Eurographics Symposium on Rendering 2025
B. Wang and A. Wilkie
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 4

Controllable Biophysical Human Faces

Minghao Liu[†2] Stephane Grabli[1] Sébastien Speierer[1] Nikolaos Sarafianos[1]
Lukas Bode[1] Matt Chiang[1] Christophe Hery[1] James Davis[2] Carlos Aliaga[1]
[1]Meta Reality Labs [2]University of California Santa Cruz

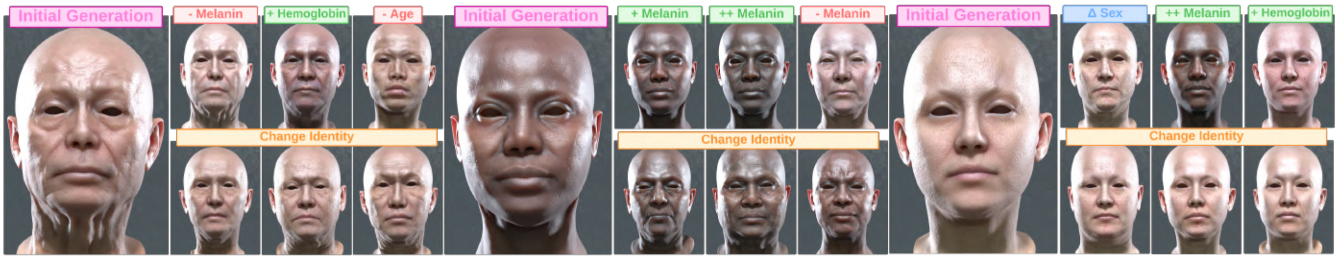

Figure 1: Our generative model can produce a wide variety of faces within a plausible biophysical manifold while enabling manipulations within realistic ranges. We showcase three generated examples conditioned on a random identity as well as random biophysical and demographic characteristics. For each example, the first row illustrates manipulations of these conditions while maintaining the same identity. In the second row, we keep the original conditions fixed and sample random identities.

We present a novel generative model that synthesizes photorealistic, biophysically plausible faces by capturing the intricate relationships between facial geometry and biophysical attributes. Our approach models facial appearance in a biophysically grounded manner, allowing for the editing of both high-level attributes such as age and gender, as well as low-level biophysical properties such as melanin level and blood content. This enables continuous modeling of physical skin properties that correlate changes in skin properties with shape changes. We showcase the capabilities of our framework beyond its role as a generative model through two practical applications: editing the texture maps of 3D faces that have already been captured, and serving as a strong prior for face reconstruction when combined with differentiable rendering. Our model allows for the creation of physically-based relightable, editable faces with consistent topology and uv layout that can be integrated into traditional computer graphics pipelines.

Keywords: biophysical face synthesis, controllable 3D generation, skin appearance modeling, facial editing, diffusion models

1. Introduction

The synthesis of photorealistic human faces has long been a challenging task in computer graphics and vision. A person's identity is defined by the combination of overall shape (geometry), features like wrinkles and pores (medium and high-frequency details), and skin tone, often presenting heterogeneities such as freckles and moles (texture). These elements are correlated and influenced by factors such as gender, age, and ethnicity. Accurately reproducing these aspects is crucial for achieving photorealism, as any decorrelation can lead to mischaracterization or ultimately result in the uncanny valley effect [MMK12, ZAJ*15].

The introduction of 3D morphable models (3DMM) [BV99] allowed for controlling the shape and appearance of a face by means of a parametric space, which has led to a vast literature on the topic [EST*20]. However, existing approaches typically focus on specific aspects of face representation, often emphasizing geometry or texture alone. Recently, a branch of neural implicit models [GKG*23, ZWS*23, KQG*23, GKR*24, KGN24, GKG*24, LDML*24, SSS*24, LKB*24] have demonstrated impressive results in capturing photorealistic faces, disentangling identity and expression, but they lack editability through meaningful parameters. Moreover, most models do not account for correlations between shape and appearance as a function of features like age, gender or skin properties. Those that attempt to model these correlations often rely on human-made labels [LBZ*20, DTP23], introducing inaccuracies that compromise the authenticity of the generated faces.

In this paper, we introduce a novel generative model that synthesizes complete 3D faces, achieving both photorealism and bio-

[†] Work done during internship at Meta Reality Labs

physical grounding. Our approach leverages real human captures to model facial geometry and skin properties, enabling the editing of high-level attributes like age and gender, as well as low-level biophysical properties such as melanin and blood concentrations. A key innovation of our approach is the continuous modeling of biophysical skin properties, estimated directly from the albedo textures [AXX*23]. This eliminates the reliance on subjective human annotations, ensuring that the generated faces and the editing space are biophysically plausible, allowing for meaningful high- and low-level manipulations that maintain correlations between shape features, skin properties, and demographic data.

At its core, our approach encodes geometry and appearance within the texture space by representing geometric deformations through texture-encoded data, alongside the appearance maps. This allows us to establish a unified framework where large scale geometry, medium and high frequency geometric details, as well as albedo and their correlations, can be learned. Specifically, we integrate a base mesh with both low- and high-frequency deformation maps, in addition to an albedo map. We then employ a latent diffusion-based generator paired with a refiner network to produce the aforementioned maps at mid-resolution and their high-resolution counterparts, respectively. This comprehensive modeling allows for the creation of face assets that are ready for integration into computer graphics pipelines.

Beyond generation, we demonstrate the usefulness of our model in two applications: a) editing the texture maps of 3D faces that have already been captured, and b) providing a strong prior to guide face reconstruction when combined with differentiable rendering techniques. In summary, our contributions are:

- A multi-modal generative method for synthesizing photorealistic, biophysically plausible faces that are ready for computer graphics engines, being relightable and riggable due to their foundation on triangle meshes with consistent topology and texture parametrization.
- Continuous modeling of skin appearance by extracting objective labels based on the biophysical properties of skin. This eliminates the reliance on subjective human annotations. In conjunction with other demographic data like age and gender, the continuous modeling of skin properties enables high- and low-level manipulations while retaining correlations between skin appearance and shape features.
- Our model serves as a robust prior in inverse rendering problems, where the appearance of the face is the result of the intricate interactions between light and complex material properties over the facial shape features. Its effectiveness is demonstrated through various examples in combination with differentiable path tracing techniques, enabling plausible edits once the face is reconstructed.

2. Related Work

The field of 3D face generation and reconstruction has been a long-standing focus of research. Due to the vast literature available, we refer readers to comprehensive surveys on 3D Morphable Models (3DMM) [EST*20] and face modeling [TLL23]. For implicit representations, we recommend recent reviews on neural rendering [PYG*24], diffusion models [PYG*24] and 3D generation [LZK*24]. In this section, we focus on closely related work that shares similarities with our approach in terms of representation, methodology, and outcome.

2.1. Biophysical Skin Models

Several biophysical skin models have been developed to simulate the complex interactions between light and human skin. Tsumura et al. [THM99, TOS*03] introduced an image-based technique for separating melanin and hemoglobin distributions in skin using independent-component analysis. Donner and Jensen [DJ06] proposed a two-layer skin model with parameters controlling oil, melanin, and hemoglobin concentrations. Later, a multi-layered model [DWd*08] featured spatially-varying absorption and scattering parameters. Jimenez et al. [JSB*10] modeled changes in skin appearance due to varying melanin and hemoglobin concentrations caused by emotional or physical states. Krishnaswamy and Baranoski [KB04, BK10] produced a seminal paper for very detailed modeling with many layers and all the chromophores involved in skin's optical properties. Iglesias Guitian [IGAJG15] simulates the effects of skin aging, while Chen et al. [CBKM15] provide a comprehensive understanding of skin's spectral characteristics through hyperspectral skin modeling.

2.2. Generative Models of Human Faces

Building on the seminal work on morphable face models by Blanz and Vetter [BV99], which used two separate linear models to capture variations in texture and geometry from approximately 200 subject scans, numerous subsequent parametric models have been developed aiming to provide more intuitive control, incorporate additional aspects such as facial expressions, and extend to more diverse identity representations using larger datasets like *BFM*, *LSFM* or *FLAME* [VBPP05, PKA*09, CWZ*14, BRZ*16, BRP*18, LBB*17]. In recent years, there has been a surge of interest in generative models of high-quality, photorealistic face assets. Chandran et al. [CBGB20] employs variational autoencoders (VAEs) to disentangle expression and identity on 224 captured subjects, modeling geometry and albedo. Subsequently, several Generative Adversarial Network (GAN)-based approaches have been proposed. Gecer et al. [GLP*20] uses a multi-branch GAN to synthesize texture and head-space geometry and low frequency normals, all defined in texture space. However, this approach lacks controllable edits and does not model medium and high frequency geometric details like bumps or normals. Exclusively focused on high-quality, high-resolution PBR maps, *GANtlitz* [GCM*24] proposed a patch-based architecture built upon *StyleGAN2* [KLA*20]. This was trained with under a hundred facial captures to generate albedo, displacement, specular, and normal maps, ensuring consistency across different map types. However, the model omits shape modeling and does not show reconstruction capabilities.

Closer to our work, *AlbedoGAN* [RGP*24] proposes a generative 3D face model that leverages *FLAME* fitting and *StyleGAN2* to achieve state-of-the-art shape reconstruction accuracy. However, this approach omits modeling PBR materials and still contains baked information in the albedo textures. Another recent work, *Dreamface* [ZQL*23], utilizes latent diffusion models and score distillation sampling (SDS) to generate and edit full heads through text prompts. However, this method has limited fine-grain control over appearance features beyond text and does not explore face inversion. In contrast, Li et al. [LBZ*20] employ *StyleGAN* to generate full faces comprising shape, middle-frequency geometry, albedo, specular, and high-frequency displacement maps learned from a large dataset of high-quality facial scans. Similarly, Deb et al. [DTP23] synthesize physically-based rendering (PBR) maps including albedo, normals, glossiness, and specular using a shape-conditioned texture generator supervised by multiple discriminators at different resolutions.

To our knowledge, we are the first generative face model to offer continuous control over skin properties and model the correlations in shape and skin properties through a unified framework where both the overall face shape, the medium and high geometric details, and the albedo can be jointly manipulated through not only high level demographic controls like age or gender, but also through low level skin properties objectively estimated from albedo textures.

2.3. Unconstrained Face Reconstruction

In face reconstruction, numerous works rely on generative models as priors to partially complete missing information and handle unseen parts, occlusions, and other ambiguities (e.g., material properties and lighting). These approaches vary in their focus, targeting various components of face representation, encompassing geometry, albedo, and physically-based materials, either individually or in combination.

A number of works leverage Generative Adversarial Networks (GANs) for this task. *GANfit* and *Fast-GANfit* [GPKZ19, GPKZ21] combine 3DMM fitting with GAN for albedo texture generation, helped by identity features extracted with *ArcFace* [DGXZ19], an off-the-shelf state-of-the-art face recognition network that has been widely employed by the community for various goals. These include estimating facial reflectance maps from single images in the wild [RDC*24], improving fairness in albedo estimation under arbitrary illumination when used within a GAN-based generative albedo model [RDM*23], or even generating new pictures of a given face under different poses and lighting [PPLM*24]. Also relying on GANs, *Avatarme* [LMG*20] presents a method for reconstructing PBR-ready faces from in-the-wild images using a combination of 3DMM fitting, GANs and super resolution. The extension *Avatarme++* [LMP*21] further enhances texture quality by synthesizing diffuse and specular colors through an image-to-image translation network trained on limited *Light Stage* [DGF15] data. Building upon *Avatarme++* data, *Fitme* [LMP*23] uses GAN inversion and proposes a BRDF generative network along with a two-stage fitting method to predict facial reflectance for unconstrained images. With a similar philosophy of utilizing 3DMM shape reconstruction and GANs for texture completion and super resolution, Dib et al. [DHG*24] incorporate differentiable shading for light normalization to reconstruct the face reflectance maps. Although all these models present incredible results, they often find trouble with baked illumination, and struggle to recover skin properties from certain ethnicities due to the unbalanced nature of their datasets, based on human-made subjective labeling.

Beyond GANs, a number of works rely on latent diffusion models. *Relightify* [PPLMZ23] reconstructs relightable 3D faces from a single image, focusing primarily on in-painting PBR material maps. It relies on off-the-shelve 3DMM-fitting techniques [BRZ*16, CKPZ18] to recover visible facial geometry and texture, completing the facial reflectance maps using a multi-modal diffusion model. Another approach, *FitDiff* [GLMZ23], shares similarities with our work, as it uses latent diffusion model to concurrently generate shape and textures from the scratch, departing from Gaussian noise. The geometry is defined by 3DMM coefficients and it produces albedo, specular, and normal maps. While the model uses *Arcface* identity conditioning to align with a reference image, it is also capable of generating random identities. However, its high and low level editing capabilities are limited.

Although these works achieve impressive results, they often focus on specific aspects of face reconstruction, such as geometry or texture, and frequently rely on third-party shape extraction methods, inheriting their limitations. Additionally, these models typically overlook the correlations between shape and appearance concerning high-level features like age or gender, and especially low-level properties like skin characteristics. This oversight limits their ability to generate faces that support bio-physical manipulations. For instance, the recent work by Li et al. [LGLG24] showcases plausible edits focused on aging, but it operates in 2D image space and only employs 3D uplifting for additional details, limiting its applicability as a generative model or as a prior for inverse rendering.

3. Biophysical face model

Our objective is to generate high-quality 3D faces by handling the face geometry and appearance jointly using a generative model conditioned on age, gender, and a set of skin biophysical properties defining the skin appearance, allowing a high level of control over the generated face. Note that these biophysical properties (e.g., melanin concentration) are tightly coupled with the identity of the generated face, including both shape and skin texture, which offer fine-grained control using objective labels rather than discrete, human-made subjective labels (e.g., ethnicity), enabling the modeling of correlations between shape and material properties when necessary, and establishing a well-bounded manifold for the generative model.

Modeling individual faces Our model builds upon a face representation consisting of a template average triangle mesh $\bar{G}$, and a set of textures $\mathbf{M}$ that encode the identity of the generated faces. For each face, a displacement texture $D \in \mathbb{R}^{3 \times h \times w}$ encodes the per-point deformation over $\bar{G}$ as a vector field of resolution $h \times w$, effectively capturing the macro-structural shape features of the face. The remaining maps in $\mathbf{M}$ are a height map $H \in \mathbb{R}^{1 \times h \times w}$, decomposed into displacement H_{disp} and bump H_{bump}, encoding medium and high frequency geometric details, and an albedo map $A \in \mathbb{R}^{3 \times h \times w}$. Thus, each face is represented by a set of three maps $\mathbf{M} = [D, H, A]$; other channels (e.g., specular or roughness map) could be added, though in this work we derive them from H (see details in Sec-

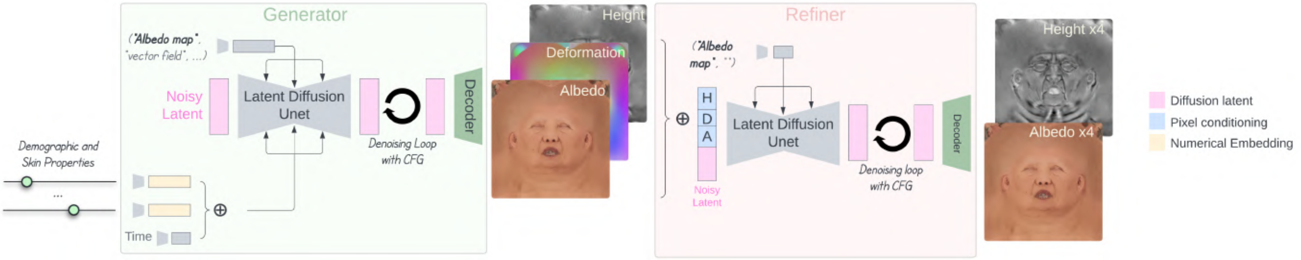

Figure 2: **Overview.** Our model generates a deformation texture D, a height map H, or an albedo map A at mid resolution depending on a text prompt $\mathbf{p}_G$ and controlled by various conditions. Then, a *Refiner* receives the concatenation of these three maps to upsample the map specified by another prompt $\mathbf{p}_R$ by ×4 resolution.

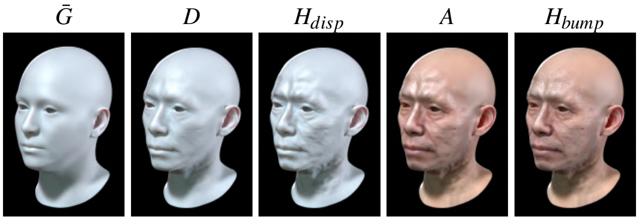

Figure 3: Progressive build-up of a face from the template mesh by adding each of the model's output maps.

tion 5). The effect of each of the maps on the final renderings is shown in Figure 3.

Parameter space of our model The parameter space Ω of our generative model is defined as

$$\Omega = \{\mathbf{z}, \mathbf{c}\}, \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^{4\times96\times96}$ is a latent noise vector, and $\mathbf{c} \in \mathbb{R}^N$ is the set of conditions. Identity emerges from $\mathbf{z}$ while biophysical attributes are controlled through $\mathbf{c}$, enabling independent manipulation of facial identity and characteristics. The conditions set is defined as $\mathbf{c} = \{a, g, m, h, o, r, e\}$, where a (age) is a demographic label and g (gender) is a continuous control variable derived from biological sex annotations in the dataset metadata. The remaining elements – m (melanin concentration), h (hemoglobin concentration), o (blood oxygenation level), r (ratio of eumelanin to pheomelanin), and e (epidermal thickness) – are aggregated pixel statistics of the biophysical low-level parameters defining the appearance of the skin and are estimated from the albedo map A [AXX*23]. Specifically, for the oxygenation level o, the melanin types ratio r, and the epidermal thickness t we use the mode of the reconstructed parameters, while for melanin m and hemoglobin h concentrations we found that using the mode and standard deviation provides a more descriptive representation, as there are faces with very even distributions and others with uneven distributions, making a total of $N = 9$ components.

Overview These parameters collectively define the input space for our method. Our model F maps the parameter space Ω to a set of high-resolution texture maps $\mathbf{M}$ as

$$F : \Omega \to \mathbf{M}. \tag{2}$$

Our model F works in two stages: First, a multi-modal generative model or *Generator* maps the parameter space Ω to the texture map set $\mathbf{M}_0$ at mid-resolution (768 × 768) (Section 3.1). Then, a super-resolution module or *Refiner* takes $\mathbf{M}_0$ and upsamples the generated maps H and A to achieve high-resolution details (Section 3.2), giving the final set of textures $\mathbf{M}$ ($2K \times 2K$). A high-level overview of the model is illustrated in Figure 2.

3.1. Generator

In this section we describe our $\Omega \to \mathbf{M}$ model or *Generator* network that produces a face in the form of a set of texture maps M conditioned by the aforementioned demographic and biophysical labels $\mathbf{c}$. Inspired by recent work in intrinsic image decomposition and synthesis [ZDG*24] we fine-tune a text-to-image latent diffusion model (LDM) [RBL*22] that operates on a downsampled latent representation, additionally conditioned on a text prompt $\mathbf{p}$ specifying the modality of the diffusion process (i.e., the map being generated, details in Section 3.1.2).

The latent representation $\mathbf{z} \in \mathbb{R}^{4\times96\times96}$ is obtained using a pre-trained variational autoencoder (VAE) consisting of an encoder-decoder pair ($\mathcal{E}$ and $\mathcal{D}$), such that for a map $M \in \mathbf{M}$, the latent vector is $\mathbf{z}_0 = \mathcal{E}(M)$, and the reconstructed map is $M' = \mathcal{D}(\mathbf{z}_0) = \mathcal{D}(\mathcal{E}(M))$.

During inference, the diffusion model iteratively denoises an initial Gaussian-noise tensor $\mathbf{z}_T \in \mathbb{R}^{4\times96\times96}$ through time steps $t \in \{0, ..., T\}$ to produce the target latent representation $\mathbf{z}_0$. Then the VAE's decoder $\mathcal{D}$ upsamples the denoised latent to a full resolution map $M = \mathcal{D}(\mathbf{z}_0)$.

For training our diffusion model, we optimize the parameters θ defining the velocity predictor $\hat{\mathbf{v}}_\theta$ by minimizing the following loss function

$$L_\theta = \|\mathbf{v}_t - \hat{\mathbf{v}}_\theta(t, \mathbf{z}_t, \mathbf{C}, \tau(\mathbf{p}))\|_2^2, \tag{3}$$

where $\tau(\mathbf{p})$ is the prompt encoded via CLIP [RKH*21], $\mathbf{C}$ is the embedded condition (see Section 3.1.1), and $\mathbf{z}_t$ is the noisy latent after adding Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at time step t. We use v-prediction L_θ during training, which has been shown [SH22] to produce better results compared to traditional noise (ϵ) prediction methods. The target velocity $\mathbf{v}_t$ at time-step t is calculated as

$$\mathbf{v}_t = \sqrt{\bar{\alpha}_t}\,\epsilon - \sqrt{1-\bar{\alpha}_t}\,\mathbf{z}_0, \tag{4}$$

where ϵ is the Gaussian noise, $\bar{\alpha} = 1 - t/T$ is a scale function of time, and $\mathbf{z}_0$ is the target latent encoded by the pre-trained encoder $\mathcal{E}$.

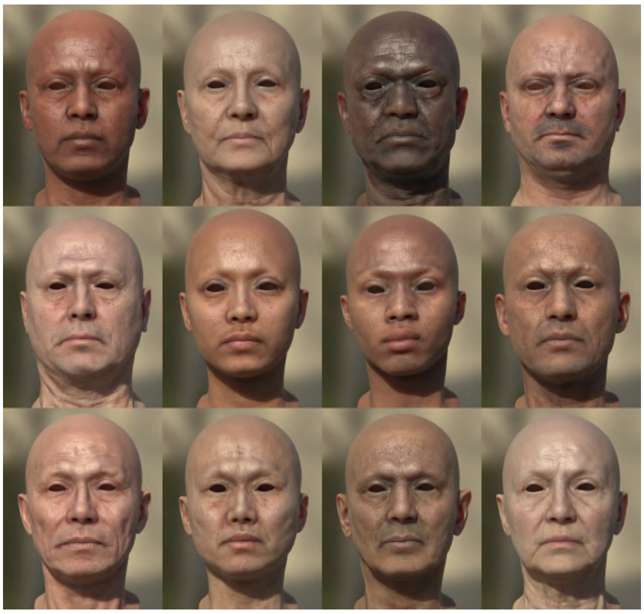

Figure 4: Faces generated by randomly sampling identity and conditions.

3.1.1. Cross-attention control with continuous-valued conditioning.

For conditioning on the demographic metadata and biophysical skin properties $\mathbf{c} \in \mathbb{R}^N$, we employ a multi-layer perceptron (MLP) of 2 hidden layers of size 1280 to encode each condition value c_i. Specifically, we encode the condition vector following

$$f_{\theta_i}(c_i) = \mathrm{MLP}(\psi(c_i,0),\dots,\psi(c_i,d-1)), \tag{5}$$

where f_{θ_i} represents the learned MLP embedding for the condition value c_i, corresponding to condition type i (e.g., melanin mode), and d is the dimension of the encoded condition vector ($d = 1280$, identical to the time t embedding). The input to the MLP is generated using a sinusoidal embedding function, similar to the timestep embedding used in diffusion models, following

$$\psi(c,j) = \begin{cases} \sin(cT^{\frac{-2j}{d}}), & \text{if } j \text{ is even} \\ \cos(cT^{\frac{-2j}{d}}) & \text{otherwise,} \end{cases} \tag{6}$$

where $T = 1000$ is a large constant representing the total number diffusion timesteps. Each condition value c is normalized to the range $[0,T]$. This approach is inspired by recent work on generative foundation models for satellite imagery [KLZ*23], which demonstrated the effectiveness of using sinusoidal embeddings for encoding numerical information, avoiding the shortcomings of text encoders when dealing with numerical information, as noted in previous work [RKH*21]. Finally, the N condition vectors are concatenated together with time embedding as $\mathbf{C} = [f_{\theta_1}(c_1) \oplus \dots \oplus f_{\theta_N}(c_N)] \oplus f_\theta(t)$. See Figure 2 for details.

3.1.2. Handling Different Modalities

Following a similar approach to *RGB2X* [ZDG*24], we utilize the text prompt $\mathbf{p}$ as a switch for modality control, producing a single map M at a time while storing only one network's weights. We use hand-picked prompts for each modality, specifically $\mathbf{p} = \{$"Vector field", "Albedo Map", "Bump and Displacement Map"$\}$, to provide optimal descriptions and enforce distinct text embeddings to generate a deformation map D, an albedo map A and a height map H, respectively. In addition, for trading off the quality and diversity of the samples generated by our model, we use classifier-free guidance (CFG) [HS22], which involves training the diffusion model for both conditional and unconditional denoising. Specifically, inspired by a recent work [BHE23], we use CFG for two conditionings: One for text prompt modality control $\mathbf{C}_p = \tau(\mathbf{p})$ and one for $\mathbf{c}$ conditioning $\mathbf{C_c}$. During training, we randomly set only $\mathbf{C}_c = \emptyset$ for 10% of examples, only $\mathbf{C}_p = \emptyset$ for 10% of examples, and both $\mathbf{C}_c = \emptyset$ and $\mathbf{C}_p = \emptyset$ for 10% of examples. We combine these score estimates during inference as follows:

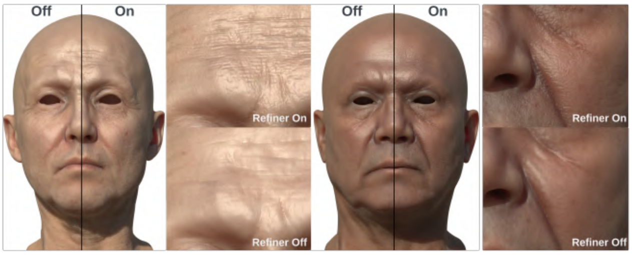

Figure 5: Effect of the *Refiner* network on two random identities.

$$\begin{aligned} \hat{\mathbf{v}}_\theta(t,\mathbf{z}_t,\mathbf{C}_c,\mathbf{C}_p) = {} & \mathbf{v}_\theta(t,\mathbf{z}_t,\emptyset,\emptyset) \\ & + s_p \cdot (\mathbf{v}_\theta(t,\mathbf{z}_t,\emptyset,\mathbf{C}_p) - \mathbf{v}_\theta(t,\mathbf{z}_t,\emptyset,\emptyset)) \\ & + s_c \cdot (\mathbf{v}_\theta(t,\mathbf{z}_t,\mathbf{C}_c,\mathbf{C}_p) - \mathbf{v}_\theta(t,\mathbf{z}_t,\emptyset,\mathbf{C}_p)), \end{aligned} \tag{7}$$

where the term $\mathbf{v}_\theta(\mathbf{z}_t,\emptyset,\emptyset)$ is the unconditional score estimate, where no conditioning is applied. The term $\mathbf{v}_\theta(\mathbf{z}_t,\mathbf{C}_c,\emptyset)$ is the score estimate conditioned on $\mathbf{C}_c$ only, and $\mathbf{v}_\theta(\mathbf{z}_t,\mathbf{C}_c,\mathbf{C}_p)$ is the score estimate conditioned on both $\mathbf{C}_c$ and $\mathbf{C}_p$. The guidance scales s_c and s_p control the influence of each conditioning. We empirically found $s_c = 5 \pm 3$ and $s_p = 5 \pm 3$ produce the best results. Tests with varying CFG values are shown in Figure 6. We demonstrates how our model generates a wide range of facial appearances while maintaining accurate correlations between shape features, skin properties, and demographic data in Figure 4.

3.2. Refiner

In this section we describe our super-resolution module or *Refiner* network, which upsamples a given height map H or an albedo A conditioned on a set of maps $\mathbf{M}$. It presents a similar architecture to the *Generator* network, also relying on text prompts $\mathbf{p}$ as switches. Unlike the *Generator* network, it takes as input every map $M \in \mathbf{M}$, downsampled in pixel space as $f_{\text{down}}(M) \in \mathbb{R}^{3\times192\times192}$, concatenated together with a latent noise vector $\mathbf{z}_T^R \in \mathbb{R}^{4\times192\times192}$ to form $\mathbf{m} = [f_{\text{down}}(D) \oplus f_{\text{down}}(H) \oplus f_{\text{down}}(A)] \oplus \mathbf{z}_T^R$. We employ latent diffusion for 4× upsampling as it provides more stable training dynamics and better detail preservation compared to standard U-Net architectures. After a number of denoising iterations, the network

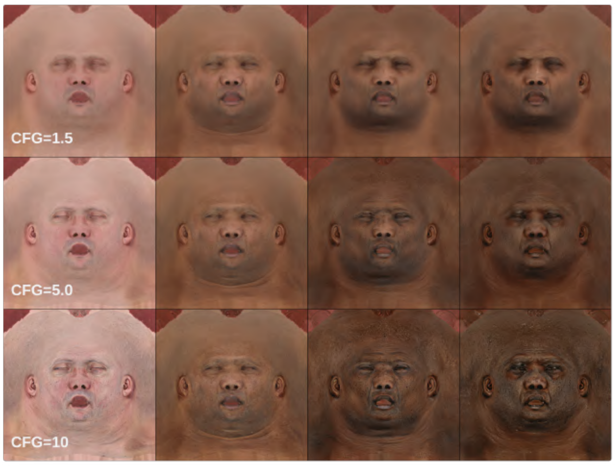

Figure 6: **Using different CFG values**. We show edits of the albedo of a given identity using different CFG values. Low values would result in subtle edits with rather minimal effect on the original albedo, failing to reach plausible human limits, while large values would cause undesirable artifacts.

produces the target latent image $\mathbf{z}_0^R = \mathcal{E}(M)$, encoding the map M by a factor of $768/192 = 4$. We minimize the following loss function:

$$L_\theta^R = \left\| \mathbf{v}_t^R - \hat{\mathbf{v}}_\theta^\mathbf{R}(t, \mathbf{m}, \mathbf{C}, \tau(\mathbf{p})) \right\|_2^2 \tag{8}$$

where $\mathbf{C}$ is the embedded condition (see Section 3.1.1). Depending on the prompt $\mathbf{p} \in$ {"Albedo Map", "Bump and Displacement Map"}, a latent $\mathbf{z}_0^R$ is generated, encoding either an albedo map A or height map H, respectively. After the final denoising step, the high-resolution map (×4) is reconstructed by the VAE decoder as $\mathcal{D}(\mathbf{z}_0^R)$. We trained this super resolution module by extracting 768×768 randomly-offset crops from the original 4K resolution images in the dataset. During inference, we generate a 2K map from the down-sampled version (512×512) of our initial 768×768 generation. The effect of adding the *Refiner* network is shown in Figure 5.

3.3. Dataset

Diffusion models [SDWMG15, RBL*22] are notorious for their high data requirements, and our high-quality 3D dataset is limited in size. To address this challenge, we train the latent diffusion model (Section 3.1) in two stages: Initially, we pre-train our *Generator* model using only uv-unwrapped face albedo maps synthesized by an existing neural model [BKZ*23] (54K faces); then, we fine-tune F using the whole set of maps $\mathbf{M}$ using a high-quality dataset of 3D assets.

Our dataset consists of 2001 high-resolution face meshes acquired in a Light Stage-like [DGF15] capture setup. Our dataset is reasonably well-balanced in certain properties, such as age and biological sex, with the latter being learned as our continuous *gender* control variable g in our model (as shown in Figure 7). However, it is important to note that our dataset excludes makeup and hair, which makes female versus male visual assessments challenging based solely on skin features. Additionally, the exclusion of minors from our capture causes the evaluation benchmark to be biased toward older faces. Despite these characteristics, the model learns the entire biophysical space and generates diverse faces as shown in Figure 4. However, we plan to extend the dataset to better represent the diversity of the population, since it remains unbalanced in other domains, like melanin concentration. Nevertheless, the model performs adequately across various biophysical properties within plausible ranges, as shown in Section 5. We made a conscious decision not to perform data augmentation through manipulations of skin properties estimated from the albedo maps to balance the dataset. This would have disrupted the realistic correlations between shape and skin material properties. By maintaining the natural variability in our dataset, we aim to model real-world face geometry and appearance more accurately.

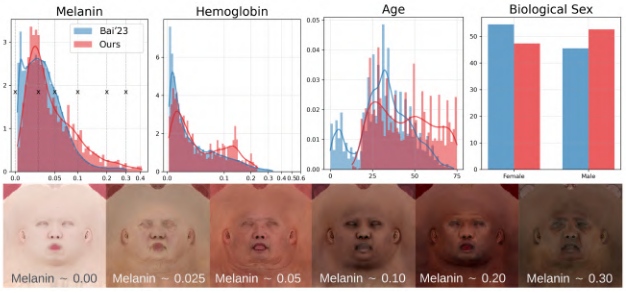

Figure 7: Distribution of labels in our dataset and Bai et al.'s model [BKZ*23], with example albedo maps showing varying melanin concentrations.

Each mesh has 379,289 vertices and is accompanied by high-quality normal and albedo maps, unwrapped in UV space at a 4K resolution. All assets share a common geometric topology and UV parametrization. We extract the base geometry $\bar{G}$ by computing the average face from all faces in the dataset, leveraging the 1-to-1 correspondence in mesh vertices. We denote the set of meshes as $\{G_i\}_{i=1}^N$, where each mesh G_i has a set of vertices $\{v_j^i\}_{j=1}^V$. The average vertex $\bar{v}_j$ is computed as $\bar{v}_j = \frac{1}{N}\sum_{i=1}^N v_j^i$. For each mesh G_i we compute the per-vertex displacement as an unormalized vector $d_j^i = v_j^i - \bar{v}_j$, which is projected in uv space, encoding D as a vector displacement field. For further details on how the remaining physically-based rendering (PBR) maps are generated and used to render the faces photorealistically, please refer to Section 5.

The nature of our deformation D and bump H_{bump} maps poses a challenge for directly using the variational autoencoder (VAE) trained on natural images, as these maps exhibit zero-mean distributions with high kurtosis and heavy tails. This results in the reconstructed map $M' = \mathcal{D}(\mathcal{E}(M))$ differing from the original map M. To address this, we enhance the contrast of the deformation and bump maps to broaden their pixel distributions, making them more similar to those of natural images. Specifically, we apply a regular normalization between 0 and 1 for the bump map, and a sigmoid-like normalization for the deformation map. For the deformation map

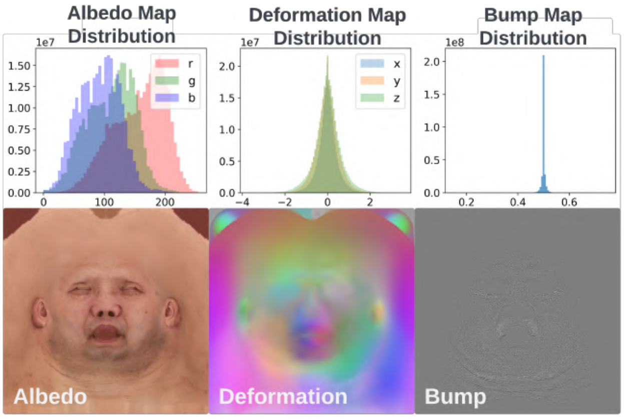

Figure 8: Normalized pixel distributions (D) and generated maps (H_{bump} scaled ×5 for visibility) for a representative identity.

$D \in \mathbb{R}^3$ (normalized into the range $[-1,1]$), we apply the transformation $D' = \frac{1}{1+e^{-s \cdot D}}$, where s is the sharpness parameter that controls the degree of emphasis around zero. We empirically found that a value of $s = 10$ performs best. Once generated, we invert this transformation using $D = -\frac{1}{s} \cdot \ln\left(\frac{1}{D'} - 1\right)$. This step is critical, as subtle artifacts in pixel values, especially in the deformation map D, can result in significant deformations of the mesh. Figure 8 shows the pixel distributions of the original maps and the actual maps.

The biophysical properties we use as control parameters are estimated using Aliaga et al. [AXX*23], which has been thoroughly validated against the Leeds spectral dataset [XYZ*17] containing measurements from diverse subjects. This validation ensures our model operates within physically plausible parameter ranges.

4. Face inversion and editing

In this section, we demonstrate two applications that showcase our model's capabilities beyond generation. First, we describe our approach to inverting and editing 3D captured faces. Then, we show how our model can be combined with differentiable rendering techniques to estimate face maps from reference images. These applications illustrate how our model serves as a powerful prior for inverse problems while maintaining physical plausibility through its biophysically-grounded parameter space.

4.1. Inverting and editing face maps

In this section, we describe the methodology for using our model to edit captured 3D faces. We begin with any given map, such as an albedo map A, and invert to its corresponding noisy latent $\mathbf{z}_T$. Then, we can run the model to regenerate this map $A' = \mathcal{D}(\mathbf{z}_0)$ or apply edits through the denoising process to create plausible edited versions of $A'' = \mathcal{D}(\mathbf{z}_0^{\text{edit}})$. It is important to note that once $\mathbf{z}_0$ or $\mathbf{z}_0^{\text{edit}}$ are reconstructed, they can also be used to produce plausible deformation D and height H maps from a single input albedo A.

To achieve this, it is necessary to invert $\mathbf{z}_0$ to its corresponding latent noise $\mathbf{z}_T$ and back to $\mathbf{z}_0$ or $\mathbf{z}_0^{\text{edit}}$. In deterministic diffusion implicit models (DDIM) [SME20], $\mathbf{z}_0 \rightarrow \mathbf{z}_T$ is typically accomplished by performing *DDIM inversion*, which assumes that the ordinary differential equation (ODE) process can be reversed in the limit of infinitesimally small steps t. However, this assumption cannot be guaranteed [JZB*23], resulting in a perturbation $\mathbf{z}_t \rightarrow \mathbf{z}_t^\star$. An additional perturbation $\mathbf{z}_t^\star \rightarrow \mathbf{z}_t'$ occurs when using *DDIM sampling* to generate a latent map $\mathbf{z}_t'$ from a random noise vector $\mathbf{z}_T^\star$. In addition, as we condition using CFG (Section 3.1.2), a further perturbation $\mathbf{z}_t' \rightarrow \mathbf{z}_t''$ arises. The accuracy of this inversion process significantly impacts the final editing outcome, influencing both the preservation of essential content from the source map, in our case the identity, and the fidelity of the edits. Consequently, following *Direct Inversion* [JZB*23], we disentangle the source and target editing branches and enable the source branch to rectify the deviation path directly. Specifically, *Direct Inversion* first computes the distance $o^{t-1} = \mathbf{z}_{t-1}^\star - \mathbf{z}_{t-1}''$, then adds the difference back to $\mathbf{z}_{t-1}''$ in the DDIM forward process. As for $\mathbf{C}$, we assume age and gender are given, and skin properties estimated using the method of Aliaga et al. [AXX*23]. The process is detailed in Algorithm 1.

Algorithm 1 Map Inversion and Editing

```
Input: A source $\mathbf{C}$, a target $\mathbf{C}_{\text{edit}}$, a source latent $\mathbf{z}_0^{\text{source}}$
Output: An edited latent $\mathbf{z}_0^{\text{edit}}$
Part I: Inverse $\mathbf{z}_0^{\text{source}}$
$\mathbf{z}_0^\star = \mathbf{z}_0^{\text{source}}$
for $t = 1$ to $T$ do
    $\mathbf{z}_t^\star \leftarrow$ DDIM Inversion$(\mathbf{z}_{t-1}^\star, t-1, C)$
end for
Part II: Perform editing on $\mathbf{z}_T^\star$ with Direct Inversion
$\mathbf{z}_T^{\text{edit}} = \mathbf{z}_T^\star$; $\mathbf{z}_T'' = \mathbf{z}_T^\star$
for $t = T$ to $1$ do
    $o^{t-1} \leftarrow \mathbf{z}_{t-1}^\star -$ DDIM Fwd$(\mathbf{z}_t'', t, \mathbf{C})$
    $\mathbf{z}_{t-1}'' =$ DDIM Forward$(\mathbf{z}_t'', t, \mathbf{C}) + o^{t-1}$
    $\mathbf{z}_{t-1}^{\text{edit}} \leftarrow$ DDIM Forward$(\mathbf{z}_t^{\text{edit}}, t, \mathbf{C}_{\text{edit}}) + o^{t-1}$
end for
return $\mathbf{z}_0^{\text{edit}}$                ▷ Return the final edited latent
```

Algorithm 2 Differentiable Rendering Guidance

```
function DRGuidance$(\mathbf{z}_t, \mathbf{C}_\mathbf{c}, \mathbf{C}_\mathbf{p}, I_{\text{ref}})$
    $\mathbf{z}_0 \leftarrow \hat{\mathbf{v}} \leftarrow \hat{\mathbf{v}}_\theta(\mathbf{z}_t, \mathbf{C}_c, \mathbf{C}_p)$        ▷ Directly estimate $\mathbf{z}_0$
                                  ▷ Stage 1: Optimize Diff. Rendering parameters
    for $i = 1$ to $N_{DR}$ do
        $I_{\text{current}} \leftarrow$ DifferentiableRenderer$(\mathbf{P})$
        $\mathcal{L}_\mathbf{P} \leftarrow \|I_{\text{ref}} - I_{\text{current}}\|_2$
        Update $\mathbf{P}$ via backpropagation with $\mathcal{L}_\mathbf{P}$
    end for
    $M_{\text{target}} \leftarrow M \leftarrow \mathbf{P}$
                                  ▷ Stage 2: Optimize velocity prediction
    for $i = 1$ to $N_\mathbf{v}$ do
        $M_{\text{current}} \leftarrow \mathcal{D}(\mathbf{z}_0)$; $\mathbf{z}_0 \leftarrow \hat{\mathbf{v}}$
        $\mathcal{L}_M \leftarrow \|M_{\text{target}} - M_{\text{current}}\|_2$
        Update $\hat{\mathbf{v}}$ via backpropagation with $\mathcal{L}_M$
    end for
     return $\hat{\mathbf{v}}$
end function
```

4.2. Face Reconstruction with Differentiable Rendering Guidance

In this section, we demonstrate how our model can be integrated with differentiable rendering (DR) to reconstruct a face from either a single view I_{ref} or multiple views $\mathbf{I}_{\text{ref}} = \{I^1_{\text{ref}}, I^2_{\text{ref}}, \dots, I^n_{\text{ref}}\}$ of a subject. Our model acts as a prior, effectively constraining the differentiable rendering optimization. The output is a set of reconstructed face maps, M, along with other unknown parameters Φ, such as environmental lighting.

The key to this approach is to incorporate an additional guidance term into the denoising process, utilizing a score computed with the assistance of the differentiable renderer. We achieve this by expanding CFG in Equation 7 to:

$$\begin{aligned}\hat{\mathbf{v}}_\theta(t, \mathbf{z}_t, \mathbf{C}_c, \mathbf{C}_p) &= \mathbf{v}_\theta(t, \mathbf{z}_t, \emptyset, \emptyset) \\ &+ s_p \cdot (\mathbf{v}_\theta(t, \mathbf{z}_t, \emptyset, \mathbf{C}_p) - \mathbf{v}_\theta(t, \mathbf{z}_t, \emptyset, \emptyset)) \\ &+ s_{\text{DR}} \cdot (\text{DRGuidance}(\mathbf{z}_t, \mathbf{C}_c, \mathbf{C}_p, I_{\text{ref}}) - \mathbf{v}_\theta(t, \mathbf{z}_t, \emptyset, \mathbf{C}_p)) \\ &+ s_c \cdot (\mathbf{v}_\theta(t, \mathbf{z}_t, \mathbf{C}_c, \mathbf{C}_p) - \text{DRGuidance}(\mathbf{z}_t, \mathbf{C}_c, \mathbf{C}_p, I_{\text{ref}})).\end{aligned} \tag{9}$$

Note that the biophysical and demographic guidance $\mathbf{C}_c$ is applied last in the CFG after *DRGuidance*, as we empirically found that it imposes stronger constraints for reconstructing plausible faces, whereas having *DRGuidance* last can overly bias the results towards faces that may fall outside our plausible biophysical space, creating artifacts specially when ambiguity is high (e.g. single view scenarios).

The velocity prediction from *DRGuidance* is obtained as follows (Algorithm 2): Through the denoising process, we estimate the final latent $\mathbf{z}_0$ directly from an intermediate time step t, similar to Galanakis et al. [GLMZ23], to obtain a map M. This map is combined with other rendering parameters Φ (e.g., lighting) to form the complete set $\mathbf{P} = (M, \Phi)$, which is used to render an image of the face I. We use the differentiable renderer to optimize the parameters $\mathbf{P}$ for N_{DR} iterations, minimizing $\mathcal{L}_\mathbf{P} \leftarrow \|I^i_{\text{ref}} - I\|_2$. The optimized map M in $\mathbf{P}$ is then used as a target M_{target} to optimize, for $N_\mathbf{v}$=10 iterations (using Adam optimizer), the velocity prediction $\mathbf{v}_t$ ultimately returned for CFG guidance.

The denoising process is executed over 50 iterations in total. The c_{DR} guidance within the CFG is initiated at iteration 5, once the modality c_p has stabilized. However, in practice, it can be applied as late as iteration 35. In our experiments, the guidance scales $s_c = 2.0$, $s_{\text{DR}} = 4.0$, and $s_p = 1.1$ produced the most optimal results. Finally, after completing the denoising process, we employ the differentiable renderer to further optimize the rendering parameters $\mathbf{P}$, refining the result to better match the reference image.

5. Experiments

We use Mitsuba 3 [JSR*22] to render the faces, using a custom skin model featuring a double GGX [WMLT07] specular lobe, and a subsurface diffuse component computed using Monte Carlo random walks on a homogeneous medium. The optical parameters of the medium are obtained using numerical albedo inversion [WVH17] from the albedo map A, as well as its mean free path and its anisotropy coefficient. The parameters associated with the specular lobes were manually adjusted through look development, and we empirically established two heuristics for their creation, derived directly from the height map H ($H_{disp} + H_{bump}$). Specifically, the specular roughness is set to $S_r \approx H$, and the specular intensity is defined as $S_i \approx 3H$.

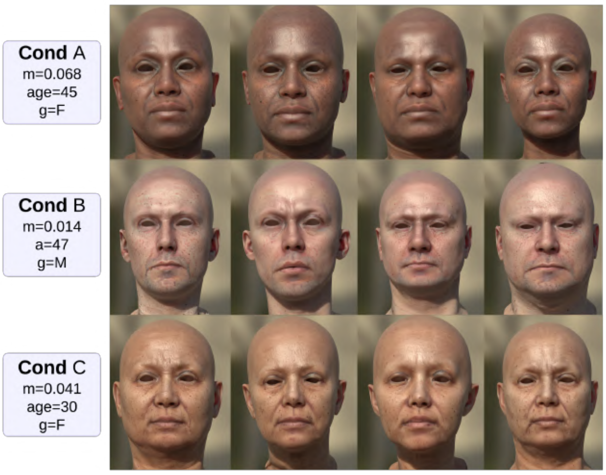

Figure 9: Random identities generated with fixed combinations of demographic and biophysical conditions. The model preserves not only the overall skin tone but also the distribution of facial heterogeneities across distinctly different identities.

Generating bio-physically plausible faces We present several results showing the biophysical space of faces and edits. Figure 4 shows different generated faces by randomly sampling both identities and conditions, showcasing how the model covers a wide range of appearances. Then we explore the space of identities and conditions (Figure 1), isolating identities (Figure 9) and conditions (Figure 15), which desmonstrates the model's capability of fine grained control over the skin tone and the distribution of facial heterogeneities. Figure 14 shows the interpolation between two different identities, shows a smooth transition between the two identities, and how the model is able to interpolate between subjects.

Comparisons Figure 10 shows comparisons with simpler heuristics over the skin property maps. Our model produces richer heterogeneities and variations, capturing not only geometric changes but also medium and high frequency details and nuances. In Figure 11, we compare our aging control with Iglesias-Guitian et al.'s model [IGAJG15]. Our model captures subtle changes in skin color as predicted by Iglesias-Guitian et al., but also includes rich heterogeneities, pronounced wrinkles, and larger geometric changes that are not present in their model.

Applications Figure 15 shows the results for 3D face asset inversion (Section 4.1), where from a random initialization $\mathbf{M}_0$ we are able to reconstruct $\mathbf{M}$ to match a given asset (second column), and to alter the condition vector $\mathbf{C}$ to provide high-level editing of the face on several different axes (columns 3 to 8). Additionally, we provide extended results showcasing a wider variety of identities and edits in the Supplemental Material. The results of our reconstruction pipeline (Section 4.2) are shown in Figures 12 and 13, for

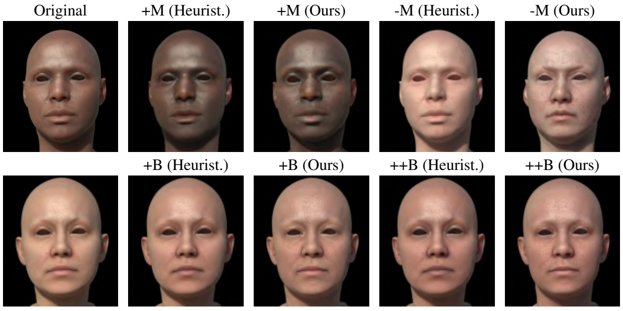

Figure 10: Comparison with a heuristic-based method, where we manipulate the properties using Aliaga et al. [AXX*23] to empirically reproduce the resulting albedo map from our model. Specifically, we increase/decrease melanin by 300%/90% and blood concentration by 150%/300% to simulate the effects of +Melanin, -Melanin, +Blood, and ++Blood, respectively.

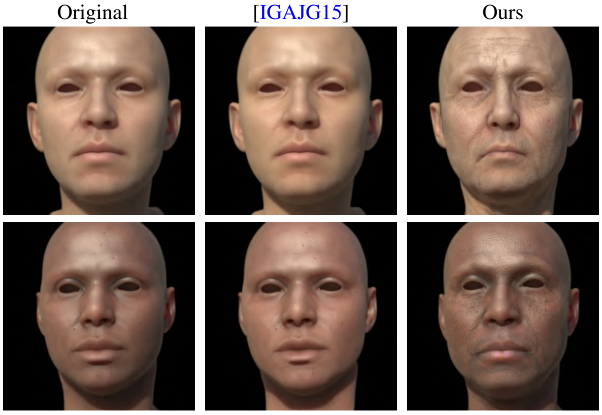

Figure 11: Comparison of our aging control with Iglesias-Guitian et al.'s aging model [IGAJG15] on two identities. We reconstruct the original albedo using Aliaga et al. [AXX*23] and simulate aging by reducing melanin concentration, blood concentration, and epidermal thickness by approximately 20%, following Iglesias et al.'s aging model (6% per decade over 3 decades).

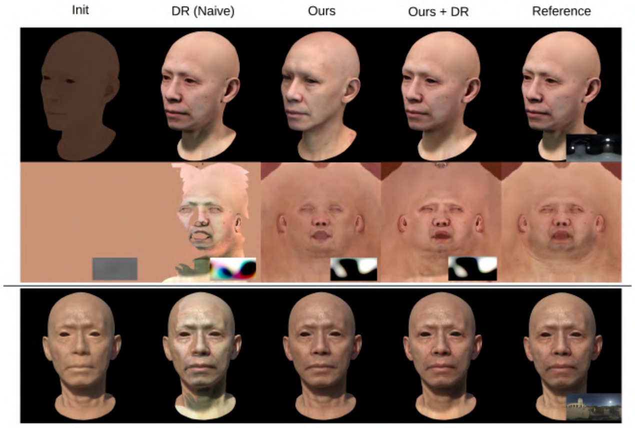

Figure 12: Reconstruction of the albedo map and environment map from a single view. Left to right: initial state; differentiable rendering only (naive); ours method; our method with differentiable rendering; and the reference image, with the environmental map as an inset. From top to bottom: renderings using the optimized albedos and environment maps for the same reference view; optimized albedo textures, with the estimated environmental maps shown as insets; and renderings using the optimized albedos, with changes in both lighting and viewpoint. Note that the reconstructed environmental maps are inherently low-frequency due to the face's reflectance acting as a low-pass filter.

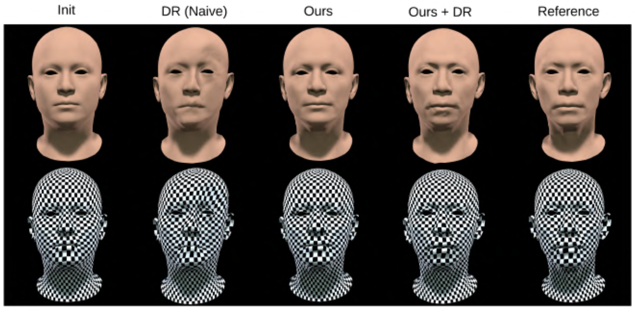

Figure 13: Reconstruction of the deformation map from a single view. From left to right: initial state; differentiable rendering only (naive); ours method; our method with differentiable rendering; and the reference face (note that we use the same reference view during the optimization as in Figure 12). We show renderings with the reconstructed deformation map and a constant albedo (first row) and a checkerboard texture to better appreciate the changes in geometry (second row).

albedo and shape inversion respectively; compared to using priorless differentiable rendering, we are able to reconstruct a closer match, which is more robust to the illumination by staying on a plausible manifold defined by the latent space, and even inpainting occluded regions in the reference. The performance of the different methods in both applications is summarized in Table 1.

6. Conclusion

We introduced a novel generative model that advances the synthesis of photorealistic faces by capturing the intricate relationships between facial geometry and biophysical skin properties. Our approach enables continuous modeling of skin properties by objectively extracting them from the albedo, eliminating the reliance on subjective human annotations. This ensures that the generated faces and the editing space are biophysically plausible. Beyond generation, the model demonstrates its versatility in applications such as editing 3D face assets and serving as a prior in inverse rendering. The model produces faces that are ready for integration into computer graphics engines, offering relightability and riggability.

However, there are several limitations that can spark future lines of work. The current model is limited to static faces and does not account for facial expressions or dynamics. The model also focuses on stationary properties like skin albedo and bump maps, without

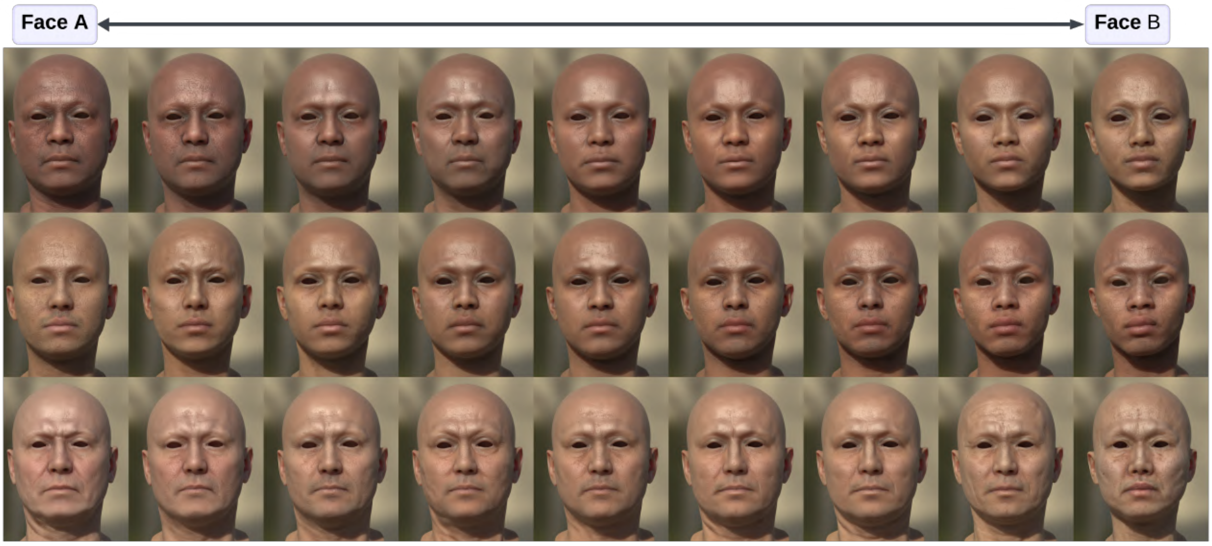

Figure 14: For the three identities in the first column, we display progressive interpolation steps in the latent space toward the three identities in the last column, demonstrating the smoothness of our biophysical space.

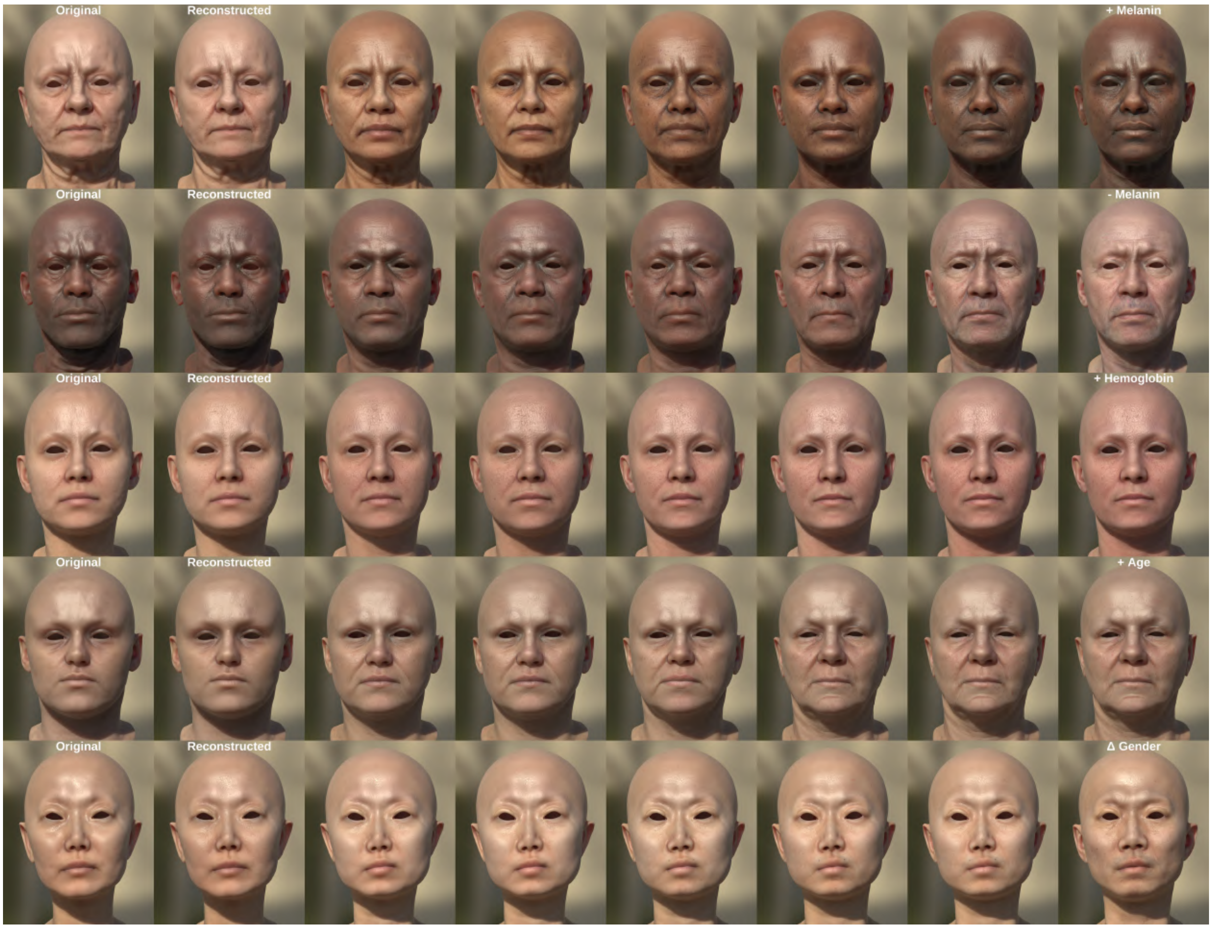

Figure 15: **Face Asset Inversion and Editing**. Each row demonstrates the reconstruction of an identity through DDIM, followed by progressive editing based on one of the conditions from left to right. The results illustrate the model's ability to effectively manipulate facial features while maintaining the identity.

Method	Metrics (MAE / MSE / PSNR)	
	Texture (Fig. 12)	Shape (Fig. 13)
init	0.0618 / 0.0065 / 21.84	0.0595 / 0.0153 / 18.16
DR	0.0920 / 0.0119 / 19.25	0.0624 / 0.0114 / 19.45
Ours	0.0426 / 0.0030 / 25.26	0.0604 / 0.0137 / 18.63
Ours DR	**0.0377 / 0.0024 / 26.28**	**0.0358 / 0.0048 / 23.19**

Table 1: Performance comparison using MAE, MSE, and PSNR metrics of different methods inverting: a) albedo and lighting jointly and b) shape . Our method outperforms naive differentiable rendering optimization, further improving when combining both through guidance.

incorporating specular maps due to their transient nature. Furthermore, at rendering time, empirical heuristics derived from production workflows are used to render the specular reflectance, not physically grounded in Fresnel theory. While effective for control, this approach could be replaced by more accurate IOR-based Fresnel reflectance models. Additionally, it lacks sub-pore level details and other facial components such as eyes and facial hair. Regarding the dataset used for finetuning, we want to improve its balance to ensure fairness across demographic groups; it has limited representation of South Asian individuals, which may impact the generalizability of our findings and highlights the need for more diverse data collection in future studies. Last, by using continuous biophysical parameters instead of explicit racial categories, we may not fully capture the nuances of certain facial features. An interesting direction for future research is to investigate how skin properties vary across different ethnicities in order to add it as a control variable.

References

[AXX*23] Aliaga C., Xia M., Xie X., Jarabo A., Braun G., Hery C.: A hyperspectral space of skin tones for inverse rendering of biophysical skin properties. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, p. e14887.

[BHE23] Brooks T., Holynski A., Efros A. A.: Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18392–18402.

[BK10] Baranoski G. V., Krishnaswamy A.: *Light and skin interactions: simulations for computer graphics applications*. Morgan Kaufmann, 2010.

[BKZ*23] Bai H., Kang D., Zhang H., Pan J., Bao L.: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 362–371.

[BRP*18] Booth J., Roussos A., Ponniah A., Dunaway D., Zafeiriou S.: Large scale 3d morphable models. *Int. J. Comput. Vision 126*, 2–4 (Apr. 2018), 233–254. URL: https://doi.org/10.1007/s11263-017-1009-7, doi:10.1007/s11263-017-1009-7.

[BRZ*16] Booth J., Roussos A., Zafeiriou S., Ponniah A., Dunaway D.: A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 5543–5552.

[BV99] Blanz V., Vetter T.: A morphable model for the synthesis of 3d faces. In *Proceedings of SIGGRAPH* (1999), p. 187–194.

[CBGB20] Chandran P., Bradley D., Gross M., Beeler T.: Semantic deep face models. In *2020 international conference on 3D vision (3DV)* (2020), IEEE, pp. 345–354.

[CBKM15] Chen T. F., Baranoski G. V., Kimmel B. W., Miranda E.: Hyperspectral modeling of skin appearance. *ACM Transactions on Graphics (TOG) 34*, 3 (2015), 1–14.

[CKPZ18] Cheng S., Kotsia I., Pantic M., Zafeiriou S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 5117–5126.

[CWZ*14] Cao C., Weng Y., Zhou S., Tong Y., Zhou K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (Mar. 2014), 413–425. URL: https://doi.org/10.1109/TVCG.2013.249, doi:10.1109/TVCG.2013.249.

[DGF15] Debevec P. E., Ghosh A., Fyffe G.: Multiview face capture using polarized spherical gradient illumination, Sept. 1 2015. US Patent 9,123,116.

[DGXZ19] Deng J., Guo J., Xue N., Zafeiriou S.: Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4690–4699.

[DHG*24] Dib A., Hafemann L. G., Got E., Anderson T., Fadaeinejad A., Cruz R. M., Carbonneau M.-A.: Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1770–1780.

[DJ06] Donner C., Jensen H. W.: A spectral bssrdf for shading human skin. *Rendering techniques 2006* (2006), 409–418.

[DTP23] Deb D., Tripathi S., Puri P.: Munch: Modelling unique'n controllable heads. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games* (2023), pp. 1–11.

[DWd*08] Donner C., Weyrich T., d'Eon E., Ramamoorthi R., Rusinkiewicz S.: A layered, heterogeneous reflectance model for acquiring and rendering human skin. *ACM transactions on graphics (TOG) 27*, 5 (2008), 1–12.

[EST*20] Egger B., Smith W. A., Tewari A., Wuhrer S., Zollhoefer M., Beeler T., Bernard F., Bolkart T., Kortylewski A., Romdhani S., et al.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG) 39*, 5 (2020), 1–38.

[GCM*24] Gruber A., Collins E., Meka A., Mueller F., Sarkar K., Orts-Escolano S., Prasso L., Busch J., Gross M., Beeler T.: Gantlitz: Ultra high resolution generative model for multi-modal face textures. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e15039.

[GKG*23] Giebenhain S., Kirschstein T., Georgopoulos M., Rünz M., Agapito L., Niessner M.: Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 21003–21012.

[GKG*24] Giebenhain S., Kirschstein T., Georgopoulos M., Rünz M., Agapito L., Niessner M.: Mononphm: Dynamic head reconstruction from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 10747–10758.

[GKR*24] Giebenhain S., Kirschstein T., Rünz M., Agapito L., Niessner M.: Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia 2024 Conference Papers* (2024), pp. 1–11.

[GLMZ23] Galanakis S., Lattas A., Moschoglou S., Zafeiriou S.: Fitdiff: Robust monocular 3d facial shape and reflectance estimation using diffusion models. *arXiv preprint arXiv:2312.04465* (2023).

[GLP*20] Gecer B., Lattas A., Ploumpis S., Deng J., Papaioannou A., Moschoglou S., Zafeiriou S.: Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)* (2020). doi:10.1007/978-3-030-58526-6_25.

[GPKZ19] Gecer B., Ploumpis S., Kotsia I., Zafeiriou S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[GPKZ21] Gecer B., Ploumpis S., Kotsia I., Zafeiriou S. P.: Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[HS22] Ho J., Salimans T.: Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[IGAJG15] Iglesias-Guitian J. A., Aliaga C., Jarabo A., Gutierrez D.: A biophysically-based model of the optical properties of skin aging. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 45–55.

[JSB*10] Jimenez J., Scully T., Barbosa N., Donner C., Alvarez X., Vieira T., Matts P., Orvalho V., Gutierrez D., Weyrich T.: A practical appearance model for dynamic facial color. In *ACM SIGGRAPH Asia 2010 papers*. 2010, pp. 1–10.

[JSR*22] Jakob W., Speierer S., Roussel N., Nimier-David M., Vicini D., Zeltner T., Nicolet B., Crespo M., Leroy V., Zhang Z.: Mitsuba 3 renderer. *URL: https://mitsuba-renderer. org 7* (2022).

[JZB*23] Ju X., Zeng A., Bian Y., Liu S., Xu Q.: Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506* (2023).

[KB04] Krishnaswamy A., Baranoski G. V.: A biophysically-based spectral model of light interaction with human skin. In *Computer graphics forum* (2004), vol. 23, Wiley Online Library, pp. 331–340.

[KGN24] Kirschstein T., Giebenhain S., Niessner M.: Diffusion-avatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 5481–5492.

[KLA*20] Karras T., Laine S., Aittala M., Hellsten J., Lehtinen J., Aila T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 8110–8119.

[KLZ*23] Khanna S., Liu P., Zhou L., Meng C., Rombach R., Burke M., Lobell D., Ermon S.: Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606* (2023).

[KQG*23] Kirschstein T., Qian S., Giebenhain S., Walter T., Niessner M.: Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–14.

[LBB*17] Li T., Bolkart T., Black M. J., Li H., Romero J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (2017), 194:1–194:17. URL: https://doi.org/10.1145/3130800.3130813.

[LBZ*20] Li R., Bladin K., Zhao Y., Chinara C., Ingraham O., Xiang P., Ren X., Prasad P., Kishore B., Xing J., et al.: Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 3410–3419.

[LDML*24] Li X., De Mello S., Liu S., Nagano K., Iqbal U., Kautz J.: Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems 36* (2024).

[LGLG24] Li X., Guarnera G. C., Lin A., Ghosh A.: Realistic facial age transformation with 3d uplifting. In *Computer graphics forum* (2024), vol. 43, Wiley Online Library, p. e15146.

[LKB*24] Lee J., Kang T., Bühler M. C., Kim M.-J., Hwang S., Hyung J., Jang H., Choo J.: Surfhead: Affine rig blending for geometrically accurate 2d gaussian surfel head avatars. *arXiv preprint arXiv:2410.11682* (2024).

[LMG*20] Lattas A., Moschoglou S., Gecer B., Ploumpis S., Triantafyllou V., Ghosh A., Zafeiriou S.: Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[LMP*21] Lattas A., Moschoglou S., Ploumpis S., Gecer B., Ghosh A., Zafeiriou S. P.: Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[LMP*23] Lattas A., Moschoglou S., Ploumpis S., Gecer B., Deng J., Zafeiriou S.: Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 8629–8640.

[LZK*24] Li X., Zhang Q., Kang D., Cheng W., Gao Y., Zhang J., Liang Z., Liao J., Cao Y.-P., Shan Y.: Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807* (2024).

[MMK12] Mori M., MacDorman K. F., Kageki N.: The uncanny valley [from the field]. *IEEE Robotics & automation magazine 19*, 2 (2012), 98–100.

[PKA*09] Paysan P., Knothe R., Amberg B., Romdhani S., Vetter T.: A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (2009), pp. 296–301. doi:10.1109/AVSS.2009.58.

[PPLM*24] Paraperas Papantoniou F., Lattas A., Moschoglou S., Deng J., Kainz B., Zafeiriou S.: Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2024).

[PPLMZ23] Paraperas Papantoniou F., Lattas A., Moschoglou S., Zafeiriou S.: Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).

[PYG*24] Po R., Yifan W., Golyanik V., Aberman K., Barron J. T., Bermano A., Chan E., Dekel T., Holynski A., Kanazawa A., et al.: State of the art on diffusion models for visual computing. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e15063.

[RBL*22] Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695.

[RDC*24] Ren X., Deng J., Cheng Y., Guo J., Ma C., Yan Y., Zhu W., Yang X.: Monocular identity-conditioned facial reflectance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 885–895.

[RDM*23] Ren X., Deng J., Ma C., Yan Y., Yang X.: Improving fairness in facial albedo estimation via visual-textual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4511–4520.

[RGP*24] Rai A., Gupta H., Pandey A., Carrasco F. V., Takagi S. J., Aubel A., Kim D., Prakash A., De la Torre F.: Towards realistic generative 3d face models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), pp. 3738–3748.

[RKH*21] Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., et al.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763.

[SDWMG15] Sohl-Dickstein J., Weiss E., Maheswaranathan N., Ganguli S.: Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (2015), PMLR, pp. 2256–2265.

[SH22] SALIMANS T., HO J.: Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512* (2022).

[SME20] SONG J., MENG C., ERMON S.: Denoising diffusion implicit models. *arXiv:2010.02502* (October 2020). URL: https://arxiv.org/abs/2010.02502.

[SSS*24] SAITO S., SCHWARTZ G., SIMON T., LI J., NAM G.: Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 130–141.

[THM99] TSUMURA N., HANEISHI H., MIYAKE Y.: Independent-component analysis of skin color image. *Journal of the Optical Society of America A 16*, 9 (1999), 2169–2176.

[TLL23] TOSHPULATOV M., LEE W., LEE S.: Talking human face generation: A survey. *Expert Systems with Applications 219* (2023), 119678. URL: https://www.sciencedirect.com/science/article/pii/S0957417423001793, doi:https://doi.org/10.1016/j.eswa.2023.119678.

[TOS*03] TSUMURA N., OJIMA N., SATO K., SHIRAISHI M., SHIMIZU H., NABESHIMA H., AKAZAKI S., HORI K., MIYAKE Y.: Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin. In *ACM SIGGRAPH 2003 Papers*. 2003, pp. 770–779.

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. *ACM Trans. Graph. 24*, 3 (July 2005), 426–433. URL: https://doi.org/10.1145/1073204.1073209, doi:10.1145/1073204.1073209.

[WMLT07] WALTER B., MARSCHNER S. R., LI H., TORRANCE K. E.: Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (Goslar, DEU, 2007), EGSR'07, Eurographics Association, p. 195–206.

[WVH17] WRENNINGE M., VILLEMIN R., HERY C.: *Path traced subsurface scattering using anisotropic phase functions and non-exponential free flights*. Tech. rep., Technical Memo, 2017.

[XYZ*17] XIAO K., YATES J. M., ZARDAWI F., SUEEPRASAN S., LIAO N., GILL L., LI C., WUERGER S.: Characterising the variations in ethnic skin colours: a new calibrated data base for human skin. *Skin Research and Technology 23*, 1 (2017), 21–29.

[ZAJ*15] ZELL E., ALIAGA C., JARABO A., ZIBREK K., GUTIERREZ D., MCDONNELL R., BOTSCH M.: To stylize or not to stylize? the effect of shape and material stylization on the perception of computer-generated faces. *ACM Trans. Graph. 34*, 6 (Nov. 2015). URL: https://doi.org/10.1145/2816795.2818126, doi:10.1145/2816795.2818126.

[ZDG*24] ZENG Z., DESCHAINTRE V., GEORGIEV I., HOLD-GEOFFROY Y., HU Y., LUAN F., YAN L.-Q., HAŠAN M.: Rgb2x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (New York, NY, USA, 2024), SIGGRAPH '24, Association for Computing Machinery. URL: https://doi.org/10.1145/3641519.3657445, doi:10.1145/3641519.3657445.

[ZQL*23] ZHANG L., QIU Q., LIN H., ZHANG Q., SHI C., YANG W., SHI Y., YANG S., XU L., YU J.: Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117* (2023).

[ZWS*23] ZHAO X., WANG L., SUN J., ZHANG H., SUO J., LIU Y.: Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Transactions on Graphics 43*, 1 (2023), 1–16.