AutoSkull: Learning-based Skull Estimation for Automated Pipelines

Aleksandar Milojevic[1], Daniel Peter[2], Niko B. Huber[2], Luis Azevedo[3], Andrei Latyshev[3], Irena Sailer[3], Markus Gross[1], Bernhard Thomaszewski[1,2], Barbara Solenthaler[1], and Baran Gözcü[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Align Technology
[3] Faculty of Medicine, University of Geneva, Switzerland

Abstract. In medical imaging, accurately representing facial features is crucial for applications such as radiation-free medical visualizations and treatment simulations. We aim to predict skull shapes from 3D facial scans with high accuracy, prioritizing simplicity for seamless integration into automated pipelines. Our method trains an MLP network on PCA coefficients using data from registered skin- and skull-mesh pairs obtained from CBCT scans, which is then used to infer the skull shape for a given skin surface. By incorporating teeth positions as additional prior information extracted from intraoral scans, we further improve the accuracy of the model, outperforming previous work. We showcase a clinical application of our work, where the inferred skull information is used in an FEM model to compute the outcome of an orthodontic treatment.

Keywords: Machine Learning · Digital Patient · Skull Estimation · Mesh Processing.

1 Introduction

With the increasing integration of 3D face and intraoral scans into medical workflows, the generation of comprehensive patient-specific head models has become more accessible. However, in many applications, the avoidance of CBCT imaging due to ionizing radiation leaves a critical gap in obtaining accurate skull information. Yet, incorporating detailed skull data into digital twin models is indispensable for a more complete modeling of anatomical structures and enabling accurate physics-based treatment simulations. Consequently, the demand for more advanced head models is apparent [1, 5, 6, 17], as these models can comprehensively capture both external soft tissue layers and intricate interior anatomical structures. Recent innovations underscore the effectiveness of head models in augmenting realism in medical simulations [21, 18, 8, 20, 10] as well as in diverse fields like entertainment animation systems [3, 19, 7, 11, 22].

This paper tackles the challenge of incorporating interior geometry into face scan models. Our objective, akin to prior works OSSO [9] and SCULPTOR [16], is to predict skull shapes from 3D facial surfaces. OSSO uses linear regression

for a coarse estimation of the skeleton, resulting in a generalized skull shape. SCULPTOR, based on the FLAME model [12], offers more accurate head modeling using medical data but lacks a non-linear component to model the skin-skull relationship

In [14], the authors present a non-parametric skull shape estimation method, using an intricate procedure that follows steps for registration, deformation, and refinement. While the method relies on multiple manual intervention steps, our work puts emphasis on components that make the method suited for automation. In [15], the authors presented a neural network to regress skin-skull-offset vectors, using skin mesh surface points and other skin mesh features as input. The result is a coarse, water-tight skull mesh, which naturally lacks skull surface details. In contrast, we aim at capturing such intricate details, making it better suited for simulation and visualization for medical applications.

Our method employs a Multilayer Perceptron (MLP) network trained on Principal Component Analysis (PCA) coefficients from CBCT-registered skin- and skull-mesh pairs. This network accurately infers skull shapes from a given skin surface, outperforming recent methods [9, 16]. Additionally, it can incorporate intraoral scans to improve accuracy in the mouth area. Our results indicate the method's potential for various applications, such as providing radiation-free bone surfaces for medical visualizations and simulating treatment outcomes. Our key contributions are as follows:

- An MLP-based deep learning solution for 3D skull inference based on medical data that outperforms previous work in terms of accuracy and simplicity.
- A method that includes a strong prior derived from teeth data to further improve accuracy.
- A demonstration of our method's applicability in a medical simulation.

2 Method

Our learning-based method aims to infer skull geometry from a standard face scan, as shown in Figure 1. It uses an MLP operating in a PCA space constructed from a CBCT head dataset. The following sections cover the preprocessing, the MLP-based learning and inference, and the incorporation of a teeth prior.

2.1 Preprocessing for Training

Using a PCA model allows us to work in a reduced dimensional space, generating smooth, natural skin and skull shapes by adjusting its parameter vector. To create this PCA space for registered meshes, we must first process the CBCT scans.

We begin by extracting the skin and skull surfaces using the Blue Sky Plan software [2]. Next, we use the landmarking software Wrap3D [4] to manually place point correspondences between the extracted surfaces and our template skin or skull mesh. To handle missing patches on the extracted surfaces, we

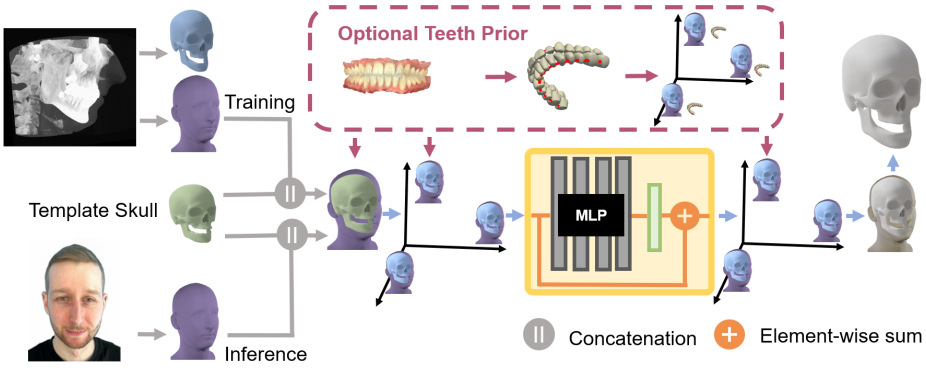

Fig. 1. Overview of our pipeline: Our model is trained on CBCT-extracted skin and skull meshes and uses a face scan mesh as input during inference. The face scan is first registered to a template skin mesh, and a template skull is incorporated. This merged mesh is then projected into a PCA space derived from CBCT skin and skull mesh pairs. The resulting PCA parameter vector is fed into an MLP network, which predicts a correction vector to adjust the input vector. This corrected vector is projected back into data space to generate the final output skull mesh. If an intraoral scan is available, it is used as a teeth prior to enhance accuracy.

exclude certain polygons on the template mesh before registering the meshes with the NICP algorithm. The final registered meshes share the Wrap3D[4] template topology and consist of N_{skin} vertices for the skin mesh and N_{skull} vertices for the skull mesh.

To construct the PCA space of the skin and skull mesh pairs, the respective mesh vertex matrices are vectorized. Each data point consists of a one dimensional vector with length $M = (N_{\text{skin}} * 3) + (N_{\text{skull}} * 3)$, giving us a data matrix D of size $N_{\text{train}} \times M$. D is then used to construct a PCA space that is spanned by the first 15 principle components of data matrix D. Consequently, a projection of a data point $d \in \mathbb{R}^{1\times M}$ into this PCA space is described by $P(d) \in \mathbb{R}^{1\times 15}$.

At inference time, only the registered face scan is available. Our PCA model, which includes both skin and skull meshes, can be directly fitted to the face scan to produce a matching skull. However, fitting the PCA parameter vector to the skin mesh alone leaves too many possibilities for valid skull shapes, often resulting in a skull mesh that doesn't resemble the true skull. To address this, we included an MLP in the pipeline, leveraging its non-linearity for a more accurate modeling of the skull and skin shape relationship.

To provide the MLP with a consistent input format as our initial guess for our PCA parameter vector, we prepare our second data matrix D_m as follows. Let $d \in \mathbb{R}^{1\times M}$ be a data point from D. Then, d can be separated into two vectors $d_f \in \mathbb{R}^{1\times(N_{\text{skin}}*3)}$ and $d_s \in \mathbb{R}^{1\times(N_{\text{skull}}*3)}$, representing the skin and skull parts of d. Consequently, let $m_s \in \mathbb{R}^{(N_{\text{skull}})*3}$ be the generic Wrap3D [4]template skull aligned to the skull part of the mean vector of D. This generic skull vector is used to construct the data point vector $d_m \in \mathbb{R}^{1\times M}$, by replacing the original

skull part of d, which is d_s, with m_s. We repeat this replacement procedure for all of the $d \in D$, resulting in the same generic skull as an initial guess for each data point, constructing our second data matrix $D_m \in \mathbb{R}^{N_{\text{train}} \times M}$.

2.2 MLP-based Learning

To facilitate learning, we employ dimensionality reduction by projecting the two data matrices into the PCA space to obtain $P(D)$, $P(D_m) \in \mathbb{R}^{N_{\text{train}} \times 15}$. Then, we train a 4-layer MLP that takes as input $P(d_m) \in P(D_m)$ and outputs y, the prediction of $P_{res}(d, d_m) = P(d) - P(d_m)$, where $P(d) \in P(D)$. Next, y can be used to construct $z = P(d_m) + y$, which in turn is the prediction for $P(d)$. Then, we can use the predicted PCA parameter vector z to consequently generate the predicted mesh vector $d' = P^{-1}(z) \in \mathbb{R}^{M \times 1}$ reshaped into the mesh $S' \in \mathbb{R}^{(N_{\text{skin}}+N_{\text{skull}}) \times 3}$.

For training we use the the following loss

$$L(S_f, S'_f, S_s, S'_s) = \lambda_m L_m(S_f, S'_f) + \lambda_{ms} L_m(S_s, S'_s), \tag{1}$$

where $S_f \in \mathbb{R}^{N_{\text{skin}} \times 3}$ and $S_s \in \mathbb{R}^{N_{\text{skull}} \times 3}$ are the ground truth skin and skull mesh, and $S'_f \in \mathbb{R}^{N_{\text{skin}} \times 3}$ and $S'_s \in \mathbb{R}^{N_{\text{skull}} \times 3}$ are the skin and skull mesh extracted from the predicted mesh vector d'. The mesh loss L_m enforces the predicted PCA parameters to project as closely as possible to the ground truth. It is defined as the average L_2 loss between the respective meshes.

2.3 Skull Inference from Face Scan

Our method is trained with medical CBCT data, but uses 3D facial scans from a smartphone at inference time. Advanced capturing systems can also generate the input facial scans. We register the face scan to our template skin mesh, similar to the procedure in subsection 2.1. After obtaining the registered skin mesh from the face scan, we vectorize the registered skin mesh into the vector d_f. The generic skull vector m_s is then concatenated to d_f, producing the vector $d_m \in \mathbb{R}^{1 \times M}$.

We then use our PCA projection to generate $x = P(d_m)$, which is used as an input to our trained MLP to generate $y = MLP(x)$. Consequently, the predicted residual parameter vector y is added to x to obtain the prediction $z = x + y$ for $P(d)$. We then project z from the PCA space into the data space, producing the mesh vector d' reshaped into the mesh S'. Discarding the skin part, we recover the skull prediction for the initial scan. This method is referred to as *AutoSkull*.

2.4 Adding a Teeth Prior

When there is additional intraoral (IO) scan data is available, *AutoSkull* can be adapted to improve the skull prediction accuracy in the mouth area. This adaptation is called *AutoSkull+*.

First, we align the IO scans to an open-mouth pose of the input face scan, utilizing a proprietary technology [4]. Next, we automatically process every IO scan mesh by identifying the gum lines, calculating the per-tooth centers at the level of these lines and then fitting a spline curve through these centers. The resulting spline curves serve as an approximation for the teeth-bone boundary to increase the accuracy of the skull estimation in these regions.

To integrate the teeth prior into our model, we sample each of the two spline curves with 3000 sample points, vectorize them and concatenate the data point d from our original model with the teeth-point vector of size $6000 * 3$. As can be seen in Figure 1, this additional modality affects data-related aspects of the pipeline, such as the input and the PCA space.

3 Evaluation and Results

In the following paragraphs, we explain our evaluation setup, detail our comparison experiments, provide an ablation study, and conclude with an application example.

3.1 Evaluation Setup

We use two datasets to construct our training and test sets for evaluation: The main dataset (1) consists of 91 CBCT and IO scan pairs. The second dataset (2) includes 8 paired CBCT, facial and IO scans. Both CBCT imaging protocols enforced a neutral pose and closed bite.

As evaluation, we conduct three experiments with *AutoSkull* and *AutoSkull+*: (A) Using (1), we train the model on 81 registered CBCT data points, and test on the other 10 registered samples. (B) We train the model on all 91 registered samples from dataset (1) and test on dataset (2), using the registered CBCT skin meshes as input. (C) We train the model on all 91 registered samples from dataset (1) and test on dataset (2), using the registered face scans as input. The registered CBCT meshes are manually aligned with corresponding face scans to obtain ground truth skulls for (C).

As our model is trained with registered meshes, we measure performance by comparing the predicted mesh to the registered mesh, not the extracted CBCT mesh counterpart.

The MLP has two hidden layers of size 410, both with residual connections and ReLU activation functions. Training spanned 40 epochs for *AutoSkull* and 80 epochs for *AutoSkull+*, with a learning rate of 0.001. These configurations were empirically determined as most effective, as depicted in the supplementary material.

Training was performed on a Laptop with an Intel Core i7-9750H CPU at 2.60 GHz, taking about 28 minutes on average. Inference on the same device takes around 2.3 seconds per instance.

[4] provided by our industrial collaborator, Align Technology

3.2 Comparisons

We compare our method with the recent methods OSSO [9] and SCULPTOR [16]. Additionally, we apply probabilistic PCA (PPCA), as seen in [13], by treating the unknown skulls as missing values of a dataset. Each compared method underwent fine-tuning tailored for our comparative analysis, optimized to predict a skull suitable for our dataset. The publicly available skeleton-inference code base of OSSO takes as input a posed SMPL / STAR body mesh and infers the respective posed skeleton. To apply the OSSO skeleton-inference to our test set, we first register each face of our test set to the face part of the SMPL / STAR body template mesh. The resulting full-body mesh serves as input and the head part of the predicted skeleton mesh is used as the final skull prediction. SCULPTOR provides publicly accessible code for their parametric model; however, the shared version is less expressive than the one used in their evaluations. Using the shared version, we implemented our own SCULPTOR skull estimation pipeline, optimizing the model parameters for a given skin mesh to obtain skull estimations for our test set.

We evaluate the error in the face and mouth areas of the skull, as detailed in the supplementary material. Table 1 shows that both *AutoSkull* and *AutoSkull+* outperform previous methods, with *AutoSkull+* providing the best performance, particularly in aligning the mouth area, as illustrated in Figure 2. In experiment (C), a performance reduction is observed due to slight differences between optical facial scans and CBCT skin surfaces. The ground truth skulls for (C) were taken from CBCT scans and manually aligned to face scans.

As shown in Figure 3, different methods use different template skulls for their predictions. The detailed surface structures in OSSO skull predictions do not appear due to a more accurate prediction but due to their presence in the template skull. The supplementary material demonstrates that the OSSO method shows less shape variability in predicted skulls compared to our approach, especially noticeable in the estimated jaw shapes.

Table 1. Average errors in mm, measured on face and mouth areas of skulls for different methods: (A) with CBCT test set coming from the same data source as train data, (B) with alternative CBCT test set, and (C) using the face scans of the alternative test set

Evaluation	Osso	Sculptor	PPCA	AutoSkull	AutoSkull+
(A)Face	3.77 ± 0.79	2.96 ± 0.55	1.55 ± 0.33	1.46 ± 0.26	**1.44 ± 0.27**
(B)Face	3.65 ± 0.72	2.99 ± 0.35	1.86 ± 0.33	1.73 ± 0.27	**1.56 ± 0.18**
(C)Face	3.47 ± 0.65	2.54 ± 0.20	2.86 ± 0.57	2.33 ± 0.15	**2.09 ± 0.24**
(A)Mouth	4.49 ± 2.19	3.52 ± 0.99	1.71 ± 0.39	1.59 ± 0.40	**1.45 ± 0.47**
(B)Mouth	4.69 ± 2.00	3.52 ± 1.18	1.86 ± 0.42	1.74 ± 0.27	**1.36 ± 0.18**
(C)Mouth	4.66 ± 2.38	2.64 ± 0.81	3.32 ± 1.3	2.22 ± 0.27	**2.12 ± 0.28**

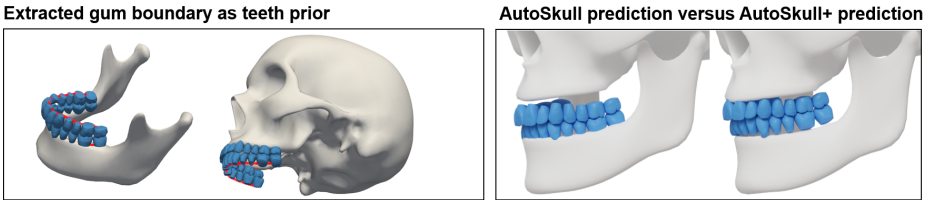

Fig. 2. We extract the gum boundaries (red) from the IO scan teeth mesh, which are very close to the bones (left). Including this teeth prior (AutoSkull+) results in higher accuracy near the mouth region (right).

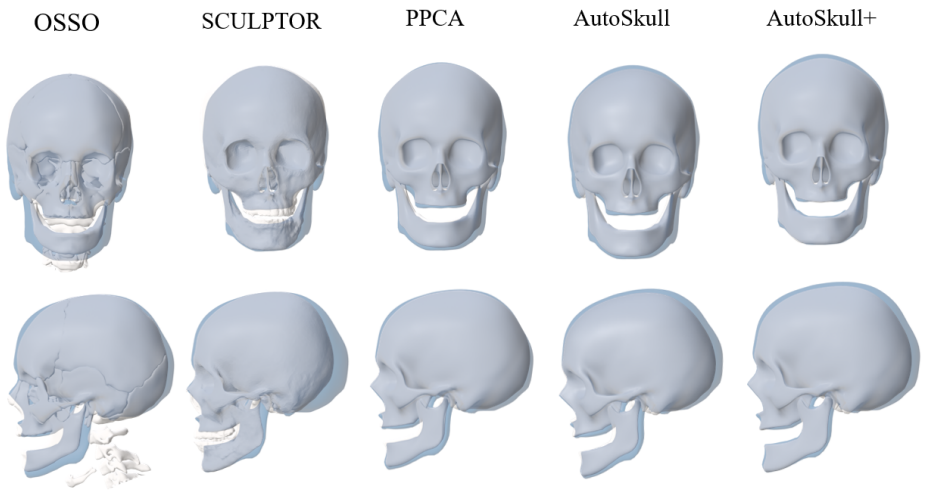

Fig. 3. A sample of the CBCT test set predictions with ground truth skull in blue and our prediction overlaid in white. We present a comparison of skull shape predictions by OSSO, Sculptor, PPCA, and our methods AutoSkull and AutoSkull+.

3.3 Ablation study

Number of PCA components We report the PCA reconstruction errors on the face area for different numbers of PCA components *(#components, error in mm)*: (5, 1.48); (15, 1.25); (25, 1.17); (35, 1.14); (45, 1.08); (55, 1.03); (65, 0.97); (75, 0.95). We opted for 15 (covering 84% of the total explained variance), which strikes a good balance between accuracy and simplicity.

Losses We evaluate the losses used during training with *AutoSkull+* and report mean errors in mm for face and mouth areas on the CBCT test set (B), calculated using vertex-to-closest-point Euclidean distance distance between the ground truth mesh and predicted mesh. We report the numbers as *(lambdas, mean error face, mean error mouth)*: (λ_m=1.0 and λ_{ms}=0.0, 1.56, 1.36); (λ_m=0.0 and

λ_{ms}=1.0, 1.63, 1.53); (λ_m=0.5 and λ_{ms}=0.5, 1.58, 1.42). The differences among the loss variations are minimal, particularly when assessed in the mouth area.

Registration quality The average Euclidean distances from registered mesh vertices to the nearest points on the corresponding extracted CBCT iso-surface from dataset (1) are 0.2mm for the skin and 0.67mm for the skull face area. Comparing the test set (B) predictions with their extracted CBCT surface counterparts as ground truth yields an average mean error of 1.49mm.

3.4 Application Example

We assess our method in a clinical context, focusing on an orthodontic treatment. As shown in Figure 4, using a simple face scan, we automatically estimate a skull shape to generate a simulation mesh for the head. This mesh is then fed to a finite element method (FEM) simulator to compute the treatment outcome. Comparing the resulting face mesh to the simulation using a CBCT-derived skull shape, our estimated skull results in an average error of only 0.14 mm (vertex-to-closest-point distances) within the soft tissue region of interest.

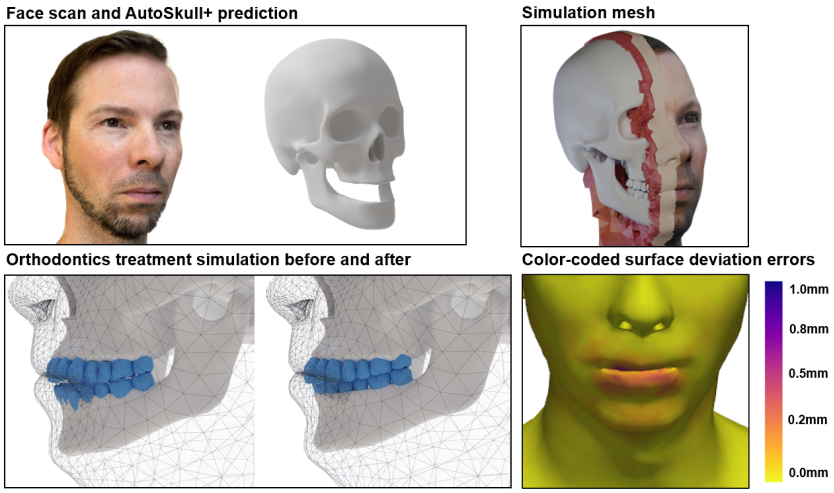

Fig. 4. From a face scan input, we estimate the person-specific skull shape and generate a simulation mesh. We then simulate an orthodontic treatment with FEM, leading to only 0.14mm average error on the deformed soft tissue compared to the simulation result that uses skull shape obtained from a CBCT scan.

4 Discussion and Conclusion

We have presented a learning-based method to reliably infer skull geometry from 3D facial surface input, which can be seamlessly integrated into automated

pipelines due to its straightforward foundational components. The primary limitation of our method is its requirement for a neutral expression input face mesh, lacking expression or pose invariance. Additionally, the generalization capability of *AutoSkull* is limited by the dataset on which it was trained. Integrating optical facial scans into training along with CBCT data could enhance the method's domain adaptation for facial scan inputs.

Future work involves enhancing our method with a fully non-linear model and exploring distinct latent spaces for the face, bones, soft tissue, and teeth. This separation can increase the variability of the spaces independently by incorporating unpaired data. Furthermore, establishing relationships between the spaces could help with generalization and decoupling shape and expression.

In conclusion, our approach facilitates the creation of digital patient twin models without the reliance on CBCT. This presents a radiation-free solution for generating comprehensive patient models and computing treatment outcomes through simulation, particularly in scenarios where the accuracy requirements are not stringent and the CBCT might not be imperative. Moreover, our models can offer flexibility by accommodating situations where only partial CBCT data is accessible, enabling the generation of complete and detailed patient models in instances of limited imaging data. This versatility positions our methodology to contribute significantly to advancements in personalized patient care.

Acknowledgments. We thank Alexandre Cavaleri for providing us with the face scan visuals. This research was funded by Align Technology.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achenbach, J., Brylka, R., Gietzen, T., Hebel, K.z., Schömer, E., Schulze, R., Botsch, M., Schwanecke, U.: A Multilinear Model for Bidirectional Craniofacial Reconstruction. In: Puig Puig, A., Schultz, T., Vilanova, A., Hotz, I., Kozlikova, B., Vázquez, P.P. (eds.) Eurographics Workshop on Visual Computing for Biology and Medicine. The Eurographics Association (2018). https://doi.org/10.2312/vcbm.20181230
2. Blue Sky Bio, LLC: Blueskyplan, https://blueskyplan.com
3. Chandran, P., Zoss, G.: Anatomically constrained implicit face models (2023)
4. Faceform, LLC: Wrap3d, https://faceform.com/
5. Gietzen, T., Brylka, R., Achenbach, J., zum Hebel, K., Schömer, E., Botsch, M., Schwanecke, U., Schulze, R.: A method for automatic forensic facial reconstruction based on dense statistics of soft tissue thickness. PLOS ONE **14**(1), 1–19 (01 2019). https://doi.org/10.1371/journal.pone.0210257, https://doi.org/10.1371/journal.pone.0210257
6. Guo, H., Planche, B., Zheng, M., Karanam, S., Chen, T., Wu, Z.: Smpl-a: Modeling person-specific deformable anatomy. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20782–20791 (2022). https://doi.org/10.1109/CVPR52688.2022.02015

7. Ichim, A.E., Kadleček, P., Kavan, L., Pauly, M.: Phace: Physics-based face modeling and animation. ACM Trans. Graph. **36**(4) (jul 2017). https://doi.org/10.1145/3072959.3073664, https://doi.org/10.1145/3072959.3073664
8. Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3d digital humans. ACM Trans. Graph. **42**(6) (dec 2023). https://doi.org/10.1145/3618381, https://doi.org/10.1145/3618381
9. Keller, M., Zuffi, S., Black, M.J., Pujades, S.: Osso: Obtaining skeletal shape from outside (2022)
10. Komaritzan, M., Wenninger, S., Botsch, M.: Inside humans: Creating a simple layered anatomical model from human surface scans. Frontiers in Virtual Reality **2** (2021). https://doi.org/10.3389/frvir.2021.694244, https://www.frontiersin.org/articles/10.3389/frvir.2021.694244
11. Kozlov, Y., Bradley, D., Bächer, M., Thomaszewski, B., Beeler, T., Gross, M.: Enriching facial blendshape rigs with physical simulation. In: Computer Graphics Forum. vol. 36, pp. 75–84. Wiley Online Library (2017)
12. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
13. Meng, D., Keller, M., Boyer, E., Black, M.J., Pujades, S.: Learning a statistical full spine model from partial observations. In: ShapeMI@MICCAI (2020), https://api.semanticscholar.org/CorpusID:222179860
14. Nguyen, T.N., Tran, V.D., Nguyen Ho, Q., Dao, T.T.: A statistical shape modeling approach for predicting subject-specific human skull from head surface (07 2020). https://doi.org/10.1007/s11517-020-02219-4
15. Nguyen Ho, Q., Nguyen, T.N., Tran, V.D., Dao, T.T.: A deep learning approach for predicting subject-specific human skull shape from head toward a decision support system for home-based facial rehabilitation. IRBM (06 2022). https://doi.org/10.1016/j.irbm.2022.05.005
16. Qiu, Z., Li, Y., He, D., Zhang, Q., Zhang, L., Zhang, Y., Wang, J., Xu, L., Wang, X., Zhang, Y., Yu, J.: Sculptor: Skeleton-consistent face creation using a learned parametric generator (2022)
17. Saito, S., Zhou, Z.Y., Kavan, L.: Computational bodybuilding: Anatomically-based modeling of human bodies. ACM Trans. Graph. **34**(4) (2015)
18. Shetty, K., Birkhold, A., Jaganathan, S., Strobel, N., Egger, B., Kowarschik, M., Maier, A.: Boss: Bones, organs and skin shape model. Computers in Biology and Medicine **165**, 107383 (2023). https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107383, https://www.sciencedirect.com/science/article/pii/S001048252300848X
19. Sifakis, E., Selle, A., Robinson-Mosher, A., Fedkiw, R.: Simulating Speech with a Physics-Based Facial Muscle Model. In: Cani, M.P., O'Brien, J. (eds.) ACM SIGGRAPH / Eurographics Symposium on Computer Animation. The Eurographics Association (2006). https://doi.org/10.2312/SCA/SCA06/261-270
20. Wagner, N., Botsch, M., Schwanecke, U.: Softdeca: Computationally efficient physics-based facial animations. In: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. MIG '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3623264.3624439, https://doi.org/10.1145/3623264.3624439
21. Wagner, N., Schwanecke, U., Botsch, M.: Neural volumetric blendshapes: Computationally efficient physics-based facial blendshapes (2023)

22. Zhang, L., Zhao, Z., Cong, X., Zhang, Q., Gu, S., Gao, Y., Zheng, R., Yang, W., Xu, L., Yu, J.: Hack: Learning a parametric head and neck model for high-fidelity animation (2023)