npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01469-6

3D convolutional deep learning for nonlinear estimation of body composition from whole body morphology

Check for updates

Isaac Y. Tian[1], Jason Liu[1], Michael C. Wong[2], Nisa N. Kelly[2], Yong E. Liu[2], Andrea K. Garber[3], Steven B. Heymsfield[4], Brian Curless[1] & John A. Shepherd[2]

Body composition prediction from 3D optical imagery has previously been studied with linear algorithms. In this study, we present a novel application of deep 3D convolutional graph networks and nonlinear Gaussian process regression for human body shape parameterization and body composition estimation. We trained and tested linear and nonlinear models with ablation studies on a novel ensemble body shape dataset containing 4286 scans. Nonlinear GPR produced up to a 20% reduction in prediction error and up to a 30% increase in precision over linear regression for both sexes in 10 tested body composition variables. Deep shape features produced 6–8% reduction in prediction error over linear PCA features for males only, and a 4–14% reduction in precision error for both sexes. All coefficients of determination (R^2) for all predicted variables were above 0.86 and achieved lower estimation RMSEs than all previous work on 10 metrics of body composition.

Total and regional body composition are correlated with many of the leading causes of death in the US and around the world[1–4]. High visceral fat deposits doubled the risk for metabolic syndrome in males with otherwise normal BMIs[4] and increased mortality rates from cancer by up to 63%[5]. 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[1,6,7]. Management of these conditions is also heavily impacted by body composition, specifically low lean mass, which predicts poor treatment outcomes[8,9] and up to 17-fold increased mortality[10,11]. However, body composition assessment historically required more advanced instruments such as dual X-ray absorptiometry (DXA) or air displacement plethysmography (ADP)[12]. 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 measurements comparable to DXA without utilizing ionizing radiation. The system should be relatively inexpensive and easy 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[13,14]. 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. 3DO scanning systems cost on the order of $10,000 and are programmed by the manufacturer to operate automatically without the need for 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[13],[15]. 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

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]University of Hawaii Cancer Center, University of Hawaii - Manoa, Honolulu, HI, USA. [3]UCSF School of Medicine, University of California—San Francisco, San Francisco, CA, USA. [4]Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA, USA. e-mail: iytian@cs.washington.edu

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 to reconstruct the original 3D mesh input through a series of graph convolutional layers. These layers are analogous to image convolution filters in 2D image networks and extract spatially localized features on a patch of the 3D body surface. A subset of these scans was captured with paired DXA scans for ground truth body composition training targets. DXA scans were captured as whole-body images in the coronal plane and contained gold-standard measurements of both total and regional body composition. 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 observed the effect of manipulating depths and dimensions of convolutional features on prediction accuracy and compared nonlinear estimates of 3D reconstruction error and body composition prediction precision and accuracy to results produced by linear methods from prior works. To our knowledge, this is the first application of a 3D graph convolutional autoencoder to body composition prediction.

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 held-out test data than previous linear methods.

Results

The Shape Up! Adults study population has been previously described[14] and summarized in Supplementary Table 1 and 2. Only this subset of the total data was used for body composition regression training and testing.

Geometric reconstruction error for both 3DAE and PCA shape models of four increasing sizes are shown in Table 1. The dimensionality d represents either the number of PCA coefficients used to parameterize shape (linear model) or the number of latent variables in the bottleneck layer of the 3DAE (nonlinear model) connecting the encoder module to the symmetric decoder. Reconstruction error was calculated as the geometric mean absolute error (MAE) between original and reconstructed vertex 3D positions. As expected, larger models 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 leveled off at just above 2 mm MAE.

Figure 1 depicts the body composition prediction results from $d = 4284$ sized models using different permutations of linear and nonlinear shape feature extraction and regression methods. Excluding the baseline column, each subsequent column represents a change of exactly one model parameter holding all others constant; i.e., OLS to GPR, PCA to 3DAE, 4824 total parameters to 400 × 64 parameters. All model permutations were trained and tested on the exact same mesh data. The smaller $d = 301$ model performed the same or worse on all error metrics and was not plotted for brevity. This result on 3DAE and GPR was consistent our observations in prior work on linear models, which showed that large parameter counts in regression models for body composition prediction did not produce observable overfitting artifacts on held-out test data.

Table 1 | 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)

	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

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.

The baseline GPR model accuracy is the root-mean-squared-error (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 refs. 14,16. This RMSE was higher than any subsequent regression model using shape features as input. The reduction of prediction error when shape information is introduced demonstrated that 3DO shape is a useful signal for body composition prediction even when nonlinear regression algorithms are employed.

PCA-OLS (PCA shape model, ordinary least squares regression model) is equivalent to the linear 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 the 3DAE.

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 relative to linear regression except arm lean on females (which was equal) and leg lean on females (which was higher by 0.02 kg, or 5%). These results indicated that GPR was a more accurate regression method than OLS when all other factors were held constant for most body composition targets.

3DAE-GPR represents the fully nonlinear pipeline where body composition is predicted from GPR using 3DAE deep features as inputs. The results for both the bottleneck layer (7 × 612) and the layer with the lowest RMSE are shown. The third layer (dimensions 400 × 64) was the most accurate feature layer for males and the first layer (dimensions 6890 × 16) 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 original 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.

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.

We rescaled the charts in Fig. 1 by normalizing each column by the RMSE of the fully linear PCA-OLS model and plotted the values in Fig. 2 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 increased in RMSE when moving from PCA-OLS to PCA-GPR. However, as previously observed most RMSEs increased in females when incrementing PCA to 3DAE. When comparing nonlinear methods to linear methods, GPR is more accurate for body composition prediction in most metrics, but 3DAE is not always a better feature extraction method for improving the accuracy of body composition prediction from shape relative to using linear PCA in females.

Test-retest precision of percent fat and visceral fat estimation is shown in Table 2. For visceral fat mass precision, Coefficient of Variation (%CV) was calculated according to the definition in Glüer et al.[17] between two scans of the same individual in the test set taken on the same day (smaller is better). 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 also decreased precision error relative to PCA in both sexes, 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.

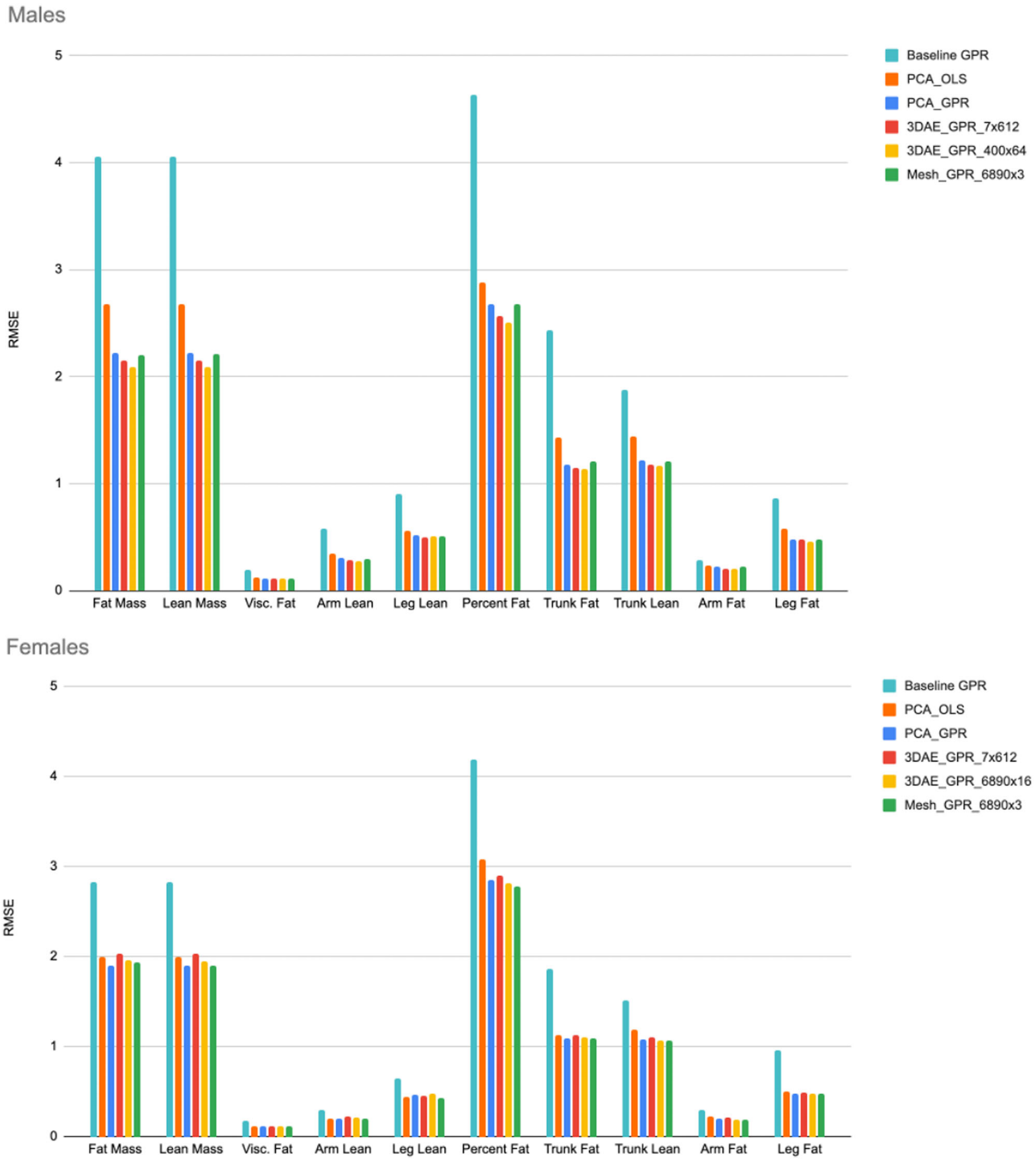

Fig. 1 | RMSE prediction errors for different model permutations, model size $d = 4284$. Every column represents exactly one modification to the pipeline parameters from the previous column (i.e., OLS to GPR) excluding the baseline column. RMSE is shown in kg except for PFAT, which is expressed as a 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.

Comparisons of our 3DAE-GPR models against prior work are shown in Table 3 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 achieved the lowest accuracy in previous works based on linear models.

GPR using the exact same PCA features of Tian et al.[14]. (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 ref. 14 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

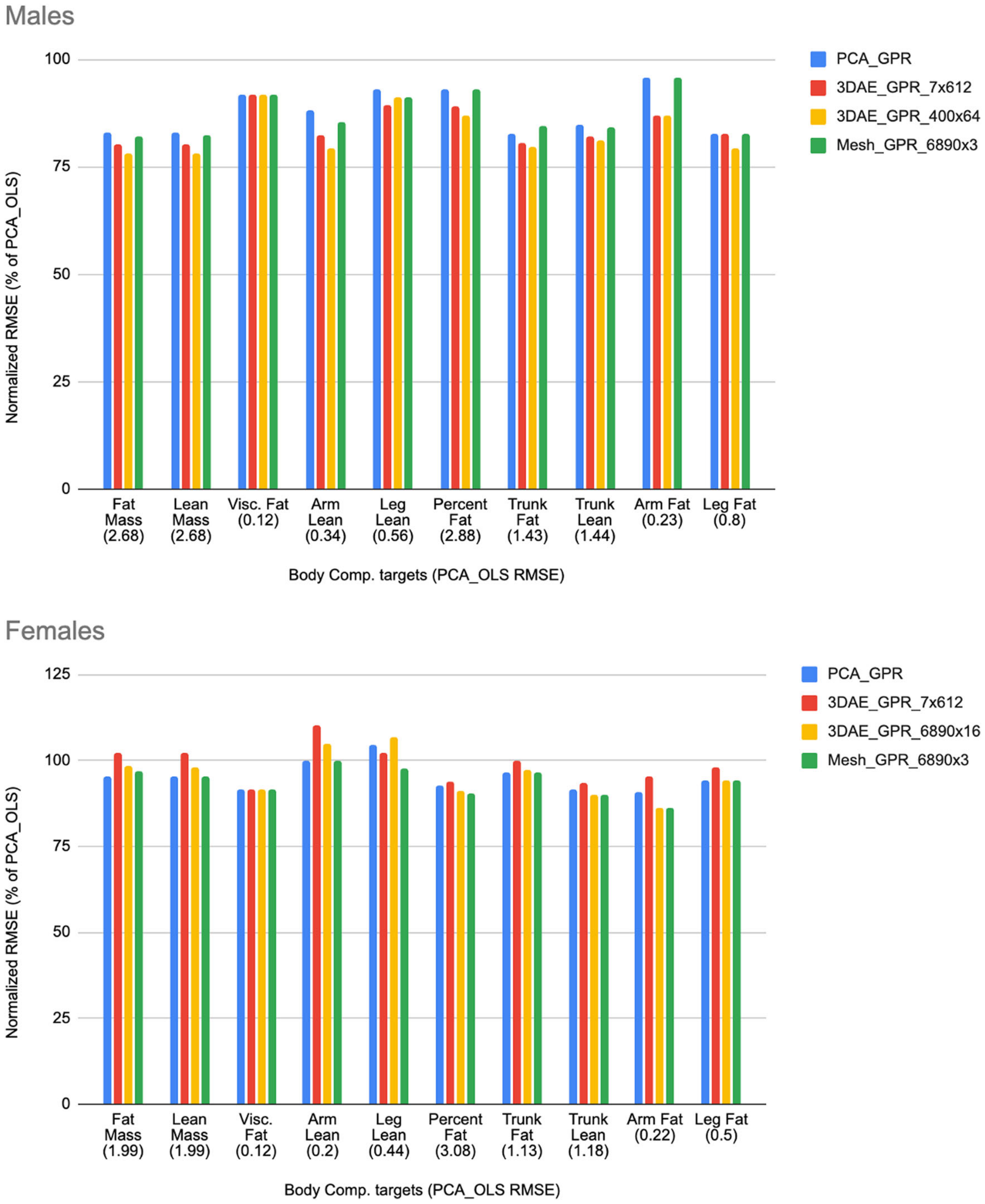

Fig. 2 | Normalized RMSE values from Fig. 3 as a percentage of the linear baseline (PCA_OLS). Values below 100% indicate lower error than the baseline model (PCA shape encoding with OLS body composition regression). Female arm lean and female leg lean were the only variables that did not outperform baseline.

Table 2 | Test-retest precision between repeat 3DO scan pairs of each participant in the SUA test set on the same scan device

		3DAE-GPR 301	3DAE-GPR 4284	PCA-GPR 4284	PCA-OLS 4284	Tian et al.[14]	DXA (Criterion)
Visceral fat % CV	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	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%

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.

Table 3 | 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 Fig. 1

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.[13]	M: 177 F: 230	M: 3.55 F: 3.88	M: 0.14 F: 0.13
Tian et al.[16]	M: 31 F: 39	M: 3.90 F: 3.29	M: 0.15 F: 0.17
Wong et al.[27]	M: 159 F: 202	M: 2.73 F: 3.46	M: 0.13 F: 0.13
Tian et al.[14]	M: 182 F: 248	M: 3.24 F: 4.22	M: 0.12 F: 0.14
PCA-GPR M391/F457[14]	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

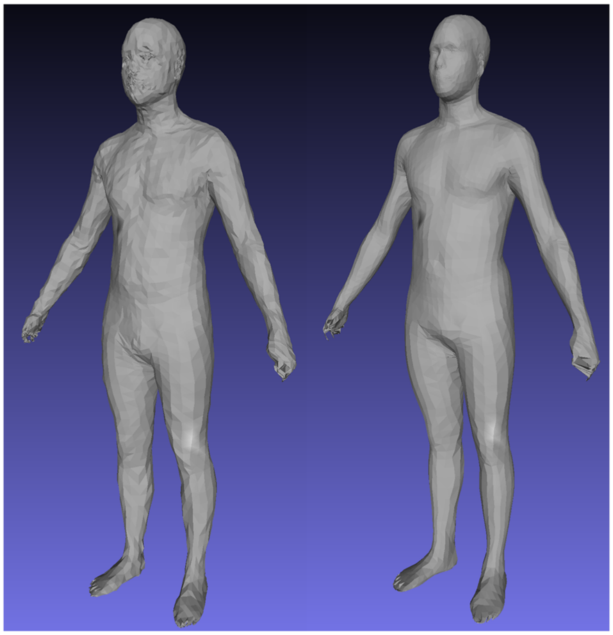

Fig. 3 | Pretraining on 40,000 DFAUST meshes improves reconstruction fidelity in single pose, multi-identity data. Left: Evaluation set mesh reconstruction after training 400 epochs from random initialization. Test set reconstruction error was 22.7 mm. Right: Same mesh reconstruction using a model pretrained first on 200 epochs with DFAUST only followed by 400 epochs of finetuning. Test set reconstruction error was 2.0 mm.

for our best performing, maximum size PCA model (PCA-GPR 4284) to demonstrate that 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. We note that compared to[14], our best 3DAE-GPR model achieved lower RMSE on all of the 10 body composition metrics measured in Fig. 2, shown in Supplementary Table 3. However, predicted metrics other than percent fat and visceral fat already showed high correlation with DXA in prior work using linear methods. Thus, we focused the comparison to prior work in Table 3 on the metabolically significant and previously underperforming predictions of percent fat and visceral fat. The test set used in this work was held the same as[14]. However, the training dataset was greatly expanded in size and scope relative to prior works.

Ablation results

Figure 3 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 22.7 mm MAE on test data reconstruction, more than 10x the error of the error show in Table 3. The pretraining steps using 40,000 DFAUST meshes was essential for creating an accurate shape model using a nonlinear 3DAE.

Figure 4 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.[18] 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 1 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.

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 (2.58 mm, 2.71 mm, 2.76 mm 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 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. 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 similar to 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.

Discussion

This study showed three 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 except female leg lean. Second, 3DAE feature extraction improved body composition prediction for males. However, deep features were not more informative than the unencoded mesh geometry in females and did not outperform PCA in geometric reconstruction at high parameter counts. Third, both 3DAE and GPR achieved higher precision than their linear counterparts in both males and females when measured on percent fat and visceral fat mass.

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

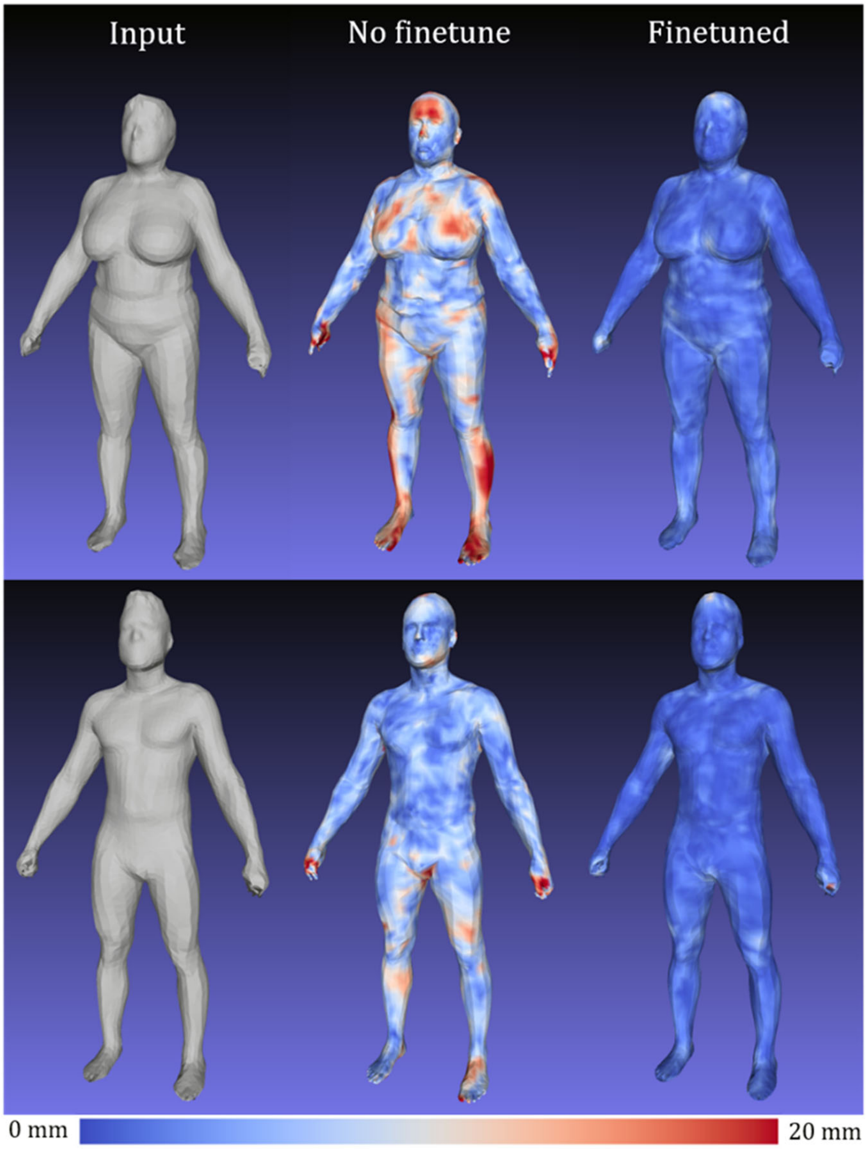

Fig. 4 | Finetune training on 4286-member multi-identity ensemble dataset reduces reconstruction error on test set meshes relative to 3DAE models trained on DFAUST only. Left: input 3DO test mesh. Center: reconstruction error heatmap using model state following 200 epochs of pretraining on DFAUST only. Test set MAE was 8.7 mm. 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.

composition prediction accuracy. The improved accuracy and precision demonstrated here builds on our prior work[13,14] and further establishes 3DO as a reliable and accessible clinical tool for the assessment of body composition. Our trained model could be packaged into the software of a commercial scanner such as Fit3D and operated end to end without clinical or computational expertize as demonstrated in Tian et al.[14,19]. The modern implementation of GPR in scikit-learn is efficient and executes in only a few seconds on our servers. Crucially, it does not need to be re-run in real time in a clinical deployment of our work as the model weights are fixed in future test executions. Such a tool has promising clinical implications for the management of chronic disease, where body composition has known associations with morbidity and mortality. The marked increase in precision of this work over prior results (up to 50% lower retest error and within 1-2x DXA precision margins) makes our nonlinear models particularly good candidates for longitudinal tracking of body composition[20] as tighter error bounds reduce the least significant change[21] observable with repeated measurements over time.

Future work can explore the relative accuracy and precision of our work compared to commercial implementations of body composition estimation from 3D scanners. Although the exact algorithm used by commercial scanners is not disclosed, the empirical accuracy of their methods can be evaluated using our paired DXA scan data and juxtaposed with our results.

Mapping 60k vertex meshes to a common 6890 vertex template inevitably reduces the smoothness and resolution of the 3D mesh surfaces. However, the objective of our experiment was to test the marginal differences between linear and nonlinear models holding all else constant. The resolution of the input meshes was not an experimental variable in our design and was necessitated by the architecture of our pretrained model and the format of the DFAUST data. We were still able to achieve higher accuracy and precision over prior work using 60k models and do not believe the mesh resolution had an effect on the conclusions of this work.

To our knowledge this was the first application of deep nonlinear autoencoder networks to 3DO body composition estimation. We explicitly chose a spatial graph convolutional network for our autoencoder approach. Other autoencoder approaches, such as implicit surface encoding[22], non-convolutional variational architectures[23], and spectral graph convolution[24] 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. Point cloud autoencoders such as PointAE[25] lacked the spatial locality property of graph convolutional networks and were not sufficiently different from PCA in their ability to encode regional part-specific features. In PointAE, the entire unconnected point cloud is fed into the network as input and shape features are extracted globally rather than locally like with convolutional filters. This makes the network functionally similar to a fully connected neural net with all-to-all correlations between vertices, which is the same weakness in PCA feature extraction that we are attempting to overcome in our design. PCA based models were highly sensitive to pose variation in prior work, and we suspect our increased precision using 3DAE is attributable to spatially local features that are insensitive to global variations in pose.

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 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[13,14,26] 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.[27] or Jiang et al.[23] 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[28]. As our dataset was transformed to adopt the mesh topology of SMPL[29] 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.[29].

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[30]. 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 2900 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. 5. 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.

The CAESAR data used in this work was built from the original 3D scan data collected by Robinette et al.[31] and not on a derived reconstruction such as the MPI CAESAR dataset produced by Pishchulin et al.[32]. 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. 6. 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

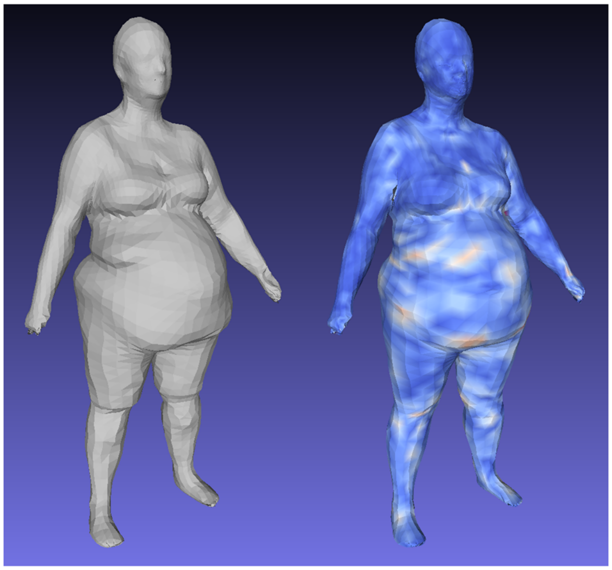

Fig. 5 | Heatmap of geometric reconstruction error for outlier body shapes. Left: an input test mesh with high BMI; right, the heat map of the reconstruction showing where the greatest errors occurred. The numerous folds in the abdomen of this individual (47.3% body fat) were unique to only a few scans in the dataset and were possibly insufficiently modeled in the latent encoding.

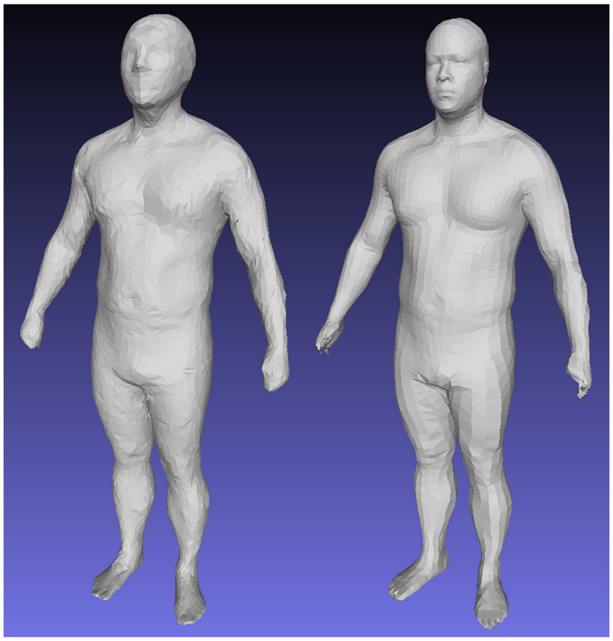

Fig. 6 | MPI CAESAR mesh compared to our remeshing of the same individual. The MPI shape was projected onto a linear PCA basis and lost considerable identifying detail.

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.

We contributed the following key findings with this work:

1. For the first time, we trained a 3D graph convolutional autoencoder (3DAE) to statistically model inherent human body shape variation independent of pose. To train this model, we assembled the largest multi-individual 3D optical body shape dataset currently in existence spanning over 4200 scans, all standardized to a common 6890 vertex mesh topology for computational compatibility.
2. We isolated the performance of linear and nonlinear statistical shape models by training 3DAE and PCA models on a common dataset to determine their effect on geometric reconstruction error and body composition prediction accuracy and precision.
3. We isolated the performance of linear and nonlinear regression models on body composition prediction by analyzing the comparative accuracy and precision of ordinary least squares (OLS) and gaussian process regression (GPR) models trained from common sets of feature vectors derived from both PCA and 3DAE.

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 10 metrics of body composition in males and for 9 in 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 10 tested body composition variables. Relative to linear PCA feature computation, 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. However, deep 3DAE features reduced precision error between 4 and 14% relative to linear features with all other variables held equal for both sexes, making their utility over linear PCA feature extraction compelling despite the lack of accuracy boost in females. 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. 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 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.

Table 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

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	-	-

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.

Methods

This work performs an ablation study 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.[18] on the composite dataset at multiple parameter sizes and evaluated its 3D reconstruction error on test data. 10 variables of body composition were studied consistent with prior work with reference values obtained with same-day DXA scans. The variables studied were:

1. Total body fat mass
2. Total body lean mass
3. Visceral fat mass
4. Arm lean mass
5. Leg lean mass
6. Percent body fat
7. Trunk fat mass
8. Trunk lean mass
9. Arm fat Mass
10. Arm lean mass

We trained nonlinear GPR models from 3DAE deep features at multiple scales to predict reference DXA values. We then trained a PCA model as a linear baseline using the same training data and parameter counts as the nonlinear model as an ablation study against prior work. From these linear PCA features we trained linear least squares regression and nonlinear GPR models to create linear-linear and hybrid linear-nonlinear models for DXA body composition prediction. This iterative modeling allowed us to quantify the impact of novel applications of nonlinear methods against linear baselines.

Experimental cohort

Four data sets were used for this study: CAESAR[31], Shape Up! Adults (SUA) (NIH R01 DK109008)[13], Shape Up! Kids (SUK) (NIH R01 DK111698)[26], and DFAUST[33]. Table 4 shows how this data was subdivided to train and test

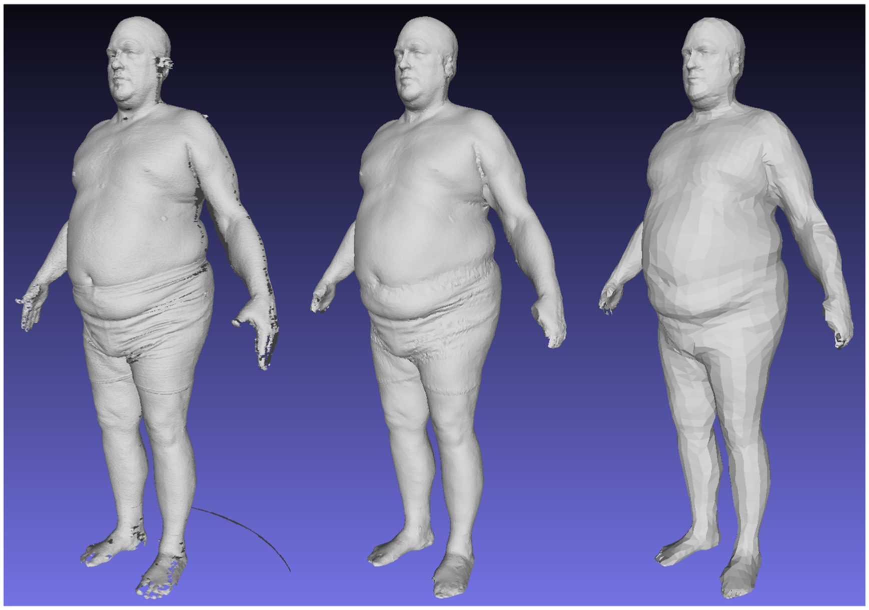

Fig. 7 | 3DO data templating pipeline from raw scan to 60k mesh to DFAUST format. Left: a raw CAESAR scan with unordered vertices and incomplete surfaces. Center: automatic 60 k template registration using the methods of Tian et al. Right: SMPL (6890) topology transformation of the 60k fit, created by multiplying the 60k mesh with a sparse nearest-neighbor matrix C and deforming the result to align with the 60k mesh.

our models. All participants gave informed consent, and the study protocols for Shape Up! Adults and Shape Up! Kids were approved by the Institutional Review Board (IRB) at Pennington Biomedical Research Center (PBRC; IRB study #2016-053, #2017-10, Federalwide Assurance #00006218); University of California, San Francisco (UCSF; IRB #15-18066, IRB #16-20197); and University of Hawaii Office of Research Compliance (UH ORC; Center of Health Sciences, #2017-01018, #24282). Please refer to refs. 31 and [33] for consent reporting for CAESAR and DFAUST.

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 Cyberware 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.[18]. 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. 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.[14] was preserved to investigate the performance of the new deep autoencoder and nonlinear GPR on the same test set. The test set was a random sample of 20% of the available Shape Up! Adults scans.

Standardizing 3D mesh topologies

To enforce topological consistency between all 3DO meshes, all scans in Table 1 were standardized to a constant topology containing 6890 vertices (referred to as the SMPL[29] topology) except for the DFAUST data, which is already in this format natively. Unorganized raw point clouds of 3D scans were converted to 60,001 vertex templates (60k) using the automated template fitting method of Tian et al.[14] and the nonrigid deformation of Allen et al. [34] 60k templated meshes were then transformed to the DFAUST format containing 6890 vertexes using a nearest-neighbor mapping matrix. A visual example of this process is shown in Fig. 7.

All input meshes needed to contain the same vertex count, order, and connectivity as the 6890 vertex DFAUST data to fit the network architecture. We used an existing automated 3D template fitting method for standardizing 3DO scans from our prior work in Tian et al.[14] to convert meshes first to a 60,001-vertex template (60k). We solved for a mapping between the 60k template and the DFAUST 6890 vertex template using a sparse matrix multiplication correlating the nearest neighbors of the vertices of a DFAUST mesh to its 60k standardized counterpart and applied this mapping to every mesh in our dataset.

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

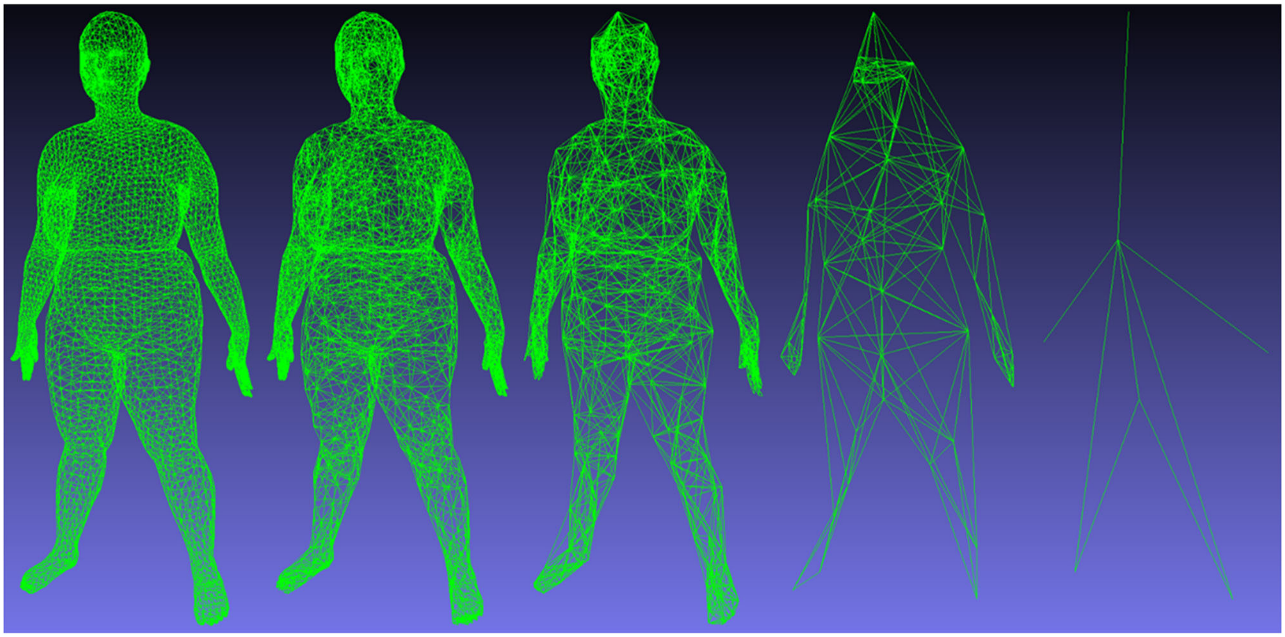

Fig. 8 | Successive convolutional downsampling of a human mesh with a radius and stride of 2. The images represent in order the input mesh (6890x3) and the intermediate graph topologies containing (# vertices x # channels): (6890 x 16), (1925 x 32), (400 x 64), (54 x 128), (7 x *f*). The final layer is the bottleneck layer, whose channel count *f* is variable in our experiments.

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.[34] 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 1.

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 data subsets 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. 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.

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.[18] 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 unpooling, 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 6890 × 3 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 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 up sampling 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. A visualization of the precomputed graph down sampling for each layer of the autoencoder is shown in Fig. 8.

Graph convolutional network implementation

The 3D convolutional network implementation was largely unchanged from Zhou et al.[18] except for channel depths in the intermediate layers. 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 subsequent 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

$$\mathbf{y}_i = \sum_{\mathbf{x}_{i,j}} \mathbf{W}_j^{\mathrm{T}} \mathbf{x}_{i,j} + \mathbf{b} \tag{1}$$

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[35] 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 $\mathbf{B}_k$:

$$\mathbf{W}_{i,j} = \sum_{k=1}^{M} \alpha_{i,j,k} \mathbf{B}_k \tag{2}$$

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.

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. To test the geometric reconstruction accuracy of the 3DAE on a range of model sizes and to check for overfitting at higher parameter counts, we trained 3DAE models with varying bottleneck channel depths of 7, 43, 90, and 612. With 7 nodes at the bottleneck layer, these channel depths defined total latent feature vector sizes of 49, 301, and 630, and 4284. The original network in Zhou et al. was trained with a bottleneck layer vector size of 63; we sampled bottleneck sizes randomly to span <1.0x and 10x the bottleneck size of the original work to investigate the correlation between model size and reconstruction accuracy. We defined the 4284 latent vector model as the upper size bound for our 3DAE bottleneck as it is the highest multiple of 7 for which there is an equivalent linear PCA shape model to compare against.

All models were trained with a kernel basis size of $M = 37$. The learning rate was set to 1e-4. Training batch size was set to at 16[18]. All model variations had the same architecture and training data aside from the bottleneck layer.

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. Prior work on estimating DXA body composition from 3D body shape used least squares linear regression, which imposed a restrictive assumption of linearity on the relationship between body shape variables and body composition. GPR was previously used in Wang et al.[36] for visceral adipose tissue estimation from body circumferences and is a nonlinear, probabilistic generalization of linear regression that relaxes the imposition of linearity on the function between body shape features and body composition. GPR is more generalized than linear regression but not as unrestricted as a multi-layer perceptron network (MLP). The limited number of data observations in our dataset made GPR a more appropriate regression method than MLPs while still relaxing tight assumptions on the function shape.

We trained a GPR model to learn a nonlinear mapping between the encoded latent vectors of the training data and the DXA measurements for 10 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 and body volume[15]. We experimented with other kernels such as the 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 is a general algorithm whose derivation details can be referenced in the Supplementary Material and in greater detail in refs. 37,38. We used the scikit-learn[38] implementation of GPR in our work. 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, we performed GPR to body composition targets. We reported the prediction accuracies for the bottleneck layer containing 4284 (7 × 612) features and for the feature layer producing the lowest RMSE in GPR prediction.

Comprehensive ablation trials vs. linear methods via model permutations

PCA is a common linear approach to dimensionality reduction for 3D human body shape due to the widespread adoption of methods such as SMPL[29] 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[14,34].

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 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.[14], the training set in this work is greatly expanded. Recreating PCA models with consistent data membership 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 same data. Due to the predefined downsampling of the 6890-vertex mesh topology shown in Fig. 8, 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 (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. We also built a 3DAE model with 301 components to test for the possibility of overfitting using 4284 parameters.

To further test our hypothesis regarding the better prediction accuracy of nonlinear GPR relative to OLS, we trained a GPR model using the same PCA basis from[13], labeled PCA-GPR M391/F457 which indicates 391 and 457 components for males and females respectively.

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 iterated through different model permutations consisting of feature extraction with both linear and nonlinear shape models followed by both linear and nonlinear regression to DXA body composition measurements. Model permutations demonstrated the marginal contributions of GPR and 3DAE to the precision and accuracy of body composition prediction relative to the baselines established by linear algorithms used in prior work.

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 estimation for four model permutations: the GPR baseline estimated using only demographic features [height, weight, age] with no shape information, the linear baseline consisting of ordinary least squares regression from PCA shape features (PCA-OLS), hybrid consisting of GPR using PCA shape features (PCA-GPR), and fully nonlinear GPR from deep shape features (3DAE-GPR). We quantified estimation error with root-mean-squared-error (RMSE) relative to reference DXA measurements and plotted the normalized RMSE of each predicted metric as a percentage of the PCA-OLS fully linear baseline. The coefficient of determination (R^2) for agreement to DXA was assessed for the 3DAE-GPR model.

We predicted body composition on test-retest data pairs to compare the precision of our model 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 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 enforce consistency with the parameter count of the PCA permutations.

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

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to the inclusion protected health information and personally identifiable information (PHI/PII). Data may be made available from the corresponding author on reasonable request pending IRB review and approval.

Code availability

The underlying code for this study [and training/validation datasets] is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author. The original 3DAE implementation from Zhou et al. is publicly available at https://github.com/papagina/MeshConvolution.

Received: 6 February 2024; Accepted: 20 January 2025;
Published online: 02 February 2025

References

1. Wilson, J. P., Kanaya, A. M., Fan, B. & Shepherd, J. A. Ratio of trunk to leg volume as a new body shape metric for diabetes and mortality. *PLoS ONE* **8**, e68716 (2013).
2. 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.* **40**, 146–152 (2021).
3. Au, P. C., Li, H. L. & Lee, G. K. Sarcopenia and mortality in cancer: a meta-analysis. *Osteoporos Sarcopenia* **7**, S28–S33 (2021).
4. Goodpaster, B. H. et al. Obesity, regional body fat distribution, and the metabolic syndrome in older men and women. *Arch. Intern. Med.* **165**, 777–783 (2005).
5. Zhang, C., Rexrode, K. M., Dam, R. M., Li, T. Y. & Hu, F. B. Abdominal obesity and the risk of all-cause, cardiovascular, and cancer mortality. *Circulation* **117**, 1658–1667 (2008).
6. Hui, W. S., Liu, Z. & Ho, S. C. Metabolic syndrome and all-cause mortality: a meta-analysis of prospective cohort studies. *Eur. J. Epidemiol.* **25**, 375–384 (2010).
7. Dulloo, A. G., Jacquet, J., Solinas, G., Montani, J.-P. & Schutz, Y. Body composition phenotypes in pathways to obesity and the metabolic syndrome. *Int. J. Obes.* **34**, S4–S17 (2010).
8. Galindo Martin, C. A. et al. The GLIM criteria for adult malnutrition and its relation with adverse outcomes, a prospective observational study. *Clin. Nutr. ESPEN* **38**, 67–73 (2020).
9. 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.* **40**, 2188–2199 (2021).
10. Marco, E., Sanchez-Rodriguez, D. & Davalos-Yerovi, V. N. Malnutrition according to ESPEN consensus predicts hospitalizations and long-term mortality in rehabilitation patients with stable chronic obstructive pulmonary disease. *Clin. Nutr.* **38**, 2180–2186 (2019).
11. Rondel, A., Langius, J. A. E., Schueren, M. A. E. & Kruizenga, H. M. The new ESPEN diagnostic criteria for malnutrition predict overall survival in hospitalised patients. *Clin. Nutr.* **37**, 163–168 (2018).
12. Fosbøl, M. Ø. & Zerahn, B. Contemporary methods of body composition measurement. *Clin. Physiol. Func. Imaging* **35**, 81–97 (2014).
13. Ng, B. K. et al. Detailed 3-dimensional body shape features predict body composition, blood metabolites, and functional strength: the shape up! studies. *Am. J. Clin. Nutr.* **110**, 1316–1326 (2019).
14. Tian, I. Y. et al. A device-agnostic shape model for automated body composition estimates from 3D optical scans. *Med. Phys.* **49**, 6395–6409 (2022).
15. Tinsley, G. M., Moore, M. L., Benavides, M. L., Dellinger, J. R. & Adamson, B. T. 3-dimensional optical scanning for body composition assessment: A 4-component model comparison of four commercially available scanners. *Clin. Nutr.* **39**, 3160–3167 (2020).
16. Tian, I. Y. et al. Predicting 3D body shape and body composition from conventional 2D photography. *Med. Phys.* **47**, 6232–6245 (2020).
17. Glüer, C.-C. et al. Accurate assessment of precision errors: How to measure the reproducibility of bone densitometry techniques. *Osteoporosis Int.* **5**, 262–270 (1995).
18. Zhou, Y. et al. Fully convolutional mesh autoencoder using efficient spatially varying kernels. In *Proc. 34th International Conference on Neural Information Processing Systems (NIPS'20)* 9251–9262 (2020).
19. Tian, I. Y. et al. Automated body composition estimation from device-agnostic 3D optical scans in pediatric populations. *Clin. Nutr.* **42**, 1619–1630 (2023).
20. Wong, M. C. et al. Monitoring body composition change for intervention studies with advancing 3D optical imaging. *Am. J. Clin. Nutr.* **117**, 802–813 (2023).
21. Shepherd, J. A & Lu, Y. A generalized least significant change for individuals measured on different DXA systems. *J. Clin. Densitom* **10**, 249–258 (2007).

22. 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).
23. Jiang, B., Zhang, J., Cai, J. & Zheng, J. Disentangled human body embedding based on deep hierarchical neural network. *IEEE Transac. Visualiz. Comput. Graphics* **26**, 2560–2575 (2020).
24. Lemeunier, C., Denis, F., Lavoué, G. & Dupont, F. Representation learning of 3D meshes using an Autoencoder in the spectral domain. *Comput. Graphics* **107**, 131–143 (2022).
25. 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), Seoul, South Korea*. 5409-5418 (2019).
26. Wong, M. C. et al. Children and adolescents' anthropometrics body composition from 3-D optical surface scans. *Obesity* **27**, 1738–1749 (2019).
27. Wong, M. C. et al. A pose-independent method for accurate and precise body composition from 3D optical scans. *Obesity* **29**, 1835–1847 (2021).
28. Wong, M. C. et al. Monitoring body composition change for intervention studies with advancing 3D optical imaging technology in comparison to dual-energy X-ray absorptiometry. *Am. J. Clin. Nutr.* **117**, 802–813 (2023).
29. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. & Black, M. J. SMPL. *ACM Transac. Graphics* **34**, 1–16 (2015).
30. Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q. & Wilson, A. G. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. *Neural Inform. Process. Syst.* **31**, 7576–7586 (2018).
31. Robinette, K. M., Blackwell, S., Daanen, H., Boehmer, M. & Fleming, S. Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, vol. 1 Summary. *Defense Technical Information Center.* (2002).
32. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C. & Schiele, B. Building statistical shape spaces for 3D human modeling. *Pattern Recognit.* **67**, 276–286 (2017).
33. Bogo, F., Romero, J., Pons-Moll, G. & Black, M. J. Dynamic FAUST: registering human bodies in Motion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
34. Allen, B., Curless, B. & Popović, Z. The space of human body shapes. *ACM Transac. Graphics* **22**, 587–594 (2003).
35. 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).
36. Wang, Q., Lu, Y., Zhang, X. & Hahn, J. K. 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).
37. Guo, S. Implement A Gaussian Process From Scratch. *Medium. Published* (2021). at https://towardsdatascience.com/implement-a-gaussian-process-from-scratch-2a074a470bce
38. Pedregosa, F., Varoquaux, G. & Gramfort, A. Scikit-learn: Machine Learning in Python. *J. Machine Learn. Res.* **12**, 2825–2830 (2011).
39. Tian, I. Y. Learning clinical body composition metrics from 2D and 3D optical imaging. (University of Washington, 2023).

Acknowledgements

Funding awards for co-authors are as follows: A.K.G.: NICHD #R01HD082166, National Institutes of Health NORC Center Grants (P30DK072476, Pennington/Louisiana and P30DK040561, Harvard); J.A.S.: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (R01DK109008 and R01DK111698). A previous version of this manuscript appeared in the doctoral dissertation "Learning clinical body composition metrics from 2D and 3D optical imaging"[39].

Author contributions

I.Y.T.: Methodology, Software, Formal Analysis, Investigation, Writing – Original. J.L.: Validation, Data Curation, Writing – R&E M.C.W.: Methodology, Investigation, Writing – R&E N.N.K.: Investigation, Writing – R&E, Data Curation, Project administration. Y.E.L.: Investigation, Writing – R&E, Data Curation. A.K.G.: Investigation, Writing – R&E, Funding Acquisition. SBH: Conceptualization, Methodology, Resources, Supervision, Funding Acquisition, Writing – R&E. B.C.: Conceptualization, Methodology, Resources, Supervision, Funding Acquisition, Writing – R&E. J.A.S.: Conceptualization, Methodology, Resources, Supervision, Funding Acquisition, Writing – R&E. All authors have read and approved this manuscript.

Competing interests

J.A.S is a scientific advisor for Styku LLC and owns stock options. B.C. is employed by Google LLC and owns stock options. S.B.H. is on the medical advisory boards of Tanita Corp. and Medifast Corp. and is a former employee of Merck & Co. S.B.H. is an Amazon Scholar and owns stock options. Styku LLC and Fit3D Inc. contributed scanning equipment in support of NIH studies.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01469-6.

Correspondence and requests for materials should be addressed to Isaac Y. Tian.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.