from this work were used to scale up the standardized 3DO datasets for additional follow up studies and could be employed in future works to support ongoing data collection protocols.
4. An automated pipeline for predicting body composition from 3DO scans of pediatric cohorts aged 5-17. Previous work only showed preliminary body composition analysis from 3DO scans of children using anthropometrics derived from 3DO meshes as regression features. This work built the first parameterized shape space and body composition prediction model from unsupervised body shape features for children in literature and demonstrated the encoding of visual markers of physical development and maturity as a function of age.
5. A systematic review of human body shape autoencoding using deep network architectures covering the major computer science and machine learning literature in the 20-year span between 2003 and 2023. Without a commonly accepted benchmark standard such as ImageNet or ShapeNet, work in 3D human shape encoding exhibits wide

variance in the standards for reporting accuracy both in the metrics used and the datasets tested. Many works omit quantitative metrics of performance altogether. This review attempted to introduce standardized conventions for judging the performance of human shape autoencoders and compiled the first known comprehensive comparison of all surveyed works with reported quantitative reconstruction errors. This review also attempted to assign structure to the current landscape of deep human body shape encoding by determining appropriate clusters for the surveyed works based on implementation and providing commentary on the observed performance differences between different model classes.

6. A comprehensive analysis and comparison of linear and nonlinear methods for shape modeling and body composition prediction from 3DO scans using the largest ensemble human shape dataset assembled for any work in clinical literature. This work was designed to thoroughly test the relative performance and marginal contributions of linear methods, represented by PCA and linear regression, and nonlinear methods, represented by a 3D graph autoencoder and GPR, under controlled experimental conditions using the exact same data for training and testing. This work challenged prior assumptions that models with large parameter counts would be overfit on test data prediction; it also validated the use of PCA as a shape encoder in prior work as the deep autoencoder did not outperform PCA in either geometric reconstruction and was only slightly more accurate in body composition prediction when GPR was used as the common regression model. This work identified GPR as the most accurate and precise method for body composition prediction and suggests future work on body composition prediction from body shape should start with GPR as the performance baseline and seek to improve its

performance through either kernel design or through the development of other nonlinear regression solutions such as MLPs and end-to-end deep regression networks from body geometry data.

9.1 Future Work

Many of the most salient directions for future work were identified in the preceding chapters; I will summarize the most important and potentially productive research directions left unanswered at the conclusion of this doctoral thesis.

9.1.1 Deep end-to-end regression networks

The works presented in this thesis as well as other contemporary work separate feature extraction from body shape and body composition regression from feature vectors into separate, disjointed tasks that are optimized independently of each other. I attempted end-to-end regression from optical image to body composition with both 2D input in Chapter 4.6.2 and with 3D mesh inputs in Chapter 8. An end-to-end deep network puts the loss function directly on the body composition prediction target and backpropagates derivatives all the way back to the initial input layer interfacing with the optical image data. Such a network may be able to optimize its deep features for body composition regression in a manner that exceeds the prediction accuracy of a separately trained regression model operating on features extracted by a shape autoencoder. Both the 2D and 3D attempts to construct this network suffered from the low quantity of training data. A useful intermediate task would be developing a deep multi-layer perceptron using an intermediate feature layer extracted with a trained autoencoder as input and tuning it to perform better than the current best results with GPR. Then, starting with the encoder network and the

separately optimized MLP as initialization weights, the two networks can be connected end to end and trained jointly.

9.1.2 Posable 3D models

The work presented in this thesis did not disentangle pose parameters from body shape. Wong et al. (2021) [136] established that even slight pose variation had negative impacts on body composition regression accuracy relative to a T-posed dataset where every mesh was reposed with a skinned model kinematic tree transformation in post processing. The work shown in Chapter 8 transformed all training and test meshes to the 6890 SMPL mesh topology. This mesh topology defines joint positions and skinning weights as a function of its vertex positions. The dataset created for the work in Chapter 8 is compatible with the animation framework defined by Loper et al. [67] in the original SMPL paper. A fully T-posed deep model could be developed from the dataset of Chapter 8 with no additional skinning or rigging effort required, assuming the registration of the vertex positions are in tight anatomical correspondence with the definitions in the SMPL model. A virtually unposed dataset would render the need for explicit pose disentanglement unnecessary as it can be assumed a priori that all pose parameters for training and test meshes are identical.

9.1.3 Interventional Studies

The work presented in this thesis only performed single-snapshot prediction of body composition prediction. Our models demonstrate good qualitative ability to visualize change in an individual in response to manipulating a target body composition variable despite having no follow ups of longitudinal change of the same participant in the training data. General trends of

shape variation in response to body composition change were learned through observation of aggregate patterns of shape variation between different individuals. Follow up studies such as that of Wong et al. (2022) [134] could be used to specifically fine tune the presented models to learn how body shape changes in an individual in response to controlled interventions such as restrictive diet or progressively overloaded physical training.

9.2 Assessment of Impact

The work presented in this thesis advanced the state of the art in body composition prediction from optical imaging of human body shape with three key impacts:

1. We presented novel methods and tools for processing optical image data inputs that increased the accessibility, utility, and practicality of body composition prediction from both 2D and 3D data sources.
2. We demonstrated increased performance metrics on body composition prediction accuracy and precision that improved upon state-of-the-art error benchmarks from prior works.
3. We introduced novel applications of deep 3D shape models and nonlinear regression to the task of body composition prediction and demonstrated its superior accuracy and precision relative to all past works in this subject, thus establishing this thesis as the state-of-the-art in body composition prediction from 3DO scans at the time of writing.

Prior to the introduction of the work in this dissertation, body composition prediction from optical images operated on 3D circumferences or 2D pixel distances measured on surface scans and image masks respectively as input shape features for regression. With few exceptions, body composition regression models from those predetermined features were almost exclusively restricted to variants of linear regression. Explorations of shape feature extraction using unsupervised learning algorithms exclusively relied on PCA, a linear autoencoder method. Experiments were conducted on small datasets restricted to a single scanning device due to the high labor overhead of manually targeting a manifold mesh template onto raw 3DO scan data. We have advanced the state-of-the-art in the field of body composition prediction from optical imaging data such that estimates can be performed from both 2D and 3D optical data with minimally restrictive, end-user friendly, device agnostic capture requirements. Our models were trained on the largest set of combined optical image and body composition data available to date enabled by the automated mesh processing tools presented in this thesis. We assert that the contributions of this dissertation advance the viability and usability of performing clinical analysis of body composition from 2D and 3D optical images of human body shape.

10. References

1. Achamrah N, Colange G, Delay J. Comparison of body composition assessment by DXA and BIA according to the body mass index: A retrospective study on 3655 measures. PLOS ONE. 2018;13(7). doi:10.1371/journal.pone.0200465

2. Affuso O, Pradhan L, Zhang C. A method for measuring human body composition using digital images. PLOS ONE. 2018;13(11). doi:10.1371/journal.pone.0206430

3. Agarwal S, Mierle K. Ceres Solver. http://ceres-solver.org.

4. Alldieck T, Xu H, Sminchisescu C. ImGHUM: Implicit generative models of 3D human shape and articulated pose. In: 2021 IEEE/CVF International Conference on Computer Vision ICCV; 2021.

5. Allen B, Curless B, Popović Z. The space of human body shapes. ACM Transactions on Graphics. 2003;22:587-594. doi:10.1145/882262.882311.

6. Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J. SCAPE. *ACM Transactions on Graphics*. 2005;24(3):408-416.

7. Au PC, Li HL, Lee GK. Sarcopenia and mortality in cancer: A meta-analysis. Osteoporos Sarcopenia. 2021;7(Suppl 1).

8. Bartol K, Bojanic D, Petkovic T, Pribanic T. A review of body measurement using 3D scanning. *IEEE Access*. 2021;9:67281-67301.

9. Ben-Hamu H, Maron H, Kezurer I, Avineri G, Lipman Y. Multi-chart generative surface modeling. *ACM Transactions on Graphics*. 2019;37(6):1-15. doi:10.1145/3272127.327505213.

10. Bigaard J, Frederiksen K, Tjønneland A. Waist circumference and body composition in relation to all-cause mortality in middle-aged men and women. International Journal of Obesity. 2005;29(7):778-784. doi:10.1038/sj.ijo.0802976

11. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. Computer Vision – ECCV 2016 Lecture Notes in Computer Science. 2016:561-578. doi:10.1007/978-3-319-46454-1_34

12. Bogo F, Romero J, Loper M, Black MJ. FAUST: Dataset and evaluation for 3D Mesh registration. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*; 2014.

13. Bogo F, Romero J, Pons-Moll G, Black MJ. Dynamic FAUST: Registering human bodies in Motion. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR*. ; 2017. doi:10.1109/cvpr.2017.591.

14. Bonnick SL. Monitoring changes in bone density. Women's Health. 2008;4(1):89-97. doi:10.2217/17455057.4.1.89

15. Bouritsas G, Bokhnyak S, Ploumpis S, Zafeiriou S, Bronstein M. Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV. ; 2019. doi:10.1109/iccv.2019.00731

16. Cahoon GD, Davison TE. Prediction of compliance with MRI procedures among children of ages 3 years to 12 years. Pediatric Radiology. 2014;44:1302-1309. doi:10.1007/s00247-014-2996-y.

17. Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nature Reviews Cancer. 2004;4(8):579-591. doi:10.1038/nrc1408

18. Chang AX, Funkhouser T, Guibas L, et al. ShapeNet: An information-rich 3D model repository. https://arxiv.org/abs/1512.03012.

19. Chen Z, Kim TK. Learning feature aggregation for Deep 3D morphable models. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*). 2021

20. Cheng ZQ, Chen Y, Martin RR, Wu T, Song Z. Parametric modeling of 3D human body shape—a survey. *Computers & Graphics*. 2018;71:88-100.

21. Chhatkuli S, Jiang I, Kamiyama K. Body composition estimation based on multimodal multi-task Deep Neural Network. https://arxiv.org/abs/2205.11031.

22. Chrysos GG, Moschoglou S, Bouritsas G, Deng J, Panagakis Y, Zafeiriou S. Deep Polynomial Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(8):4021-4034,. doi:10.1109/TPAMI.2021.3058891.

23. Dai H, Shao L. Pointae: Point Auto-encoder for 3D statistical shape and texture modelling. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV*). 2019.

24. Deng B, Lewis JP, Jeruzalski T, et al. NASA neural articulated shape approximation. *Computer Vision – ECCV*. Published online 2020:612-628.

25. Dulloo AG, Jacquet J, Solinas G, Montani JP, Schutz Y. Body composition phenotypes in pathways to obesity and the metabolic syndrome. International Journal of Obesity. 2010;34. doi:10.1038/ijo.2010.234.

26. Eckel RH, Kahn SE, Ferrannini E. Obesity and Type 2 Diabetes: What Can Be Unified and What Needs to Be Individualized? The Journal of Clinical Endocrinology & Metabolism. 2011;96(6):1654-1663. doi:10.1210/jc.2011-0585

27. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004;32(2). doi:10.1214/009053604000000067

28. Fan Z, Chiong R, Hu Z, Keivanian F, Chiong F. Body fat prediction through feature extraction based on anthropometric and laboratory measurements. PLOS ONE. 2022;17(2). doi:10.1371/journal.pone.0263333

29. Farina G, Spataro F, Lorenzo AD, Lukaski H. A Smartphone Application for Personal Assessments of Body Composition and Phenotyping. Sensors. 2016;16(12). doi:10.3390/s16122163

30. Fosbøl MØ, Zerahn B. Contemporary methods of body composition measurement. Clinical Physiology and Functional Imaging. 2014;35:81-97. doi:10.1111/cpf.12152.

31. Foti S, Koo B, Stoyanov D, Clarkson MJ. 3D shape variational autoencoder latent disentanglement via mini-batch feature swapping for bodies and faces. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

32. Freidenberg G, Reichart D, Olefsky J, Henry R. Reversibility of defective adipocyte insulin receptor kinase activity in non-insulin-dependent diabetes mellitus. Effect of weight loss Journal of Clinical Investigation. 1988;82(4).

33. Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119-139.

34. Furukawa Y, Ponce J. Accurate, dense, and robust multi-view stereopsis. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007.

35. Galindo Martín C, Aportela Vázquez V, Becerril Hernández F, et al. The GLIM criteria for adult malnutrition and its relation with adverse outcomes, a prospective observational study. Clin Nutr ESPEN. 2020;38:67-73. doi:10.1016/j.clnesp.2020.06.015.

36. Gao L, Lai YK, Yang J, Zhang LX, Xia S, Kobbelt L. Sparse data driven mesh deformation. IEEE Transactions on Visualization and Computer Graphics. 2021;27(3):2085-2100. doi:10.1109/tvcg.2019.2941200

37. Gardner JR, Pleiss G, Bindel D, Weinberger KQ, Wilson AG. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. Neural Information Processing Systems. 2018;31:7576-7586.

38. Genova K, Cole F, Sud A, Sarna A, Funkhouser T. Local deep implicit functions for 3D shape. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

39. Glüer CC, Blake G, Lu Y, Blunt BA, Jergas M, Genant HK. Accurate assessment of precision errors: How to measure the reproducibility of bone densitometry techniques. Osteoporosis International. 1995;5(4):262-270. doi:10.1007/bf01774016

40. Gong S, Chen L, Bronstein M, Zafeiriou S. SpiralNet++: A fast and highly efficient mesh convolution operator. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019

41. Goodpaster BH, Krishnaswami S, Harris TB, Katsiaras A, Kritchevsky SB, Simonsick EM. Obesity, regional body fat distribution, and the metabolic syndrome in older men and women. Archives of Internal Medicine. 2005;165(777). doi:10.1001/archinte.165.7.777.

42. Grigorev A, Iskakov K, Ianina A, et al. StylePeople: A generative model of fullbody human avatars. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

43. Guan P, Weiss A, Balan AO, Black MJ. Estimating human shape and pose from a single image. In: 2009 IEEE 12th International Conference on Computer Vision. ICCV; 2009. doi:10.1109/iccv.2009.5459300

44. Guo S. Implement A Gaussian Process From Scratch. Medium Published January. 2021;26. https://towardsdatascience.com/implement-a-gaussian-process-from-scratch-2a074a470bce

45. Hahner S, Garcke J. Mesh convolutional Autoencoder for semi-regular meshes of different sizes. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022. doi:10.1109/wacv51458.2022.00240.

46. Hanocka R, Hertz A, Fish N, Giryes R, Fleishman S, Cohen-Or D. Meshcnn. ACM Transactions on Graphics. 2019;38(4):1-12. doi:10.1145/3306346.3322959

47. Hasler N, Stoll C, Sunkel M, Rosenhahn B, Seidel HP. A statistical model of human pose and body shape. *Computer Graphics Forum*. 2009;28(2):337-346.

48. He W, Jiang Z, Zhang C, Sainju AM. CurvaNet: Geometric Deep Learning based on Directional Curvature for 3D Shape Analysis. In: *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. doi:10.1145/3394486.3403272

49. Hu S, Shum HP, Mucherino A. DSPP: Deep Shape and pose priors of humans. *Motion, Interaction and Games*. Published online 2019.

50. Hu SM, Liu ZN, Guo MH. Subdivision-Based Mesh Convolution Networks. *ACM Transactions on Graphics*. 2022;41(3):1-16. doi:10.1145/3506694

51. Hui WS, Liu Z, Ho SC. Metabolic syndrome and all-cause mortality: A meta-analysis of prospective cohort studies. European Journal of Epidemiology. 2010;25:375-384. doi:10.1007/s10654-010-9459-z.

52. Imboden MT, Swartz AM, Finch HW, Harber MP, Kaminsky LA. Reference standards for lean mass measures using GE dual energy x-ray absorptiometry in Caucasian adults. Plos One. 2017;12(4). doi:10.1371/journal.pone.0176161

53. Ingole AL, Karande KJ. Automatic age estimation from face images using facial features. In: 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN. ; 2018. doi:10.1109/gcwcn.2018.8668629

54. Jiang B, Zhang J, Cai J, Zheng J. Disentangled human body embedding based on deep hierarchical neural network. IEEE Transactions on Visualization and Computer Graphics. 2020;26(8):2560-2575. doi:10.1109/tvcg.2020.2988476

55. Kaashki NN, Hu P, Munteanu A. Deep learning-based automated extraction of anthropometric measurements from a single 3-D scan. *IEEE Transactions on Instrumentation and Measurement*. 2021;70:1-14.

56. Klarqvist MDR, Agrawal S, Diamant N. Silhouette images enable estimation of body fat distribution and associated cardiometabolic risk. npj Digital Medicine. 2022;5(1). doi:10.1038/s41746-022-00654-1

57. Kootaka Y, Kamiya K, Hamazaki N. The GLIM criteria for defining malnutrition can predict physical function and prognosis in patients with cardiovascular disease. Clin Nutr. 2021;40(1):146-152.

58. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional Neural Networks. *Communications of the ACM*. 2017;60(6):84-90.

59. Kuk JL, Katzmarzyk PT, Nichaman MZ, Church TS, Blair SN, Ross R. Visceral Fat Is an Independent Predictor of All-cause Mortality in Men*. Obesity. 2006;14(2):336-341. doi:10.1038/oby.2006.43

60. Lahav A, Tal A. Meshwalker. *Meshwalker ACM Transactions on Graphics*. 2020;39(6):1-13.

61. Lee H, Chung HS, Kim YJ, et al. Association between body shape index and risk of mortality in the United States. *Scientific Reports*. 2022;12(1).

62. Lemeunier C, Denis F, Lavoué G, Dupont F. Representation learning of 3D meshes using an Autoencoder in the spectral domain. Computers & Graphics. 2022;107:131-143.

63. Li J, Xu K, Chaudhuri S, Yumer E, Zhang H, Guibas LJ. GRASS. *ACM Transactions on Graphics*. 2017;36(4):1-14. doi:10.1145/3072959.3073637

64. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of Convolutional Neural Networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*. 2022;33(12):6999-7019.

65. Liang Y, Zhao S, Yu B, Zhang J, He F. MeshMAE: Masked Autoencoders for 3D mesh data analysis. *Lecture Notes in Computer Science*. 2022:37-54. doi:10.1007/978-3-031-20062-5_3.

66. Lombardi S, Yang B, Fan T, et al. Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In: *2021 International Conference on 3D Vision (3DV)*. 2021.

67. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: A Skinned Multi-Person Linear Model. ACM Transactions on Graphics. 2015;34(6):1-16. doi:10.1145/2816795.2818013

68. Lu Y, Hahn JK, Zhang X. 3D shape-based body composition inference model using a Bayesian network. IEEE Journal of Biomedical and Health Informatics. 2020;24(1):205-213. doi:10.1109/jbhi.2019.2903190

69. Lu Y, Mathur AK, Blunt BA. Dual X-ray absorptiometry quality control: Comparison of visual examination and process-control charts. Journal of Bone and Mineral Research. 2009;11(5):626-637. doi:10.1002/jbmr.5650110510.

70. Lu Y, McQuade S, Hahn JK. 3D shape-based body composition prediction model using machine learning. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018. doi:10.1109/embc.2018.8513261

71. Lu Y, Zhao S, Younes N, Hahn JK. Accurate nonrigid 3D human body surface reconstruction using commodity depth sensors. Computer Animation and Virtual Worlds. 2018;29(5). doi:10.1002/cav.1807

72. Macek P, Biskup M, Terek-Derszniak M, et al. Optimal body fat percentage cut-off values in predicting the obesity-related cardiovascular risk factors: A cross-sectional cohort study. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy. 2020;13:1587-1597. doi:10.2147/dmso.s248444

73. Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black M. AMASS: Archive of motion capture as surface shapes. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.2019.

74. Marco E, Sanchez-Rodriguez D, Davalos-Yerovi VN. Malnutrition according to ESPEN consensus predicts hospitalizations and long-term mortality in rehabilitation patients with stable chronic obstructive pulmonary disease. Clin Nutr. 2019;38(5):2180-2186.

75. Maron H, Galun M, Aigerman N. Convolutional neural networks on surfaces via seamless toric covers. *ACM Transactions on Graphics*. 2017;36(4):1-10. doi:10.1145/3072959.3073616

76. Mihajlovic M, Zhang Y, Black MJ, Tang S. LEAP: Learning articulated occupancy of people. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

77. Mramba L, Ngari M, Mwangome M. A growth reference for mid upper arm circumference for age among school age children and adolescents, and validation for mortality: growth curve construction and longitudinal cohort study. Bmj. Published online March 2017. doi:10.1136/bmj.j3423

78. Muhammad ZUD, Huang Z, Khan R. A review of 3D human body pose estimation and mesh recovery. *Digital Signal Processing*. 2022;128:103628.

79. Muja M, Lowe D. Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP International Conference on Computer Vision Theory and Applications, ICCV; 2009.

80. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society. 2004;18(6):275-285.

81. Ng BK, Hinton BJ, Fan B, Kanaya AM, Shepherd JA. Clinical anthropometrics and body composition from 3D whole-body surface scans. European Journal of Clinical Nutrition. 2016;70(11):1265-1270. doi:10.1038/ejcn.2016.109

82. Ng BK, Liu YE, Wang W, Kelly TL, Wilson KE, Schoeller DA. Validation of rapid 4-component body composition assessment with the use of dual-energy X-ray absorptiometry and bioelectrical impedance analysis. The American Journal of Clinical Nutrition. 2018;108:708-715. doi:10.1093/ajcn/nqy158.

83. Ng BK, Sommer MJ, Wong MC, Pagano I, Nie Y, Fan B. Detailed 3-dimensional body shape features predict body composition, blood metabolites, and functional strength: The shape up! studies. The American Journal of Clinical Nutrition. 2019;110:1316-1326. doi:10.1093/ajcn/nqz218

84. Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. DEEPSDF: Learning continuous signed distance functions for shape representation. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

85. Pathak P, Panday SB, Ahn J. Artificial Neural Network model effectively estimates muscle and fat mass using simple demographic and anthropometric measures. Clinical Nutrition. 2022;41(1):144-152. doi:10.1016/j.clnu.2021.11.027

86. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AA, Tzionas D, et al. Expressive body capture: 3D hands, face, and body from a single image. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019. doi:10.1109/cvpr.2019.01123.

87. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825-2830.

88. Phong BT. Illumination for computer generated pictures. Communications of the ACM. 1975;18(6):311-317. doi:10.1145/360825.360839

89. Pishchulin L, Insafutdinov E, Tang S. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR. ; 2016. doi:10.1109/cvpr.2016.533

90. Pishchulin L, Wuhrer S, Helten T, Theobalt C, Schiele B. Building statistical shape spaces for 3d human modeling. Pattern Recognition. 2017;67:276-286. doi:10.1016/j.patcog.2017.02.018

91. Pons-Moll G, Romero J, Mahmood N, Black MJ. Dyna. *ACM Transactions on Graphics*. 2015;34(4):1-14.

92. Poulenard A, Ovsjanikov M. Multi-directional geodesic neural networks via equivariant convolution. *ACM Transactions on Graphics*. 2018;37:1-14. doi:10.1145/3272127.3275102.

93. Price GM, Uauy R, Breeze E, Bulpitt CJ, Fletcher AE. Weight, shape, and mortality risk in older persons: elevated waist-hip ratio, not high body mass index, is associated with a greater risk of death. The American Journal of Clinical Nutrition. 2006;84(2):449-460. doi:10.1093/ajcn/84.2.449

94. Qatanani M, Lazar MA. Mechanisms of obesity-associated insulin resistance: many choices on the menu. Genes & development. 2007;21(12):1443-1455.

95. Qi Y. Random forest for bioinformatics. In: Ensemble machine learning. Springer; 2012:307-323.

96. Ranjan A, Bolkart T, Sanyal S, Black MJ. Generating 3D faces using convolutional mesh autoencoders. *Computer Vision – ECCV*. Published online 2018:725-741.

97. Robinette KM, Blackwell S, Daanen H, Boehmer M, Fleming S. Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report. Volume 1. Summary. Defense Technical Information Center. 2002;1. doi:10.21236/ada406704.

98. Rogers SL, Broadbent R, Brown J, Fraser A, Speelman C. Realistic motion avatars are the future for social interaction in virtual reality. Published online 2021.

99. Romero J, Tzionas D, Black MJ. Embodied hands. *ACM Transactions on Graphics*. 2017;36(6):1-17.

100. Rondel A, Langius JAE, Schueren MAE, Kruizenga HM. The new ESPEN diagnostic criteria for malnutrition predict overall survival in hospitalised patients. Clin Nutr. 2018;37(1):163-168.

101. Rother C, Kolmogorov V, Blake A. Grabcut": Interactive foreground extraction using iterated graph cuts. ACM Trans Graph. 2004;23(3):309-314. doi:10.1145/1186562.1015720

102. Sanchez-Rodriguez D, Locquet M, Bruyere O. Prediction of 5-year mortality risk by malnutrition according to the GLIM format using seven pragmatic approaches to define the criterion of loss of muscle mass. Clin Nutr. 2021;40(4):2188-2199.

103. Santos LP, Ong KK, Day F. Body shape and size in 6-year old children: Assessment by three-dimensional photonic scanning. International Journal of Obesity. 2016;40(6):1012-1017. doi:10.1038/ijo.2016.30

104. Schneider V, Oganov V, LeBlanc A, et al. Space Flight Bone Loss and Change In Fat and Lean Body Mass. Journal of Bone and Mineral Research. 1992;7(Supp. 1).

105. Scholz M, Vigário R. Nonlinear PCA: a new hierarchical approach. InEsann 2002 Apr. 24:439-444.

106. Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. Journal of Mathematical Psychology. 2018;85:1-16. doi:10.1016/j.jmp.2018.03.001

107. Sharp N, Attaiki S, Crane K, Ovsjanikov M. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics*. 2022;41:1-16. doi:10.1145/3507905.

108. Shepherd JA, Lu Y. A generalized least significant change for individuals measured on different DXA Systems. Journal of Clinical Densitometry. 2007;10(3):249-258. doi:10.1016/j.jocd.2007.05.002

109. Shepherd JA, Wang L, Fan B. Optimal monitoring time interval between DXA measures in children. Journal of Bone and Mineral Research. 2011;26(11):2745-2752. doi:10.1002/jbmr.473

110. Simmonds M, Llewellyn A, Owen CG, Woolacott N. Predicting adult obesity from childhood obesity: A systematic review and meta-analysis. Obesity Reviews. 2015;17(2):95-107. doi:10.1111/obr.12334

111. Simoni P, Guglielmi R, Gómez MP. Imaging of body composition in children. Quantitative Imaging in Medicine and Surgery. 2020;10(8):1661-1671. doi:10.21037/qims.2020.04.06

112. Sims EA, E D Jr, Horton ES, Bray GA, Glennon JA, Salans LB. Endocrine and metabolic effects of experimental obesity in man. Recent Prog Horm Res. 1973;29:457-496. doi:10.1016/b978-0-12-571129-6.50016-6.

113. Smirnov D, Solomon J. Hodgenet. *ACM Transactions on Graphics*. 2021;40(4):1-11.

114. Su FG, Lin CS, Wang YCF. Learning interpretable representation for 3D point clouds. In: *25th International Conference on Pattern Recognition (ICPR)*. 2020.

115. Sun R, Wang G, Zhang W, Hsu LT, Ochieng WY. A gradient boosting decision tree based GPS signal reception classification algorithm. Applied Soft Computing. 2020;86:105942.

116. Tan Q, Zhang LX, Yang J, Lai YK, Gao L. Variational Autoencoders for Localized Mesh Deformation Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Published online 2021:1-1.

117. Tian IY, Ng BK, Wong MC, Kennedy S, Hwaung P, Kelly N. Predicting 3D body shape and body composition from conventional 2D photography. Medical Physics. 2020;47:6232-6245. doi:10.1002/mp.14492.

118. Tian IY, Wong MC, Kennedy S, Kelly NN, Liu YE, Garber AK, et al. A device-agnostic shape model for automated body composition estimates from 3D optical scans. Medical Physics 2022;49:6395–409. doi:10.1002/mp.15843.

119. Tian IY, Wong MC, Nguyen WM, Kennedy S, McCarthy C, Kelly NN, et al. Automated body composition estimation from device-agnostic 3D optical scans in pediatric populations. Clinical Nutrition 2023;42:1619–30. doi:10.1016/j.clnu.2023.07.012.

120. Tinsley GM, Moore ML, Benavides ML, Dellinger A JR, B.T. 3-dimensional optical scanning for body composition assessment: A 4-component model comparison of four commercially available scanners. Clinical Nutrition. 2020;39(10):3160-3167. doi:10.1016/j.clnu.2020.02.008

121. Tretschk E, Tewari A, Zollhöfer M, Golyanik V, Theobalt C. DEMEA: Deep mesh autoencoders for non-rigidly deforming objects. *Computer Vision – ECCV*. Published online 2020:601-617.

122. Varol G, Romero J, Martin X, et al. Learning from synthetic humans. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

123. Vollmer J, Mencl R, Muller H. Improved laplacian smoothing of noisy surface meshes. Computer Graphics Forum. 1999;18(3):131-138. doi:10.1111/1467-8659.00334

124. Wang PS, Liu Y, Tong X. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics*. 2022;41(4):1-15.

125. Wang Q, Lu Y, Zhang X, Hahn JK. A novel hybrid model for visceral adipose tissue prediction using shape descriptors. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2019. doi:10.1109/embc.2019.8857092.

126. Wang Q, Xue W, Zhang X, Jin F, Hahn J. Pixel-wise body composition prediction with a multi-task conditional generative adversarial network. Journal of Biomedical Informatics. 2021;120(103866). doi:10.1016/j.jbi.2021.103866

127. Wang Z, Wang J, Zeng X, et al. Prediction of garment fit level in 3D virtual environment based on Artificial Neural Networks. *Textile Research Journal*. 2021;91(15-16):1713-1731.

128. Ward LC. Bioelectrical impedance analysis for Body Composition Assessment: Reflections on accuracy, Clinical Utility, and standardisation. European Journal of Clinical Nutrition. 2018;73(2):194-199. doi:10.1038/s41430-018-0335-3

129. Weber DR, Leonard MB, Zemel BS. Body Composition Analysis in the Pediatric Population. Pediatr Endocrinol Rev. 2012;10(1):130-139.

130. Wells JC, Stocks J, Bonner R. Acceptability, precision and accuracy of 3D photonic scanning for measurement of body shape in a multi-ethnic sample of children aged 5-11 years: The Slic Study. PLOS ONE. 2015;10(4). doi:10.1371/journal.pone.0124193

131. Wenninger S, Achenbach J, Bartl A, Latoschik ME, Botsch M. Realistic virtual humans from smartphone videos. In: 26th ACM Symposium on Virtual Reality Software and Technology 2020. doi:10.1145/3385956.3418940.

132. Wiersma R, Eisemann E, Hildebrandt K. CNNs on surfaces using rotation-equivariant features. *ACM Transactions on Graphics*. 2020;39. doi:10.1145/3386569.3392437.

133. Wilson JP, Kanaya AM, Fan B, Shepherd JA. Ratio of trunk to leg volume as a new body Shape metric for diabetes and mortality. PLoS ONE. 2013;8(7). doi:10.1371/journal.pone.0068716

134. Wong MC, Bennett JP, Leong LT, Tian IY, Liu YE, Kelly NN, et al. Monitoring body composition change for intervention studies with advancing 3D optical imaging technology in comparison to dual-energy X-ray absorptiometry. The American Journal of Clinical Nutrition,Volume. 117. doi:10.1016/j.ajcnut.2023.02.006.

135. Wong MC, Ng BK, Kennedy SF, Hwaung P, Liu EY, Kelly NN, et al. Children and adolescents' anthropometrics body composition from 3-D optical surface scans. Obesity. 2019;27:1738-1749. doi:10.1002/oby.22637.

136. Wong MC, Ng BK, Tian I, et al. A pose-independent method for accurate and precise body composition from 3D optical scans. Obesity. 2021;29:1835-1847. doi:10.1002/oby.23256.

137. Woodham RJ. Photometric Method For Determining Surface Orientation From Multiple Images. Optical Engineering. 1980;19(1). doi:10.1117/12.7972479

138. Wu Y, Wang Y, Jung S, Hoermann S, Lindeman RW. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and Attraction. *Frontiers in Virtual Reality*. 2021;2.

139. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*. 2021;32(1):4-24.

140. Xanthopoulos P, Pardalos PM, Trafalis TB. Linear discriminant analysis. In: Robust data mining. Springer; 2013:27-33.

141. Xu H, Bazavan EG, Zanfir A, Freeman WT, Sukthankar R, Sminchisescu C. GHUM & GHUML: Generative 3D human shape and articulated pose models. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

142. Yang J, Gao L, Tan Q, Huang YH, Xia S, Lai YK. Multiscale mesh deformation component analysis with attention-based autoencoders. *IEEE Transactions on Visualization and Computer Graphics*. 2023;29(2):1301-1317.

143. Yuan YJ, Lai YK, Yang J, Duan Q, Fu H, Gao L. Mesh variational autoencoders with edge contraction pooling. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020.

144. Zhang C, Rexrode KM, Dam RM, Li TY, Hu FB. Abdominal obesity and the risk of all-cause, cardiovascular, and cancer mortality. Circulation. 2008;117(13):1658-1667. doi:10.1161/circulationaha.107.739714

145. Zhang Y, Hassan M, Neumann H, Black MJ, Tang S. Generating 3D people in scenes without people. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

146. Zhou Y, Wu C, Li Z, et al. Fully Convolutional Mesh Autoencoder using Efficient Spatially Varying Kernels. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20); 2020:9251-9262.

147. Demographics of Mobile Device Ownership and Adoption in the United States. https://www.pewresearch.org/internet/fact-sheet/mobile/.