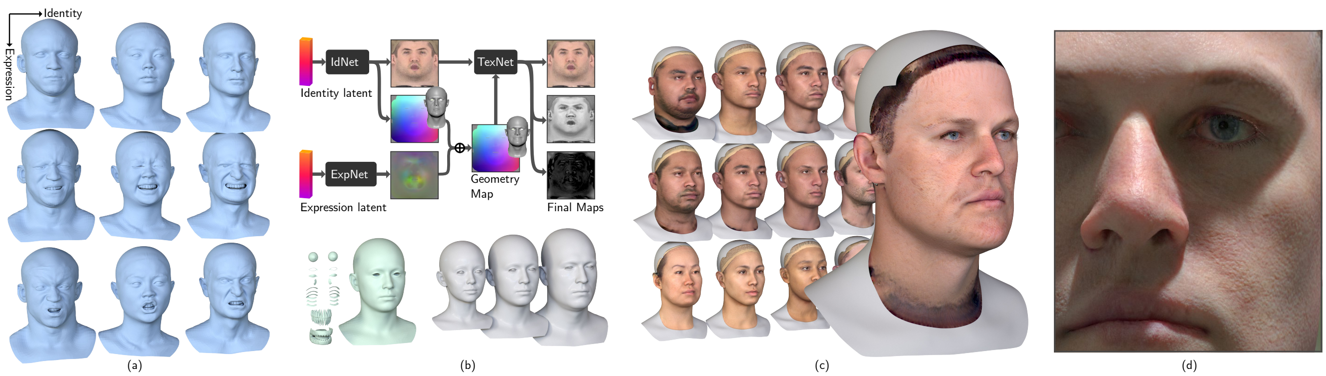
Learning Formation of Physically-Based Face Attributes

Ruilong Li[1,2*] Karl Bladin[1*] Yajie Zhao[1*] Chinmay Chinara[1] Owen Ingraham[1]
Pengda Xiang[1,2] Xinglei Ren[1] Pratusha Prasad[1] Bipin Kishore[1] Jun Xing[1] Hao Li [1,2,3]

[1]USC Institute for Creative Technologies *[2]University of Southern California* *[3]Pinscreen*



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.

*Joint first authors

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, $< 1mm$) 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 analisys-by-synthesis problems, while highly constrained data capture remains percise but expensive [6]. 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 encorporates *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 beforementioned 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 stereo-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

Figure 1: Capture system and camera setup. Left: Light Stage capturing system. Right: camera layout.

	(a)	(b)	(c)	(d)	(e)
LS	$4k \times 4k$	$3.9M$	$4k \times 4k$	79	26
TG	$8k \times 8k$	$3.5M$	N/A	99	20

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.

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.

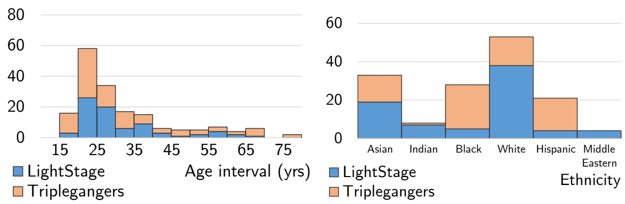
GAN-based Image-to-Image frameworks [22] have proven to be powerful for high-quality detail synthesis, such as the coarse [44], medium [36] or even mesoscopic [21] scale facial geometry inferred directly from images. Beside geometry, Yamaguchi *et al*. [47] presented a comprehensive method to infer facial 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.



(a) Age distribution (b) Ethnicity distribution

Figure 2: Distribution of age (a) and ethnicity (b) in the data sets.

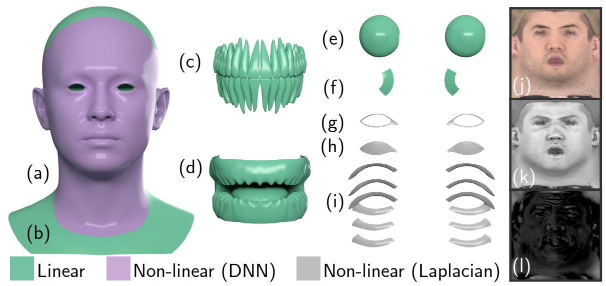


Figure 3: Our generic face model consists of multiple geometries constrained by different types of deformation. In addition to face (a), head, and neck (b), our model represents teeth (c), gums (d), eyeballs (e), eye blending (f), lacrimal fluid (g), eye occlusion (h), and eyelashes (i). Texture maps provide high resolution (4K) albedo (j), specularity (k), and geometry through displacement (l).

Processing Pipeline. Starting from the multi-view imagery, a neutral scan base mesh is reconstructed using MVS. Then a linear PCA model in our topology (See Fig.3) based on a combination and extrapolation of two existing models (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 $\lceil \sqrt{2 \times m} \rceil^2 = 155 \times 155$, 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.

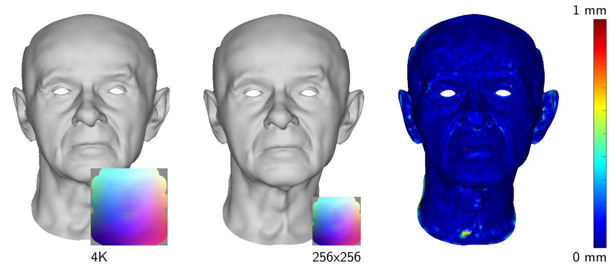


Figure 4: Comparison of base mesh geometry resolutions. Left: Base geometry reconstructed in $4K$ resolution. Middle: Base geometry reconstructed in 256×256 resolution. Right: Error map showing the Hausdorff distance in the range $(0mm, 1mm)$, with a mean error of $0.068mm$.

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($< \pm 0.5$ 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 synthesize the corresponding $1K$ specular and displacement maps. Finally, all the maps except for the middle frequency geometry map are upscaled to $4K$ using *Super-resolution* [24], as we observed that 256×256 pixels are sufficient to represent the details of

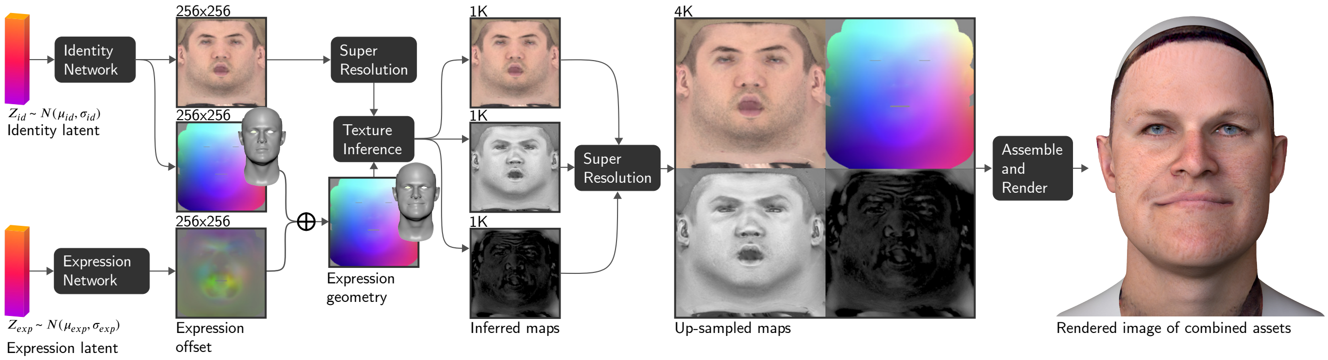


Figure 5: Overview of generative pipeline. Latent vectors for identity and expression serve as input for generating the final face model.

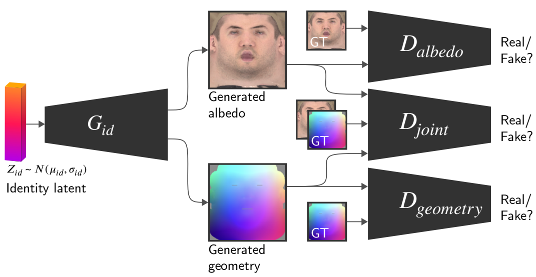


Figure 6: *Identity* generative network. The identity generator G_{id} produces albedo and geometry which get checked against ground truth (GT) data by the discriminators, D_{albedo}, D_{joint}, and $D_{geometry}$ during training.

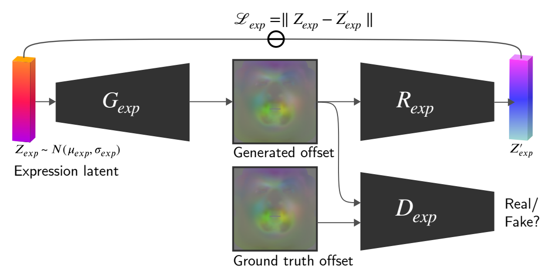


Figure 7: *Expression* generative network. The expression generator G_{exp} generates offsets which get checked against ground truth offsets by the discriminator D_{exp}. The regressor R_{exp} produces an estimate of the latent code Z'_{exp} so that the L1 loss $\mathcal{L}_{exp}$ can be modeled.

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:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = \min_{G_{id}} \max_{D_{joint}} \; & \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})} \left[\log D_{joint}(\boldsymbol{\mathcal{A}}) \right] + \\ & \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})} \left[\log \left(1 - D_{joint}(G_{id}(\boldsymbol{z}))\right) \right]. \end{aligned} \tag{1}$$

where $p_{\text{data}}(\boldsymbol{x})$ and $p_{\boldsymbol{z}}(\boldsymbol{z})$ represent the distributions of real paired albedo and geometry $\boldsymbol{x}$ and noise variables $\boldsymbol{z}$ in the domain of $\boldsymbol{\mathcal{A}}$ 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}. We then force the generator to understand the input latent code Z_{exp} under the perspective of the pre-trained expression regression network. As a result, each dimension of the latent code Z_{exp} will control the corresponding expression defined in the original blendshape set. The loss we introduce here is:

$$\mathcal{L}_{exp} = \| Z_{exp} - Z'_{exp} \| \tag{2}$$

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

$$\mathcal{L} = \mathcal{L}_{l_2}^{exp} + \beta_1 \mathcal{L}_{adv}^{exp} + \beta_2 \mathcal{L}_{exp} \tag{3}$$

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

4.3. Inference and Super-resolution

Similar to [47]; upon obtaining albedo and geometry maps (256×256), 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 in to 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_{exp} which enables full control of the generated results. The regression module (R_{exp}-block in Fig.7) has the same structure as the discriminator D_{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_{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.

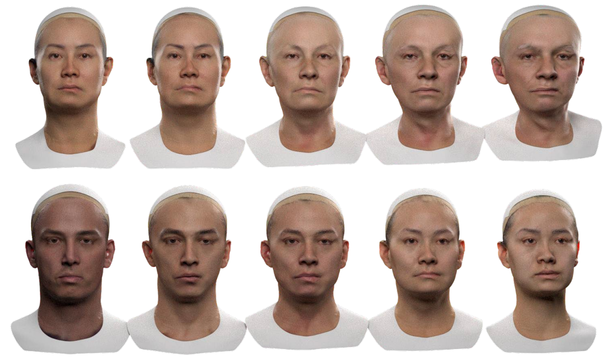


Figure 8: Non-linear identity interpolation between generated subjects. Age (top) and gender (bottom) are interpolated from left to right.

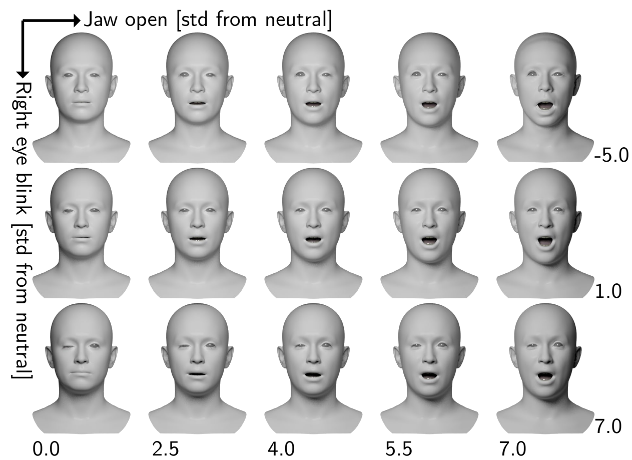


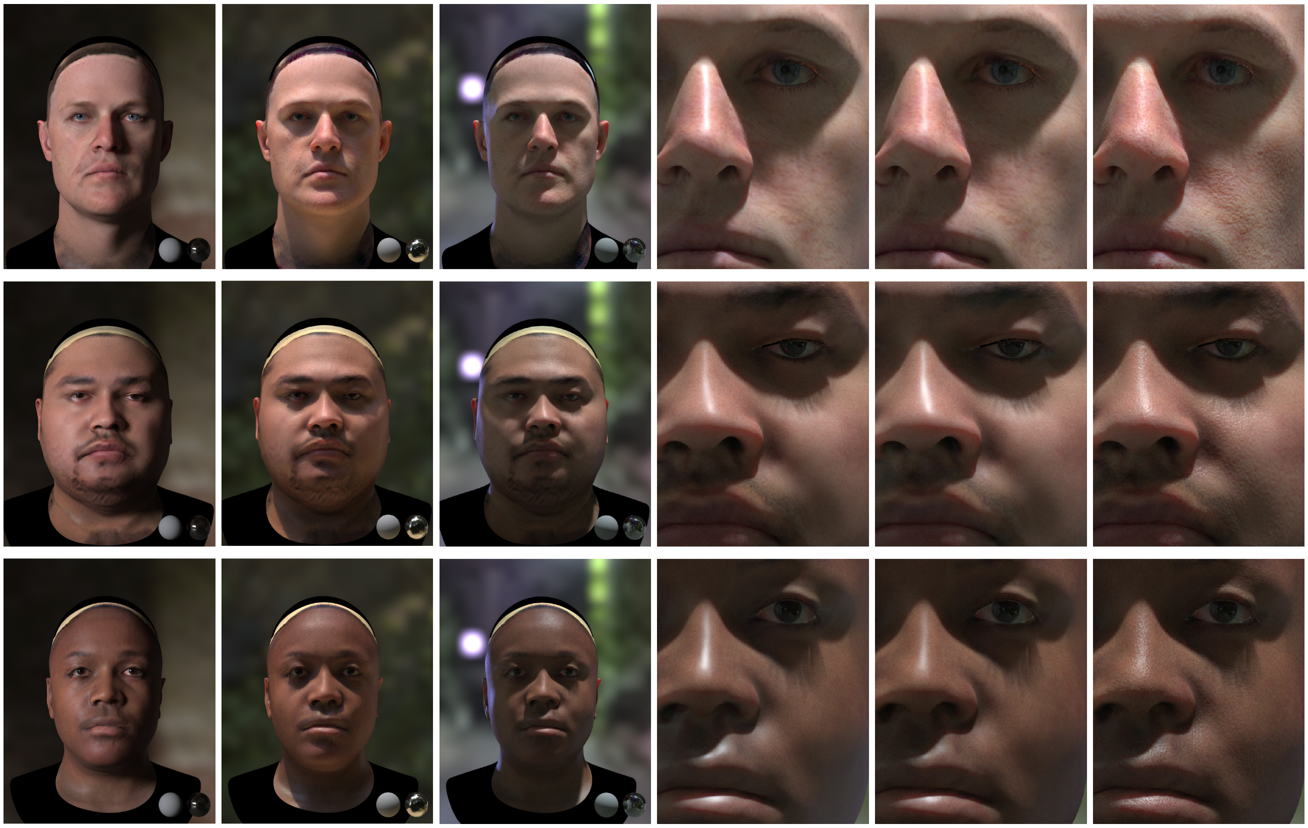
Figure 9: Non linear expression interpolation using generative expression network. Combinations of two example shapes are displayed in a grid where the number of standard deviations from the generic neutral model define the extent of an expression shape.

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.



(a) (b) (c) (d) (e) (f)

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.

Generation Method	IS↑	FID↓
independent	2.22	23.61
joint	**2.26**	**21.72**
groud truth	2.35	-

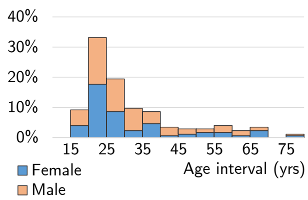
Table 2: Evaluation on our *Identity* generation. Both IS and FID are calculated on images rendered with independently/jointly generated albedo and geometry.

6.3. Quantitative Evaluation

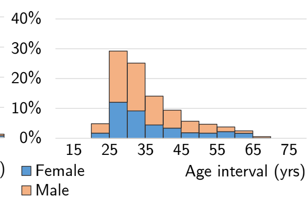
We evaluate the effectiveness of our identity network's *joint* generation in Table 2 by computing Frechet Inception Distances (FID) and Inception-Scores (IS) on rendered images of three categories: randomly paired albedo and geometry, paired albedo and geometry generated using our model, and ground 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



(a) Training data



(b) Generated data

Figure 12: The age distribution of the training data (a) *VS.* randomly generated samples (b).

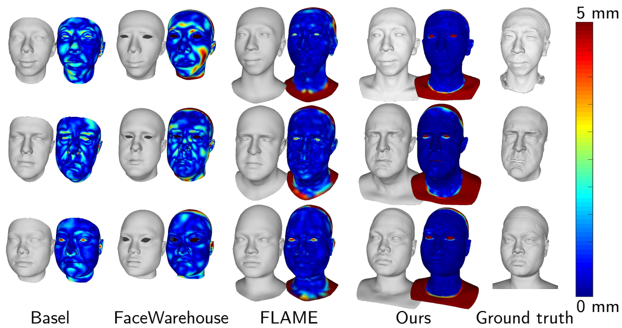


Figure 13: Comparison of 3D scan fitting with Basel [7], FaceWareHouse [12], and FLAME [27]. Error maps are computed using Hausdorff distance between each fitted model and ground truth scans.

our model's reconstructions. On average, our method's MSE is $1.2mm$ while FaceWarehouse's is $2.4mm$. This shows that for expression fitting, our non-linear model numerically outperforms a linear model of the same dimensionality.

To demonstrate our generative identity model's coverage of the training data, we show the gender, and age distributions of the original training data and 5000 randomly generated samples in Fig.12. The generated distributions are well aligned with the source.

6.4. Applications

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.

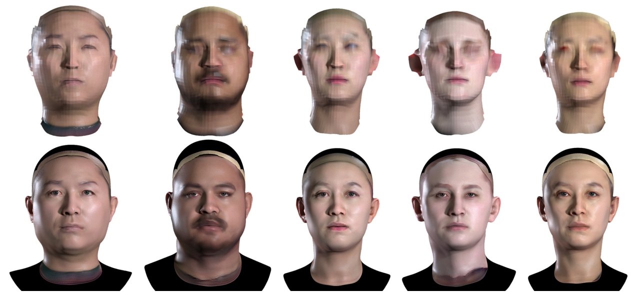


Figure 14: Low-quality data domain transfer. Top row: Models with low resolution geometry and albedo. Bottom row: Enhancement result using our model.

7. Conclusion and Limitations

Conclusion We have introduced the first published use of a high-fidelity face database, with physically-based marerial 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 prerecorded 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

[1] 3d scan store: Male and female 3d head model 48 x bundle. https://www.3dscanstore.com/3d-head-models/female-retopologised-3d-head-models/male-female-3d-head-model-48xbundle. Online; Accessed: 2019-11-22. 7

[2] Triplegangers. https://triplegangers.com/. Online; Accessed: 2019-11-22. 3

[3] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH Courses*, 2009. 2

[4] Thabo Beeler, Bernd Bickel, Paul A. Beardsley, Bob Sumner, and Markus H. Gross. High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (TOG)*, 2010. 2

[5] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul A. Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Groß. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, 2011. 1

[6] Ayush Tewari Stefanie Wuhrer Michael Zollhoefer Thabo Beeler Florian Bernard Timo Bolkart Adam Kortylewski Sami Romdhani Christian Theobalt Volker Blanz Thomas Vetter Bernhard Egger, William A. P. Smith. 3d morphable face models–past, present and future. *arXiv preprint arXiv:1909.01815*, 2019. 1, 2

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM Transactions on Graphics (TOG)*, SIGGRAPH '99, 1999. 1, 2, 3, 8

[8] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126:233–254, 2017. 2

[9] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, 2016. 2

[10] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[11] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (TOG)*, 34(4), July 2015. 3

[12] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Zhou Kun. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 2014. 1, 2, 3, 4, 7, 8

[13] Paul Ekman and Wallace V. Friesen. Facial action coding system: a technique for the measurement of facial movement. In *Consulting Psychologists Press*, 1978. 3

[14] Pablo Garrido, Levi Valgaert, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics (TOG)*, 32(6), Nov. 2013. 2

[15] Baris Gecer, Alexander Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. *arXiv preprint arXiv:1909.02215*, 2019. 2

[16] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018. 2

[17] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul E. Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)*, 30:129, 2011. 1, 2, 3

[18] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas A. Funkhouser. A statistical model for synthesis of detailed facial geometry. *ACM Transactions on Graphics (TOG)*, 25:1025–1034, 2006. 3

[19] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM transactions on graphics (TOG)*, 30(4):70, 2011. 4

[20] Antonio Haro, Irfan A. Essa, and Brian K. Guenter. Real-time photo-realistic physically based rendering of fine scale human skin structure. In *Rendering Techniques*, 2001. 3

[21] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul E. Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 6

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2016. 3

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5, 6

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 4, 6

[25] Chloe LeGendre, Kalle Bladin, Bipin Kishore, Xinglei Ren, Xueming Yu, and Paul Debevec. Efficient multispectral facial capture with monochrome cameras. In *Color and Imaging Conference*, volume 2018, pages 187–202, 2018. 2, 3

[26] Hao Li, Robert W. Sumner, and Mark Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Proceedings of the Symposium on Geometry Processing*, SGP '08, pages 1421–1430, Aire-la-Ville, Switzerland, Switzerland, 2008. Eurographics Association. 4

[27] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194, 2017. 2, 3, 8

[28] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017. 8

[29] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Félix Chabert, Malte Weiss, and Paul E. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Rendering Techniques*, 2007. 2

[30] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. volume 38, July 2019. 8

[31] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37:258:1–258:12, 2018. 3

[32] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus A. Magnor, and Christian Theobalt. Sparse localized deformation components. *ACM Transactions on Graphics (TOG)*, 32:179:1–179:10, 2013. 3

[33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 2, 4

[34] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[35] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2326–2335, 2016. 3

[36] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1585–1594, 2017. 3

[37] Yeongho Seol, Wan-Chun Ma, and J. P. Lewis. Creating an actor-specific facial rig from performance capture. In *Proceedings of the 2016 Symposium on Digital Production*, DigiPro '16, 2016. 1

[38] Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *arXiv preprint arXiv:1901.06551*, 2019. 2

[39] Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33:222:1–222:13, 2014. 2

[40] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):79, 2019. 8

[41] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: real-time face capture and reenactment of rgb videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[42] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1126–1135, 2019. 2

[43] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[44] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. Face normals "in-the-wild" using fully convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 340–349, 2017. 3

[45] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, 2005. 2

[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 6

[47] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37:162:1–162:14, 2018. 3, 6

[48] G. Zaal. HDRI Haven. `https://hdrihaven.com/hdris/`. Online; Accessed: 2019-11-22. 7

[49] Yajie Zhao, Qingguo Xu, Weikai Chen, Chao Du, Jun Xing, Xinyu Huang, and Ruigang Yang. Mask-off: Synthesizing face images in the presence of head-mounted displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 267–276. IEEE, 2019. 4