Analyzing and Improving the Skin Tone Consistency and Bias in Implicit 3D Relightable Face Generators

Libing Zeng
Texas A&M University
College Station, USA
libingzeng@tamu.edu

Nima Khademi Kalantari
Texas A&M University
College Station, USA
nimak@tamu.edu

Abstract

With the advances in generative adversarial networks (GANs) and neural rendering, 3D relightable face generation has received significant attention. Among the existing methods, a particularly successful technique uses an implicit lighting representation and generates relit images through the product of synthesized albedo and light-dependent shading images. While this approach produces high-quality results with intricate shading details, it often has difficulty producing relit images with consistent skin tones, particularly when the lighting condition is extracted from images of individuals with dark skin. Additionally, this technique is biased towards producing albedo images with lighter skin tones. Our main observation is that this problem is rooted in the biased spherical harmonics (SH) coefficients, used during training. Following this observation, we conduct an analysis and demonstrate that the bias appears not only in band 0 (DC term), but also in the other bands of the estimated SH coefficients. We then propose a simple, but effective, strategy to mitigate the problem. Specifically, we normalize the SH coefficients by their DC term to eliminate the inherent magnitude bias, while statistically align the coefficients in the other bands to alleviate the directional bias. We also propose a scaling strategy to match the distribution of illumination magnitude in the generated images with the training data. Through extensive experiments, we demonstrate the effectiveness of our solution in increasing the skin tone consistency and mitigating bias.

1. Introduction

In recent years, there has been a growing interest in learning 3D generative models of faces from 2D images [2, 3, 6, 13, 24, 37, 40] through a combination of generative adversarial networks (GAN) [12] and neural radiance fields [22]. However, utilizing such 3D digital humans in various applications, such as virtual/augmented reality and gaming, necessitates full control over different image formation factors such as geometry, appearance, and lighting.

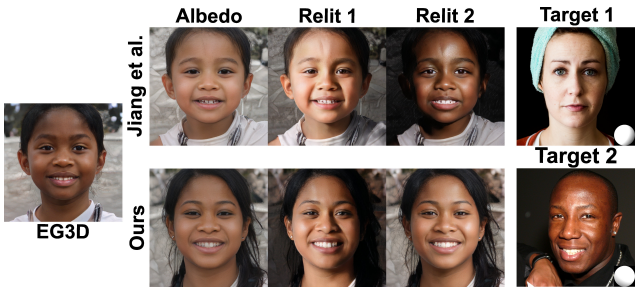

Figure 1. On the top, we show two relit images produced by NeRFFaceLighting (Jiang et al.) [17], using the lighting extracted from images of individuals with fair and dark skin tones (shown on the right). As seen, NeRFFaceLighting produces relit images with inconsistent skin tones. Additionally, when distilling the EG3D triplane, NeRFFaceLighting tends to produce albedo maps that are biased towards lighter skin colors. Our method mitigates this bias and improves the consistency of the skin tone in relit images. Note that even though we use the same latent vector to generate the results with EG3D, NeRFFaceLighting, and ours, there are variation in the images as the backbone EG3D network is fine-tuned separately in NeRFFaceLighting and ours.

Several approaches [4, 25, 30, 35] attempt to address this entanglement problem by incorporating analytical lighting models, such as Phong, into the 3D face generators. Specifically, the generator estimates the shading components, such as diffuse and specular, which are then used along with the lighting information in a shading model to produce the final images. Because of using analytical models, these approaches can produce consistent relit images that match the input illumination. However, due to computational costs, these techniques use simple shading models that ignore realistic effects such as subsurface scattering and interreflection. As a result, they often produce images that do not have the intricate shading details of real faces.

The state-of-the-art method of Jiang et al. [17] (NeRFFaceLighting) tackles this issue through an implicit lighting representation. The key idea is to distill the original single triplane of EG3D [2] into shading and albedo triplanes. The

shading triplane is estimated through an adapter network that is built on top of the EG3D generator and additionally takes the lighting information in form of spherical harmonics (SH) [29] as the input. Training is performed using the original EG3D discriminator, along with another discriminator that is additionally conditioned on the SH coefficients. The extra conditional discriminator forces the generator to produce relit images that are consistent with the lighting. Due to the flexibility of the implicit lighting model, NeRFFaceLighting produces 3D portraits with more realistic lighting compared to methods using explicit lighting models.

Despite producing impressive results, NeRFFaceLighting suffers from two major drawbacks. First, this technique is not able to produce relit images with consistent skin tones. Specifically, when the target lighting is obtained from images with dark skin, NeRFFaceLighting produces portraits with a dark skin tone, regardless of the albedo's skin color. As shown in Fig. 1, the relit image corresponding to the target lighting from a dark skin individual (Target 2), has a significantly darker skin color than the albedo. Second, NeRFFaceLighting distills EG3D into an albedo representation that is biased towards lighter skin tones (see Figs. 1 and 7).

Our key contribution is to conduct an analysis to pinpoint the root causes of these issues, and present a simple, yet effective, strategy to mitigate the problems. We observe that the two drawbacks are connected and the core issue lies in the bias of existing lighting estimation methods [32], which are used to generate the SH coefficients for training images. These techniques tend to estimate SH coefficients that exhibit correlation with skin color. The discriminator observes the correlation in real data and forces the generator to produce relit images with dependency between the skin color and lighting. Because of using implicit lighting representation, the generator in NeRFFaceLighting has the flexibility to change the skin color through shading. Furthermore, since the generator can produce individuals with dark skin color through shading, it opts for synthesizing albedo images with lighter skin tone.

Following this observation, we conduct an analysis and demonstrate that the bias manifests itself not only in band 0 (DC term) of the estimated SH coefficients, but also in the other bands. Specifically, estimated SH coefficients of individuals with darker skin have smaller DC term, indicating that existing light estimation methods are biased towards albedos with lighter skin and compensate the difference by estimating dimmer lighting. Moreover, the coefficients in other bands of images with dark skin form a distinct cluster, indicating directional bias in the estimated coefficients.

We propose to address the magnitude and directional biases by utilizing normalized and statistically-aligned SH coefficients as input for both the generator and discriminator during training. Specifically, we normalize the coefficients by the DC term to ensure a constant average for all lightings, thereby alleviating the bias in illumination magnitude. To address the bias in the other bands, we make a key observation that the illumination in the training data is independent of the skin tone. As such, the illumination of the dark and non-dark skin tones should be statistically similar. Based on this observation, we propose to statistically align the SH coefficients of the individuals with dark skin tones to the ones corresponding to images with non-dark skin colors.

We perform the training using these normalized and statistically-aligned SH coefficients. One potential problem is that the generator takes normalized SH coefficients, and thus should produce relit images with constant illumination magnitude. However, the illumination magnitude of the training images vary significantly. To address this issue, we utilize the linearity of light and propose to scale the generated relit images, thereby scaling their lighting, to match the distribution of the training data.

Once trained, our generator can produce relit images consistent with the direction of illumination, although it is limited to handling lighting with a constant magnitude. To accommodate illuminations with arbitrary magnitudes, we use our generator to produce results under normalized lighting and then directly scale the relit images to the appropriate magnitude. We demonstrate that these modifications significantly improve the consistency of skin tone in relit images under varying lighting conditions and alleviate the albedo generation bias toward lighter skin tones (see Fig. 1). Moreover, we demonstrate that our approach can be used to improve the skin tone consistency of other relighting approaches like, DiFaReli [27]. *We will release the source code upon publication.*

2. Related Work

In this section, we discuss the closely related work on 3D generative networks, 3D relightable face generators, and portrait relighting approaches. We also provide a brief review of the algorithms that address bias in light/albedo estimation.

2.1. 3D Generative Networks

Generative Adversarial Networks (GAN) [12], in particular StyleGAN [18–20], are capable of producing results that are virtually indistinguishable from real images. A large number of techniques [2, 3, 6, 13, 24, 37, 40] combine GAN with neural radiance field (NeRF) [22] to synthesize 3D consistent high-quality images. In particular, EG3D [2], among the widely used 3D generators, integrates NeRF into StyleGAN [18–20] by introducing tri-plane representation. Built on the success of EG3D, several approaches [11, 33, 36] use it as a prior for controllable 3D-aware facial image manipulation. In particular, Tang et al. [36] incorporate guidance from a parametric head model into the generator to control illumination. However, the quality of their results is limited by the expressiveness of the parametric model.

2.2. 3D Relightable Face Generators

In recent years, several approaches [4, 25, 30, 35] propose 3D relightable face generators by incorporating analytical shading models into their generators. For example, Pan et al. [25] estimate shading using the Phong illumination model and multiplies it with the synthesized albedo to reconstruct the relit image. Deng et al. [4] utilize a more complex shading model and handle visibility. The shading models used by these approaches, however, are not able to handle effects such as subsurface scattering which are necessary for faithful reproduction of the appearance of facial skin. To address this issue, Jiang et al. [17] (NeRFFaceLighting) propose to implicitly model the shading by distilling EG3D into two separate triplane representations. However, their approach has difficulty in maintaining the consistency of skin color in relit images and synthesizes albedo images that are biased towards lighter skin tones. The goal of our paper is to determine and address the underlying causes of these issues.

2.3. Portrait Relighting

A large number of approaches [14, 15, 23, 26, 27, 34, 38, 39] perform relighting from a single 2D image. These approaches typically need to estimate the reflectance and geometry, and thus rely on either lab-captured or synthetic images for training. Unfortunately, such a reliance hampers the generalization capabilities of these algorithms. Furthermore, the 3D structure of the portrait is often not fully taken into consideration during the relighting, leading to suboptimal results, particularly in challenging lighting conditions.

2.4. Bias in Light/Albedo Estimation

Estimating the image formation factors, particularly albedo and illumination, from a single image is a highly ambiguous task. A couple of recent techniques [9, 31], demonstrate that current techniques [1, 7–10, 16, 31] are biased towards estimating albedo images with lighter skin colors, and propose various ways to mitigate the problem. For example, Feng et al. [9] propose to use the full scene to disambiguate lighting and appearance. Despite producing promising results, we cannot utilize their method for unbiased light estimation as we only have access to portrait images. Ren et al. [31] utilize text to image models to force the estimated albedo and input images have similar skin tones. However, they only focus on albedo estimation and do not present a strategy for unbiased light estimation.

Notably, the work by Legendre et al. [21] focuses on estimating high frequency lighting from portrait images, striving for enhanced precision and reduced bias in light estimation. To achieve this, they use light stage data of 70 subjects to train a light estimation network. Unfortunately, such a small scale dataset inherently lacks diversity in subjects, expressions, and accessories. Consequently, the performance of this approach on real portrait images with diverse characteristics could be suboptimal.

3. Method

Built upon the pre-trained 3D face generator model by Chen et al. [2] (EG3D), NeRFFaceLighting [17] distills the fused appearance and lighting information in the original triplane, into two triplanes. In this approach, one triplane encodes the geometry and albedo information, while the other is only responsible for producing the shading. Given a random latent vector $\mathbf{z}$, camera pose $\mathbf{v}$, and lighting condition $\mathbf{l}$, NeRFFaceLighting produces a relit image $G(\mathbf{z}, \mathbf{v}, \mathbf{l})$, through the product of a synthesized shading $S(\mathbf{z}, \mathbf{v}, \mathbf{l})$ and albedo $A(\mathbf{z}, \mathbf{v})$, i.e., $G(\mathbf{z}, \mathbf{v}, \mathbf{l}) = S(\mathbf{z}, \mathbf{v}, \mathbf{l})A(\mathbf{z}, \mathbf{v})$. Note that only the shading is conditioned on the lighting as appearance (albedo) is independent of the illumination.

Training the entire system is done using two discriminators: one exactly follows EG3D and is only conditioned on the camera pose, while the other is additionally conditioned on lighting to ensure the relit images are consistent with the lighting condition. This additional discriminator, necessitates extracting lighting information from the real images in the training data. NeRFFaceLighting does so using SfSNet [32] which represents the light with nine 2^{nd} order spherical harmonics (SH) coefficients, $\mathbf{l} \in \mathbb{R}^9$.

As discussed, although the implicit lighting representation allows this approach to model intricate shading details, this flexibility introduces two major issues. First, when lighting condition comes from individuals with dark skin tones, this approach produces relit images with dark skin, regardless of the albedo's skin color. As shown in Fig. 5, the shading image corresponding to the target lighting from a dark skin individual (Target 2), exhibits darkening of the eyes and bright highlights on the cheeks, forehead, and lips, resulting in a relit image with an altered skin color. Second, NeRFFaceLighting exhibits a significant bias towards albedo maps with lighter skin tones, as shown in Fig. 7.

In generative adversarial networks (GANs), the generator attempts to follow the distribution of the training data. Since, NeRFFaceLighting uses two conditional discriminators, the generator is forced produce relit images that match the following two data distributions: $p_{\text{data}}(\mathbf{I}|\mathbf{l}, \mathbf{v})$ and $p_{\text{data}}(\mathbf{I}|\mathbf{v})$. Our key observation is that the generator produces dark skin tones for certain lightings in an attempt to match the training data distribution $p_{\text{data}}(\mathbf{I}|\mathbf{l}, \mathbf{v})$, i.e., the SH coefficients and skin tone for training images of individuals with dark skin are correlated. This correlation stems from the existence of bias in the estimated SH coefficients by SfSNet [32].

Note that this also explains the reason behind NeRFFaceLighting's bias towards albedos with lighter skin tones. The generator should produce relit images that follow the distribution $p_{\text{data}}(\mathbf{I}|\mathbf{v})$. That means the skin tone distribution of generated and the training images should be similar. Since

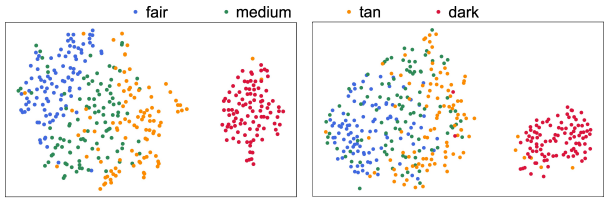

Figure 2. We visualize the 2nd order SH coefficients estimated using SfSNet [32] and DECA [10] from 400 images with different skin colors (100 in each category). We use t-SNE to visualize the coefficients in 2D. The coefficients extracted from images with dark skin form a distinct cluster in both cases.

the generator can produce images of individuals with dark skin tone through shading, it can match the distribution of the training data without resorting to producing albedo's with dark skin tone. Therefore, addressing the skin tone consistency of relit images will automatically fix the albedo's bias toward lighter skin colors.

In the following sections, we perform an analysis to better understand the bias and present our solution.

3.1. Analysis

We begin by conducting an experiment to verify our observation that the problems originate from the bias in predicted SH coefficients. To do so, we randomly select 50,000 images from FFHQ [19] and classify them based on the skin tone. To classify the images, we utilize the CLIP model [28] and evaluate the similarity between a set of four texts, describing the skin tone, and the input image. Specifically, we use the text "a photo of a person with $\{c\}$ skin tone", where $c \in \{\text{fair, medium, tan, dark}\}$. We pick the text with the highest similarity as the skin tone of the individual.

Subsequently, we randomly sample 100 images from each class and extract their SH coefficients using SfSNet [32], following the methodology of NeRFFaceLighting [17]. We then visualize these 400 9-dimensional SH coefficients in a two-dimensional space using t-SNE, which is a nonlinear dimensionality reduction technique. As depicted in Fig. 2 (left), the SH coefficients extracted from images with dark skin tones (shown in red) are distinctly clustered away from the other non-dark SH coefficients. Because of this bias, the discriminator can easily find significant correlation between the clustered SH coefficients and darker skin tones and force the generator to produce similar results. Note that this bias is not unique to SfSNet and exists in other portrait light estimation techniques as well. To demonstrate this, we perform the same analysis, but using the SH coefficients estimated by DECA [10]. As shown in Fig. 2 (right), the SH coefficients from dark skin tones once again form a separate cluster.

Given this observation, a question naturally arises: how do these SH coefficients differ from those extracted from non-dark faces? Estimating albedo and lighting from a single image is a highly ambiguous task with many plausible solutions. For example, an image can be explained by a dark albedo and bright light, as well as a light albedo and dim light. Recent studies [9, 31] have shown that the current albedo/light estimation methods are biased towards producing albedos with lighter skin tones; they compensate this bias by estimating dimmer lighting. To verify this, we visualize the SH coefficients in band 0 (corresponding to the lighting magnitude) in Fig. 3 (left). Note that to better visualize the one dimensional band 0 coefficients, we add an extra dimension, filled with random values between 0 and 1, and show the scatter plot of these 2D points. As seen in Fig. 3 (left), the coefficients corresponding to dark skin tones have generally smaller DC terms (left is smaller values), and thus represent dimmer lightings.

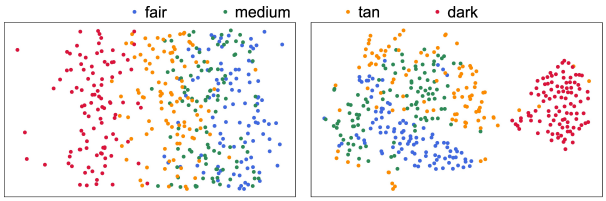

Figure 3. We visualize the SH coefficients, estimated by SfSNet, in band 0 and other bands. We augment the one dimensional coefficients in band 0 with an additional randomly filled dimension for better visualization. For other bands, however, we use t-SNE to reduce the dimensions from eight to two. As seen, the bias is not limited to the magnitude of the lighting (band 0) and appears in other higher order SH coefficients as well.

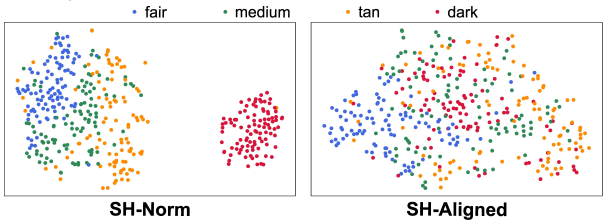

Figure 4. On the left, we visualize the SH coefficients after normalization. On the right, we showcase the normalized SH coefficients with statistical alignment of the coefficients of dark to non-dark skin tones. This approach effectively mitigates bias in the estimated SH coefficients.

We further visualize the SH coefficients in other bands to determine if the bias in band 0 fully accounts for the bias in the SH coefficients. Fig. 3 (right) show this visualization where we use t-SNE to reduce the dimension of other bands from eight to two. Surprisingly, other bands show even stronger bias towards individuals with dark skin tones. This demonstrates that the bias not only manifests itself in the magnitude of lighting, but is also encoded in the directions provided by higher order SH coefficients.

3.2. Mitigating Bias

To address the problem of skin tone consistency and bias in relightable generators, we need to tackle the source of the issue, i.e., use unbiased SH coefficients during training. The obvious solution is to use an unbiased light estimation method. However, as demonstrated in Sec. 3.1, even the

widely used technique by Feng et al. [10], estimates biased SH coefficients. Note that while there have been a couple of recent attempts to mitigate the skin tone bias [9, 21, 31], these techniques either only focus on albedo estimation, require the full scene (not just portrait images), or could have difficulty generalizing to real images with diverse subjects, expression, and poses.

Therefore, we opt for using the existing biased light estimation methods, but propose a simple strategy to mitigate their bias and prevent the discriminator from finding correlation between the coefficients and skin tone. Specifically, to combat the bias in band 0, we propose to normalize the SH coefficients by their DC term, i.e. $\mathbf{l}_n = \mathbf{l}[0:9]/\mathbf{l}[0]$. Normalizing the coefficients by the DC term completely removes the bias from the coefficients in band 0.

To address the bias in the other bands, we make a key observation that the illumination of the scene in the training data is independent of the individual's skin color. Therefore, the statistics (mean and standard deviation) of the SH coefficients of the images with dark skin colors, should be similar to the ones with non-dark skin tones. Armed by this observation, we propose to compute the mean and standard deviation of the SH coefficients corresponding to dark and non-dark individuals over the entire training data. We use the CLIP-based strategy presented in Sec. 3.1 to place the images into dark and non-dark (fair, medium, and tan) categories. We then adjust the SH coefficients of images with dark skin tone to match the statistics of the non-dark individuals as follows:

$$\mathbf{l}_{\text{nsa}}[i] = \frac{\mathbf{l}_n[i] - \mu_d[i]}{\sigma_d[i]}\sigma_{\text{nd}}[i] + \mu_{\text{nd}}[i], \tag{1}$$

where $\mathbf{l}_{\text{nsa}}[i]$ is the i^{th} normalized and statistically-aligned SH coefficient. Moreover, μ_d, σ_d, μ_{nd}, and σ_{nd} are the mean and standard deviation of the normalized SH coefficient of dark and non-dark images, respectively. Note that for images with non-dark individuals we use the normalized SH coefficients without any other modifications, i.e., $\mathbf{l}_{\text{nsa}} = \mathbf{l}_n$.

As shown in Fig. 4, the 2^{nd} order coefficients after normalization still display noticeable clusters, whereas the normalized coefficients after statistical-alignment show no discernible clusters, indicating effective bias mitigation.

Special care should be taken when using the normalized SH coefficients during training. To minimize the GAN loss, the generated relit images (input to the discriminator) should match the distribution of training data with diverse illumination magnitude. On the other hand, our generator takes normalized (and statistically-aligned) SH coefficients as the input and is expected to produce relit images with constant illumination magnitude. Training the system without considering this mismatch will force the generator to produce images with diverse illumination magnitude, even with normalized SH coefficients as the input.

To address this issue, we propose to randomly adjust the magnitude of the illumination of the generated images during training. Because of the linearity of the light, adjusting the magnitude of the illumination can be done in the image domain as follows:

$$\hat{I} = (G(\mathbf{z}, \mathbf{v}, \mathbf{l}_{\text{nsa}})^{\gamma} \times s)^{\frac{1}{\gamma}} \tag{2}$$

where $\gamma = 2.2$ in our implementation and s is the scaling factor. Here, we first take the generated image with normalized illumination into the linear domain with gamma expansion. We then transform the scaled image into tonemapped domain with gamma compression.

The key idea is to force the generator to produce relit images with an illumination magnitude equal to the average light intensity over the training data by carefully setting the distribution of the scaling factor s. This can be achieved by calculating s as follows:

$$s = \frac{m(I_i)}{\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} m(I_n)}, \tag{3}$$

where I_i is a randomly selected training image, $m(I_i)$ is the illumination magnitude of the image, and $\mathcal{N}$ refers to the set of all training images. Since s is computed based on the ratio of magnitudes, we propose to approximate the illumination magnitude using average of the pixel intensities in the facial area. The only remaining caveat is that the ratio should be computed on images with similar skin tones. Our final formulation is thus as follows:

$$s = \frac{m(I_i)}{\frac{1}{|\mathcal{N}_c|}\sum_{n\in\mathcal{N}_c} m(I_n)}, \text{ where } m(I_i) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} I_i[p]. \tag{4}$$

Here, $\mathcal{N}_c$ refers to the subset of training images with the same skin tone (fair, medium, tan, and dark) as I_i. Moreover, $\mathcal{P}$ is a subset of pixels on the face. We demonstrate the impact of this magnitude scaling scheme in Sec. 4.4.

Once trained, our generator can produce relit images with constant illumination. To produce images with varying lighting magnitude, we simply use Eq. 2 to adjust the illumination scale.

4. Results

Throughout this section we analyze the performance of different versions of NeRFFaceLighting and compare them against our proposed strategy. Specifically, these variations include:

- **NeRFFaceLighting**: The officially released checkpoint, provided by the authors.
- **NeRFFaceLighting-DECA**: Using SH coefficients estimated by DECA [10].
- **SH-Norm**: Using normalized SH coefficients.
- **Ours**: Using normalized and statictically-aligned SH coefficients.

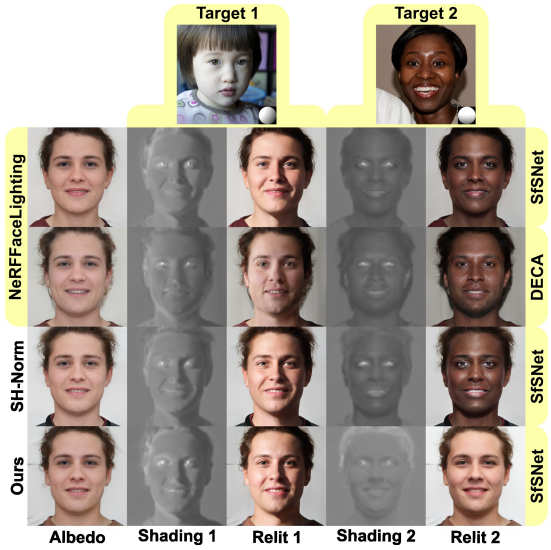

Figure 5. We compare different approaches by producing relit images using two target lightings. Our approach produces results with consistent skin tone for both lightings.

We train all the variations on FFHQ [19] for 5M images, following the training strategy of NeRFFaceLighting. We use a batch size of 28 and perform the training on 4 A100 GPUs. Note that, in all cases except **NeRFFaceLighting-DECA**, we use SfSNet [32] to obtain the SH coefficients.

4.1. Relighting Sampled Images

In Fig. 5, we compare how different variations preserve the skin tone in relit images. Specifically, we perform relighting using the SH coefficients extracted from two target images with fair (Target 1) and dark (Target 2) skin tones. As seen, while the sampled albedo has a fair skin tone, **NeRFFaceLighting** produces a relit image with dark skin for the second target. Similarly, **NeRFFaceLighting-DECA** is not able to produce relit images that are consistent with albedo's skin color. This demonstrates that both light estimation techniques (DECA and SfSNet) produce biased SH coefficients. Moreover, normalizing the SH coefficients (**SH-Norm**) alone is not sufficient. On the other hand, our approach using normalized and statistically-aligned SH coefficients produces relit images that have fair skin color and are consistent with the albedo. It is worth noting that, although the second target exhibits strong specular highlights, our approach correctly produces a relit image with a more diffuse appearance. This is because the effect of subsurface scattering is stronger in fair skin tones, and thus fair skins under the same illumination have a more diffuse appearance.

We further numerically evaluate the effectiveness of different variations in maintaining skin color consistency after relighting. Specifically, we randomly sample 100 latent codes corresponding to albedo maps with a fair skin color. We then relight each albedo with 100 randomly sampled lightings and measure the skin tone consistency between each relit image and its corresponding albedo. To do so, we first use the CLIP model [28] and follow the process described in Sec. 3.1 to obtain a set of 4 similarity scores between each image and four skin tone categories (fair, medium, tan, dark). We then compute the cosine similarity between the scores for the relit image and its corresponding albedo as the measure of skin tone consistency. A value close to one corresponds to perfect skin tone consistency. Table 1 shows the average (Avg.), standard deviation (STD), and minimum scores. Our method produces results with higher average and lower standard deviation, indicating superior performance in preserving facial color in relit images. Notably, our worst case scenario, indicated by the minimum score, is significantly better than NeRFFaceLighting.

Table 1. Quantitative comparison against the other approaches in terms of skin tone consistency metric.

	Avg. ↑	STD ↓	Minimum ↑
NeRFFaceLighting	0.9704	0.0386	0.4404
NeRFFaceLighting-DECA	0.9622	0.0591	0.1065
SH-Norm	0.9652	0.0389	0.5912
Ours	**0.9745**	**0.0221**	**0.6388**

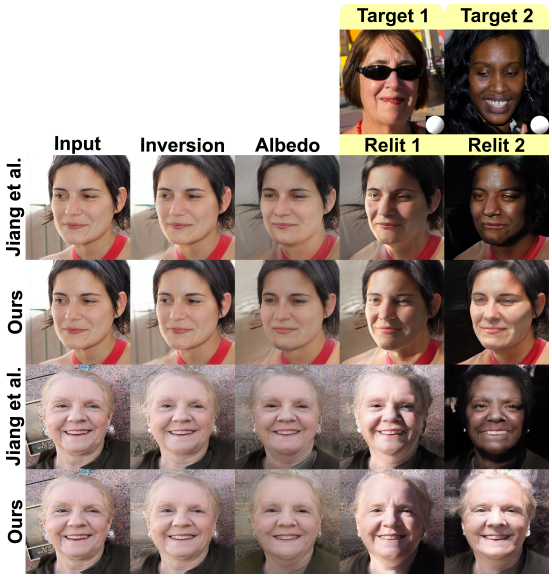

Figure 6. Comparison against NeRFFaceLighting [Jiang et al.] on real images. Our relit images have consistent skin colors, while NeRFFaceLighting produces results with significantly darker skin for the second target.

4.2. Relighting Real Images

We compare the ability of different variations in preserving the consistency of skin tone when relighting real images. To do so, we follow NeRFFaceLighting [17] and project the input images into the latent space of the generator. We then relight the inverted images using lighting estimated from two target images with fair (Target 1) and dark (Target 2) skin tones. As shown in Fig. 6, NeRFFaceLighting produces relit images with different skin tones, while our relit images have consistent skin colors.

Table 2. KL divergence between the distribution of the skin colors of generated albedo images by different approaches and EG3D.

	KL divergence ↓
NeRFFaceLighting	0.0043
NeRFFaceLighting-DECA	0.0047
SH-Norm	0.0040
Ours	**0.0029**

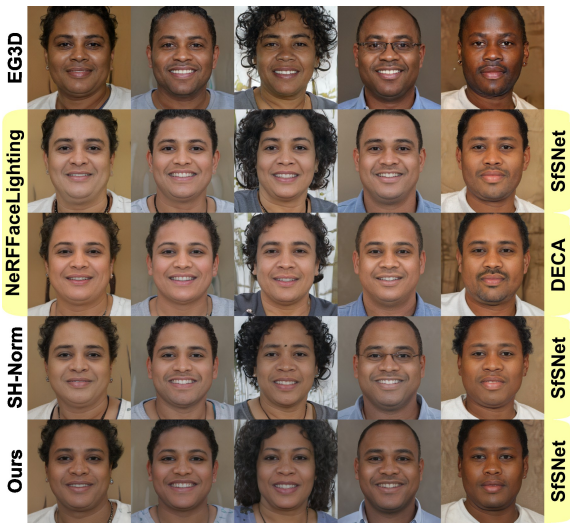

Figure 7. We show several randomly sampled images from EG3D and the corresponding albedo images produced by different variations of NeRFFaceLighting and ours. When distilling EG3D, both versions of NeRFFaceLighting tend to produce albedos with lighter skin colors. On the other hand, our solution produces albedo images that better represents the skin color of the original EG3D samples.

4.3. Albedo's Skin Tone Bias

We begin by showing a few synthesized albedo images with dark skin in Fig. 7. When distilling the EG3D triaplane, NeRFFaceLighting trained with SH coefficients from both SfSNet and DECA has bias towards producing albedos with lighter skin colors, as evident by the results. Our solution, however, mitigates this bias and produces albedo images with darker skin tones, resembling the corresponding samples from EG3D. Note that although the images in each column are produced by sampling the same latent vector, there are small variations in appearance. These variations, however, are expected as the models are trained independently. While we start with the EG3D generator in each case, fine-tuning the model using the adversarial loss leads to small variations in different attributes.

Next, we numerically compare the ability of different methods in properly distilling the appearance (albedo) from EG3D. To do so, we randomly sample 50,000 latent codes and generate the corresponding albedo for each method. We then use the CLIP-based method, described in Sec. 3.1, to assign one of the four skin color categories to each albedo image. Following this, we obtain the distribution of generated albedos for each method in terms of the four categories. We also perform the same process and obtain the distribution of skin colors for EG3D. We then find the KL divergence between the skin color distribution of each method and EG3D. Smaller KL divergence means the approach is not introducing skin color bias on top of EG3D. As shown in Table 2, our results have the smallest KL divergence, demonstrating the effectiveness of our solution in mitigating the bias.

Figure 8. Evaluation of the impact of magnitude scaling scheme during training. When randomly sampling images with the same lighting condition, our results demonstrate constant illumination magnitude. In contrast, the generator trained without the scaling technique displays varying illumination magnitudes.

Table 3. Illumination Magnitude Consistency.

	Standard deviation ↓
w/o Scaling	0.2349
Ours	**0.1011**

4.4. Effect of Magnitude Scaling

Here, we investigate the impact of the magnitude scaling scheme, discussed in Sec. 3.2 (Eq. 2), by comparing our approach against a variant of our technique that is trained without scaling. In Fig. 8, we show images obtained by randomly sampling the latent code, but with fixed camera and lighting condition. As seen, the approach without scaling produces images with slight variation in the lighting magnitude, while our method exhibits constant illumination.

We further evaluate the impact of this component numerically. We do so by first generating 1,000 relit images by randomly sampling the latent code, camera, and lighting (normalized SH coefficients). We then use DECA [10] to estimate the illumination magnitude (DC of the SH coefficients) and report their standard deviation in Table 3. Note that, to avoid DECA's bias, we ensure all the 1,000 generates images are of individuals with fair skin tones using the CLIP-based method, described in Sec. 3.1. As seen, our generator trained with scaling exhibits smaller magnitude variation.

4.5. Direct Applications

Our approach statistically mitigates the bias in SH coefficients, making it applicable for improving skin tone consistency in any relighting technique that uses SH coefficients as a condition. To demonstrate this, we apply our approach to another GAN-based 3D relightable generator, FaceLit [30], and a recent diffusion-based relighting method, DiFaReli [27]. Both of these methods use SH coefficients of lighting extracted via DECA [10]. Note that here, we di-

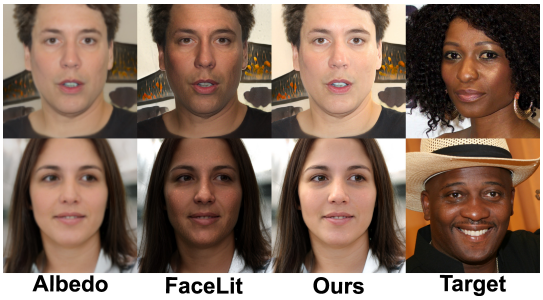

Figure 9. Comparisons against FaceLit [30] on sampled albedo images.

Table 4. Quantitative comparison against FaceLit [30] in terms of the consistency of skin tone and identity.

	Skin tone consistency			Identity similarity		
	Avg. ↑	STD ↓	Minimum ↑	Avg. ↑	STD ↓	Minimum ↑
FaceLit	0.9651	0.0439	0.7886	0.7542	0.1007	0.4716
Ours	**0.9708**	**0.0381**	**0.8052**	**0.7908**	**0.0967**	**0.5272**

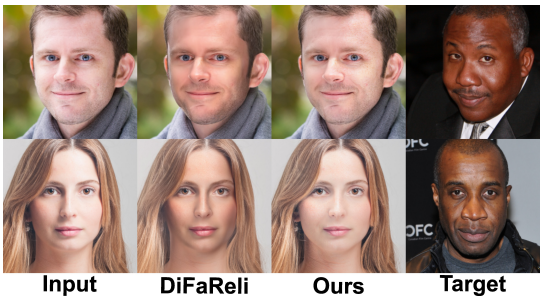

Figure 10. Comparisons against DiFaReli [27] on real images.

rectly apply our approach to FaceLit [30] and DiFaReli [27] without fine-tuning.

FaceLit [30], similar to NeRFFaceLighting, produces relightable 3D portraits by incorporating lighting condition into EG3D [2]. We follow their recommended approach to generate albedo images by rendering them before the superresolution module under constant illumination. We then relight the albedo images using SH coefficients extracted via DECA [10]. As shown in Fig. 9, with our statistically aligned SH coefficeints, the relit images exhibit improved skin tone consistency. To further numerically compare our approach with FaceLit [30], we randomly sample 100 albedo images and relight each with 100 random lighting conditions from the FFHQ dataset. We then compute skin tone consistency score, as introduced in Sec. 4.1, and identity similarity score using ArcFace [5], a face recognition method, between the sampled albedo and relit images. As shown in Table 4, our method performs better across all metrics, with higher averages, lower standard deviations, and larger minimum values in terms of both skin tone consistency and identity similarity scores.

We further visually compare our approach against DiFaReli [27], a relighting method that utilizes a conditional

Table 5. Quantitative comparison against DiFaReli [27] in terms of the consistency of skin tone and identity.

	Skin tone consistency			Identity similarity		
	Avg. ↑	STD ↓	Minimum ↑	Avg. ↑	STD ↓	Minimum ↑
DiFaReli	0.9846	0.0257	0.6973	0.7622	0.1568	0.1913
Ours	**0.9894**	**0.0205**	**0.8096**	**0.8170**	**0.1280**	**0.3332**

diffusion model, in Fig. 10. As seen, this technique produces results with inconsistent skin tone using the SH coefficients estimated by DECA [10] from the target images with dark skin tones. However, using our normalized and statistically aligned SH coefficients, it is able to produce results with better skin tone consistency. We further numerically demonstrate the effectiveness of our approach in preserving skin tone consistency and identity similarity in Table 5. As seen, our method produces better results with better skin tone consistency and identity similarity, compared to DiFaReli with the original DECA coefficients.

5. Conclusion, Limitations, and Future Work

We have presented a comprehensive analysis and solution to the problem of the skin tone inconsistency and bias in the relightable face generator with implicit lighting representation. We observe that the issue stems from the bias in estimated lighting, which presents itself not only in the magnitude of illumination (band 0), but also in the other higher order bands of the spherical harmonics coefficients. Based on this observation we suggest performing the training using normalized and statistically-aligned SH coefficients. We demonstrate that this simple solution is highly effective in mitigating the bias and preserving the skin tone consistency of relit images, produced by a few GAN-based and diffusion-based relighting methods, highlighting the broad applicability of our method.

Our method relies on proper estimation of the skin tone category using the CLIP model. While we demonstrate that our solution is highly effective, there may be cases where the CLIP model incorrectly detects the skin category, resulting in an incorrect alignment. We believe that mitigating this issue by investigating an unbiased light estimation method would be an interesting future research direction.

Moreover, although our solution significantly reduces the bias, and improves the quality compared to NeRFFaceLighting, there is still a small gap in image quality between our approach and the backbone EG3D (see Table 1 of the supplementary document). In the future, we would like to explore potential approaches to further reduce this gap.

6. Acknowledgements

We sincerely thank the anonymous reviewers for their valuable feedback and constructive suggestions. Additionally, portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

References

[1] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093, 2012. 3

[2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 8

[3] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[4] Boyang Deng, Yifan Wang, and Gordon Wetzstein. Lumigan: Unconditional generation of relightable 3d human faces. In *International Conference on 3D Vision (3DV)*, 2024. 1, 3

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 8

[6] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[7] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[8] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision (IJCV)*, 126:1269–1287, 2018. 3

[9] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 5

[10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 3, 4, 5, 7, 8

[11] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. VIVE3D: Viewpoint-independent video editing using 3D-aware GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 1, 2

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 1, 2

[14] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4217–4226, 2022. 3

[15] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14719–14728, 2021. 3

[16] Guosheng Hu, Pouria Mortazavian, Josef Kittler, and William Christmas. A facial symmetry prior for improved illumination fitting of 3d morphable model. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013. 3

[17] Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. Nerffacelighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics (TOG)*, 42, 2023. 1, 3, 4, 6

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. 34:852–863, 2021. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 4, 6

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[21] Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. Learning illumination from diverse portraits. In *SIGGRAPH Asia 2020 Technical Communications*, pages 1–4. 2020. 3, 5

[22] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[23] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5124–5133, 2020. 3

[24] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[25] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3

[26] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 3

[27] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22646–22657, 2023. 2, 3, 7, 8

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 4, 6

[29] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. 2001. 2

[30] Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, and Oncel Tuzel. Facelit: Neural 3d relightable faces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 7, 8

[31] Xingyu Ren, Jiankang Deng, Chao Ma, Yichao Yan, and Xiaokang Yang. Improving fairness in facial albedo estimation via visual-textual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, June 2023. 3, 4, 5

[32] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 6

[33] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 2

[34] 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):1–12, 2019. 3

[35] Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with hdri relighting. *ACM SIGGRAPH*, 2022. 1, 3

[36] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. 3dfaceshop: Explicitly controllable 3d-aware portrait generation. *IEEE Transactions on Visualization & Computer Graphics*, (01):1–18, 2023. 2

[37] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2195–2205, October 2023. 1, 2

[38] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6):1–21, 2022. 3

[39] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7194–7202, 2019. 3

[40] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. 2021. 1, 2