VoLux-GAN: A Generative Model for 3D Face Synthesis with HDRI Relighting

Feitong Tan[1,2,*] Sean Fanello[1] Abhimitra Meka[1] Sergio Orts-Escolano[1] Danhang Tang[1]
Rohit Pandey[1] Jonathan Taylor[1] Ping Tan[2] Yinda Zhang[1]
[1] Google [2] Simon Fraser University

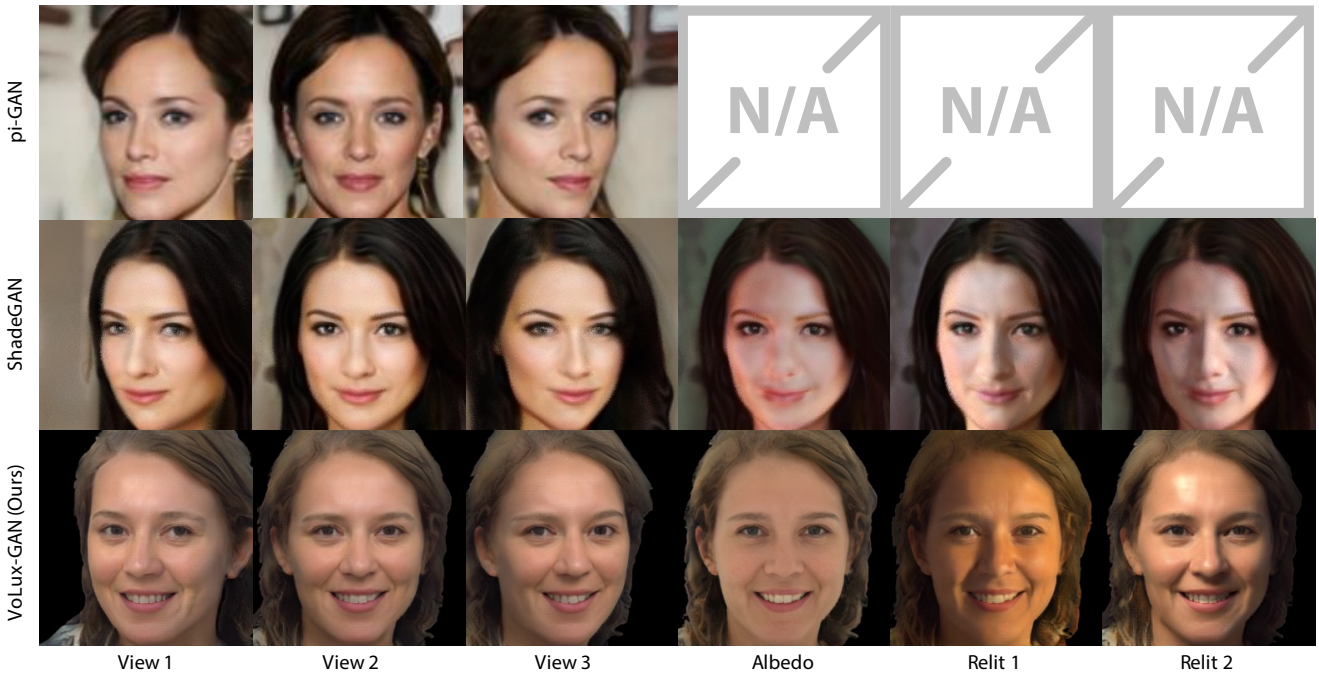

Figure 1. We propose VoLux-GAN, a 3D-aware generator that produces faces with full HDRI relighting capability. Here we show a comparison of images generated by VoLux-GAN and related work pi-GAN [9] (which does not support relighting) and ShadeGAN [39].

Abstract

We propose VoLux-GAN, a generative framework to synthesize 3D-aware faces with convincing relighting. Our main contribution is a volumetric HDRI relighting method that can efficiently accumulate albedo, diffuse and specular lighting contributions along each 3D ray for any desired HDR environmental map. Additionally, we show the importance of supervising the image decomposition process using multiple discriminators. In particular, we propose a data augmentation technique that leverages recent advances in single image portrait relighting to enforce consistent geometry, albedo, diffuse and specular components. Multiple experiments and comparisons with other generative frameworks show how our model is a step forward towards photorealistic relightable 3D generative models.

1. Introduction

Generating synthetic novel human subjects with convincing photorealism is one of the most desired capabilities for automatic content generation and pseudo ground truth synthesis for machine learning. Such data generation engines can thus benefit many areas including the gaming and movie industries, telepresence in mixed reality, and computational photography. In order to achieve realism and flexibility when delivered in specific applications, the generated

*Work done while the author was an intern at Google.

images should: 1) be enriched in details, *e.g*. with high resolution; 2) support free viewpoint rendering to deliver immersive 3D experiences; 3) adapt to novel environmental illumination for realism; 4) synthesize novel identities for scalable data diversity.

Motivated by these principles, in this paper we propose a neural human portrait generator, which delivers compelling rendering quality on arbitrary camera viewpoints and under any desired illumination. With the success of Neural Radiance Field (NeRF) on volumetric rendering [32] and Generative Adversarial Networks (GAN) on image generation [25], 3D-aware generators [9, 19, 39] have been proposed as a promising solution, which combine the merits of both. By learning from a collection of portrait images, these methods are able to generate NeRF models from randomly sampled latent codes, which result in impressive free viewpoint rendering capabilities despite arguable underlying geometry quality and multi-view consistency. Concurrent work proposed by Pan et al. [39] adds a shading model to enforce multi-lighting constraints during training, however the method shows substantial limitations in terms of photorealism and does not allow for full HDRI relighting.

In this work, we propose a 3D aware generative model with HDRI relighting supervised by adversarial losses. To overcome the limitations of prior arts, we identified and contributed to two main aspects:

Volumetric HDRI Relighting. We propose a novel approach of the volumetric rendering function that naturally supports efficient HDRI relighting. The core idea relies on the intuition that diffuse and specular components can be efficiently accumulated per-pixel when pre-filtered HDR lighting environments are used [18, 44]. This was successfully applied to single image portrait relighting [40], and here we introduce an alternative formulation to allow for volumetric HDRI relighting. Differently from [35, 40, 58] that predict surface normals and calculate the shading with respect to the light sources (for a given HDR environment map), we propose to directly integrate the diffuse and specular components at each 3D location along the ray according to their local surface normal and viewpoint direction. Simultaneously, an albedo image and neural features are accumulated along the 3D ray. Finally, a neural renderer combines the generated outputs to infer the final image.

Supervised Image Decomposition. Though producing impressive rendering quality, the geometry from 3D-aware generators is often incomplete or inaccurate [9, 19]. As a result, the model tends to bias the image quality for highly sampled camera views (e.g. front facing), but starts to show unsatisfactory multi-view consistency and 3D perception, breaking the photorealism when rendered from free-viewpoint camera trajectories. Additionally, high quality geometry is particularly important for relighting since any underlying reflectance models rely on accurate surface normal directions in order to correctly accumulate the light contributions from the HDR environment map.

Similarly, decomposing an image into albedo, diffuse and specular components without explicit supervision could lead to artifacts and inconsistencies, since, without any explicit constrains, the network could encode details in any channel even though it does not follow light transport principles. For instance in Fig. 1, the albedo image generated by previous methods [39] contains clear shading information, whereas the expected albedo (*i.e*. flat lit image) should be closer to ours. At the same time, such supervision is not available for in-the-wild datasets like FFHQ [25].

Motivated by this, and inspired by other works that apply pseudo-groundtruth labels [11] or synthetic renderings [48, 53, 59, 68] for in-the-wild tasks, we propose a data augmentation technique to explicitly supervise the image decomposition in geometry, albedo, diffuse and specular components. In particular, we employ the work of Pandey et al. [40] to generate albedo, geometry, diffuse, specular and relit images for each image of the dataset, and have additional discriminators guide the intrinsic decomposition during the training. This technique alone, however, would guide the generative model to synthesize images that are less photorealistic since their quality upper bound would depend on the specific image decomposition and relighting algorithm used as supervision (*e.g*. [40]). In order to address this, we also add a final discriminator on the original images, which will guide the network towards real photorealism and higher order light transport effects such as specular highlights and subsurface scattering.

We summarize the contributions of this paper: 1) We propose a novel approach to generate HDRI relightable 3D faces with a volumetric rendering framework. 2) Supervised adversary losses are leveraged to increase the geometry and relighting quality, which also improves multi-view consistency. 3) Exhaustive experiments demonstrated the effectiveness of the framework for image synthesis and relighting.

2. Related Work

2D Image Generation. Generating convincing renderings of humans is a very active trend in the field of neural rendering [55]. Here, we consider works that rely on a generative adversarial framework [17] to synthesize photorealistic portraits. High quality results have been demonstrated by multiple early works [15, 34, 62] and since the groundbreaking work of StyleGAN [25], the community has made tremendous progress in synthesizing photorealistic and high resolution images [6, 12, 24, 26] with methods focusing on addressing most of the common issues with GANs including stability [26], resolution [6] and aliasing [24]. These

approaches generate impressive photorealistic images, but results typically lack free-viewpoint rendering and/or multi-view consistency.

3D Aware Generation. Many recent approaches incorporated the use of geometry and its multi-view consistency to allow for 3D aware synthesis. [1, 9, 10, 19, 27, 36, 37, 43, 66, 67]. Past works rely on voxels [16, 36, 43, 67], meshes [51], face models [7] or shape primitives [27] as the 3D representation for image generation, but the majority have been limited to low resolution image generation. Inspired by the success of NeRF [32], methods [9, 37, 49] adopt implicit volumetric rendering framework, and require only unconstrained images for 3D GAN training, but these architectures are computational consuming, which limit the training for high-resolution image generation. Concurrently, StyleNeRF [19], CIPS-3D [66], StyleSDF [38] adopt the two-stage rendering strategy to reduce the computation for high-resolution image generation. EG3D [8] introduces triplane representation for fast and scalable rendering, and GRAM [14] proposes to render radiance manifolds first to produce high quality images. However, all these concurrent methods lack controllable relighting capabilities.

Relightable Generative Models. Relightable NeRF models [4, 5, 63, 65] have shown that full image decomposition is possible when explicit multi-view imagery is provided as supervision. As for generative networks, the concurrent work of Pan et al. [39] is, to the best of our knowledge, the first at enabling relightability into generative model in a volumetrics 3D framework. The method enforces both multi-view and multi-lighting consistency to allow controllable viewpoint and illumination. This approach, however, adopts a simplified Lambertian model and only supports one specific light direction at the time and extending it to full HDR relighting is computationally prohibitive. HeadNeRF [21] propose a NeRF-based parametric head model which can control the illumination by adjusting latent code. However, limited by insufficient coverage of illumination in their dataset, the method cannot control the image shading like continuously moving light source position. Also, they cannot explicitly control the illumination by adjusting latent code, while it is achievable in HDRI relighting by given a desired HDR map.

Intrinsic Image Decomposition. Decomposing an image into albedo, geometry and reflectance components has achieved using model-fitting techniques [2, 29] and deep learning based approaches [22, 30, 45, 60] that attempt at inferring image properties from one or multiple images. Very recently, state-of-art image based portrait relighting methods [35, 40, 52, 58] have shown impressive results by predicting explicit surface normals, albedo and shading information to formulate the interaction between light sources and geometry. These approaches usually rely on a specific shading model (e.g. Phong) and a neural renderer to synthesize the final image.

Our Approach. In contrast, we propose a volumetric generative model that supports full HDR relighting. We show how we can efficiently aggregate albedo, diffuse and specular components within the 3D volume. Thanks to the explicit supervision in our adversarial losses, we demonstrate that the method can perform such a full image component decomposition for novel face identities, starting from a randomly sampled latent code.

3. VoLux-GAN Framework

In this section, we introduce our neural generator that produces novel faces that can be rendered at free camera viewpoints and relit under an arbitrary HDR environment light map. Our method starts from a neural implicit field that takes a randomly sampled latent vector as input and produces an albedo, volume density, and reflectance properties for any queried 3D location. These outputs are then aggregated via volumetric rendering to produce low resolution albedo, diffuse shading, specular shading, and neural feature maps. These intermediate outputs are then upsampled to high resolution and fed into a neural renderer to produce relit images. The overall framework is depicted in Figure 2.

3.1. Preliminaries: Neural Volumetric Rendering.

To aid the reader, we first briefly introduce the neural volumetric rendering framework originally presented in Mildenhall *et al*. [32]. There, the 3D appearance of an object of interest is encoded into a neural implicit field implemented using a multilayer perceptron (MLP), which takes a 3D coordinate $x \in R^3$ (mapped through a sinusoidal function based positional encoding [32, 54]) and viewing direction $\mathbf{d} \in S^2$ as inputs and outputs a volume density $\sigma \in R^+$ and view-dependent color $\mathbf{c} \in R^3$. To render an image, the pixel color $\mathbf{C}$ is accumulated along each camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ as

$$\mathbf{C}(\mathbf{r}, \mathbf{d}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt, \qquad (1)$$

where $T(t) = exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds)$ and bounds t_n and t_f. Compared to surface based rendering, volumetric rendering more naturally handles translucent materials and regions with complex geometry such as thin structures.

3.2. Generative Neural Implicit Intrinsic Field.

Similar to other state-of-the-art 3D-aware generators [9, 19, 39], we train a MLP-based neural implicit field conditioned on a latent code z sampled from a Gaussian distribution $N(0, I)^d$ and extend it to support HDRI relighting. We

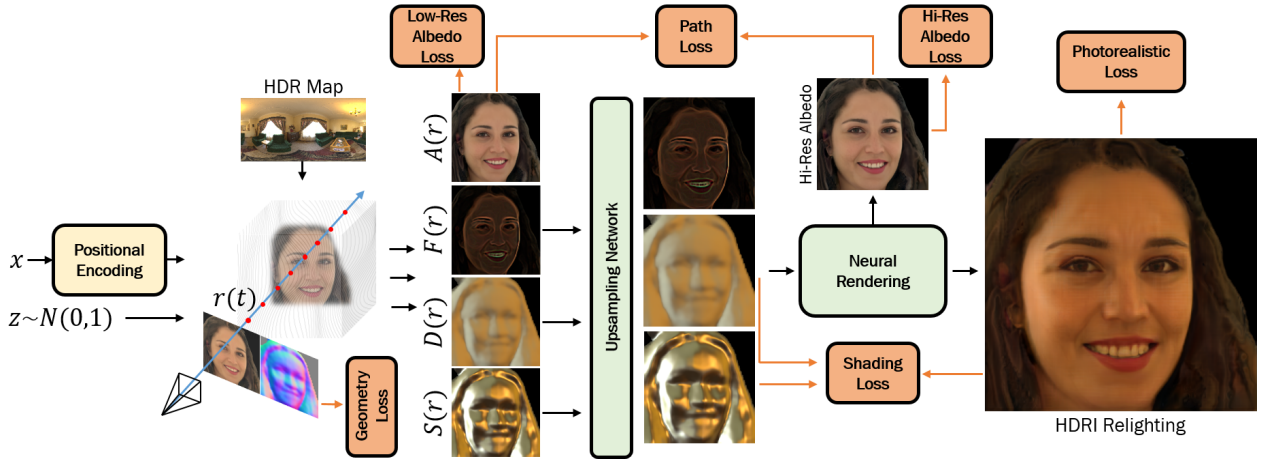

Figure 2. VoLux-GAN Framework. Starting from a latent code we can efficiently accumulate albedo $A(\mathbf{r})$, surface normals $N(\mathbf{r})$, diffuse $D(\mathbf{r})$, specular components $S(\mathbf{r})$, and a feature map $F(\mathbf{r})$ along the 3D ray $\mathbf{r}(t)$ for any given HDR map. An upsampling strategy and a neural renderer synthesize the final relit image.

adopt a Phong shading model [42], where the illumination of each point is determined by albedo, diffuse, and specular component. Therefore, instead of having the network predict per-point radiance and directly obtaining a color image (via Eq. 1), our network produces per-point albedo (α), density (σ) and reflectance properties from separate MLP heads. The normal directions are obtained via the spatial derivative of the density field, which are used together with HDR illumination to compute diffuse and specular shading. Similar to [40], rather than explicitly using the Phong model for the final rendering, we feed the albedo, diffuse and specular components to a lightweight neural renderer, which can also model higher order light transport effects.

Efficient Shading Computation. Concurrent work [39] assumes Lambertian shading from a single light source. Extending this to support full HDR illumination would require the integration of the shading contribution from multiple positional lights, making the approach computationally prohibitive, especially when performed at training time for millions of images. Inspired by the success of recent image based portrait relighting work [40], we adopt a method designed for real-time shading rendering under HDR illumination [18, 44]. The core idea is to approximate the diffuse and specular components using a preconvolved HDRI map. Specifically, we first preconvolve the given HDRI map ($\mathbf{H}$) into light maps ($L_{n_i}, i = 1, 2, \cdots, N$) with cosine lobe functions corresponding to a set of pre-selected Phong specular exponents ($n_i, i = 1, 2, \cdots, N$) [33]. The diffuse shading D is the first light map (*i.e.* $n = 1$ above) following the surface normal direction, and the specular shading is defined as a linear combination of all light maps indexed by the reflection direction. In order to capture possible diverse material properties of the face, we let the network estimate the blending weights (ω_i) with another MLP branch, which are then used for the specular component S.

Volumetric Shading Rendering. Typically, a reflectance model is defined on a surface [42] and relighting methods [35, 40, 58] explicitly estimate surface normals from a single image. Here, we propose a volumetric formulation to compute albedo, diffuse and view-dependent specular shading maps as:

$$
\begin{aligned}
A(\mathbf{r}) &= \int_{t_n}^{t_f} T(t)\sigma\left(\mathbf{r}(t)\right)\alpha(\mathbf{r}(t))dt \\
D(\mathbf{r}) &= \int_{t_n}^{t_f} T(t)\sigma\left(\mathbf{r}(t)\right)L_{n=1}(\mathbf{n}\left(t\right))dt \\
S(\mathbf{r}) &= \int_{t_n}^{t_f} T(t)\sigma\left(\mathbf{r}(t)\right)\sum_{i}^{N}\omega_i L_{n_i}(\mathbf{n}(t),\mathbf{d})dt \\
F(\mathbf{r}) &= \int_{t_n}^{t_f} T(t)\sigma\left(\mathbf{r}(t)\right)f(\mathbf{r}(t))dt
\end{aligned}
\tag{2}
$$

where $\mathbf{n}(t)$ is the normal direction estimated via $\nabla\sigma(\mathbf{r}(t))$, $L_{n=1}(\mathbf{n}\left(t\right))$ is the diffuse light map indexed by the normal direction $\mathbf{n}(t)$, and $L_{n_i}(\mathbf{n}(t),\mathbf{d})$ is the specular component n_i indexed by the inbound reflection direction depending on the local normal and viewing direction $\mathbf{d}$. Finally, α, σ, ω, and a per-location feature f are the network outputs conditioned on the sampled latent code z. We restrict the albedo to be view and lighting independent and encourage multi-view consistency. Note that in addition to rendering components such as albedo, diffuse and specular components, we let our network accumulate additional features $F(\mathbf{r})$, so that it can capture high frequency details and material properties in an unsupervised fashion.

Volumetric Model Network Architecture We extend the architecture from concurrent work proposed by Gu *et al.*

[19] for our neural implicit field. Rather than explicitly use the low resolution albedo $A(\mathbf{r})$ following Eq. 2, our network produces a feature vector $f(\mathbf{r}(t)) \in R^{256}$ via 6 fully-connected layers from the positional encoding on the 3D coordinates. A linear-layer is attached to the output of the 4-th layer to produce the volume density, and an additional two-layer MLP is attached to 6-th layer to produces the albedo and reflectance properties. Diffuse component D and Specular Component S are estimated following Eq. 2, where the blending weights ω_i are estimated by the network.

Neural Rendering Network To reduce the memory consumption and computation cost, we render albedo, diffuse, and specular shading in low resolution and upsample them to high resolution for relighting. The specific low and high resolutions depend on the dataset used and details can be found in the Section 4. To generate the high resolution albedo, we upsample the feature map $F(\mathbf{r})$ and enforce it's first 3 channels to correspond to the albedo image, similar to some other works in the literature [31, 56]. Each upsampling unit consists of two 1×1 convolutions modulated by the latent code z, a pixelshuffle upsampler [50] and a BlurPool [64] with stride 1. The low resolution albedo $A(\mathbf{r})$ is still used to enforce consistency with the upsampled high resolution albedo (see Section 3.3). For shading maps, we directly apply bilinear upsampling.

Finally, a relighting network takes as input the albedo map A, the diffuse map D, the specular component map S and the features F and generates the final I_{relit} image. The architecture of Relighting Network is a shallow U-Net [46].

3.3. Supervised Adversarial Training

Here we introduce the scheme to train our pipeline from a collection of unconstrained in-the-wild images. While it is possible to train the full pipeline with a single adversarial loss on the relit image, we found empirically that adding additional supervision on intermediate outputs significantly improves the training convergence and rendering quality.

Pesudo Ground Truth Generation. Large scale in-the-wild images provides great data diversity, which is critical for training a generator. However, the groundtruth labels for geometry and shading are usually missing. In our case, we are particularly interested in having "real examples" of the albedo and geometry to supervise our method. To this end, we resort to the state-of-the-art image based relighting algorithm, Total Relighting [40], to produce pseudo ground truth albedo and normals and to also further increase the data diversity. Specifically, for each image in our training set, we randomly select an HDRI map from a collection of 400 maps sourced from public repository [61], apply a random rotation, and run Total Relighting to generate the albedo, surface normal and a relit image with the associated light maps (diffuse and specular components). Example images from the CelebA dataset [28] augmented with this technique are shown in Figure 3.

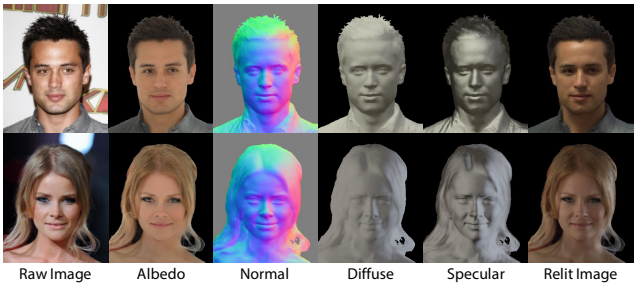

Figure 3. Relighting augmentation on CelebA [28]. We run Total Relighting [40] to generate albedo, normal, shading, and relit images, which supervise the training via adversarial losses.

Albedo Adversarial Loss $\mathcal{L}_A$: $D_A(A(\mathbf{r})) + D_A(A_{hi-res})$ We supervise the output albedo images in both low and high resolution with adversarial loss using the pseudo ground truth generated with [40]. A standard non-saturating logistic GAN loss with R1 penalty is applied to train the generator and discriminator. The discriminator architecture D_* for all the losses follows the one proposed in [26].

Geometry Adversarial Loss $\mathcal{L}_G$: $D_G(\nabla\sigma(\mathbf{r}(t)))$ We also supervise the geometry as it is crucial for multi-view consistent rendering and relighting realism. While the density σ is the immediate output from the network that measures the geometry, we find it is more convenient to supervise the surface normals computed via $\nabla\sigma(\mathbf{r}(t))$. Therefore, we add an adversarial loss between the volumetric rendered normal from the derivative of the density and the pseudo ground truth normal obtained from [40].

Shading Adversarial Loss $\mathcal{L}_S$: $D_S(D(\mathbf{r}), S(\mathbf{r}), I_{relit})$ To enforce that the Relight Network faithfully integrates shading with albedo, we apply a conditional adversarial loss on the relit image. This is achieved by adding a discriminator D_S that takes the concatenation of the relit image I_{relit}, diffuse map $D(\mathbf{r})$ and specular map $S(\mathbf{r})$ as the inputs and discriminate if the group is fake, *i.e.* from our model, or true, *i.e.* from [40]. The training gradients are only allowed to back-propagate to the relit image but not the other inputs (*i.e.* set to zero) as they are the data to be conditioned on.

Photorealistic Adversarial Loss $\mathcal{L}_P$: $D_P(I_{relit})$ A downside of the Shading Adversarial Loss is that the model performance is upper-bounded by the specific algorithm used to generate pseudo-groundtruth labels, in our case [40]. As a result, inaccuracies in the relighting examples, *e.g.* overly smoothed shading and lack of specular highlights, may affect our rendering quality. To enhance the photorealism, we add one final adversarial loss directly on the generated relit images with the original images from the dataset.

Path Loss $\mathcal{L}_{path}$: $\ell_1(A(\mathbf{r}), A_{hi-res})$ Following StyleNeRF [19], we add a path loss to ensure the consistency between the albedo maps in low and high resolutions. Specifically,

we downsample the high resolution to the low resolution, and add a per-pixel ℓ_1 loss.

The final loss function is a weighted sum of all above mentioned terms: $\mathcal{L} = \lambda_1\mathcal{L}_A + \lambda_2\mathcal{L}_G + \lambda_3\mathcal{L}_S + \lambda_4\mathcal{L}_P + \lambda_5\mathcal{L}_{path}$, where for our experiments we empirically determined these weights to be $1.0, 0.5, 0.25, 0.75, 0.5$.

3.4. Implementation Details

To ensure stable training, we first train the neural implicit field and the upsampling network for high quality albedo and geometry. We only enable albedo and geometry adversarial loss and adopt progressive growing training strategy [23] with the path loss to train the upsampling network. Once the network converges, we enable all the loss terms and train the whole network end-to-end. For the optimization, we use the Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.99$. The batch size is set as 24 and the learning rates for the generator and the discriminator are set empirically to 0.0022 and 0.0025 respectively. We train the VoLux-GAN model for 1 million iterations with albedo and geometry adversarial losses, then it is trained for additional 500k iterations to generate the relit images.

4. Experiments

In this section, we compare our rendering quality and relighting performance with state-of-the-art methods. We also provide ablation study showing the contribution of major system design choice to the final performance.

Datasets. We train our model on CelebA dataset [28] which is widely used for such comparisons, and on the FFHQ [25] where a comparison of high resolution results can be made. On CelebA, our model produces volumetric renderings at 64×64 and final outputs at 128×128. On FFHQ, the volumetric renderings and final resolution are 64×64 and 256×256 respectively.

Baseline Methods. We show qualitative and quantitative comparison with ShadeGAN [39] since, to the best of our knowledge, it is the only 3D-aware generator that supports relighting. In addition, we consider an alternative strong baseline where we use pi-GAN [9] to render multi-view images and then run a single image based portrait relighting method [40] for HDR relighting.

Metrics. Many evaluation metrics relying on perception features have been proposed to measure the rendering quality [3, 20]. While these metrics indeed measure the similarity between two collections of images, they are very sensitive to implementation details such as training image resolution, image post-processing *e.g.* cropping, or the choice of training dataset, as has been shown in literature [41]. As a result, these metrics are not suitable to evaluate our model since our pipeline is not trained directly on publicly available datasets but on a specifically tailored augmented dataset for good rendering and relighting performance.

Instead, we evaluate our pipeline and other methods by measuring the perceptual impact of view-synthesis and relighting on a subject's identity. Specifically, we use a similarity metric based on the embedding space of a state-of-the-art face recognition network [13]. This stability metric indicates how well the subject identity is preserved when we synthesize novel views and novel light renderings for a synthesized face.

4.1. Relightable Face Generation

In this section we demonstrate the capabilities of our framework. In Fig. 4, we show one subject randomly sampled from latent space z $N(0, I)$ trained on the FFHQ dataset. The first row shows the faces rendered at different camera poses. Our network successfully renders consistent faces even under a large yaw angle (*e.g.* $45°$) thanks to better geometry supervised by the geometry adversarial loss. The second row shows the same subject rendered under a rotating HDR map. Note how the specularities and shading on the face respond correctly to the HDR environment maps.

Our latent space also supports interpolation. In Fig. 5, we linearly interpolate between two randomly sampled latent codes, and show relit images of each subject under three HDR lighting conditions. As seen, the appearance of the subject transitions smoothly and the intermediate identities are successfully relit. Note also the consistent relighting, where view dependent effects and specularities are successfully transferred between different latent codes.

4.2. Comparisons with State-of-the-Art

We compare to ShadeGAN [39], pi-GAN [9] coupled to an image based relighting [40], and show the qualitative results in Fig. 6. For a fair comparison in terms of image resolution, we trained our model on CelebA, like ShadeGAN and pi-GAN. In each row, we show the albedo map and color images rendered under two different lightings from three camera viewpoints.

Our method produces significantly better albedo than ShadeGAN thanks to our supervised albedo adversarial loss. Moreover, our results contain more high frequency specular components and can respond to more diverse global illumination. pi-GAN coupled with [40] can produce plausible relighting results but occasionally with inconsistent shading (*e.g.* in the 3rd row middle, the cheek is dark in one view but bright in another). In contrast, our results show more consistency across views and lights thanks to the proposed volumetric relighting formulation, which is also reflected by the quantitative metric below.

Figure 4. Our synthesized images under rotating camera or rotating lighting. Note the relighting consistency and view-dependent effects.

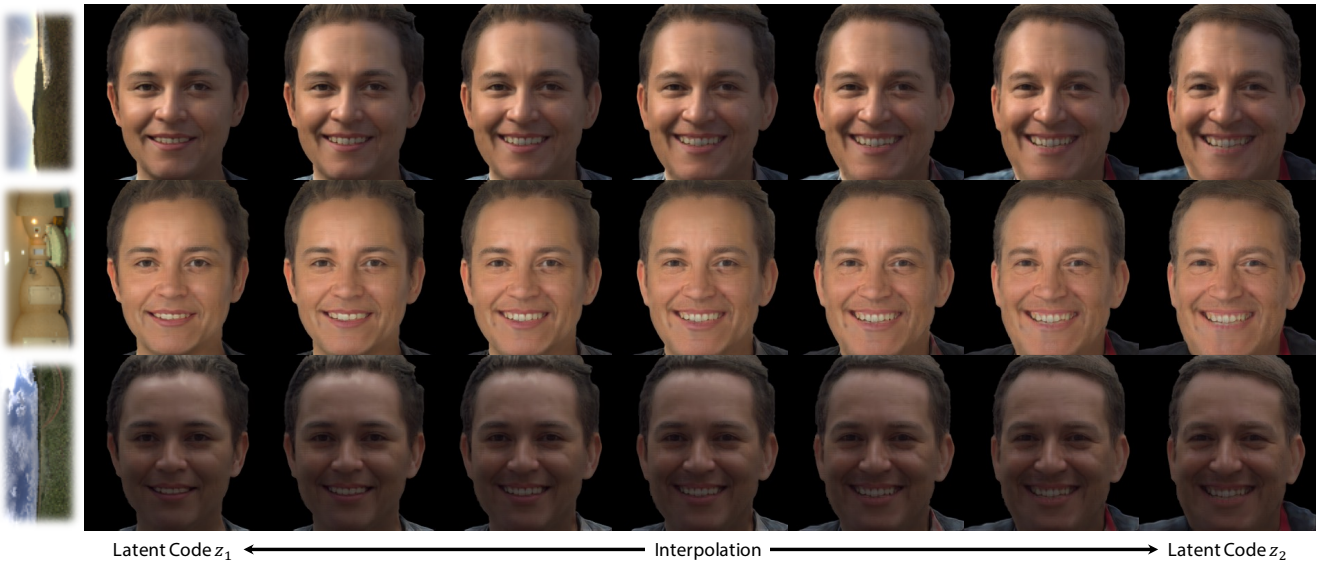

Figure 5. Our result with under interpolated latent code and relighting results. Note the smooth transition across users where relighting effects are correctly transferred.

4.2.1 Quantitative Results

We evaluate our method with quantitative metrics. The goal of the following experiments is to demonstrate that our method is able to synthesize images that are consistent across views in terms of geometry and relighting.

Geometry Consistency. To demonstrate the geometry consistency, we render a fixed random latent code to multiple views. We then compute the similarity score [13] of yaw-posed renderings with the frontal facing rendering, and average it over 100 such randomly sampled latent codes. We compute the score for both relit images and the high-res albedo images. We also show the score computed with the same scheme for ShadeGAN [39] and the baseline of pi-GAN [9] + portrait relighting [40]. The results are showed in Table 1. Note how our method consistently outperforms the other state-of-art approaches, demonstrating better multi-view consistency for each identity on generated albedo and relit images.

Relighting Consistency. Similarly, we also evaluate the stability of our relighting. Following the geometry consistency experiment, we use the embedding space of a face recognition network [13] to generate the identity similarity score between the albedo and 3 renderings under different environment maps (as shown in Fig. 5). We report an average score over 100 randomly sampled latent codes. A stable relighting method should give a high similarity score, since relighting does not change the identity.

The results are reported in Table 2, showing that our approach is also able to generate relit images that are more consistent with the original albedo identity. At the same time our relit images look more photorealistic as shown in Figure 6, where we better capture higher order light trans-

Method	**Relit image identity similarity↑**					**Albedo image identity similarity↑**				
	-0.5 rad	-0.25 rad	0 rad	0.25 rad	0.5 rad	-0.5 rad	-0.25 rad	0 rad	0.25 rad	0.5 rad
ShadeGAN [39]	0.4814	0.7513	-	0.7628	0.4997	0.4818	0.7582	-	0.7702	0.5091
pi-GAN [9] + TR [40]	0.5215	0.7472	-	0.7419	0.4981	0.5135	0.7378	-	0.7438	0.4898
VoLux-GAN+Surface Relighting	0.4471	0.7065	-	0.7388	0.4959	0.5531	0.7611	-	0.7796	0.5585
VoLux-GAN - $\mathcal{L}_P$	0.4827	0.6487	-	0.6721	0.5156	0.5467	0.7809	-	0.8151	0.5901
VoLux-GAN - $\mathcal{L}_S$	0.4071	0.6996	-	0.7788	0.4718	0.4886	0.7292	-	0.8095	0.5581
VoLux-GAN - $\mathcal{L}_G$	0.4776	0.7182	-	0.7564	0.5015	0.4652	0.7278	-	**0.8099**	0.5398
VoLux-GAN	**0.6064**	**0.7736**	-	**0.7997**	**0.5985**	**0.6389**	**0.7919**	-	0.7863	**0.6162**

Table 1. Identity consistency across camera poses around the yaw axis. The scores indicate the similarity calculated as the dot product between normalized embeddings from a state-of-the-art face recognition network [13] (higher is better). While our method performs comparably or marginally better at small view changes, we significantly outperform the state-of-the-art at more extreme viewpoints.

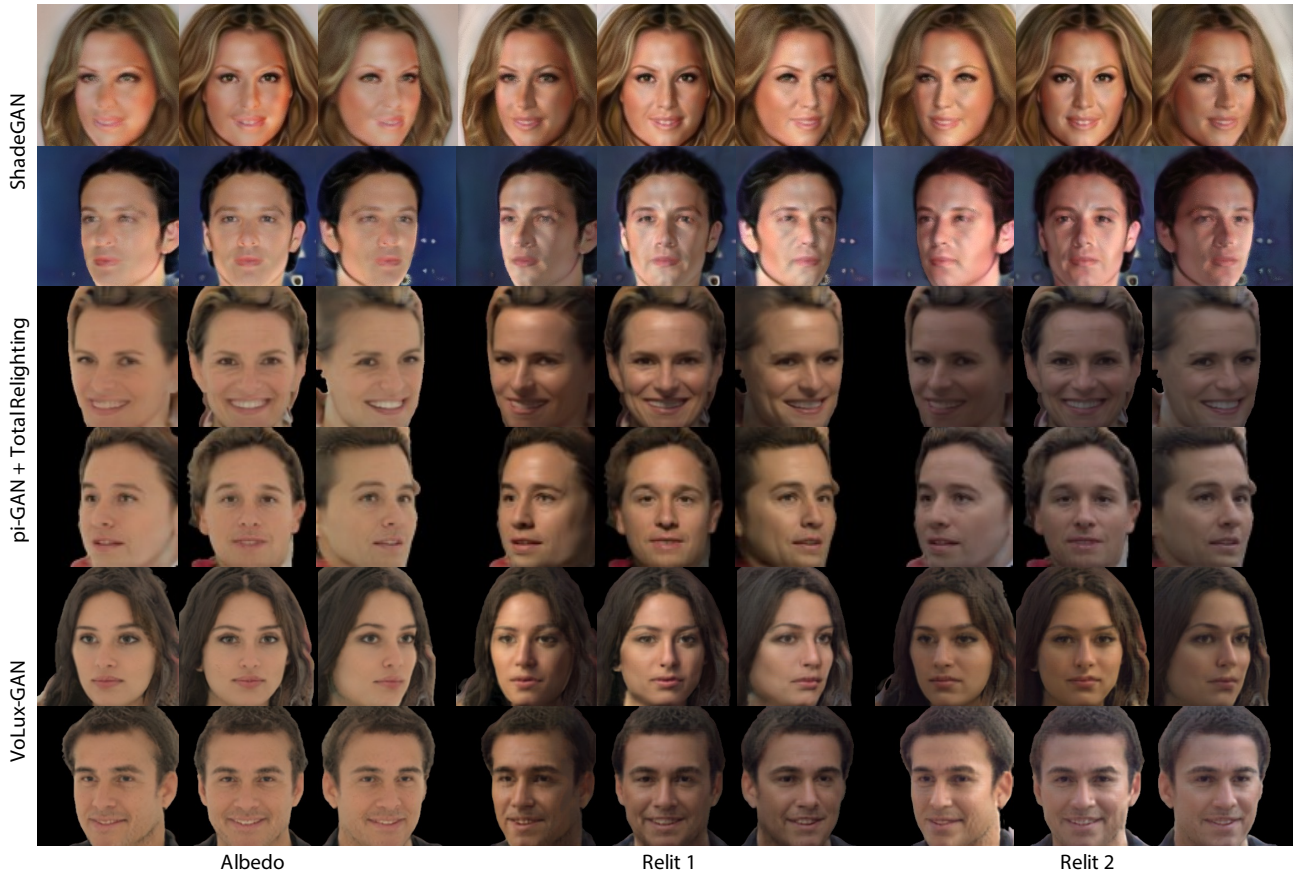

Figure 6. Qualitative comparisons on CelebA with [39] and pi-GAN [9] + portrait relighting [40]: note how our method produces more consistent albedo and relighting results across multiple views.

port effects.

Ablation Study. We report an ablation study of the loss functions used for training out pipeline and Table 2. As demonstrated by these quantitative results, the full framework and the proposed supervised adversarial losses are all contributing to the final rendering quality. In particular, we show that removing the shading loss $\mathcal{L}_S$, or geometry loss $\mathcal{L}_G$ or photorealistic $\mathcal{L}_P$ all lead to lower metrics. Similarly, when we first accumulate the surface geometry and then perform image based relighting, the results are unsatisfactory (see comparison with VoLux-GAN+Surface Relighting in Table 1 and Table 2).

Method	HDR map 1	HDR map 2	HDR map 3
ShadeGAN [39]	0.5806	0.6486	0.6559
pi-GAN [9] + TR [40]	0.7014	0.8796	0.7677
VoLux-GAN+Surface Relighting	0.5510	0.5548	0.5349
VoLux-GAN - $\mathcal{L}_P$	0.4132	0.4471	0.4712
VoLux-GAN - $\mathcal{L}_S$	**0.8917**	0.8846	0.7387
VoLux-GAN - $\mathcal{L}_G$	0.5331	0.5864	0.6158
VoLux-GAN	0.7600	**0.8900**	**0.8082**

Table 2. Relighting consistency. Our method is able to better preserve the original identity (albedo) under different illuminations. Note that ShadeGAN is evaluated over three different positional light sources instead of HDR maps.

5. Discussion

We proposed a generative model of face images, that internally leverages a volumetric representation to facilitate multi-view generation and full HDRI relighting. Of particular note is that we have shown how to efficiently perform the aggregation of albedo, specular and diffuse components that helps to preserve the identity. Furthermore, a proposed supervised adversarial framework guides the network to generate the right intrinsic properties of faces. Our results prove the effectiveness of the approach for synthesizing novel identities. Future work could explore the use of semantic information to allow for expression control similar to StyleGAN [25]. While far from the intent of this work, we do recognize that generative models could be misused to synthesize fake content (see [57] for an exhaustive survey). We believe that in order to address this, the community should prioritize open-sourcing pre-trained models to encourage the development of forgery detection methods. Great steps in that direction have been made thanks to the availability of datasets such as FaceForensics++ [47]. Additionally, in order to mitigate misuses, researchers could put more emphasis on the adversarial models (*i.e*. discriminators) and make them publicly available when releasing a generative model. Enforcing higher importance to the discriminator loss, while fixing the generator could provide an effective method to detect misuse of the specific generative model.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Andreas Geiger, and Carsten Rother. Geometric image synthesis. *ACCV*, 2018. 3

[2] Jonathan T. Barron and Jitendra Malik. Shape, Illumination, and Reflectance From Shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 3

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. *ICCV*, 2020. 3

[5] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, 2021. 3

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[7] Marcel C. Buehler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[8] 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 *arXiv*, 2021. 3

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 1, 2, 3, 6, 7, 8, 9

[10] Xuelin Chen, Daniel Cohen-Or, Baoquan Chen, and Niloy J. Mitra. Towards a neural graphics pipeline for controllable image generation. *Computer Graphics Forum*, 40(2), 2021. 3

[11] George Chogovadze, Rémi Pautrat, and Marc Pollefeys. Controllable data augmentation through deep relighting, 2021. 2

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains, 2020. 2

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6, 7, 8

[14] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *arXiv preprint arXiv:2112.08867*, 2021. 3

[15] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks, 2017. 2

[16] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 3

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

[18] Ned Greene. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications*, 6(11):21–29, 1986. 2, 4

[19] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware genera-

tor for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2, 3, 5

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[21] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. *arXiv preprint arXiv:2112.05637*, 2021. 3

[22] Yoshihiro Kanamori and Yuki Endo. Relighting Humans: Occlusion-Aware Inverse Rendering for Full-Body Human Images. *ACM Transactions Graphics (Proc. SIGGRAPH Asia)*, 2018. 3

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 2

[25] 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, 6, 9

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 2, 5

[27] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5, 6

[29] Abhimitra Meka, Gereon Fox, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. Live User-Guided Intrinsic Video for Static Scene. *IEEE Transactions on Visualization and Computer Graphics*, 2017. 3

[30] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. LIME: Live Intrinsic Material Estimation. In *Proc. Computer Vision and Pattern Recognition*, 2018. 3

[31] Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe LeGendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. Deep relightable textures: Volumetric performance capture with neural rendering. *ACM Transactions on Graphics*, 2020. 5

[32] 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*, pages 405–421. Springer, 2020. 2, 3

[33] Gene S Miller and CR Hoffman. Illumination and reflection maps. In *ACM SIGGRAPH*, 1984. 4

[34] Gonçalo Mordido, Haojin Yang, and Christoph Meinel. Dropout-gan: Learning from a dynamic ensemble of discriminators, 2018. 2

[35] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M. Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, 2020. 2, 3, 4

[36] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *The IEEE International Conference on Computer Vision (ICCV)*, Nov 2019. 3

[37] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3

[38] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *arXiv e-prints*, pages arXiv–2112, 2021. 3

[39] 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 *NeurIPS*, 2021. 1, 2, 3, 4, 6, 7, 8, 9

[40] 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. 2, 3, 4, 5, 6, 7, 8, 9

[41] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 6

[42] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 4

[43] Thu Nguyen Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy J Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *NeurIPS 2020: Conference on Neural Information Processing Systems*, 2020. 3

[44] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 2, 4

[45] Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. Image Based Relighting Using Neural Networks. *ACM Transactions on Graphics*, 2015. 3

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[47] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV 2019*, 2019. 9

[48] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis, 2020. 3

[50] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[51] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 3

[52] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation, 2021. 3

[53] Feitong Tan, Danhang Tang, Dou Mingsong, Guo Kaiwen, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, Ping Tan, and Yinda Zhang. Humangps: Geodesic preserving feature for dense human correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[54] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. 2021. 3

[55] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B Goldman, and Michael Zollhoefer. State of the art on neural rendering. In *Eurographics*, 2020. 2

[56] Justus Thies, Michael Zollhöfer, and Matthias Niessner. Deferred neural rendering: Image synthesis using neural textures. *SIGGRAPH and ACM TOG*, 2019. 5

[57] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 2020. 9

[58] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM SIGGRAPH Asia and Transactions on Graphics*, 2020. 2, 3, 4

[59] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone, 2021. 2

[60] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics*, 2018. 3

[61] Greg Zaal, Sergej Majboroda, and Andreas Mischok. Hdri haven. https://www.hdrihaven.com/, 2020. Accessed: 2021-11-13. 5

[62] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, 2019. 2

[63] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 3

[64] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 5

[65] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv preprint arXiv:2106.01970*, 2021. 3

[66] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3

[67] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in Neural Information Processing Systems*, 2018. 3

[68] Tyler Zhu, Per Karlsson, and Christoph Bregler. Simpose: Effectively learning densepose and surface normals of people from simulated data. In *European Conference on Computer Vision*, pages 225–242. Springer, 2020. 2