International Journal of Computer Vision (2024) 132:555–580
https://doi.org/10.1007/s11263-023-01870-2

Building 3D Generative Models from Minimal Data

Skylar Sutherland[1,2] · **Bernhard Egger**[2,3] · **Joshua Tenenbaum**[2]

Received: 3 March 2022 / Accepted: 28 July 2023 / Published online: 13 September 2023

Abstract
We propose a method for constructing generative models of 3D objects from a single 3D mesh and improving them through unsupervised low-shot learning from 2D images. Our method produces a 3D morphable model that represents shape and albedo in terms of Gaussian processes. Whereas previous approaches have typically built 3D morphable models from multiple high-quality 3D scans through principal component analysis, we build 3D morphable models from a single scan or template. As we demonstrate in the face domain, these models can be used to infer 3D reconstructions from 2D data (inverse graphics) or 3D data (registration). Specifically, we show that our approach can be used to perform face recognition using only a single 3D template (one scan total, not one per person). We extend our model to a preliminary unsupervised learning framework that enables the learning of the distribution of 3D faces using one 3D template and a small number of 2D images. Our approach is motivated as a potential model for the origins of face perception in human infants, who appear to start with an innate face template and subsequently develop a flexible system for perceiving the 3D structure of any novel face from experience with only 2D images of a relatively small number of familiar faces.

Keywords Generative models · 3D morphable models · Face recognition · Inverse graphics · Unsupervised learning · Low-shot learning

1 Introduction

3D generative models of objects are used in many computer vision and graphics applications. Present methods for constructing such models typically require either significant amounts of 3D data processed through specialized pipelines, substantial manual annotation, or extremely large amounts of 2D data (Chaudhuri et al., 2020; Egger et al., 2020). We explore a novel approach that could provide a means to build generative models from very limited data: a single 3D object template (such as a single 3D scan of a face, or the average face in some population, or simply a hand-built coarse face mesh). Our initial model is built using simple preprogrammed heuristics. We then show that it can be improved using an unsupervised wake-sleep-like algorithm which learns statistical distributions of objects based on 2D observations, without relying on pretrained networks for feature point detection or face recognition. The models we build are 3D morphable models (3DMMs) ((Blanz & Vetter, 1999; Egger et al., 2020)), a type of generative model which creates samples by applying randomized shape and albedo deformations to a reference mesh. Traditionally, 3DMMs (e.g. (Paysan et al., 2009; Gerig et al., 2018; Li et al., 2017; Booth et al., 2018)) are built through principal component analysis (PCA) applied to datasets of 50 to 10,000 3D meshes produced by specialized (and expensive) 3D scanners (Egger et al., 2020). Furthermore, a registration step is required to align the scans to a common topology. In contrast, we use only a single scan or template, and so can eschew registration, an intrinsically ill-posed problem.

Our approach uses the provided scan as our generative model's mean and smoothly deforms the scan as a surface in physical (3D) space and color (RGB) space using Gaussian processes. Our shape deformation model follows that of Lüthi

Communicated by Matteo Poggi.

✉ Bernhard Egger
bernhard.egger@fau.de

[1] Department of Psychology, Yale University, 1 Hillhouse Ave, New Haven, CT 06511, USA

[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[3] Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstraße 11, 91058 Erlangen, Bavaria, Germany

et al. (2017). We define the albedo deformations by combining analogous smooth albedo deformations on the mesh with smooth deformations on the surface defined by considering the mesh as a shape in RGB-space, with each vertex's location determined by its albedo rather than its position. We initially define very generic Gaussian processes and add domain-specificity through correlation between color channels and bilateral symmetry. Our models are fully compatible with PCA-based 3DMMs; the only difference is that our models' covariances are constructed through Gaussian processes rather than PCA. As our 3DMMs use the same format and support the same operations as PCA-based 3DMMs, they can be used in existing pipelines to perform downstream tasks. They can additionally be used to augment PCA-based 3DMMs (Lüthi et al., 2017).

This is, to the best of our knowledge, the most data-efficient procedure currently extant for constructing 3D generative models, and the sole procedure that only uses a single datapoint. While the performance of our models is significantly poorer than that of PCA-based 3DMMs, they nevertheless perform surprisingly well given their data-efficiency. While 3DMMs are a common prior in computer vision systems, their scalability is limited because their synthesis involves careful capture and modeling with category-specific domain knowledge. Our method's data-efficiency obviates the need for large amounts of data capture, while our method's generality enables its use for any object class. Finally, our approach minimizes the amount of sensitive personal data required to construct face 3DMMs.

We also prototype an extension of our single-scan approach to a multi-scan setting by constructing mixture models of separate single-scan 3DMMs. This can be seen as a generalization of kernel density estimation (KDE). While performing inference with such a mixture model is more computationally expensive than with a PCA-based 3DMM, we demonstrate that the reconstruction quality obtained is much higher if the number of scans used in the models is low. Furthermore, constructing this type of KDE-based model does not require correspondence between scans.

In addition to extending our approach to multi-scan settings, we prototype a method for extending our single-scan 3DMMs on the basis of unsupervised low-shot learning from 2D images. We do this by using our single-scan 3DMMs to perform analysis-by-synthesis (Yuille & Kersten, 2006) on the 2D images, yielding a dataset of 3D reconstructions. A new 3DMM can then be produced from this dataset through PCA. Analysis-by-synthesis is performed using a four-stage pipeline: first, a CNN trained on synthetic data generated by the 3DMM is used to regress pose and lighting; second, a Markov chain Monte Carlo (MCMC) method is used to reconstruct the object within the single-scan 3DMM's eigenspaces; and third, a shape-from-shading strategy using the 3DMM as a source of regularization is used to reconstruct fine details, finally a new model is learned from the inference results. We demonstrate in the face domain that this approach can greatly improve the 3DMMs' visual quality using only a few hundred images. Furthermore, we do not implicitly rely on supervision in the form of pre-trained feature-point detectors or face recognition networks trained on labeled data; rather, we also bootstrap face alignment during learning. Our approach is therefore unsupervised besides the single template and the heuristics that generate the initial model.

We believe this approach to enhancing our 3DMM through the incorporation of unlabeled 2D data has applications not just for computer vision but also as a potential computational model of the development of face perception in infants. The visual perception of infants is an area of key interest in cognitive science (Kellman & Arterberry, 2007) which is typically studied through psychophysical experiments, and rarely through computational models. Building such a computational model was the initial motivation behind this project and we show that learning a statistical model is feasible in this way but we do not provide any further evidence that human visual development works in such a way. There are several theories explaining the development of face-selective areas in the visual system and the preference very young infants have for faces (Slater et al., 1998). One theory posits an innate subcortical face template, with the average face a likely candidate for such a template (Powell et al., 2018). Studies of imitation in infants suggest some sort of basic face model might be present at birth (Meltzoff & Moore, 1989). Psychophysical experiments with aftereffects in face perception also indicate the presence in the brain of a linear model of the space of faces that 3DMM's can quantitatively model (Leopold et al., 2001; Egger et al., 2020). Furthermore, the adult brain's representation of face space seems to be refined over the course of development (Valentine et al., 2016). Recent neurological analysis of the face system in macaque monkeys supports analysis-by-synthesis using a 3DMM as a plausible underlying mechanism of face perception (Yildirim et al., 2020). There is also strong evidence that human face perception improves drastically over the course of long-term development; despite the strong improvements in early vision, our face perception capabilities grow substantially in adulthood and peak in the 30 s (Germine et al., 2011).

In the context of human cognitive development, our research also seeks to identify the minimal inductive biases that learning systems require to derive a 3DMM-based face perception framework from 2D images through unsupervised learning. Our proposed unsupervised approach to learn a 3DMM is based on an inverse rendering framework along with a minimal 3D seed (which we demonstrate can be simpler than a full face scan) which we use to produce a weak generative model. We argue that these mechanisms are

plausibly innate in the human brain, with the average face representing a minimal innate template whose existence is indicated by infant experiments (Powell et al., 2018). We then demonstrate that our framework can learn a rich 3DMM from 2D data. This is the first fully unsupervised method for learning a 3DMM from 2D data, and the resulting model reaches broadly similar quality to existing 3DMMs learned from 2D data in a highly supervised manner.

Although we demonstrate the applicability of our approach to other object categories, we focus our experiments on faces. This is mainly because 3DMMs have historically been built for face modeling, so we can better compare our models to prior work in a face setting, and do so through well-established pipelines. Our results with other object categories are harder to interpret, since our method is unique not only in its ability to generalize from a single datapoint, but also in its flexibility of object category. However, our paper's methods may have their greatest relevance in domains outside of face perception since in the face domain high-quality 3DMMs built from 3D data are already widely extant.

The main contributions of this work are the following:

1. We offer a novel albedo deformation model by combining surface-based and color-space-based kernels.
2. We introduce a framework for 3DMM construction from a single 3D scan by extending an existing framework to build statistical shape models (Lüthi et al., 2017) with our albedo deformation model.
3. We evaluate our model on three downstream tasks, namely inverse rendering (2D to 3D registration), face recognition, and 3D to 3D registration, as well as the direct quality measures of specificity, generalization against compactness (Styner et al., 2003). We compare its performance with that of the 2019 (or, where relevant, 2017) Basel Face Model (Gerig et al., 2018), a state-of-the-art 3DMM produced from 200 3D scans.
4. We build a prototype KDE-based face model from 10 face scans, and demonstrate that on a face recognition task it outperforms a PCA-based 3DMM built from the same 10 scans.
5. We extend our framework on the basis of unsupervised low-shot learning from 2D images to enrich our simple model with image observations, demonstrating the feasibility of fully unsupervised learning of statistical 3DMMs.

This paper is organized as follows: We first review the related literature in Sect. 2, followed by the methods part. In the methods part in Sect. 3 we introduce the basic ideas of our generative model including the design of our shape and albedo covariance kernels in Sects. 3.1 and 3.2. We then propose the idea how such kernels could potentially be used in a kernel density-based generative model in Sect. 3.3. In Sect. 3.4 we then dive into the details how we learn and improve on our simple shape and albedo models based on few 2D observations. We provide an overview over all involved hyperparameters in Sect. 3.5. In the experiments Sect. 4 we show inverse rendering results and present the quality of the learned model using qualitative and quantitative measures.

2 Related Work

Our methods combines ideas from diverse prior work. In the following we summarize the state-of-the-art in the areas most close and relevant. We start with different ideas for statistical shape modeling, focusing on Gaussian Process based models, then summarize how 3D models have previously been learned from 2D data and how they are then applied to retrieve a 3D reconstruction from a 2D image. Finally we show similarities of ideas in our work to shape-from-template based approaches.

2.1 Gaussian Processes for Shape Modelling

The idea of building an axiomatic shape deformation model using Gaussian processes was previously explored in Lüthi et al. (2017), which used such a deformation model as a prior for 3D registration tasks. We extend this approach to include albedo along with shape by building Gaussian processes in RGB-space as well as physical space. This enables its use as prior in an inverse graphics setting, and allows us to take albedo into account during registration. Kemelmacher-Shlizerman and Basri (2010) presented a method for the 3D reconstruction of faces from 2D images through axiomatic deformation of a single 3D scan. However, unlike our approach, this paper did not produce a generative model, and performed 3D reconstruction through shape-from-shading rather than probabilistic inference, using the 3D scan as purely as regularizer. Tegang et al. (2020) applied a Gaussian process intensity model in medical imaging for co-registration of CT and MRI images and for data augmentation. Other shape representation strategies (e.g. (Kilian et al., 2007)) incorporate geodesic distances instead of Euclidean distances; while geodesic distances are beneficial in modeling motion and expression, since they are not easily transferable to color spaces we here focus on Euclidean distance. Ovsjanikov et al. (2011) proposes a method for modeling variability in 3D datasets without correspondence by deforming a single template mesh. However, unlike our work, Ovsjanikov et al. (2011) learns a nonlinear deformation model from a significant number of (unregistered) 3D scans through dimensionality reduction techniques, and so is inapplicable given only a single scan. Furthermore, they only study 3D-to-3D reconstruction and it is unclear how their approach could be applied in a computer vision setting.

2.2 Learning 3DMMs from 2D Data

While classically 3DMMs have been built from a collection of 3D scans, there are also several approaches that start from 2D data or combine 2D and 3D data. Building a 3DMM solely from 2D data was first explored by Cashman and Fitzgibbon (2012). Although they, like us, also start from a 3D mean shape as an initial template, their work neglects albedo. Recently, methods to improve 3DMMs through 2D observations were proposed (Tewari et al., 2018; Tran & Liu, 2019). While they seek to build 3DMMs from 2D data, their approaches start with a full 3DMM built from 3D scans, and primarily refine the appearance model to increase flexibility. Neither method offers a way to derive this initial model other than capturing 3D data and establishing correspondence between scans. Tran et al. (2019) further extended these ideas to incorporate nonlinear models so as to overcome the limitations inherent in the linearity of classical 3DMMs. In contrast to these works aiming to build a 3DMM from a large collection of 2D data and an initial 3DMM, our work focuses on building a 3DMM from just a single 3D scan. Such a model could be used as an initial model for the 2D learning strategies discussed above.

Additionally, some recent work has focused on the problem of the unsupervised learning of 3D generative models from a large 2D training corpus (Szabó et al., 2019; Wu et al., 2020) or from depth data (Abrevaya et al., 2018). The 3D generative models learned by these approaches do not disentangle illumination and albedo (or neglect albedo entirely, as in Abrevaya et al. (2018)), and do not preserve correspondence, making them difficult to interpret. Furthermore, this means that they are incompatible with existing 3DMM-based pipelines; in contrast, generative models produced through our approach can be used interchangeably with PCA-based 3DMMs.

Tewari et al. (2020) was the first to propose a complete 3DMM learned from 2D images and video data through self-supervised learning, using an average 3D face for initialization. This paper is more directly comparable to our work. However, we show that the average face is already sufficient to produce a usable 3DMM, without any 2D data. While we do prototype extensions of our 3DMMs on the basis of 2D data, we do so using far less data than Tewari et al. (2020): we use only static images, use many orders of magnitude fewer images, and do not rely on pre-trained face detectors, feature point detectors or face recognition networks. Their model however also incorporates facial expressions through video supervision which are omitted in our proof of concept.

Other works have focused on extending 3D morphable models beyond a linear latent space (Ranjan et al., 2018; Bouritsas et al., 2019; Tran et al., 2019). In contrast, we use a traditional linear latent space and rather focus on how such latent spaces can be learned. For additional work on applications of 3DMMs and shape and albedo representations used with 3DMMs, we refer to Egger et al. (2020).

2.3 3D from 2D Through 3DMMs

Our analysis-by-synthesis method presented in Sect. 3.4 is closely related to a number of prior works. The first stage of our method, a CNN trained on 3DMM-generated synthetic data for pose and lighting regression, is similar to the Efficient Inverse Graphics network of Yildirim et al. (2020) and previous work on regressing 3DMM parameters directly from images (Tuan Tran et al., 2017). The second stage, an MCMC method for shape and albedo regression within a 3DMM's shape and albedo subspaces, is similar to the MCMC method presented in Schönborn et al. (2017) (and we use Schönborn et al. (2017)'s method directly in other parts of the paper). The third stage, a 3DMM-regularized shape-from-shading strategy, is loosely similar to the approaches of Kemelmacher-Shlizerman and Basri (2010) and Patel and Smith (2012); however, the specific combination of these approaches, and their use case, is original to our paper.

2.4 Shape-from-Template Approaches

In addition to 3D morphable models, our work can also be connected with *shape-from-template* approaches to 3D vision. These approaches typically address the following problem: given a reference mesh, an input image, and a set of dense (pixel-level) correspondences between the input image and a rendering of the mesh (Bartoli et al., 2012; Östlund et al., 2012; Brunet et al., 2011; Malti et al., 2011; Moreno-Noguer et al., 2010; Salzmann, & Fua, 2011) or with the mesh directly (Moreno-Noguer et al., 2009; Salzmann et al., 2008), deform the mesh to match the input image. Restrictions on the allowed deformations (e.g. isometry or conformality) make this problem well-posed and sometimes solvable analytically. This framing means that shape-from-template approaches are rarely applicable without dense 2D correspondence annotations and generally ignore albedo. Shape-from-template approaches that do not require dense 2D correspondence have typically previously relied on additional 3D or video data and still do not fully model albedo (Yu et al., 2015; Salzmann et al., 2008; Shaji et al., 2010). In contrast, our approach can infer 3D reconstructions from single images reliably using only a small set of landmarks (sparse 2D correspondence) for localization, and our unsupervised learning approach is capable of fully unsupervised (albeit much less reliable) 3D face reconstruction. We furthermore separate albedo and illumination and fully incorporate albedo in our deformations. Furthermore, shape-from-template approaches have no way to incorporate statistical information about the distribution of 3D objects likely to be observed, whereas we demonstrate our approach can incorporate statistical learning.

This work extends our previous conference paper to incorporate unsupervised learning (Sutherland et al., 2021).

3 Methods

A 3DMM consists of a shape model and an albedo model; samples from a 3DMM are meshes with a common topology, with the position and albedo of each vertex generated by the 3DMM's shape and albedo models, respectively (Egger et al., 2020). Our framework represents samples from the shape and albedo models as deformations of a vertex-colored mesh that defines both the topology of all samples and the mean of the shape and albedo distributions. Our approach uses the 3D scan as the mean of the resulting 3DMM. We define the shape and albedo models in terms of Gaussian processes, each consisting of a mean and a covariance kernel (Lüthi et al., 2017; Rasmussen, 2003). While this method's performance depends on the choice of mean mesh, PCA-based 3DMMs face the same issue since registration likewise requires a choice of common topology.

We define a Gaussian process g as a pair (μ, Σ), where μ is the mean of the Gaussian process and Σ is the covariance kernel of the Gaussian process; μ is a function from A to $\mathbb{R}^n$ for some set A and constant n, and Σ is a positive-definite function from A^2 to $\mathbb{R}^{n\times n}$, where $\mathbb{R}^{n\times n}$ is the space of n-by-n matrices. In our case, for both the shape and albedo models, A is the set of mesh vertices, and $n = 3$. A sample from the shape model maps A to positions in $\mathbb{R}^3$, whereas a sample from the albedo model maps A to RGB values, represented as vectors in $\mathbb{R}^3$. We represent our shape and albedo kernels using Mercer decomposition computed through the Nyström method for computational efficiency to make the calculation of tractable, for details we refer to Rasmussen (2003); Lüthi et al. (2017).

3.1 Shape Covariance Kernels

We follow the approach of Lüthi et al. (2017): defining covariance kernels which give a high correlation between nearby points and a low correlation between distant points. The most straightforward way to do this is with physical distance. Our shape kernels are based on radial basis function kernels (Rasmussen, 2003; Lüthi et al., 2017).

A function $f : A^2 \to \mathbb{R}$ is positive-definite if the matrix M defined by $M_{i,j} = f(x_i, x_j)$ is positive-semidefinite for any $x_1, \ldots, x_n \in \mathbb{R}$ Mercer (1909). This definition can be extended to matrix-valued kernels by letting $M_{i,j}$ represent a block submatrix of M instead of an entry of M Rasmussen (2003); Lüthi et al. (2017). Since the set of positive-semidefinite matrices is closed under addition and positive scalar multiplication (Horn, 2012), so are matrix-valued kernels. In order to create a kernel with a coarse-to-fine structure, possessing strong short-range correlations and weaker long-range correlations, we define our shape kernel as a linear combination of radial basis function kernels. Letting $\Sigma_{s,\sigma}$ represent the radial basis function kernel (subscript s for shape) defined using physical distance as its metric and scale σ (in millimeters), we define the family of scalar kernels $\Sigma_{\mathtt{std}}(a, b, c, A, B, C) = a\Sigma_{s,A} + b\Sigma_{s,B} + c\Sigma_{s,C}$. We here let $\Sigma_0 = \Sigma_{\mathtt{std}}(a_s, b_s, c_s, A_s, B_s, C_s)$, where a_s, b_s, c_s, A_s, B_s, and C_s are hyperparameters (listed in Sect. 3.5). For an intuition regarding our shape covariance kernels and their combination we refer to Fig. 1 as well as for the mathematical details to Lüthi et al. (2017).

In order to represent 3D deformations, we must multiply scalar kernels by 3-by-3 matrices. Since we wish for deformations in x, y, and z to be uncorrelated, we simply multiply by I_3, the 3-by-3 identity matrix. Thus, our standard shape kernel is $K_s = I_3\Sigma_0$. One limitation of this kernel is that it does not encode bilateral symmetry. Many object categories, including faces, are bilaterally symmetric. In order to add symmetry to this kernel, we wish to make the deformations applied to points on opposite sides of the object closely correlated in the up-down and forward-back axes and strongly anticorrelated in the left-right axis (Morel-Forster, 2016). To define such kernels, let $\Phi_m \in \mathbb{R}^{3\times 3}$ be the matrix which, considered as a linear transformation applied to points in physical space, negates a point's left-right component (where left and right are defined relative to the scan). Then our symmetric shape kernel is defined as $K_s^{\mathtt{sym}} = I_3\Sigma_0(x, y) + \alpha\Phi_m\Sigma_0(x, \Phi_m(y))$, where $\Phi_m(y)$ denotes applying Φ_m as a linear transformation to y's position in $\mathbb{R}^3$, and α is a hyperparameter (listed in Sect. 3.5) Morel-Forster (2016).

3.2 Albedo Covariance Kernels

What we principally desire in an albedo kernel is that deformations applied to different areas should be highly correlated if and only if the areas are related. Unlike shape deformations, albedo deformations in general need not be spatially continuous, and so a global notion of similarity is needed in addition to physical proximity. We measure the similarity of mesh vertices by combining their distance in physical space with their distance in albedo space.

Physical distance is a straightforward way of assessing similarity. We define a physical distance-based albedo kernel similarly to K_s. Specifically, we define $K_{a,\mathtt{xyz}} = I_3\Sigma_{\mathtt{xyz}}$, where $\Sigma_{\mathtt{xyz}} = \Sigma_{\mathtt{std}}(a_a, b_a, c_a, A_a, B_a, C_a)$, with hyperparameters a_a, b_a, c_a, A_a, B_a, and C_a listed in Sect. 3.5 (subscript a for albedo). A_a, B_a, and C_a are interpretable since they are again in millimeters. Samples from $K_{a,\mathtt{xyz}}$ represent deformations in RGB-space, not position. However, this kernel neglects some kinds of similarity. For instance, in a human face, a point on a lip is more similar to another point on

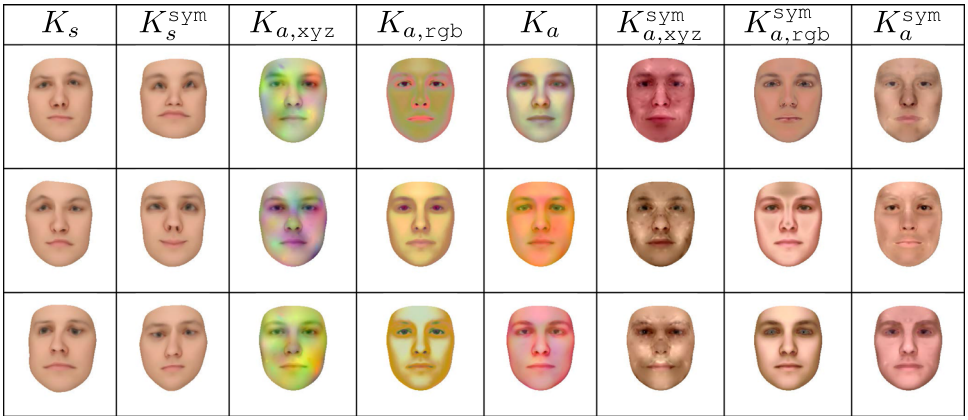

Fig. 1 Three random samples from each of the shape and albedo kernels applied to the mean of the 2019 Basel Face Model (Gerig et al., 2018) and rendered under ambient illumination. The first two columns are the two shape kernels, while the remaining eight columns are the albedo kernels

a lip than it is to an equidistant point on a cheek; more generally, many objects exhibit part-based similarity in addition to distance-based similarity. Color-space distance provides us with an estimate of part-based similarity that does not depend on explicit part annotations. Just as the distances between mesh points in physical space (in the mean) constitute a metric on the set of mesh points, so do the Euclidean distances between mesh points' albedos, represented as RGB values and considered as points in $\mathbb{R}^3$. Using this alternate metric, we may define another family of radial basis function kernels, which we term $\Sigma_{a,\sigma}$ for $\sigma \in \mathbb{R}$. We then define the alternate albedo kernel $K_{a,\mathrm{rgb}} = I_3 \Sigma_{\mathrm{rgb}}$, where $\Sigma_{\mathrm{rgb}} = d\Sigma_{a,D}$, with hyperparameters d and D (listed in Sect. 3.5), where D correspond to RGB intensity values between [0, 1]..

To use both local and global information, we average these kernels. A core contribution here is the combined kernel $K_a = 0.5(K_{a,\mathrm{xyz}} + K_{a,\mathrm{rgb}})$. This kernel takes into account both the differences in position and differences in albedos between points on the mesh, and can thus relatively robustly assess whether different parts of the object are parts of the same component.

As stated, all three of our albedo kernels are products of a scalar-valued kernel with I_3. Multiplying by a different matrix enables us to incorporate domain knowledge about an object category's common albedos by correlating the different color channels (red, green, and blue). In particular, as a very rough approximation to human skin tones, we introduce additional kernels $K_{a,\mathrm{xyz}}^{\mathrm{cor}}$ and $K_{a,\mathrm{rgb}}^{\mathrm{sym}}$, depending, respectively, on physical and RGB-space distance. Letting

$$M_x = \begin{bmatrix} 1 & x & x \\ x & 1 & x \\ x & x & 1 \end{bmatrix} \tag{1}$$

we define $K_{a,\mathrm{xyz}}^{\mathrm{cor}} = M_\beta \Sigma_{\mathrm{xyz}}$ and $K_{a,\mathrm{rgb}}^{\mathrm{sym}} = M_\gamma \Sigma_{\mathrm{rgb}}$, where β and γ are hyperparameters (listed in Sect. 3.5).

To add further domain knowledge we create additional albedo kernels that incorporate bilateral symmetry. The idea behind this is that color of mouth, eyes or the cheeks will likely change following bilateral symmetry, the effect of the proposed symmetry kernels is depicted in Fig. 1. Since the albedo of a member of a bilaterally symmetric object class is essentially bilaterally symmetric, $K_{a,\mathrm{rgb}}^{\mathrm{sym}}$ is already symmetric in practice. However, the physical-distance-based albedo kernels can be symmetrized via a process analogous to that used for the shape kernels in Sect. 3.1, with the difference that we do not wish to negate left-right deformations located on opposite sides of the object. We choose to consider color channel correlations and symmetry simultaneously, and so define $K_{a,\mathrm{xyz}}^{\mathrm{sym}}(x, y) = K_{a,\mathrm{xyz}}^{\mathrm{cor}}(x, y) + \alpha K_{a,\mathrm{xyz}}^{\mathrm{cor}}(x, \Phi_m y)$, and define $K_a^{\mathrm{sym}} = 0.5(K_{a,\mathrm{rgb}}^{\mathrm{sym}} + K_{a,\mathrm{xyz}}^{\mathrm{sym}})$.

To attempt to separate the roles played by symmetry and color-channel correlation, in the Appendix we also present results with albedo kernels that have correlated color channels but lack symmetry.

In Fig. 1, we show samples from our various shape and albedo kernels, applied to the mean of the 2019 Basel Face Model (Gerig et al., 2018). While these samples are clearly non-naturalistic, this does not invalidate the results of Section 3 of our main paper. We make no claim that our initial 3DMMs based on those kernels accurately model the distribution of human faces; rather, we claim that they are of sufficient quality to be useful in a machine vision context and that we can improve their quality based on a few 2D observations.

3.3 Kernel Density Estimation

One limitation of our approach is that it provides no way to leverage the information present in multiple scans. However, an extension of our approach can be used in a setting where multiple scans are available. To construct a model from multiple scans, we create single-scan 3DMMs for each scan separately, and those individual models together then represent a mixture model. This joint model over different 3DMMs built from each individual scan results in a non-parametric 3DMM-based model. This essentially amounts to an exten-

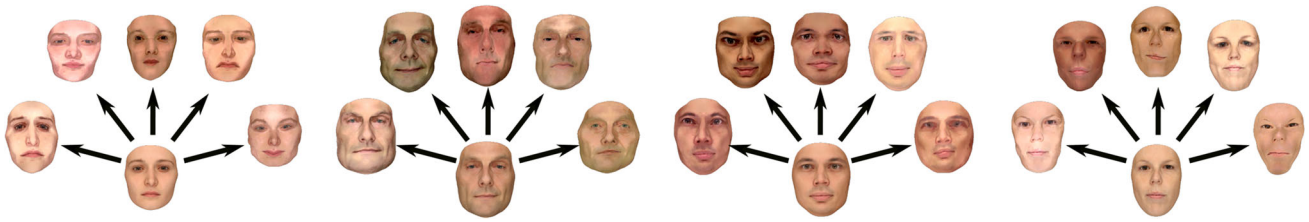

Fig. 2 Visualization of our KDE-based model. It consists of multiple independent 3DMM-based models which jointly build the KDE model. We draw the mean for each model in the center and resulting model samples around each individual single-scan 3DMM. The samples for the indidivual models are samples from the joint model. To perform inference we have to perform inference for each individual 3DMM to then pick the one where the reconstruction would be most likely

sion of kernel density estimation (KDE), where Gaussian processes replace uniform Gaussian distributions in the definitions of each mixture component, providing a non-uniform noise model. The idea is visualized in Fig. 2.

An advantage of this kernel density estimation approach is that, unlike PCA, it does not require dense correspondence between scans. This could enable the creation of 3DMM-based generative models of object categories where many 3D scans exist but where establishing dense correspondence is impossible (e.g. chairs). However, the non-parametric nature of a KDE-based model means that, unlike a PCA-based 3DMM, the amount of computation required to perform inference grows with the number of scans as in practice we need to fit every single model to the target image. In case of 10 scans the computational complexity increases by a factor of 10 which makes large models unfeasible without novel inference techniques that could handle such a model in a smarter way. We demonstrate the potential of this idea on a face recognition task in Sect. 4.1.

3.4 Learning from 2D Data

The single-scan 3DMMs constructed above, while usable in some downstream tasks, remain very far from object categories' true distributions. In the face domain, we further demonstrate how our 3DMMs can be augmented through unsupervised low-shot learning from 2D data. This is done by first producing 3D reconstructions of faces from 2D images through inverse graphics using the initial single-scan 3DMM, and then applying PCA to the resulting dataset. This cannot be done using the analysis-by-synthesis method of Schönborn et al. (2017) (which we use in Sect. 4) for two reasons. Firstly, Schönborn et al.'s method relies on manual landmark annotations. Secondly, it can only produce reconstructions which lie within the support of the single-scan 3DMM, and applying PCA to a dataset of such reconstructions will yield a 3DMM whose support is equal to or a subset of that of the single-scan 3DMM. For these reasons we produce reconstructions using a new unsupervised analysis-by-synthesis method, which is outlined below.

We produce 3D face reconstructions from 2D images based on an algorithm in a wake-sleep style (Hinton et al., 1995). The algorithm is originally motivated by human learning: during the day or when awake we collect observations, we however not only collect but also process them. During the night or when sleeping we further process what we have seen and might update some models based on experience during the wake phase. Our particular implementation is as follows: in the wake phase we process new observations with a fixed feed-forward network and with a fixed model, in the sleep phase we update the model and retrain the feed-forward network with samples from this model. We split our wake and sleep phases into a four-stage inference process. The first and last stage are part of the sleep cycle of the algorithm, the second and third stage are part of the wake cycle. Steps that are performed when observing new data are part of the wake cycle (inference), steps that do training, updating or finetuning based on previous observations are part of the sleep cycle. In our particular implementation the inference network and the model is fixed during the wake phase and does not update, during the sleep phase we are not processing new input and are only processing the results from the observations we had during the day to update our underlying model and inference (CNN, but also MCMC through new model). In the following, we explain the individual components in more detail and mention why they are part of the wake or sleep cycle.

First, as illustrated in Fig. 3, a convolutional neural network (CNN) trained on synthetic data regresses the face position and orientation and scene lighting, similarly to Yildirim et al. (2020) but with non-face scene parameters regressed rather than a 3DMM's principal components. We use a ResNet50-v2 network (He et al., 2016) which was pretrained on ImageNet (Deng et al., 2009). The network is trained with synthetic data that incorporates variation in both shape and albedo as well as camera and illumination parameters. For shape and albedo we sample according to the prior from the respective Gaussian Process Model, for the pose we sampled the full range from -90 to $90°$ of yaw, -30 to $30°$ for pitch and -60 to $60°$ for roll and illumination parameters we sampled point light sources with random color and position. We sampled 400k images spanning a wide range of poses and illumination conditions. Examples of the training data can be seen in Fig. 3. The aim of this first step is to roughly align the face by estimating the rotation matrix for

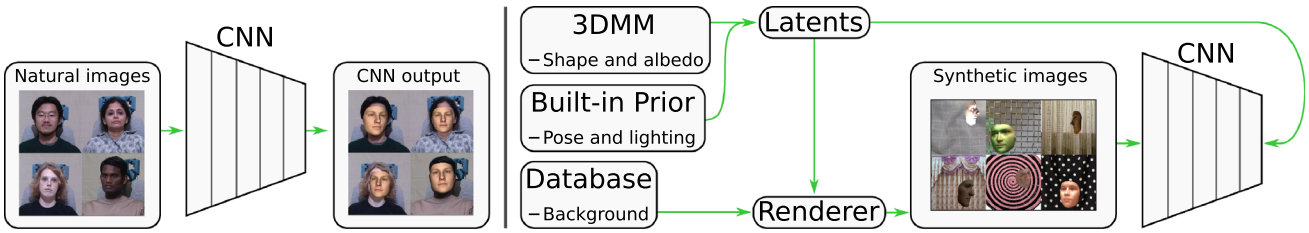

Fig. 3 The first stage of our inference pipeline. A convolutional neural network is trained on synthetic data rendered from our single scan model to regress the location and pose of faces in the input images

pose, x and y position, scale and estimate the illumination. The network is trained in a fully supervised fashion applying an L2 loss on parameters scaled to have uniform variance.

This stage of the algorithm is part of the sleep stage, as we pretrain or refine our network based on a fixed model. There are no new observations or data fed at this sleep stage of the algorithm. In the following two stages two and three there are new observations fed into the network but the inference pipeline and model remains fixed—those are part of the wake cycle.

Next, as illustrated in Fig. 4, these initial estimates of the camera and illumination parameters parameters are used to initialize a MCMC process broadly similar to that of Schönborn et al. (2017) (without landmarks and with a somewhat different proposal distribution) that produces a 3D reconstruction of the face's shape and albedo within the 3DMM's subspaces while also inferring lighting and refining the estimated pose. The precise hyperparameters used in the proposal generation distribution are slightly altered due to the use of a different underlying computational framework, and unlike in the case of Schönborn et al. (2017), we incorporate a canonical prior on pose (beta distributions between -90 to $90°$ of yaw, -30 to $30°$ for pitch and -60 to $60°$ which prefer frontal poses) and lighting parameters (uniform distribution for light direction in front of the face and normal distributions for color components centered around ambient, which are also used to generate synthetic training data). Such additional priors are incorporated using likelihood functions in the same fashion as the existing priors for shape and albedo proposed by Schönborn et al. (2017). However, the basic structure of Schönborn et al. (2017)'s proposal distribution—a coarse-to-fine mixture model of Gaussian drift hypotheses—is preserved. We use initial n_1 initial MCMC steps to estimate the lighting parameters only, followed by n_2 MCMC steps to estimate the other parameters (along with refined lighting parameters). We perform a rough estimation of the lighting parameters as lighting is dominating appearance and this helps to guide inference in the right direction. n_1 and n_2 are hyperparameters listed in Sect. 3.5.

Third, as illustrated in Fig. 5, a shape-from-shading strategy is used to reconstruct the face's fine details outside of the 3DMM's shape and albedo subspaces. Since shape-from-shading is an ill-posed problem, shape-from-shading approaches inherently require some source of regularization (Zhang et al., 1999). We use our 3DMM as a source of regularization, penalizing reconstructions both based on their distance from the 3DMM's shape and albedo subspaces, and the probability the 3DMM assigns the reconstructions' projections into those subspaces. The probability of the shape in the 3DMM model is derived from the probabilistic interpretation of PCA which enables to use the model as a prior. We assume the illumination estimation of the MCMC inference to be correct and optimize for the normals and albedo of the surface. The 3DMM prior is applied to both, normals (shape) and albedo. Optimization is performed using gradient descent, using n_3 sequential gradient descent steps, where n_3 is a hyperparameter listed in Sect. 3.5. This makes the shape-from-shading process similar to maximum *a posteriori* optimization using differentiable rendering, and is based on to the approaches of Kemelmacher-Shlizerman and Basri (2010) and Patel and Smith (2012).

Finally in the fourth step, once a dataset of detailed 3D reconstructions has been produced, a new 3DMM is constructed by applying PCA to this dataset. This part is again part of the sleep cycle of the algorithm as the model is updated. After this step the learning process repeats in the next iteration where in the first stage the inference pipeline is updated to reflect the new model. Before PCA is applied, a denoising step is applied to the shape of each mesh to remove any spikes introduced by the shape-from-shading process, and an alignment step using the algorithm of Umeyama (1991) is applied. Once a new 3DMM has been produced, the CNN used to initialize pose and lighting estimation is finetuned through retraining on new synthetic data. The new synthetic dataset used to train the CNN does not consist solely of samples from the newly constructed 3DMM, but rather is a mixture of samples from all the 3DMMs produced throughout the learning process. Improvements in the 3DMM also improve the accuracy of the MCMC and shape-from-shading steps, since these are both model-in-the-loop processes.

Quality control at each of the steps in our framework is essential for the learning of a 3DMM. A single bad reconstruction can introduce significant errors and severe artifacts into the learned model. We therefore implemented simple quality control heuristics that ensure that only the very best

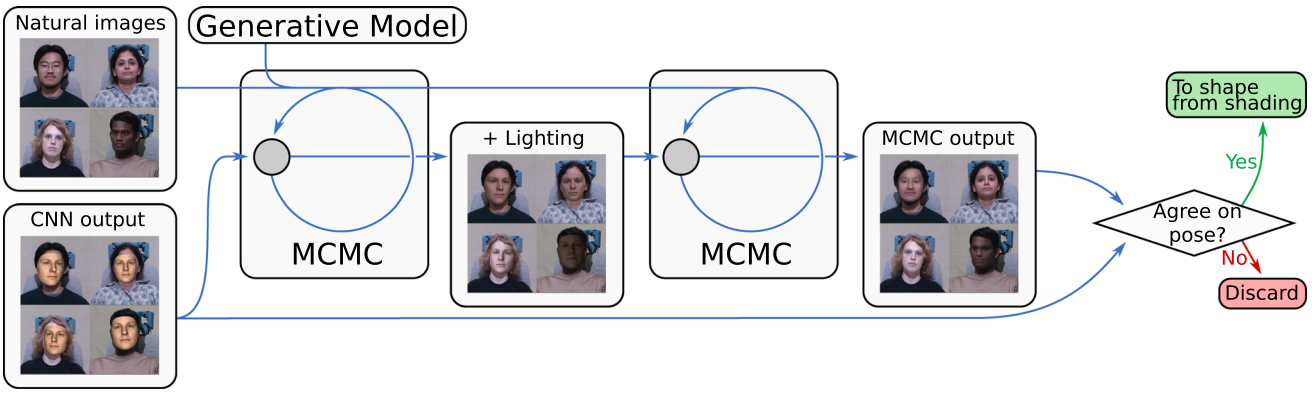

Fig. 4 The second stage of our inference pipeline. A Markov chain Monte Carlo (MCMC) process is used to infer the shape and albedo of the face along with environmental illumination, as well as refine the regressed pose. If the MCMC process does not approximately preserve the 2D silhouette of the rendered face the face reconstruction is discarded

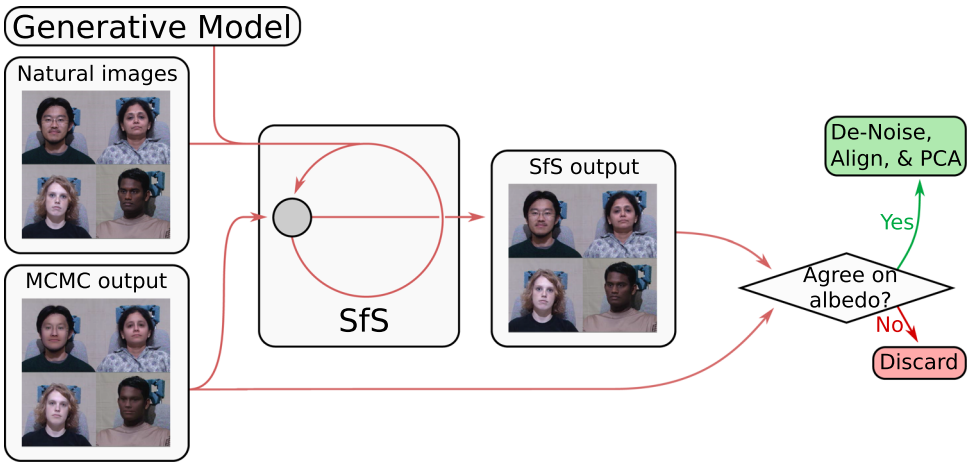

Fig. 5 The third stage of our inference pipeline. A shape-from-shading strategy is used to infer the fine details of the shape and albedo of the face along with refined environmental illumination. Once the shape-from-shading process is complete, denoising and alignment steps are applied as post-processing. If the shape-from-shading process does not approximately preserve albedo, the reconstruction is discarded

reconstructions end up in the resulting model. Misalignments of the face in the image by the CNN can yield highly inaccurate MCMC reconstructions, and similar misalignments during the MCMC process can yield gross errors in the shape-from-shading reconstruction. For this reason, we discard poor fits during the process using simple heuristics. Specifically, if the MCMC process does not approximately preserve the 2D silhouette of the face reconstruction, or if the shape-from-shading strategy does not approximately preserve the albedo of the face reconstruction, the reconstruction is discarded as a probable failure. We would now like to describe those two heuristics in more detail: The quality control heuristic for the second stage aims at the end of the MCMC process, the 2D silhouettes of the rendered fits are compared and if the ratio of pixels in both silhouettes to pixels in either silhouette is not at least r_1 (a hyperparameter), the fit is discarded. The basic intuition behind this quality metric is that the CNN and MCMC have to agree on the rough pose, otherwise the result is rejected. Similarly as quality control heuristic after the third stage of the algorithm, at the end of the shape-from-shading process, we compute the average distance between the albedo of each vertex in the mesh before and after the shape-from-shading process, and if this exceeds a threshold $r_2 - nr_3$ the fit is discarded, where r_2 and r_3 are hyperparameters and n is the number of previously performed wake-sleep iterations. The intuition behind this step is that the shape-from-shading estimate has to be somewhat similar to the MCMC estimate—otherwise something went wrong and we reject the result. While the precise number of reconstructions that pass these quality control steps varies, it is generally very low, and the quality control steps can be viewed as selecting only very best reconstructions to ensure that the model is only built from good reconstructions. We chose conservative threshold values as a few bad fitting results can lead to a bad new model and the whole model learning can diverge.

3.5 Choices of Hyperparameters

We present an overview over our hyperparameters in Table 1.

Importantly, these parameters have generally not been extensively tuned, which is reflected in the fact that we use

Table 1 Hyperparameters used for our method

Parameter	Value	Unit	Description
a_s	7	–	Amplitude of coarse shape kernel
b_s	5	–	Amplitude of medium shape kernel
c_s	3	–	Amplitude of fine shape kernel
A_s	100	Millimeters	Bandwidth of the coarse shape kernel
B_s	50	millimeters	Bandwidth of the medium shape kernel
C_s	10	millimeters	Bandwidth of the fine shape kernel
a_a	0.02	–	Amplitude of coarse albedo kernel in shape space
b_a	0.01	–	Amplitude of medium albedo kernel in shape space
c_a	0.01	–	Amplitude of fine albedo kernel in shape space
A_a	500	millimeters	Bandwidth of the coarse albedo kernel in shape space
B_a	20	millimeters	Bandwidth of the medium albedo kernel in shape space
C_a	2	millimeters	Bandwidth of the fine albedo kernel in shape space
d	0.015	–	Amplitude of the albedo in color space
D	0.15	color [0,1]	Amplitude of the albedo in color space
α	0.7	–	Strength of symmetry constraint
β	0.9375	–	Strength of color channel correlation heuristic for albedo kernel in shape space
γ	0.95	–	Strength of color channel correlation heuristic for albedo kernel in color space
n_1	1000	–	Number of MCMC samples to initially estimate lighting
n_2	10000	–	Number of MCMC samples for all parameters
n_3	5000	–	Number of gradient steps for shape-from-shading
r_1	0.625	–	Ratio of silhouette that has to overlap for quality control
r_2	8	–	Initial quality control value for shape-from-shading
r_3	0.5	–	Decay of quality control value for shape-from-shading

the same kernels for faces, birds and fish. Our core idea is simply to combine radial basis function kernels at three different scales and magnitudes so as to incorporate global as well as local flexibility. We are aware that a different set of hyperparameters would very likely lead to better performance, however tuning those parameters would be performed on data and since we would like to build a model based on minimal data we did not tune the parameters beyond having selected different scales. Some of our hyperparameters are interpretable and therefore easy to set by intuition since they are representing physical distances (namely A_s, B_s, C_s, A_a, B_a, and C_a) have units of millimeters. We represent RGB values as points in $[0, 1]^3$, and D, r_2 and r_3 represent distances or magnitudes in color space using this unit system.

4 Experiments

We produce a set of 3DMMs from our kernels using the average face of the 2019 Basel Face Model (Gerig et al., 2018) as our reference mesh. These are listed with their corresponding kernels in Table 2. These 3DMMs have the same mean as the 2019 Basel Face Model, and so comparing their performance with that of the 2019 Basel Face Model constitutes a direct comparison of our axiomatic Gaussian process-based covariance kernels with the learned covariance model of the 2019 Basel Face Model. We also produce 3DMMs by combining our kernels with face scans provided with the 2009 Basel Face Model (Paysan et al., 2009). We assess these models' performance on downstream tasks where 3DMMs are often used, namely inverse graphics (in Sect. 4.1) and registration (in Sect. 4.4), and directly compare these 3DMM's specificity and generalization on real faces with that of the 2017 Basel Face Model (Gerig et al., 2018) in Sect. 4.3. In Sect. 4.2 we show samples and face reconstructions from 3DMMs learned from 2D data using our wake-sleep approach and compare them with the LeMoMo model of Tewari et al. (2020). In

Table 2 Our Gaussian processes for modeling faces

Name	Shape kernel	Albedo kernel
standard-full	K_s	K_a
standard-RGB	K_s	$K_{a,\mathtt{rgb}}$
standard-XYZ	K_s	$K_{a,\mathtt{xyz}}$
symmetric-full	$K_s^{\mathtt{sym}}$	$K_a^{\mathtt{sym}}$
symmetric-RGB	$K_s^{\mathtt{sym}}$	$K_{a,\mathtt{rgb}}^{\mathtt{sym}}$
symmetric-XYZ	$K_s^{\mathtt{sym}}$	$K_{a,\mathtt{xyz}}^{\mathtt{sym}}$

Fig. 6 Reconstructions produced from natural images using various 3DMMs. The first row shows the natural images used as input, while the remaining rows show the reconstructions inferred using different 3DMMs. The standard-full and symmetric-full models were produced using the mean of the 2019 Basel Face Model (Gerig et al., 2018) as template

Sect. 4.5 we experiment with simple 3DMMs of birds and fish.

4.1 Inverse Rendering

One of the most direct ways to assess the value of our model is to apply it in an analysis-by-synthesis setting (Yuille & Kersten, 2006). Using our 3DMMs as priors on 3D meshes, we can perform inverse rendering to reconstruct 3D meshes from 2D images through approximate posterior inference (Schönborn et al., 2017). We use a spherical harmonics lighting model, as in Zivanov et al. (2013), and a pinhole camera model, as in Blanz and Vetter (1999). Since no synthetic image will ever exactly match a natural image, we treat foreground pixels as subject to Gaussian noise and background pixels as sampled from the input image, following the method of Schönborn et al. (2015).

To perform inference, we use the MCMC method presented in Schönborn et al. (2017). Specifically, we use Gaussian drift proposals to update pose, perform closed-form estimation of illumination, and use Gaussian drift proposals applied in the 3DMM's low-dimensional eigenspaces to update the mesh itself. In order to locate the face in the image we constrain pose using landmark annotations provided with each image. Although we generated these landmark annotations manually, they could also have been obtained automatically using existing tools (e.g. OpenPose (Cao et al., 2019)).

One analysis-by-synthesis task is to reconstruct 3D face meshes from natural images, render the results and compare them with said natural images. We here perform this task on images from the Labeled Faces in the Wild dataset (Huang et al., 2008) and show in Fig. 6 the reconstructions produced using the standard-full and symmetric-full 3DMMs (as defined in Table 2), as well as the reconstructions that our inverse graphics pipeline produces using the 2019 Basel Face Model (Gerig et al., 2018). As Fig. 6 demonstrates, all of these 3DMMs produce plausible reconstructions with the BFM reconstructions being slightly superior—full results are shown in Fig. 20.

In addition to the models we define using the mean of the 2019 Basel Face Model, we construct additional 3DMMs using the symmetric kernel and ten scans provided with the 2009 Basel Face Model (Paysan et al., 2009) as different means. We name these models symmetric-x, where x is the ID number of the scan (`001`, `002`, `006`, `014`, `017`, `022`, `052`, `053`, `293`, or `323`). Reconstructions produced by these 3DMMs can be found in the Appendix, along with side views of our reconstructions. To assess our 3DMMs' performance in an inverse graphics setting where the choice of prior gains importance, the Appendix also includes reconstructions of partially occluded faces produced with the occlusion-aware MCMC method described in Egger et al. (2018). All models again yield similar reconstruction quality.

In our second experiment, we use the inverse rendering used above to perform face recognition, as outlined in Schönborn et al. (2017); Gerig et al. (2018); Blanz and Vetter (2003). By reconstructing the shape and albedo latents from a gallery of reference images $\{f_1, \ldots, f_n\}$ (with one image per identity), we can obtain latents $(c_{s,i}, c_{a,i})$ for each reference image f_i. Faces in a novel image f_0 are then identified by reconstructing shape and albedo latents $(c_{s,0}, c_{a,0})$ from said image and determining the reference image with the maximum cosine angle in the joint shape-albedo latent space, as in Blanz and Vetter (2003). We conduct face recognition on images from the CMU Multi-PIE database (Gross et al., 2010). The results are presented in Table 3.

Table 3 illustrates that the 3DMMs with albedo kernels that combine RGB-space and physical-space distance information perform face recognition significantly more accurately on all image types than do 3DMMs with albedo kernels that only make use of one type of distance metric. Furthermore, we may observe that the performance of the symmetric model is better on all image types than that of the BU3D-FE model (Gerig et al., 2018), a 3DMM built from 100 3D scans. Table 3 also illustrates that 3DMMs defined using the mean of the 2019 Basel Face Model have better performance than those defined using individual face scans. This is particularly true on images with a yaw angle over 15°, since as the yaw angle increases, the prior (in this case the 3DMM) plays a larger role in generating the reconstruction.

Table 3 Face recognition results for images from the Multi-PIE database (Gross et al., 2010)

angle	15°	30°	45°
probe id	140_16	130_16	080_16
standard-full	84.7	69.9	54.2
standard-RGB	76.3	57.8	28.9
standard-XYZ	77.1	62.7	35.7
symmetric-full	**93.2**	**85.9**	**72.3**
symmetric-RGB	73.5	61.4	40.2
symmetric-XYZ	73.1	58.6	44.2
symmetric-001	78.7	62.2	49.8
symmetric-002	78.7	70.7	48.6
symmetric-006	77.5	63.9	38.2
symmetric-014	71.9	59.0	47.8
symmetric-017	**88.0**	72.3	50.6
symmetric-022	85.9	73.5	59.4
symmetric-052	85.9	71.5	55.0
symmetric-053	84.3	**76.7**	55.4
symmetric-293	85.5	74.3	**59.8**
symmetric-323	87.6	76.3	55.4
10-scan PCA	86.0	65.9	42.2
10-scan KDE	**94.0**	**85.9**	**71.5**
BU3D-FE Gerig et al. (2018)	90.4	82.7	68.7
BFM '17 Gerig et al. (2018)	**98.8**	**98.0**	**90.0**

Each column represents the accuracy for a set of probe images with a common yaw angle given in the first row. The second row gives the common ending of the IDs in the Multi-PIE dataset of the probe images with a given yaw angle. The gallery is constructed from the images of all 249 identities with a yaw angle of 0° (dataset IDs ending in 051_16). Chance rate is 0.4. The 3DMMs in the second box (standard-full to symmetric-XYZ) were produced using the mean of the 2019 Basel Face Model (Gerig et al., 2018), while the 3DMMs in the third box (symmetric-001 to symmetric-323) were produced using the 3D scans provided with the 2009 Basel Face Model (Paysan et al., 2009). BFM '17 refers to the 2017 Basel Face Model (Gerig et al., 2018)

The previously presented face recognition results relied on the mean of the 2019 Basel Face Model (Gerig et al., 2018). The performances of the 3DMMs built using individual face scans (symmetric-001 to symmetric-323) are also listed in Table 3. The performances of these 3DMMs are clearly significantly lower than that of the symmetric-full 3DMM. However, by combining the information present in the 10 scans through our KDE approach (examples of those models can be found in Fig. 2, we can produce a new model that achieves performance comparable to that of the symmetric-full 3DMM. To perform face recognition with this non-parametric model, we perform inference for each mixture component separately on both the probe image and each gallery image. We then compute the cosine-angle in latent space between the probe reconstruction and all gallery reconstructions for each mixture component, and classify the probe image based on which 3DMM and gallery image yields the smallest cosine-angle.

The performance of this mixture model on our face recognition task is listed in Table 3 as "10-scan KDE". As Table 3 shows, this approach offers face recognition performance comparable to that achieved by the symmetric-full 3DMM, and outperforms the BU3D-FE model on all yaw levels, despite using only 10 scans. To provide a more direct comparison between our novel KDE approach and PCA-based 3DMMs, we also produced a 3DMM by performing PCA with the 10 scans. The face recognition performance of this 3DMM is listed in Table 3 as "10-scan PCA". Table 3 demonstrates that this PCA-based 3DMM has far poorer face recognition performance than our KDE-based model. In fact, the performance of the 10-scan PCA-based 3DMM is comparable to that of the 3DMMs produced from a single individual face scan (symmetric-001 to symmetric-323).

4.2 Learned Models

Despite performing well on face recognition tasks, samples from our single-scan 3DMMs are nevertheless highly non-naturalistic, as shown in Fig. 1. Using the learning approach outlined in Sect. 3.4, we augmented our standard-full and symmetric-full 3DMMs using 200 images from the Multi-PIE dataset (Gross et al., 2010). We used images of 50 distinct individuals shown in a frontal perspective, from a 15° angle, from a 30° angle, and from 45° angle. However, no identity or pose annotations were used; our learning algorithm treated each image as if it was an image of a novel individual in an unknown pose. We augmented these 200 images by adding alternate versions of each image that were flipped left-to-right, so our learning algorithm used 400 images total.

We ran five iterations of our wake-sleep-like procedure, creating ten new 3DMMs total; here we show only four of them, namely the 3DMMs produced after one iteration, and the 3DMMs produced after all five iterations. We name these 3DMMs "standard-1", "standard-5", "symmetric-1", and "symmetric-5". Random samples from these 3DMMs shown in frontal and side views are shown in Fig. 8.

We additionally wanted to see if our learning procedure could reconstruct the mean face even when initialized with a simplified face scan. To do so, we applied a blur filter to (separately) the shape and albedo of the mean of the 2019 Basel Face Model (Gerig et al., 2018), resulting in the simplified meshes shown in Fig. 7, and constructed analogues of the symmetric-full models using these simplified meshes, yielding two new 3DMMs. We then reran our learning algorithm using these new 3DMMs as initializations, yielding 10 new learned 3DMMs; we again show only four of them, namely those produced after one iteration and after all five iterations. We name the 3DMMs produced using a simplified albedo "smooth-albedo-1" and

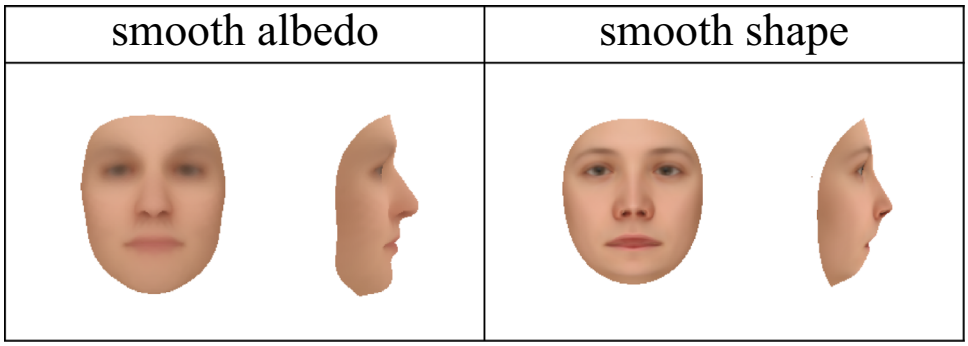

Fig. 7 Simplified 3D face templates used to initialize the learning process used to construct the smooth-albedo-1, smooth-albedo-5, smooth-shape-1, and smooth-shape-5 models shown in Figs. 8 and 9. These 3D templates were constructed by applying blur transformations to the mean of the 2019 Basel Face Model (Gerig et al., 2018)

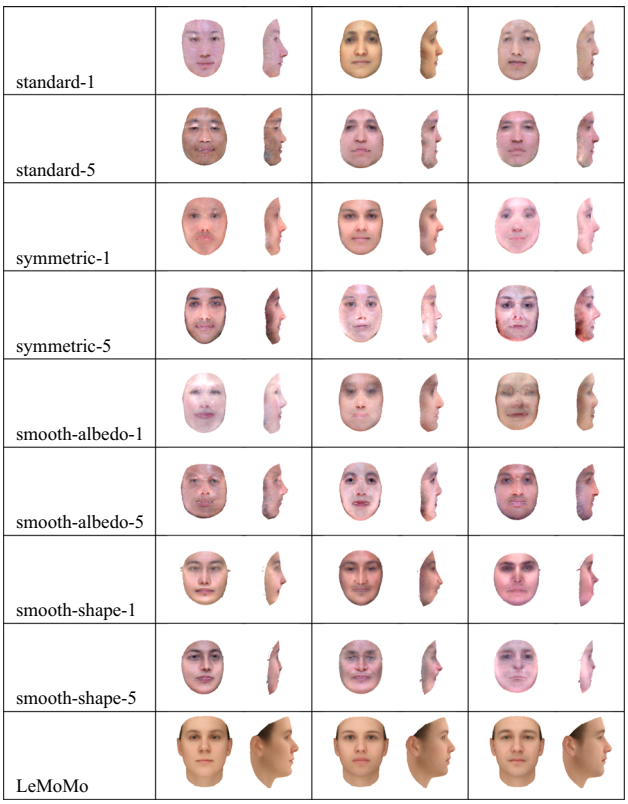

Fig. 8 Samples from the models learned through the learning process outlined in Sect. 3.4, using as initial 3DMMs the standard-full (standard-1 and standard-5) and symmetric-full (symmetric-1 and symmetric-5) 3DMMs, as well as analogues of the symmetric-full 3DMMs built using initial scans with simplified albedo (smooth-albedo-1 and smooth-albedo-5) and simplified shape (smooth-shape-1 and smooth-shape-5), and the LeMoMo model developed by Tewari et al. (2020). Learning was performed for one iteration (standard-1, symmetric-1, smooth-albedo-1, and smooth-shape-1) or for five iterations (standard-5, symmetric-5, smooth-albedo-5, smooth-shape-5)

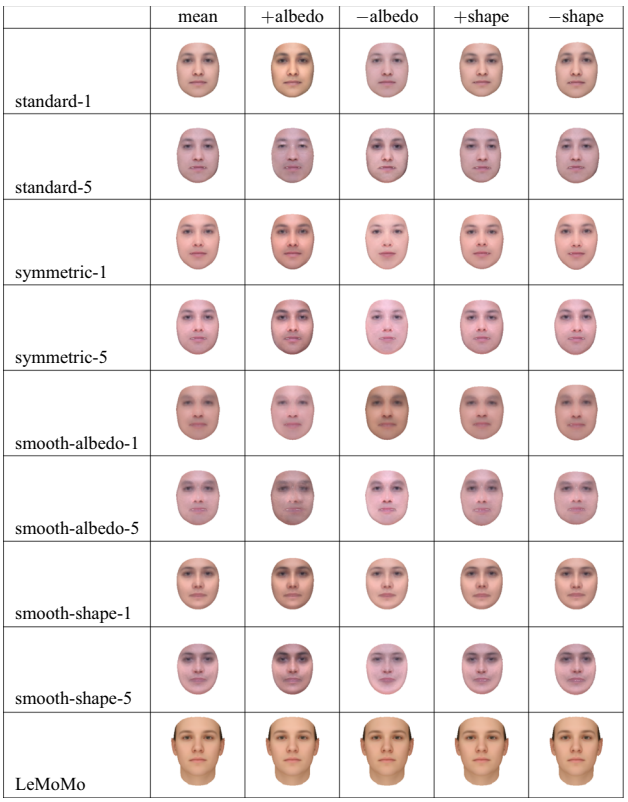

Fig. 9 The mean, as well as the mean offset by ±1 times the first principal component of the albedo ("+albedo", "−albedo") or shape ("+shape", "−shape") models of the standard-1, standard-5, symmetric-1, symmetric-5, smooth-albedo-1, smooth-albedo-5, smooth-shape-1, and smooth-shape-5 models, and the LeMoMo model developed by Tewari et al. (2020)

"smooth-albedo-5", and those produced using a simplified shape "smooth-shape-1" and "smooth-shape-5". In Fig. 8, we also show random samples (in frontal and side views) from the analogues of standard-1, standard-5, symmetric-1, and symmetric-5 3DMMs produced using the simplified initial meshes. In addition to showing samples from these distributions, we can also examine their means (that is, the means of the learned distributions, not the initial scan used to build the models) and their first principal components. In Fig. 9, we show the means of the standard-1, standard-5, symmetric-1, symmetric-5, smooth-albedo-1, smooth-albedo-5, smooth-shape-1, and smooth-shape-5 3DMMs, in front and side views. Additionally, Fig. 9 also shows these mean altered by adding their respective 3DMMs' first shape or albedo principal components.

Fig. 10 Reconstructions produced from natural images using several of our learned 3DMMs. As in Fig. 6, the first row shows the images used as input, and the subsequent rows show 3D reconstructions produced using our 3DMMs

Finally, Fig. 10 shows qualitative 3D face reconstructions produced from images from the Labeled Faces in the Wild dataset (Huang et al., 2008) using the standard-1, standard-5, symmetric-1, and symmetric-5 models, and the inference method of Schönborn et al. (2017). This directly mirrors Fig. 6, with the only difference being that the reconstructions are produced using different 3DMMs. While the reconstructions in Fig. 10 are not visually much better than those in Fig. 6—likely because the increased realism of the learned models comes with a reduction in flexibility—they nevertheless demonstrate that our learned models can likewise be used in basic inverse graphics settings.

While the resulting face 3DMMs are still non-naturalistic in some ways they seem a clear improvement over our initial 3DMMs, as can be seen by comparing Fig. 8 with Fig. 1. For instance, the standard-1 and standard-5 models have clearly learned naturalistic face tones as well as approximate facial symmetry, which were lacking from the initial standard-full 3DMM. Figure 9 demonstrates that although the smooth-shape-1 and smooth-shape-5 models do not appear to be able to learn a realistic face shape given a highly unrealistic template mesh, the smooth-albedo-1 and smooth-albedo-5 do appear to be able to learn a realistic mean face albedo even when the initial template mesh possesses a non-naturalistic albedo. While our learned 3DMMs show a significant visual improvement over the standard-full and symmetric-full 3DMMs, quantitatively demonstrating an improvement has proven difficult. Face recognition performance as measured in Table 3 is significantly lower with the learned models than with our initial standard-full and symmetric-full 3DMMs, and shape specificity and generalization (as shown in Fig. 11) are also far lower, while albedo specificity and generalization are comparable or somewhat poorer. This may be at least partially due to artifacts of the learning process; in particular, for the face recognition results, the number of principal components is much lower in the learned models because it is limited by the number of reconstructions which pass the quality control process, while the vastly lower shape specificity might be a byproduct of the alignment process applied during learning.

4.3 Specificity, Generalization, Against Compactness

Figure 11 shows plots of the specificity, generalization, against compactness (Styner et al., 2003) of our 3DMMs and the 2017 Basel Face Model (Gerig et al., 2018); specifically, it shows the specificity and generalization of the shape and albedo models of each 3DMM as a function of the number of principal components included. We compare the 2017 Basel Face Model ("BFM 2017") and versions of the standard-full ("standard"), symmetric-full ("symmetric"), and correlated-full ("correlated") models built using the mean of the 2017 Basel Face Model as template. The correlated-full model, presented in the Appendix, is analogous to the symmetric-full model but lacks bilateral symmetry. We use as our dataset the ten scans provided with the 2009 Basel Face Model (Paysan et al., 2009). We also include the symmetric-x models, where x is a scan ID number (`001`, `002`, `006`, `014`, `017`, `022`, `052`, `053`, `293`, or `323`); for these models we exclude the scan used to build the model. We report results averaged across the symmetric-x models as "single-scan". We measure specificity and generalization using 1, 2, 5, 10, 20, 50, 100, and all (199) principal components. We indicate the specificity and generalization of the mean of the 2017 Basel Face Model, considered as a 3DMM with zero principal components, with a black line.

We may observe that, for all numbers of principal components, the generalization of our 3DMMs' shape models is

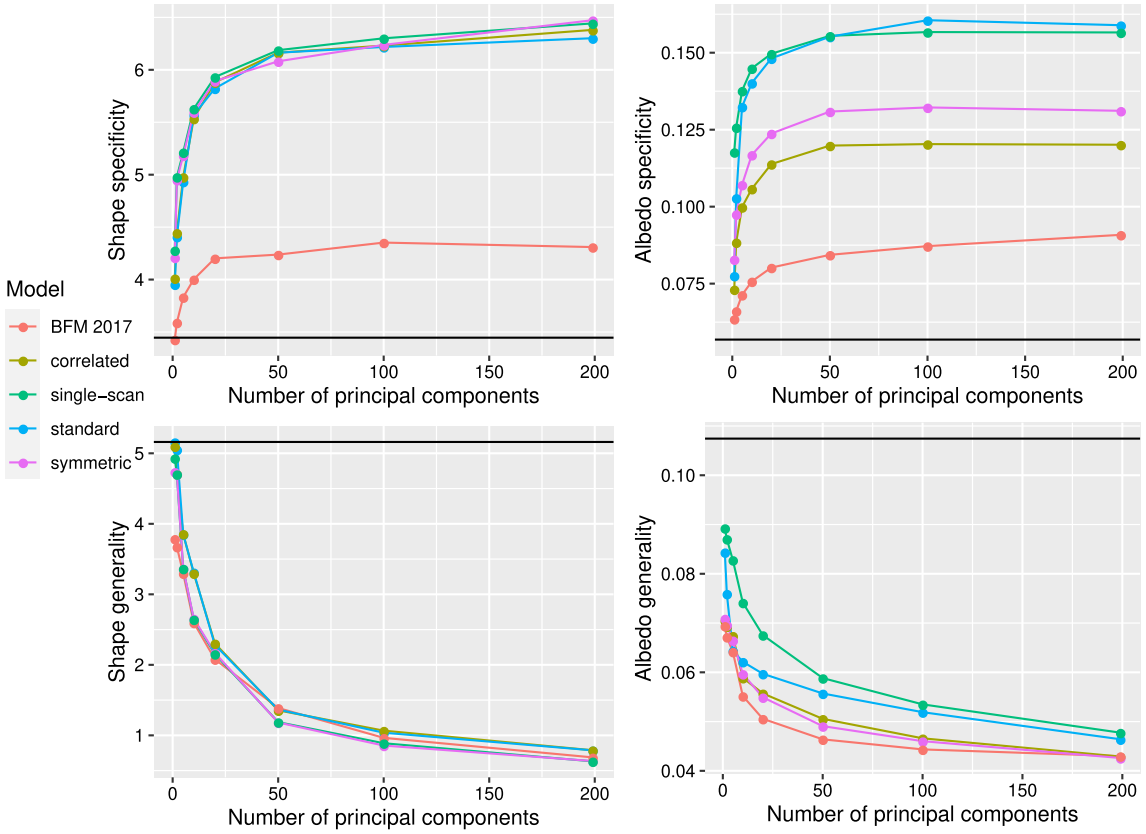

Fig. 11 A plot of the specificity and generalization in relation to the number of principal components (compactness) (Styner et al., 2003) of our 3DMMs' relative to the 2017 Basel Face Model (Gerig et al., 2018). "standard", "correlated", and "symmetric" refer to versions of the standard-full, correlated-full, and symmetric-full models built using the mean of the 2017 Basel Face Model, while the "single-scan" results are an average of the performance of the various symmetric-x models. The scans included with the 2009 Basel Face Model (Paysan et al., 2009) were used as a dataset; for the symmetric-x 3DMMs, the scan used to construct the 3DMM was excluded. See Sect. 4.3 for more details

comparable to that of the 2017 Basel Face Model, while the generalization of our 3DMMs' albedo models is in fact superior to that of the 2017 Basel Face Model. The specificity of our 3DMMs' shape and albedo models, is, of course, inferior to that of the 2017 Basel Face Model. This is unavoidable as our models' were constructed using far less data than the 2017 Basel Face Model. We may additionally observe that our single-scan models perform comparably to the standard-full model across all conditions.

4.4 Registration Tasks

Registration is another task for which 3DMMs may be used. In this task we wish to transform an arbitrary face mesh into a mesh with a given topology while preserving the face as closely as possible. Prior work has nearly exclusively relied on shape information to compute such a transformation (Egger et al., 2020). However, albedo information also provides important constraints on face registration. For instance, the eyebrows and the pupils of the eyes are almost entirely defined by albedo.

To perform registration tasks with our 3DMMs, we adapted the inverse rendering approach of Schönborn et al. (2017) to minimize the Chamfer distance between the model mesh and the target mesh while simultaneously minimizing the pixel error between the rendered model instance and the rendered target mesh. We achieve this by combining an image-based reconstruction likelihood, which constrains 2D appearance, with a shape-based likelihood, which enforces 3D shape consistency as measured by Chamfer distance. This minimizes shape distance and induces albedo consistency while establishing correspondence with the topology of our 3DMM's template. While both those ideas are often applied in isolation, they are rarely combined in registration tasks (or only combined as post-processing). We roughly align the meshes to initialize the pose, but, unlike typical approaches, do not use landmarks during registration. Instead, the location of facial features is constrained by the albedo component of the evaluation. As post-processing we eliminate any net

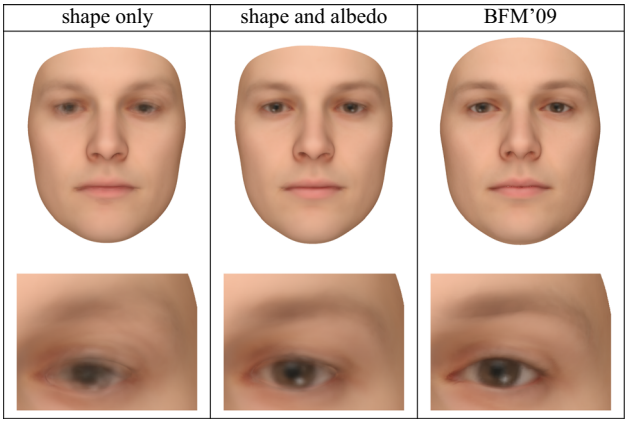

Fig. 12 The average of the registration results produced by the MCMC methods using both shape and albedo information ("shape and albedo") or shape information only ("shape only") on all ten scans, along with the average of the corresponding registered meshes produced in the construction (strongly reliant on manual landmark annotations) of the 2009 Basel Face Model (Paysan et al., 2009) ("BFM'09"). Close-ups of the left eye and eyebrow are provided, illustrating that the eyebrows and the pupils of the eyes are far less clearly defined in the shape-only condition

translation using the transformation parameters estimation method based on least-squares of Umeyama (1991) and set each vertex's albedo by projecting vertex normals onto the scan as performed in the Basel Face Pipeline to extract texture information from 3D scans (Gerig et al., 2018).

This process enables us to make use of both shape and albedo information in registration. We compare the result of doing so with the analogous registration result produced using only shape information in our MCMC process. We apply both registration methods to the unprocessed meshes for face scans `001`, `002`, `006`, `014`, `017`, `022`, `052`, `053`, `293`, and `323`. To do so we use the standard-full 3DMM with the mean of the highest point-count version of the 2019 Basel Face Model (Gerig et al., 2018) as reference. To evaluate our registration we build a 3DMM from the registration results using PCA.

We compare these results with the registration used by the 2009 Basel Face Model (Paysan et al., 2009), which used shape information along with manual landmark annotations. Figure 12 demonstrates that by using shape and albedo information our registration process produces a sharp and stable albedo reconstruction whose quality is comparable to that of the 2009 Basel Face Model's registration, and far superior to that produced using shape information alone. This performance is impressive, since the 2009 Basel Face Model heavily relied on human-provided landmark annotations in its registration pipeline, whereas our approach requires no annotations.

The Appendix contains a quantitative assessment of our shape registration performance, and shows that including albedo information in registration slightly increases the shape error. This is unsurprising, as the shape-only reconstruction is optimized to produce the lowest shape error possible, and the shape and albedo reconstruction by definition cannot have less than the minimum shape error. However, as Fig. 12 demonstrates, the shape and albedo reconstruction has far higher quality overall.

4.5 Constructing 3DMMs for Other Objects

We have thus far focused on 3DMM for faces; we now demonstrate that analogous methods can be used to build 3DMMs for other object categories. Specifically, we construct single-scan 3DMMs for fish and birds using as references synthetic meshes with simple manual coloring.[1] These meshes are simple artistic models and were constructed without 3D scanning. We can build 3DMMs from each of these references using the same kernels as used in standard-full and symmetric-full, i.e. K_s and K_a in the first case, and $K_s^{\mathtt{sym}}$ and $K_a^{\mathtt{sym}}$ in the second. This produces two new 3DMMs for each mesh, which we term the standard and symmetric models for each object category. As our reference meshes lack many details that 3D scans possess, the performance of these 3DMMs is likely much lower than that of single-scan 3DMMs built from 3D scans. Results with additional fish and bird 3DMMs produced with the $K_{a,\mathtt{xyz}}$ and $K_{a,\mathtt{xyz}}^{\mathtt{sym}}$ albedo kernels are presented in the Appendix.

We seek to model a wide range of birds, but restrict ourselves to simple standing poses. We restrict ourselves to the *Acanthurus* genus of fish, which possesses a wide range of color variability but lack the fine details (such as scales) that many other fish possess. In Fig. 13, we show qualitative reconstruction results along with samples from our bird and fish models and the reference meshes used to construct them. While these reconstructions are not as accurate as those in Fig. 6, they do capture some rough features. We suggest that three main factors make birds a more difficult object category than faces: birds have a much more complex albedo, including high-frequency components that our models capture poorly; birds have a well-defined silhouette, whereas faces have somewhat arbitrary boundaries; and color-correlation, while beneficial in modeling faces, impedes the symmetric bird model's ability to model birds. Our standard model's performance on fish seems somewhat better, likely due to the lack of high-frequency components. The symmetric model does much more poorly on fish, likely because the correlation of its color channels impedes its ability to model the regional color variation of fish. It is important to keep in mind, however, that our method's performance is

[1] Our bird and fish reference meshes are obtained from, respectively, https://www.blendswap.com/blend/11752 and https://www.turbosquid.com/3d-models/free-tail-animation-3d-model/368484.

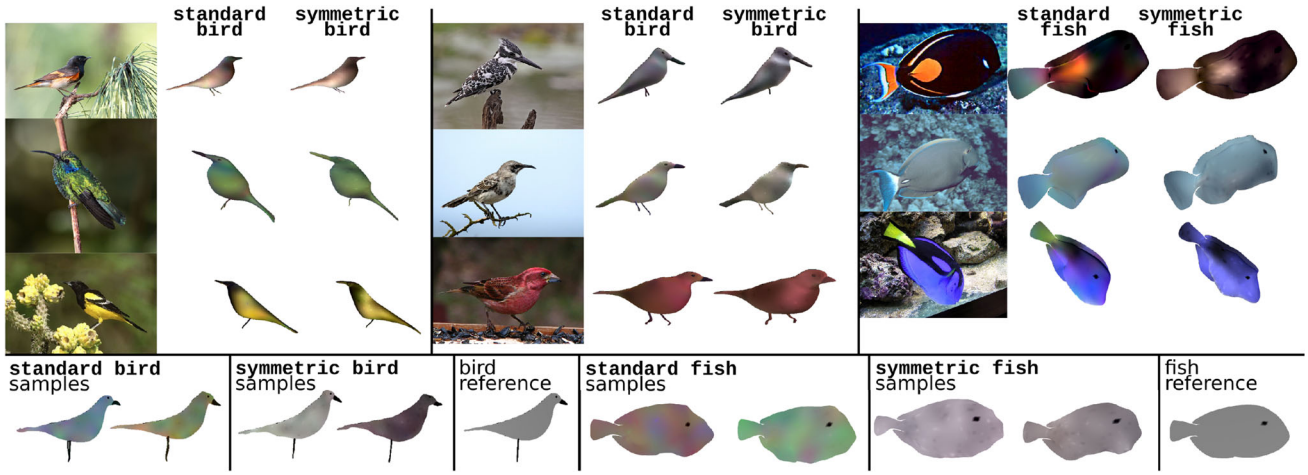

Fig. 13 On the upper left and middle: the reconstructions produced by the standard and symmetric bird models on six images taken from the Caltech-UCSD Birds 200 dataset (Wah et al., 2011). On the upper right: the reconstructions produced by the standard and symmetric fish models on three public-domain images taken from Wikipedia. On the bottom: samples from the standard and symmetric bird and fish models, shown in side views, as well as the reference meshes used to build these 3DMMs

not directly comparable to that of other, less data-efficient approaches.

5 Conclusion

Our research demonstrates that we can build a simple 3DMM from a single template through the application of Gaussian process-based deformations. Although the result is of lower-quality than 3DMMs produced from high-quality 3D scans, our simple models can still be used in many contexts where hand-produced 3DMMs have previously been required, and can be constructed using far less data and far simpler pipelines. We demonstrate a preliminary unsupervised learning method for a 3DMM of faces from 2D images based solely on a single template without any supervision. For object categories where the number of available scans is extremely limited or where dense correspondence between scans cannot be easily obtained, this procedure thus offers a promising method for building 3DMMs. Additionally, our results demonstrate the high value of fully integrating albedo into the 3DMM pipeline, and show that this can be done by combining covariance kernels which produce spatially continuous deformations with kernels that produce color-space-continuous deformations. In addition to the results demonstrated in this paper, we believe our method can be highly beneficial in addressing dataset bias, a limitation of all currently available 3DMMs.

In addition to its relevance in a computer vision context, our paper further demonstrates that a statistical model of faces can be learned from a initial simple template and limited unsupervised 2D data similar to what a human infant has access to. This was motivated by interest in computationally modeling the cognitive development of human face perception in infants, and we hope that in the future our approach may inspire novel computational models of the development of human face perception.

Funding This work was funded by the DARPA Learning with Less Labels (LwLL) program (Contract No: FA8750-19-C-1001), the DARPA Machine Common Sense (MCS) program (Award ID: 030523-00001) and by the Center for Brains, Minds and Machines (CBMM) (NSF STC award CCF-1231216). B. Egger was supported by a PostDoc Mobility Grant, Swiss National Science Foundation P400P2_191110. Open Access funding enabled and organized by Projekt DEAL.

Declarations

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

Appendix A Color-Correlated Asymmetric 3DMMs

In our main paper we build 3DMMs using Gaussian processes that include symmetry and color-channel correlation heuristics. To assess the effects of these heuristics individually, we can also build 3DMMs that include only one of these heuristics. Specifically, we experimented with constructing 3DMMs whose albedo models have correlated color channels but which lack symmetry. This enables us to compare the relative importance in an analysis-by-synthesis setting of the symmetry and color-correlation heuristics of our symmetric 3DMMs.

Our main paper defined albedo kernels $K^{\mathtt{cor}}_{a,\mathtt{xyz}}$ and $K^{\mathtt{sym}}_{a,\mathtt{rgb}}$. These kernels possess a color-channel correlation heuristic but lack an explicit symmetry heuristic ($K^{\mathtt{sym}}_{a,\mathtt{rgb}}$ is symmetric, but this is only because we assume that the reference face is symmetric). We may average these kernels to produce an albedo kernel $K^{\mathtt{cor}}_{a} = 0.5(K^{\mathtt{sym}}_{a,\mathtt{rgb}} + K^{\mathtt{cor}}_{a,\mathtt{xyz}})$ which combines physical-space and RGB-space distance information. By combining these albedo kernels with our shape kernel K_a, we can construct 3DMMs which possess a color-channel correlation heuristic but which lack an explicit symmetry heuristic. We list these 3DMMs in Table 4.

We repeat the face recognition experiment presented in Section 3.1.2 of our main paper with these 3DMMs (once again using the mean of the 2019 Basel Face Model (Gerig et al., 2018) as our reference mesh). The results of this experiment are shown in Table 5, along with a copy of the results with the standard and symmetric 3DMMs that were shown in the main paper. Table 5 demonstrates that the color-correlated asymmetric 3DMMs perform comparably to the symmetric (and color-correlated) 3DMMs on faces with 15° and 30° yaw angles. On faces with 45° yaw angles, they are significantly worse, indicating that (unsurprisingly) a symmetry prior becomes more important as the yaw angle increases. Nevertheless, in general the color-correlated asymmetric 3DMMs perform quite well. This indicates that in an inverse graphics context the color-channel correlation heuristic is more important to our symmetric 3DMMs than the symmetry heuristic is, at least for input images with a low yaw angle.

Table 4 Our Gaussian processes for modeling faces with color-correlated but asymmetric kernels

Name	Shape kernel	Albedo kernel
Correlated-full	K_s	$K^{\mathtt{cor}}_{a}$
Correlated-RGB	K_s	$K^{\mathtt{cor}}_{a,\mathtt{rgb}}$
Correlated-XYZ	K_s	$K^{\mathtt{cor}}_{a,\mathtt{xyz}}$

Table 5 Face recognition results on images from the Multi-PIE database (Gross et al., 2010)

angle	15°	30°	45°
probe id	140_16	130_16	080_16
standard-full	84.7	69.9	54.2
standard-RGB	76.3	57.8	28.9
standard-XYZ	77.1	62.7	35.7
correlated-full	92.4	**87.1**	66.7
correlated-RGB	71.5	58.6	37.8
correlated-XYZ	76.7	61.4	45.8
symmetric-full	**93.2**	85.9	**72.3**
symmetric-RGB	73.5	61.4	40.2
symmetric-XYZ	73.1	58.6	44.2

Each column represents the accuracy for a set of probe images with a common yaw angle given in the first row. The second row gives the common ending of the IDs in the Multi-PIE dataset of the probe images with a given yaw angle. The gallery is constructed from images with a yaw angle of 0° (dataset IDs ending in 051_16)

Appendix B Additional Bird and Fish Models

In our paper we presented bird and fish 3DMMs created analogously to the standard-full and symmetric-full 3DMMs, i.e. with K_s and K_a, and with $K_s^{\mathtt{sym}}$ and $K_a^{\mathtt{sym}}$, respectively. We can also define similar bird and fish 3DMMs using albedo kernels that only rely on physical distance; i.e., using $K_{a,\mathtt{xyz}}$ and $K^{\mathtt{sym}}_{a,\mathtt{xyz}}$ instead of K_a and $K_a^{\mathtt{sym}}$. This produces two new 3DMMs for each reference mesh, which for space reasons are listed as listed as "XYZ standard" and "XYZ symmetric". Figure 14 shows samples from these two bird 3DMMs, as well as reconstructions produced with these models of the bird images that were used in the main paper. Figure 15 shows analogous samples and reconstructions for the two new physical-distance-based fish 3DMMs. The results are close to those produced in the main paper; this is unsurprising given that the 3D mesh used to build these 3DMMs does not include complex coloration, and instead has near-piecewise-constant albedo.

In both Fig. 15 and in the main paper, we obtain our input natural fish images from Wikipedia.[2]

[2] The links are: https://en.wikipedia.org/wiki/File:Acanthurus_achilles1.jpg, https://en.wikipedia.org/wiki/File:Acanthurus_dussumieri.jpg, and https://en.wikipedia.org/wiki/File:Paracanthurus_hepatus_(Regal_Tang).jpg.

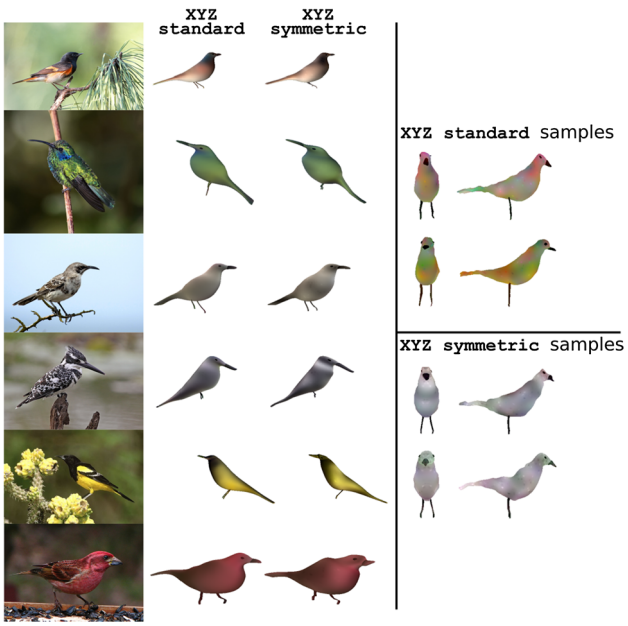

Fig. 14 On the left: the reconstructions produced by the two bird models built using only physical distance information on six images taken from the Caltech-UCSD Birds 200 dataset (Wah et al., 2011). On the right: samples from these models, shown in frontal and side views

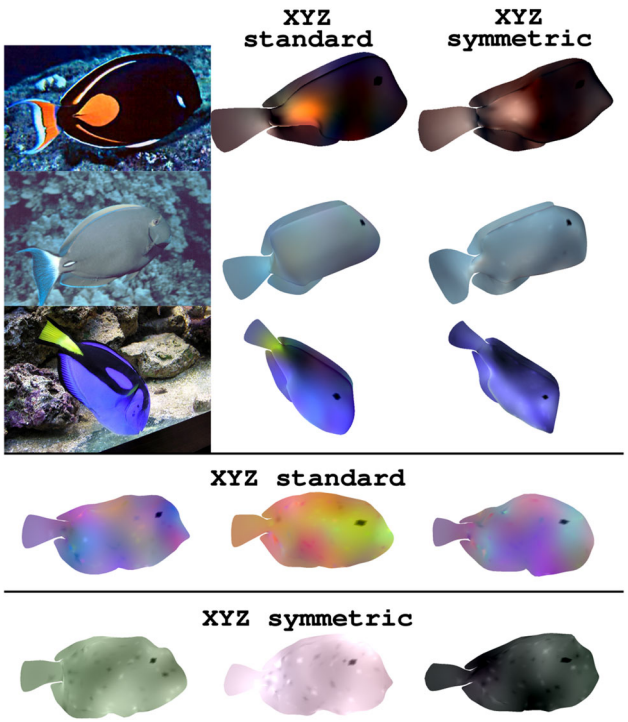

Fig. 15 On the top: the reconstructions produced by the two fish models built using only physical distance information on three natural images of fish. On the bottom: samples from these models, shown in side views

Appendix C Additional Registration Results

Figures 17, 16, 19, and 18 offer a variety of quantitative metrics of the shape error of the registered meshes produced in our paper's registration tasks. As in the main paper, in all figures the "shape and albedo" option refers to meshes registered using both shape and albedo information in the MCMC method, while the "shape only" option refers to meshes registered using only shape information in the MCMC method. These figures do not take into account the stability of the reconstruction or the albedo error.

We estimate shape error through Hausdorff distance (Figs. 17 and 19) and Chamfer distance (Figs. 16 and 18) between either the vertices of the registered meshes and the corresponding scans (Figs. 17 and 16) or a sparse set of landmarks on the registered meshes and corresponding scans (Figs. 19 and 18). We obtained landmark information by using the landmark annotations given in Paysan et al. (2009) for each of the 10 input meshes and the landmark annotations provided with the 2019 Basel Face Model (Gerig et al., 2018) for the registered meshes (since these have the same topology as the 2019 Basel Face Model).

Figures 17 and 16 demonstrate that including albedo information along with shape information slightly increases the shape reconstruction error. As noted in the main text, this is to be expected; the shape-only reconstruction is optimized to produce the lowest shape error possible, whereas the reconstruction produced using both shape and albedo is also optimized to produce a low albedo error, and by definition cannot have a lower shape error than the reconstruction

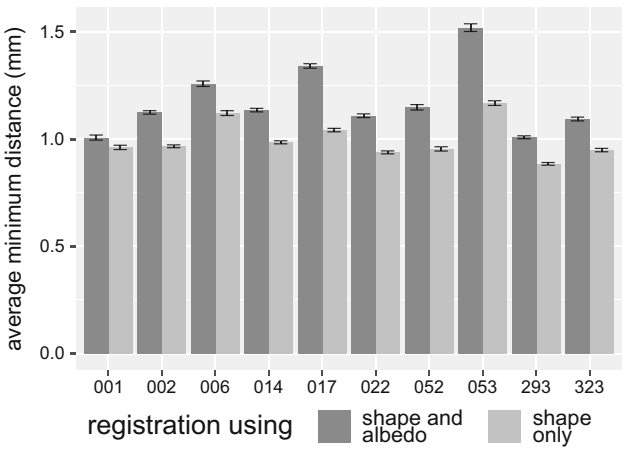

Fig. 16 The average distance between each vertex in each of the registered meshes and the closest point in the corresponding face scan, with error bars (±1.96 standard error)

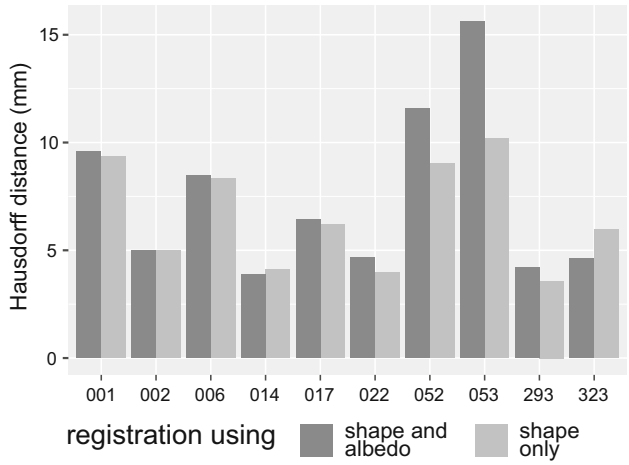

Fig. 17 The Hausdorff distance between the vertices of each of the registered meshes and the vertices of the corresponding face scan

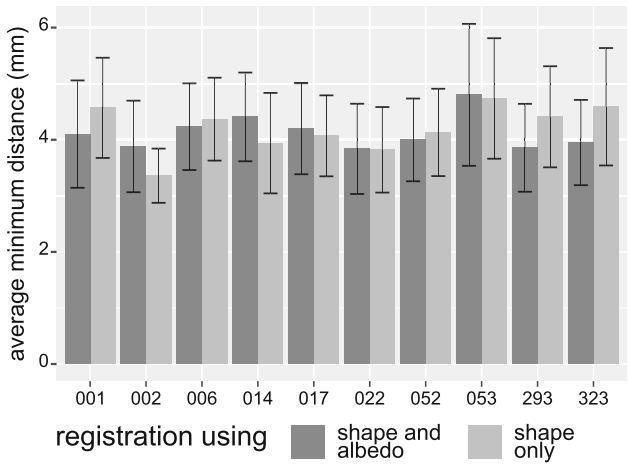

Fig. 18 The average distance between each landmark in each of the registered meshes and the closest landmark in the corresponding face scan, with error bars (±1.96 standard error)

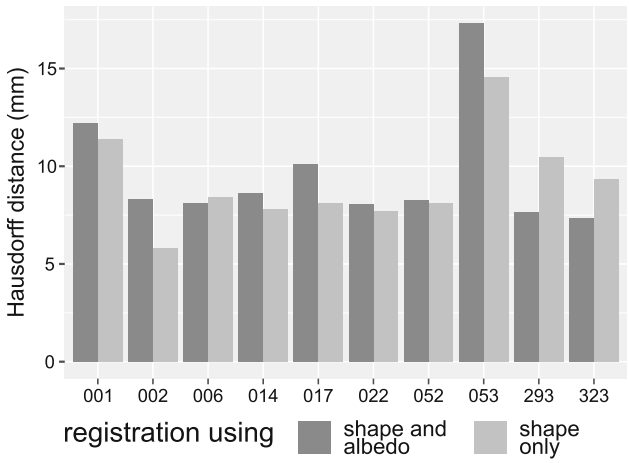

Fig. 19 The Hausdorff distance between the landmarks of each of the registered meshes and the landmarks of the corresponding face scan

with the minimum shape error. However, the increase in shape error is not very large. Furthermore, Figs. 19 and 18 demonstrate that including albedo in registration does not significantly affect the shape error of landmarks. This suggests that the incorporation of albedo information does not reduce the registration quality of the *important aspects* of face shape.

Appendix D Qualitative Reconstructions

Figures 20, 21, 22, and 23 provide additional qualitative reconstruction results. Figures 20 and 21 present qualitative reconstructions (in frontal and side views, respectively) of images from the Labeled Faces in the Wild dataset (Huang et al., 2008) produced using all the 3DMMs constructed using the mean of the 2019 Basel Face Model (Gerig et al., 2018). Figure 22 presents qualitative reconstructions of the same images produced using 3DMMs built from the scans included with the 2009 Basel Face Model (Paysan et al., 2009). These reconstructions are significantly lower-quality, because a significant portion of the shape of the template mesh is preserved during the MCMC process. Figure 23 presents qualitative reconstructions of different images from the Labeled Faces in the Wild dataset that contain significant occlusion. Figure 23 includes both 3D reconstructions as well inferred occlusion masks. Figure 23's results were produced using the occlusion-aware MCMC method described in Egger et al. (2018).

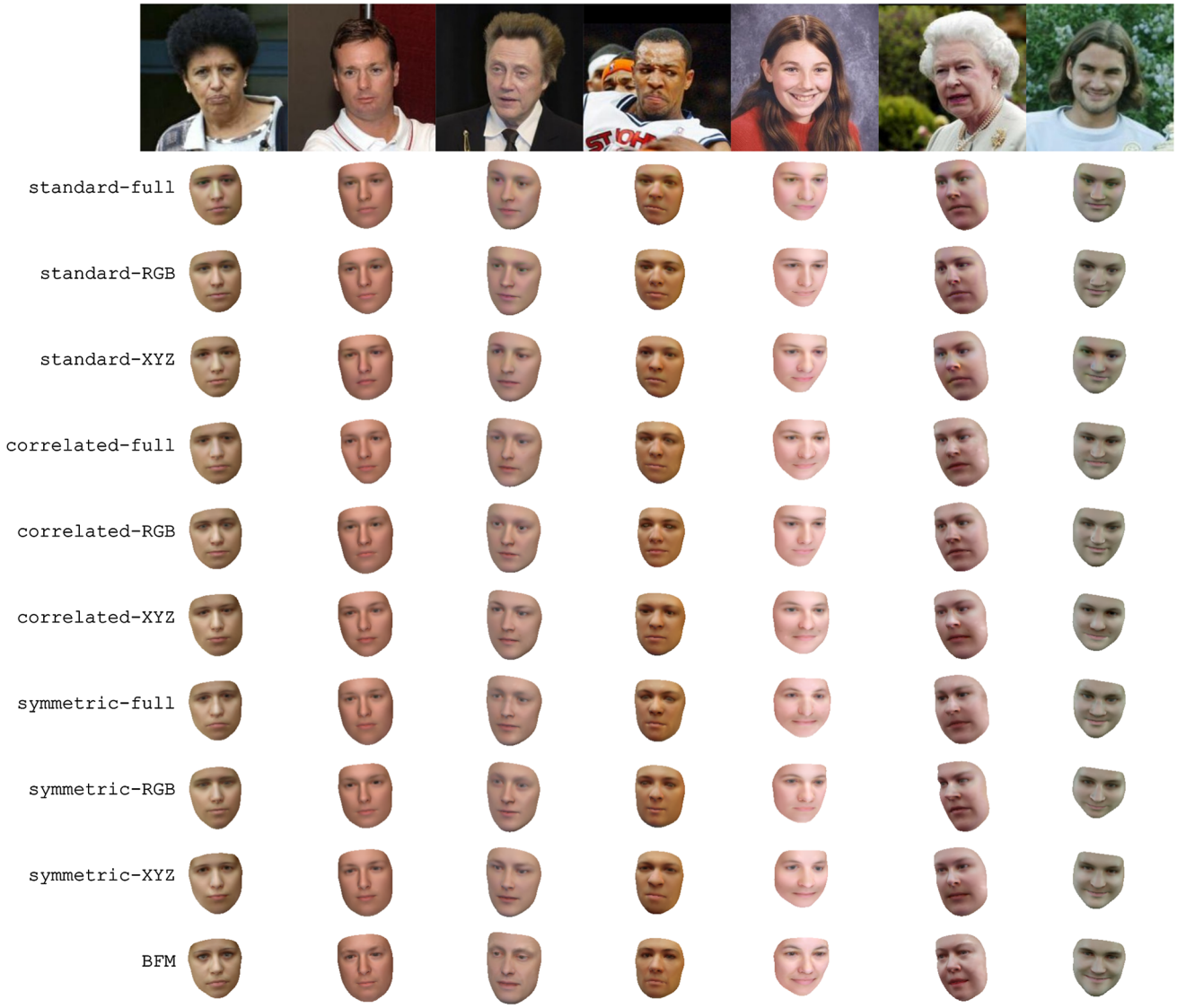

Fig. 20 The face reconstructions produced from all the 3DMMs built using the mean of the 2019 Basel Face Model (Gerig et al., 2018) on natural images from the Labeled Faces in the Wild dataset (Huang et al., 2008), as well as the reconstructions produced using the 2019 Basel Face Model itself ("BFM")

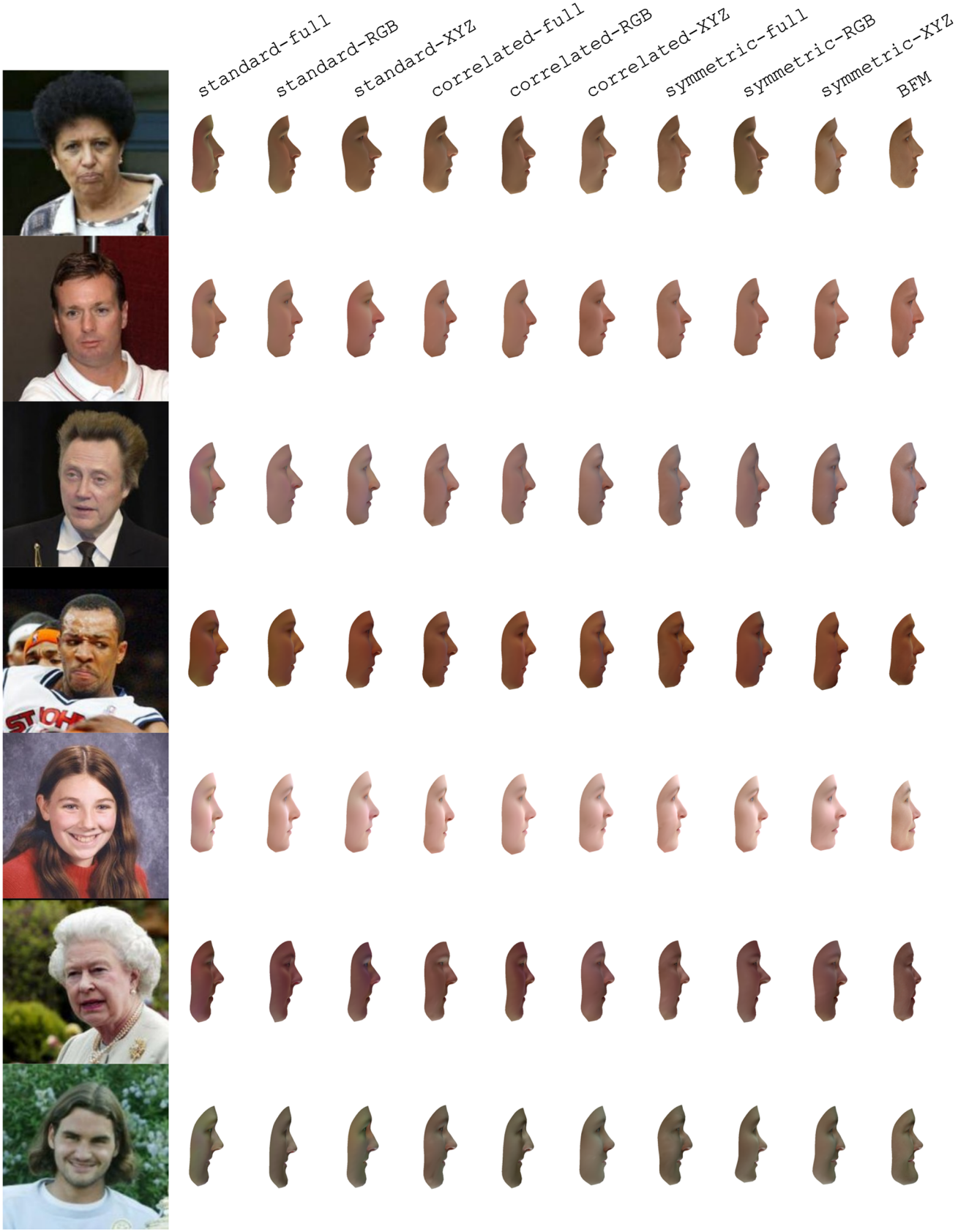

Fig. 21 Side views of the reconstructions presented in Fig. 20

Fig. 22 The face reconstructions produced from all 3DMMs built using the symmetric-full kernel and scans included with the 2009 Basel Face Model (Paysan et al., 2009) on natural images from the Labeled Faces in the Wild dataset (Huang et al., 2008)

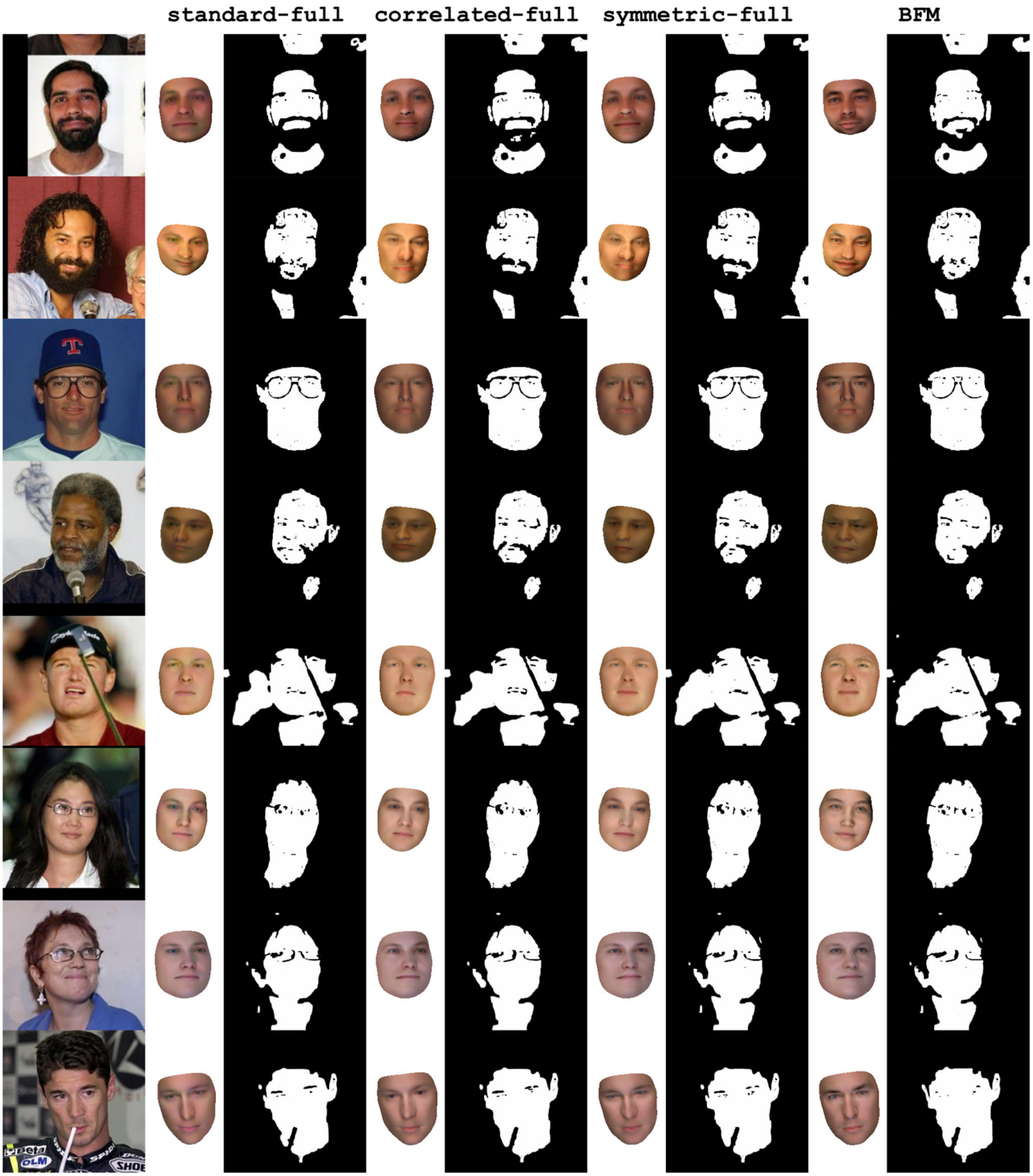

Fig. 23 The face reconstructions produced by our standard-full, correlated-full, and symmetric-full models, as well as the 2019 Basel Face Model (Gerig et al., 2018) ("BFM"), on images from the Labeled Faces in the Wild dataset (Huang et al., 2008), produced using the occlusion-aware MCMC method described in Egger et al. (2018). Both the segmentation masks and face reconstructions were inferred purely with top-down inference

References

Abrevaya, V. F., Wuhrer, S., & Boyer, E. (2018). Multilinear autoencoder for 3D face model learning. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 1–9). https://doi.org/10.1109/WACV.2018.00007.

Bartoli, A., Gérard, Y., Chadebecq, F., & Collins, T. (2012). On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2026–2033). https://doi.org/10.1109/CVPR.2012.6247906.

Blanz, V. & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194).

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(9), 1063–1074. https://doi.org/10.1109/TPAMI.2003.1227983

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., & Zafeiriou, S. (2018). Large scale 3D morphable models. *International Journal of Computer Vision, 126*(2), 233–254. https://doi.org/10.1007/s11263-017-1009-7

Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., & Zafeiriou, S. (2019). Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In *2019 IEEE/CVF international conference on computer vision (ICCV)*. https://doi.org/10.1109/ICCV.2019.00731.

Brunet, F., Hartley, R., Bartoli, A., Navab, N., & Malgouyres, R. (2011). Monocular template-based reconstruction of smooth and inextensible surfaces. In R. Kimmel, R. Klette, & A. Sugimoto (Eds.), *Computer vision: ACCV 2010. Lecture notes in computer science* (pp. 52–66). Springer. https://doi.org/10.1007/978-3-642-19318-7_5

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. arXiv:1812.08008 [cs]. Accessed 2020-07-26.

Cashman, T. J., & Fitzgibbon, A. W. (2012). What shape are dolphins? Building 3D morphable models from 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 232–244.

Chaudhuri, S., Ritchie, D., Wu, J., Xu, K., & Zhang, H. (2020). Learning generative models of 3D structures. In *Computer graphics forum* (Vol. 39, pp. 643–666). Wiley Online Library.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR09*.

Egger, B., Schönborn, S., Schneider, A., Kortylewski, A., Morel-Forster, A., Blumer, C., & Vetter, T. (2018). Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision, 126*(12), 1269–1287.

Egger, B., Smith, W. A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al. (2020). 3D morphable face models-past, present, and future. *ACM Transactions on Graphics (TOG), 39*(5), 1–38.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Lüthi, M., Schönborn, S. & Vetter, T. (2018). Morphable face models—An open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 75–82). IEEE.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118*(2), 201–210.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. *Image and Vision Computing, 28*(5), 807–813. https://doi.org/10.1016/j.imavis.2009.08.002

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645). Springer.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268*(5214), 1158–1161.

Horn, R. A. (2012). *Matrix analysis* (2nd ed.). Cambridge University Press.

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in thewild:Adatabase for studying face recognition in unconstrained environments. In Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Oct 2008, Marseille, France. https://inria.hal.science/inria-00321923

Kellman, P. J. & Arterberry, M. E. (2007). Infant Visual Perception. In W. Damon, R. M. Lerner, D. Kuhn & R. Siegler (Eds.). Handbook of Child Psychology Wiley. https://doi.org/10.1002/9780470147658.chpsy0203

Kemelmacher-Shlizerman, I., & Basri, R. (2010). 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(2), 394–405. https://doi.org/10.1109/TPAMI.2010.63

Kilian, M., Mitra, N. J., & Pottmann, H. (2007). Geometric modeling in shape space. In *ACM SIGGRAPH 2007 papers* (p. 64).

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*(1), 89–94. https://doi.org/10.1038/82947

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, 36*(6), 194–119417. https://doi.org/10.1145/3130800.3130813

Lüthi, M., Forster, A., Gerig, T., & Vetter, T. (2017). Shape modeling using Gaussian process morphable models. In G. Zheng, S. Li, & G. Székely (Eds.), *Statistical shape and deformation analysis* (pp. 165–191). Academic Press. https://doi.org/10.1016/B978-0-12-810493-4.00008-0

Lüthi, M., Gerig, T., Jud, C., & Vetter, T. (2017). Gaussian process morphable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(8), 1860–1873.

Malti, A., Bartoli, A., & Collins, T. (2011). A pixel-based approach to template-based monocular 3D reconstruction of deformable surfaces. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 1650–1657). https://doi.org/10.1109/ICCVW.2011.6130447.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology, 25*(6), 954.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 209*, 415–446.

Morel-Forster, A. (2016). Generative shape and image analysis by combining Gaussian processes and mcmc sampling. PhD thesis, University of Basel.

Moreno-Noguer, F., Salzmann, M., Lepetit, V., & Fua, P. (2009). Capturing 3D stretchable surfaces from single images in closed form. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 1842–1849). https://doi.org/10.1109/CVPR.2009.5206758.

Moreno-Noguer, F., Porta, J. M., & Fua, P. (2010). Exploring ambiguities for monocular non-rigid shape estimation. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer vision: ECCV 2010. Lecture notes in computer science* (pp. 370–383). Springer. https://doi.org/10.1007/978-3-642-15558-1_27

Östlund, J., Varol, A., Ngo, D. T., & Fua, P. (2012). Laplacian meshes for monocular 3D shape recovery. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer vision: ECCV*

2012. Lecture notes in computer science (pp. 412–425). Springer. https://doi.org/10.1007/978-3-642-33712-3_30

Ovsjanikov, M., Li, W., Guibas, L., & Mitra, N. J. (2011). Exploration of continuous variability in collections of 3D shapes. *ACM Transactions on Graphics, 30*(4), 33–13310. https://doi.org/10.1145/2010324.1964928

Patel, A., & Smith, W. A. P. (2012). Driving 3D morphable models using shading cues. *Pattern Recognition, 45*(5), 1993–2004. https://doi.org/10.1016/j.patcog.2011.11.013

Paysan, P., Knothe, R., Amberg, B., Romdhani, S. & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE international conference on advanced video and signal based surveillance* (pp. 296–301). IEEE.

Powell, L. J., Kosakowski, H. L., & Saxe, R. (2018). Social origins of cortical face areas. *Trends in Cognitive Sciences, 22*(9), 752–763.

Ranjan, A., Bolkart, T., Sanyal, S., & Black, M. J. (2018). Generating 3D faces using convolutional mesh autoencoders. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision: ECCV 2018. Lecture notes in computer science* (pp. 725–741). Springer. https://doi.org/10.1007/978-3-030-01219-9_43

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*. Springer.

Salzmann, M., & Fua, P. (2011). Linear local models for monocular reconstruction of deformable surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(5), 931–944. https://doi.org/10.1109/TPAMI.2010.158

Salzmann, M., Urtasun, R., & Fua, P. (2008). Local deformation models for monocular 3D shape recovery. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1–8). https://doi.org/10.1109/CVPR.2008.4587499.

Salzmann, M., Moreno-Noguer, F., Lepetit, V., & Fua, P. (2008). Closed-form solution to non-rigid 3D surface registration. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer vision: ECCV 2008. Lecture notes in computer science* (pp. 581–594). Springer. https://doi.org/10.1007/978-3-540-88693-8_43

Schönborn, S., Egger, B., Forster, A., & Vetter, T. (2015). Background modeling for generative image models. *Computer Vision and Image Understanding, 136*, 117–127.

Schönborn, S., Egger, B., Morel-Forster, A., & Vetter, T. (2017). Markov chain Monte Carlo for automated face image analysis. *International Journal of Computer Vision, 123*(2), 160–183.

Shaji, A., Varol, A., Torresani, L., & Fua, P. (2010). Simultaneous point matching and 3D deformable surface reconstruction. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 1221–1228). https://doi.org/10.1109/CVPR.2010.5539827.

Slater, A., Von der Schulenburg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S., & Samuels, C. (1998). Newborn infants prefer attractive faces. *Infant Behavior and Development, 21*(2), 345–354.

Styner, M. A., Rajamani, K. T., Nolte, L.-P., Zsemlye, G., Székely, G., Taylor, C. J., & Davies, R. H. (2003). Evaluation of 3d correspondence methods for model building. In *Biennial international conference on information processing in medical imaging* (pp. 63–75). Springer.

Sutherland, S., Egger, B., & Tenenbaum, J. (2021). Building 3d morphable models from a single scan. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2514–2524).

Szabó, A., Meishvili, G., & Favaro, P. (2019). Unsupervised generative 3d shape learning from natural images. arXiv preprint arXiv:1910.00287.

Tegang, N. H. N., Fouefack, J.-R., Borotikar, B., Burdin, V., Douglas, T. S., & Mutsvangwa, T. E. (2020). A Gaussian process model based generative framework for data augmentation of multi-modal 3d image volumes. In *International workshop on simulation and synthesis in medical imaging* (pp. 90–100). Springer.

Tewari, A., Seidel, H.-P., Elgharib, M., & Theobalt, C., et al. (2020). Learning complete 3d morphable face models from images and videos. arXiv preprint arXiv:2010.01679.

Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., & Theobalt, C. (2018). Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2549–2559).

Tran, L., Liu, F., & Liu, X. (2019). Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1126–1135).

Tran, L., & Liu, X. (2019). On learning 3D face morphable model from in-the-wild images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 157–171. https://doi.org/10.1109/TPAMI.2019.2927975.

Tuan Tran, A., Hassner, T., Masi, I., & Medioni, G. (2017). Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5163–5172).

Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(4), 376–380. https://doi.org/10.1109/34.88573

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology, 69*(10), 1996–2019.

Wah, C., Branson, S.,Welinder, P., Perona, P., & Belongie, S. J. (2011). The Caltech-UCSD Birds-200-2011 dataset. California Institute of Technology, No. CNS-TR-2011-001. https://www.vision.caltech.edu/datasets/cub_200_2011/

Wu, S., Rupprecht, C., & Vedaldi, A. (2020). Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1–10).

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances, 6*(10), 5979.

Yu, R., Russell, C., Campbell, N. D. F., & Agapito, L. (2015). Direct, dense, and deformable: Template-based non-rigid 3D reconstruction from RGB video. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 918–926). https://doi.org/10.1109/ICCV.2015.111.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.

Zhang, R., Tsai, P.-S., Cryer, J. E., & Shah, M. (1999). Shape-from-shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(8), 690–706. https://doi.org/10.1109/34.784284

Zivanov, J., Forster, A., Schönborn, S., & Vetter, T. (2013). Human face shape analysis under spherical harmonics illumination considering self occlusion. In *2013 International conference on biometrics (ICB)* (pp. 1–8). https://doi.org/10.1109/ICB.2013.6612967.