We introduce a comprehensive framework for learning physically based face models from highly constrained facial scan data. Our deep learning based approach for 3D morphable face modeling seizes the fidelity of nearly 4000 high resolution face scans encompassing expression and identity separation (a). The model (b) combines a multitude of anatomical and physically based face attributes to generate an infinite number of digitized faces (c). Our model generates faces at pore level geometry resolution (d).

Abstract

Based on a combined data set of 4000 high resolution facial scans, we introduce a non-linear morphable face model, capable of producing multifarious face geometry of pore-level resolution, coupled with material attributes for use in physically-based rendering. We aim to maximize the variety of the participant’s face identities, while increasing the robustness of correspondence between unique components, including middle-frequency geometry, albedo maps, specular intensity maps and high-frequency displacement details. Our deep learning based generative model learns to correlate albedo and geometry, which ensures the anatomical correctness of the generated assets. We demonstrate potential use of our generative model for novel identity generation, model fitting, interpolation, animation, high fidelity data visualization, and low-to-high resolution data domain transferring. We hope the release of this generative model will encourage further cooperation between all graphics, vision, and data focused professionals, while demonstrating the cumulative value of every individual’s complete biometric profile.

1. Introduction

Graphical virtual representations of humans are at the center of many endeavors in the fields of computer vision and graphics, with applications ranging from cultural media such as video games, film, and telecommunication to medical, biometric modeling, and forensics [6].

Designing, modeling, and acquiring high fidelity data for face models of virtual characters is costly and requires specialized scanning equipment and a team of skilled artists and engineers [17, 5, 37]. Due to limiting and restrictive data policies of VFX studios, in conjunction with the absence of a shared platform that regards the sovereignty of, and incentives for the individuals’ data contributions, there is a large discrepancy in the fidelity of models trained on publicly available data, and those used in large budget game and film production. A single, unified model would democratize the use of generated assets, shorten production cycles and boost quality and consistency, while incentivizing innovative applications in many markets and fields of research.

The unification of a facial scan data set in a 3D morphable face model (3DMM) [7, 12, 41, 6] promotes the favorable property of representing facial scan data in a compact form, retaining the statistical properties of the source
without exposing the characteristics of any individual data point in the original data set.

Previous methods, including traditional methods [7, 12, 27, 34, 16, 9], or deep learning [42, 38] to represent 3D face shapes; lack high resolution (sub-millimeter, < 0.5 mm) geometric detail, use limited representations of facial anatomy, or forgo the physically based material properties required by modern visual effects (VFX) production pipelines. Physically based material intrinsics have proven difficult to estimate through the optimization of unconstrained image data due to ambiguities and local minima in analysis-synthesis problems, while highly constrained data capture remains precise but expensive [5]. Although variations occur due to different applications, most face representations used in VFX employ a set of texture maps of at least 4096 × 4096 (4K) pixels resolution. At a minimum, this set incorporates diffuse albedo, specular intensity, and displacement (or surface normals).

Our goal is to build a physically-based, high-resolution generative face model to begin bridging these parallel, but in some ways divergent, visualization fields; aligning the efforts of vision and graphics researchers. Building such a model requires high-resolution facial geometry, material capturing and automatic registration of multiple assets. The handling of said data has traditionally required extensive manual work, thus scaling such a database is non-trivial. For the model to be light weight these data need to be compressed into a compact form that enables controlled reconstruction based on novel input. Traditional methods such as PCA [7] and bi-linear models [12] – which are limited by memory size, computing power, and smoothing due to inherent linearity – are not suitable for high-resolution data.

By leveraging state-of-the-art physically-based facial scanning [17, 25], in a Light Stage setting, we enable acquisition of diffuse albedo and specular intensity texture maps in addition to 4K displacement. All scans are registered using an automated pipeline that considers pose, geometry, anatomical morphometrics, and dense correspondence of 26 expressions per subject. A shared 2D UV parameterization data format [15, 43, 38], enables training of a non-linear 3DMM, while the head, eyes, and teeth are represented using a linear PCA model. Hence, we propose a hybrid approach to enable a wide set of head geometry assets as well as avoiding the assumption of linearity in face deformations.

Our model fully disentangles identity from expressions, and provides manipulation using a pair of low dimensional feature vectors. To generate coupled geometry and albedo, we designed a joint discriminator to ensure consistency, along with two separate discriminators to maintain their individual quality. Inference and up-scaling of before-mentioned skin intrinsics enable recovery of 4K resolution texture maps.

Our main contributions are:

- The first published upscaling of a database of high resolution (4K) physically based face model assets.
- A cascading generative face model, enabling control of identity and expressions, as well as physically based surface materials modeled in a low dimensional feature space.
- The first morphable face model built for full 3D real time and offline rendering applications, with more relevant anatomical face parts than previously seen.

2. Related Work

Facial Capture Systems Physical object scanning devices span a wide range of categories; from single RGB cameras [14, 39], to active [3, 17], and passive [4] light stereo capture setups, and depth sensors based on time-of-flight or stereo re-projection. Multi-view photogrammetry (MVS) [4] is the most readily available method for 3D face capturing. However, due to its many advantages over other methods (capture speed, physically-based material capturing, resolution), polarized spherical gradient illumination scanning [17] remains state-of-the-art for high-resolution facial scanning. A mesoscopic geometry reconstruction is bootstrapped using an MVS prior, utilizing omni-directional illumination, and progressively finalized using a process known as photometric stereo [17]. The algorithm promotes the physical reflectance properties of dielectric materials such as skin; specifically the separable nature of specular and subsurface light reflections [29]. This enables accurate estimation of diffuse albedo and specular intensity as well as pore-level detailed geometry.

3D Morphable Face Models The first published work on morphable face models by Blanz and Vetter [7] represented faces as dense surface geometry and texture, and modeled both variations as separate PCA models learned from around 200 subject scans. To allow intuitive control; attributes, such as gender and fullness of faces, were mapped to components of the PCA parameter space. This model, known as the Basel Face Model [33] was released for use in the research community, and was later extended to a more diverse linear face model learnt from around 10,000 scans [9, 8].

To incorporate facial expressions, Vlasic et al. [45] proposed a multi-linear model to jointly estimate the variations in identity, viseme, and expression, and Cao et al. [12] built a comprehensive bi-linear model (identity and expression) covering 20 different expressions from 150 subjects learned from RGBD data. Both of these models adopt a tensor-based method under the assumption that facial expressions can be modeled using a small number of discrete
poses, corresponded between subjects. More recently, Li et al. [27] released the FLAME model, which incorporates both pose-dependent corrective blendshapes, and additional global identity and expression blendshapes learnt from a large number of 4D scans.

To enable adaptive, high level, semantic control over face deformations, various locality-based face models have been proposed. Neumann et al. [32] extract sparse and spatially localized deformation modes, and Brunton et al. [10] use a large number of localized multilinear wavelet modes. As a framework for anatomically accurate local face deformations, the Facial Action Coding System (FACS) by Ekman [13] is widely adopted. It decomposes facial movements into basic action units attributed to the full range of motion of all facial muscles.

Morphable face models have been widely used for applications like face fitting [7], expression manipulation [12], real-time tracking [41], as well as in products like Apple’s ARKit. However, their use cases are often limited by the resolution of the source data and restrictions of linear models causing smoothing in middle and high frequency geometry details (e.g. wrinkles, and pores). Moreover, to the best of our knowledge, all existing morphable face models generate texture and geometry separately, without considering the correlation between them. Given the specific and varied ways in which age, gender, and ethnicity are manifested within the spectrum of human life, ignoring such correlation will cause artifacts; e.g. pairing an African-influenced albedo to an Asian-influenced geometry.

Image-based Detail Inference To augment the quality of existing 3DMMs, many works have been proposed to infer the fine-level details from image data. Skin detail can be synthesized using data-driven texture synthesis [20] or statistical skin detail models [18]. Cao et al. [11] used a probability map to locally regress the medium-scale geometry details, where a regressor was trained from captured patch pairs of high-resolution geometry and appearance. Saito et al. [35] presented a texture inference technique using a deep neural network-based feature correlation analysis.

GAN-based Image-to-Image frameworks [22] have proven to be powerful for high-quality detail synthesis, such as the coarse [44], medium [36] or even mesoscopic [21] scale facial geometry inferred directly from images. Beside geometry, Yamaguchi et al. [47] presented a comprehensive method to infer facial reflectance maps (diffuse albedo, specular intensity, and medium- and high-frequency displacement) based on single image inputs. More recently, Nagano et al. [31] proposed a framework for synthesizing arbitrary expressions both in image space and UV texture space, from a single portrait image. Although these methods can synthesize facial geometry or/and texture maps from a given image, they don’t provide explicit parametric controls of the generated result.

3. Database

3.1. Data Capturing and Processing

Data Capturing Our Light Stage scan system employs photometric stereo [17] in combination with monochrome color reconstruction using polarization promotion [25] to allow for pore level accuracy in both the geometry reconstruction and the reflectance maps. The camera setup (Fig.1) was designed for rapid, database scale, acquisition by the use of Ximea machine vision cameras which enable faster streaming and wider depth of field than traditional DSLRs [25]. The total set of 25 cameras consists of eight 12MP monochrome cameras, eight 12MP color cameras, and nine 4MP monochrome cameras. The 12MP monochrome cameras allow for pore level geometry, albedo, and specular reflectance reconstruction, while the additional cameras aid in stereo base mesh-prior reconstruction.

To capture consistent data across multiple subjects with maximized expressiveness, we devised a FACS set [13] which combines 40 action units to a condensed set of 26 expressions. In total, 79 subjects, 34 female, and 45 male, ranging from age 18 to 67, were scanned performing the 26 expressions. To increase diversity, we combined the data set with a selection of 99 Triplegangers [2] full head scans; each with 20 expressions. Resolution and extent of the two data sets are shown in Table 1. Fig. 2 shows the age and ethnicity (multiple choice) distributions of the source data.

![Figure 1: Capture system and camera setup. Left: Light Stage capturing system. Right: camera layout.](image)

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>4k × 4k</td>
<td>3.9M</td>
<td>4k × 4k</td>
<td>79</td>
<td>26</td>
</tr>
<tr>
<td>TG</td>
<td>8k × 8k</td>
<td>3.5M</td>
<td>N/A</td>
<td>99</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 1: Resolution and extent of the datasets. (a) Albedo resolution. (b) Geometry resolution. (c) Specular intensity resolution. (d) # of subjects. (f) # of expressions per subject.
and non-rigid ICP new scans. We fit expressions using generic blendshapes were immediately added to the PCA basis for registering sockets and skull shapes. The resulting set of neutral scans fitted to retain the fitting accuracy of the eyeballs, mouth els (Basel mesh. Next, Laplacian deformation is applied to deform on a combination and extrapolation of two existing mod-
els (Basel [33] and Face Warehouse [12]) is used to fit the mesh. Next, Laplacian deformation is applied to deform the face area to further minimize the surface-to-surface error. Cases of inaccurate fitting were manually modeled and fitted to retain the fitting accuracy of the eyeballs, mouth sockets and skull shapes. The resulting set of neutral scans were immediately added to the PCA basis for registering new scans. We fit expressions using generic blendshapes and non-rigid ICP [26]. Additionally, to retain texture space and surface correspondence, image space optical flow from neutral to expression scan is added from 13 different virtual camera views as additional dense constraint in the final Laplacian deformation of the face surface.

3.2. Training Data Preparation

Data format. The full set of the generic model consists of a hybrid geometry and texture maps (albedo, specular intensity, and displacement) encoded in 4K resolution, as illustrated in Fig. 3. To enable joint learning of the correlation between geometry and albedo, 3D vertex positions are rasterized to a three channel HDR bitmap of 256 × 256 pixels resolution. The face area (pink in Fig. 3) used to learn the geometry distribution in our non-linear generative model consists of m = 11892 vertices, which, if evenly spread out in texture space, would require a bitmap of resolution greater or equal to \(\sqrt{2 \times m} = 155 \times 155\), according to Nyquist’s resampling theorem. As shown in Fig. 4, the proposed resolution is enough to recover middle-frequency detail. This relatively low resolution base geometry representation enables great simplification in training data load.

Data Augmentation Since the number of subjects is limited to 178 individuals, we apply two strategies to augment the data for identity training: 1) For each source albedo, we randomly sample a target albedo within the same ethnicity and gender in the data set using [49] to transfer skin tones of target albedos to source albedos (these samples are restricted to datapoints of the same ethnicity), followed by an image enhancement [19] to improve the overall quality and remove artifacts. 2). For each neutral geometry, we add a very small expression offset using FaceWarehouse expression components with a small random weights(< ±0.5 std) to loosen the constraints of “neutral”. To augment the expressions, we add random expression offsets to generate fully controlled expressions.

4. Generative Model

An overview of our system is illustrated in Fig. 5. Given a sampled latent code \(Z_{id} \sim N(\mu_{id}, \sigma_{id})\), our Identity network generates a consistent albedo and geometry pair of neutral expression. We train an Expression network to generate the expression offset that can be added to the neutral geometry. We use random blendshape weights \(Z_{exp} \sim N(\mu_{exp}, \sigma_{exp})\) as the expression network’s input to enable manipulation of target semantic expressions. We upscale the albedo and geometry maps to 1K, and feed them into a transfer network [46] to synthetize the corresponding 1K specular and displacement maps. Finally, all the maps except for the middle frequency geometry map are upcaled to 4K using Super-resolution [24], as we observed that 256 × 256 pixels are sufficient to represent the details of
the base geometry (Section 3.2). The details of each component are elaborated on in Section 4.1, 4.2, and 4.3.

4.1. Identity Network

The goal of our Identity network is to model the cross correlation between geometry and albedo to generate consistent, diverse and biologically accurate identities. The network is built upon the Style-GAN architecture [23], that can produce high-quality, style-controllable sample images. To achieve consistency, we designed 3 discriminators as shown in Fig.6, including individual discriminators for albedo ($D_{albedo}$) and geometry ($D_{geometry}$), to ensure the quality and sharpness of the generated maps, and an additional joint discriminator ($D_{joint}$) to learn their correlated distribution. $D_{joint}$ is formulated as follows:

$$
L_{adv} = \min_{G_{id}} \max_{D_{joint}} \mathbb{E}_{x \sim p_{data}(x)} [\log D_{joint}(A)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D_{joint}(G_{id}(z)))]
$$

(1)

where $p_{data}(x)$ and $p_z(z)$ represent the distributions of real paired albedo and geometry $x$ and noise variables $z$ in the domain of $A$ respectively.

4.2. Expression Network

To simplify the learning of a wide range of diverse expressions, we represent them using vector offset maps, which also makes the learning of expressions independent from identity. Similar to the Identity network, the expression network adopts Style-GAN as the base structure. To allow for intuitive control over expressions, we use the blendshape weights, which correspond to the strength of 25 orthogonal facial activation units, as network input. We introduce a pre-trained expression regression network $R_{exp}$ to predict the expression weights from the generated image, and force this prediction to be similar to the input latent code $Z_{exp}$ so that the L1 loss $L_{exp}$ can be modeled.

$$
L_{exp} = \| Z_{exp} - Z_{exp}^{l1} \|
$$

(2)

This loss, $L_{exp}$, will be back propagated during training to enforce the orthogonality of each blending unit. We minimize the following losses to train the network:

$$
\mathcal{L} = \mathcal{L}_{exp}^{l_2} + \beta_1 \mathcal{L}_{adv} + \beta_2 \mathcal{L}_{exp}
$$

(3)

where $\mathcal{L}_{exp}^{l_2}$ is the $L_2$ reconstruction loss of the offset map and $\mathcal{L}_{adv}$ is the discriminator loss.
4.3. Inference and Super-resolution

Similar to [47]; upon obtaining albedo and geometry maps (256 x 256), we use them to infer specular and displacement maps in 1K resolution. In contrast to [47], using only albedo as input, we introduce the geometry map to form stronger constraints. For displacement, we adopted the method of [47, 21] to separate displacement into individual high-frequency and low-frequency components, which makes the problem more tractable. Before feeding the two inputs into the inference network [46], we up-sample the albedo to 1K using a super-resolution network similar to [24]. The geometry map is super-sampled using bilinear interpolation. The maps are further up-scaled from 1K to 4K using the same super-resolution network structure. Our method can be regarded as a two step cascading up-sampling strategy (256 to 1K, and 1K to 4K). This makes the training faster, and enables higher resolution in the final results.

5. Implementation Details

Our framework is implemented using Pytorch and all our networks are trained using two NVIDIA Quadro GV100s. We follow the basic training schedule of Style-GAN [23] with several modifications applied to the Expression network, like by-passing the progressive training strategy as expression offsets are only distinguishable on relatively high resolution maps. We also remove the noise injection layer, due to the input latent code \( Z_{\text{exp}} \) which enables full control of the generated results. The regression module (\( R_{\text{exp}} \)-block in Fig.7) has the same structure as the discriminator \( D_{\text{exp}} \), except for the number of channels in the last layer, as it serves as a discriminator during training. The regression module is initially trained using synthetic unit expression data generated with neutral expression and FaceWarehouse expression components, and then fine-tuned on scanned expression data. During training, \( R_{\text{exp}} \), is fixed without updating parameters. The Expression network is trained with a constant batch size of 128 on 256x256-pixel images for 40 hours. The Identity network is trained by progressively reducing the batch size from 1024 to 128 on growing image sizes ranging from 8x8 to 256x256 pixels, for 80 hours.

6. Experiments And Evaluations

6.1. Results

In Fig.10, we show the quality of our generated model rendered using Arnold. The direct output of our generative model provides all the assets necessary for physically-based rendering in software such as Maya, Unreal Engine, or Unity 3D. We also show the effect of each generated component.

6.2. Qualitative Evaluation

We show identity interpolation in Fig.8. The interpolation in latent space reflects both albedo and geometry. In contrast to linear blending, our interpolation generates subjects belonging to a natural statistical distribution.

In Fig.9, we show the generation and interpolation of our non-linear expression model. We pick two orthogonal blendshapes for each axis and gradually change the input weights. Smooth interpolation in vector space will lead to a smooth interpolation in model space.

We show nearest neighbors for generated models in the training set in Fig.11. These are found based on point-wise Euclidean distance in geometry. Albedos are compared to prove our ability to generate new models that are not merely recreations of the training set.
Figure 10: Rendered images of generated random samples. Column (a), (b), and (c) show images rendered under novel image-based HDRI lighting [48]. Column (c), (d), and (e), show geometry with albedo, specular intensity, and displacement added one at the time.

Figure 11: Nearest neighbors for generated models in training set. Top row: albedo from generated models. Bottom row: albedo of geometrically nearest neighbor in training set.

Table 2: Evaluation on our Identity generation. Both IS and FID are calculated on images rendered with independently/jointly generated albedo and geometry.

<table>
<thead>
<tr>
<th>Generation Method</th>
<th>IS↑</th>
<th>FID↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>independent</td>
<td>2.22</td>
<td>23.61</td>
</tr>
<tr>
<td>joint</td>
<td>2.26</td>
<td>21.72</td>
</tr>
<tr>
<td>ground truth</td>
<td>2.35</td>
<td>-</td>
</tr>
</tbody>
</table>

6.3. Quantitative Evaluation

We evaluate the effectiveness of our identity network’s joint generation in Table 2 by computing Frechet Inception Distances (FID) and Inception-Scores (IS) on rendered images of three categories: randomly paired albedo and geometry, paired albedo and geometry generated using our model, and ground truth pairs. Based on these results, we conclude that our model generates more plausible faces, similar to those using ground truth data pairs, than random pairing.

We also evaluate our identity network’s generalization to unseen faces by fitting 48 faces from [1]. The average Hausdorff distance is 2.8mm, which proves that our model’s capacity is not limited by the training set.

In addition, to evaluate the non-linearity of our expression network in comparison to the linear expression model of FaceWarehouse [12], we first fit all the Light Stage scans using FaceWarehouse, and get the 25 fitting weights, and expression recoveries, for each scan. We then recover the same expressions by feeding the weights to our expression network. We evaluate the reconstruction loss with mean-square error (MSE) for both FaceWarehouse’s and
Figure 12: The age distribution of the training data (a) VS. randomly generated samples (b).

Figure 13: Comparison of 3D scan fitting with Basel [7], FaceWarehouse [12], and FLAME [27]. Error maps are computed using Hausdorff distance between each fitted model and ground truth scans.

To test the extent of our identity model’s parameter space, we apply it to scanned mesh registration by reversing the GAN to fit the latent code of a target image [28]. As our model requires a 2D parameterized geometry input, we first use our linear model to align the scans using landmarks, and then parameterize it to UV space after Laplacian morphing of the surface. We compare our fitting results with widely used (linear) morphable face models in Fig.13. This evaluation does not prove the ability to register unconstrained data but shows that our model is able to reconstruct novel faces by the virtue of it’s non-linearity, to a degree unobtainable by linear models.

Another application of our model is transferring low-quality scans into the domain of our model by fitting using both MSE loss and discriminator loss. In Fig.14, we show examples of data enhancement of low resolution scans.

7. Conclusion and Limitations

Conclusion We have introduced the first published use of a high-fidelity face database, with physically-based material attributes, in generative face modeling. Our model can generate novel subjects and expressions in a controllable manner. We have shown that our generative model performs well on applications such as mesh registration and low resolution data enhancement. We hope that this work will benefit many analysis-by-synthesis research efforts through the provision of higher quality in face image rendering.

Limitations and Future work In our model, expression and identity are modeled separately without considering their correlation. Thus the reconstructed expression offset will not include middle-frequency geometry of an individual’s expression, as different subjects will have unique representations of the same action unit. Our future work will include modeling of this correlation. Since our expression generation model requires neural network inference and re-sampling of 3D geometry it is not currently as user friendly as blendshape modeling. Its ability to re-target pre-recorded animation sequences will have to be tested further to be conclusive. One issue of our identity model arises in applications that require fitting to 2D imagery, which necessitates an additional differentiable rendering component. A potential problem is fitting lighting in conjunction with shape as complex material models make the problem less tractable. A possible solution could be an image-based relighting method [40, 30] applying a neural network to convert the rendering process to an image manipulation problem. The model will be continuously updated with new features such as variable eye textures and hair as well as more anatomically relevant components such as skull, jaw, and neck joints by combining data sources through collaborative efforts. To encourage democratization and wide use cases we will explore encryption techniques such as federated learning, homomorphic encryption, and zero knowledge proofs which have the effect of increasing subjects’ anonymity.
References


