Lite2Relight: 3D-aware Single Image Portrait Relighting

Pramod Rao
prao@mpi-inf.mpg.de
MPI for Informatics, SIC &
VIA Research Center
Saarbrücken, Germany

Gereon Fox
gfox@mpi-inf.mpg.de
MPI for Informatics
Saarbrücken, Germany

Abhimitra Meka
abhim@google.com
Google Inc.
San Fransisco, USA

Mallikarjun B R
mbr@mpi-inf.mpg.de
MPI for Informatics
Saarbrücken, Germany

Fangneng Zhan
fzhan@mpi-inf.mpg.de
MPI for Informatics
Saarbrücken, Germany

Tim Weyrich
tim.weyrich@fau.de
Friedrich-Alexander-Universität
Erlangen-Nürnberg (FAU)
Nürnberg, Germany

Bernd Bickel
bernd.bickel@ist.ac.at
IST Austria
Klosterneuburg, Austria
ETH Zürich
Zürich, Switzerland

Hanspeter Pfister
pfister@g.harvard.edu
Harvard University
Cambridge, USA

Wojciech Matusik
wojciech@csail.mit.edu
MIT
Cambridge, USA

Mohamed Elgharib
elgharib@mpi-inf.mpg.de
MPI for Informatics
Saarbrücken, Germany

Christian Theobalt
theobalt@mpi-inf.mpg.de
MPI for Informatics, SIC &
VIA Research Center
Saarbrücken, Germany

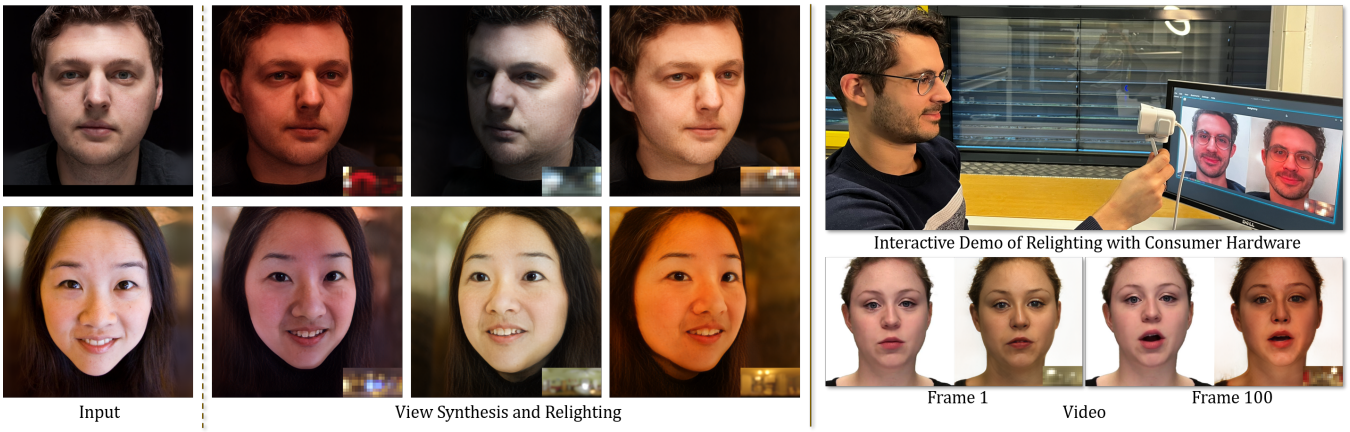

Figure 1: We present Lite2Relight, a method that can relight monocular portrait images given HDRI environment maps. Our method demonstrates strong generalization to in-the-wild images, maintains 3D consistent pose synthesis of the subjects and performs physically accurate relighting. Moreover, courtesy of our lightweight architecture, Lite2Relight can relight subjects captured by a live webcam at interactive rates. Image credits to Flickr.

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657470

ABSTRACT

Achieving photorealistic 3D view synthesis and relighting of human portraits is pivotal for advancing AR/VR applications. Existing methodologies in portrait relighting demonstrate substantial limitations in terms of generalization and 3D consistency, coupled with inaccuracies in physically realistic lighting and identity preservation. Furthermore, personalization from a single view is difficult

to achieve and often requires multiview images during the testing phase or involves slow optimization processes. This paper introduces Lite2Relight , a novel technique that can predict 3D consistent head poses of portraits while performing physically plausible light editing at interactive speed. Our method uniquely extends the generative capabilities and efficient volumetric representation of EG3D, leveraging a lightstage dataset to implicitly disentangle face reflectance and perform relighting under target HDRI environment maps. By utilizing a pre-trained geometry-aware encoder and a feature alignment module, we map input images into a relightable 3D space, enhancing them with a strong face geometry and reflectance prior. Through extensive quantitative and qualitative evaluations, we show that our method outperforms the state-of-the-art methods in terms of efficacy, photorealism, and practical application. This includes producing 3D-consistent results of the full head, including hair, eyes, and expressions. Lite2Relight paves the way for large-scale adoption of photorealistic portrait editing in various domains, offering a robust, interactive solution to a previously constrained problem.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; *Reflectance modeling; Volumetric models; Image-based rendering*.

KEYWORDS

Faces, Relighting, Volumetric Representation, Generative Modeling

ACM Reference Format:
Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. 2024. Lite2Relight: 3D-aware Single Image Portrait Relighting . In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3641519.3657470

1 INTRODUCTION

Photorealistic editing and compositing of human portrait images is a technical challenge underlying various graphics applications: computationally enhanced photography [Sun et al. 2019], content generation [Fried et al. 2020] and immersive telepresence [Cao et al. 2022] are some examples. Large-scale adoption of such applications is limited by the challenge of solving this interactively with minimal computing power using sparse data input, usually an image or video from a single camera. The underlying technical challenge consists of modeling the very diverse range of 3D geometry and reflectance of human heads from such sparse data and achieving a perceivable high degree of photorealism.

An interactive photorealistic 3D portrait editing and relighting solution that can generalize to unseen subjects from a single in-the-wild input image does not exist due to multiple challenges. Solving this under-constrained optimization problem with sparse data, like a single 2D image, requires strong priors on the image formation model. Volumetric generative models of faces learnt from large image datasets [Bühler et al. 2023; Chan et al. 2020, 2021; Deng et al. 2022; Karras et al. 2020; Tewari et al. 2022a] have been successful in modeling high-frequency detail including skin pores and strand-level semi-transparent hair and provide a rich latent space capable of modeling any arbitrary novel identity. They also enable semantic editing such as adding accessories (like glasses) or adding or removing wrinkles etc. However, these methods do not natively allow for relighting, which is required for accurately compositing the face into different backgrounds or environments. Some extensions [Deng et al. 2023; Jiang et al. 2023; Pan et al. 2022, 2021; Ranjan et al. 2023] aim to disentangle the geometry and reflectance of the face from the environmental lighting by implicitly learning a subspace of intrinsic components like albedo, specularity, and normals, but, do not model the light transport accurately enough with ground truth disentangled data. In-the-wild images have a low dynamic range, non-linear photometric effects due to saturation and colored lighting, and different camera response curves. Hence, in-the-wild images are incapable of disentangling the accurate dynamic range of face reflectance, leading to dampened and inaccurate relighting results. However, these methods lack groundtruth supervision during training, thus, despite following physically-based rendering principles, they are physically inaccurate. This is indicated in pink in Sec. 1.

Specialized hardware with controlled lighting, such as lightstages [Debevec et al. 2000], have been used for physically accurate disentanglement of geometry and reflectance. Particularly, 2D image-based rendering (IBR) using such data has been efficiently used to achieve physically-accurate relighting of portrait images [Meka et al. 2019a; Pandey et al. 2021; Yeh et al. 2022]. Several methods have attempted to learn a 3D generative model of the face geometry and reflectance using lightstage datasets. Therefore, the inductive bias of the trained networks of the above methods is learnt from a physically accurate dataset. NeLF [Sun et al. 2021] relies on synthetically rendered "virtual" lightstage images for training, but at inference time suffers from domain gap issues with real data, and still requires at least 5 input views captured for the same time frame, making it impractical for casual capture applications. VoLux-GAN [Tan et al. 2022] and VoRF [Rao et al. 2023, 2022] use real lightstage datasets to learn a 3D generative model of faces that can be relit under target environment maps. While VoRF does so by decomposing the input image into an "OLAT" reflectance basis, VoLux-GAN decomposes the image into intrinsic components and renders "shading" images under a target environment map that are passed through a neural rendering network to generate the relit outputs. While both methods enable photorealistic and consistent view-synthesis and physically accurate relighting, they have limited generative capacity to convert in-the-wild input images into a reflectance basis, and are weighed down at test time by the additional step of inversion and finetuning [Roich et al. 2021] for the given image, thus, preventing their application for interactive use cases. To circumvent the additional inversion step, several fast feed-forward approaches have been proposed in the generative models literature, that train an encoder to directly predict the latent code or features given an input image [Richardson et al. 2021; Trevithick et al. 2023a], but hasn't been extended to the relighting task.

We propose a novel technique Lite2Relight that takes an in-the-wild portrait image or video and synthesizes 3D-consistent head poses, physically plausible light editing at interactive frame rates as outlined in Sec. 1. Our method uses a lightstage dataset to extend

the generative capabilities of EG3D [Chan et al. 2021] and learns an efficient volumetric representation to implicitly disentangle face reflectance and perform relighting given a target HDRI environment map. We represent the input image as a combination of a low-dimensional latent vector and a feature image that lies in the latent manifold of EG3D by using a pre-trained geometry-aware encoder and feature alignment module by following the GAN inversion process [Yuan et al. 2023]. To achieve relighting, we design a simple MLP network that transforms the input latent vector to the desired illumination space by conditioning the network on a target environment map. We utilize the lightstage data [Weyrich et al. 2006] to synthesize an illumination dataset and embed it in the EG3D latent manifold to enable ground truth supervision of our pipeline. We train the relighting network with the illumination dataset to transform the inverted latent code into the desired illumination space. Since it is challenging to encode all 3D information in the low-dimension relit latent code, similar to [Yao et al. 2022], we refine the generator convolution layer with the combination of inverted and relit feature codes. This allows our method to learn physically accurate relighting in a 3D consistent manner, thus enabling rendering a given portrait from a novel viewpoint and performing various semantic edits made possible by the EG3D latent space. In summary, we present:

- A lightweight technique that enforces a strong face geometry and reflectance prior to lift 2D images to a relightable 3D space. We achieve this by leverging a pre-trained 3D generative model of faces in combination with a lightstage capture dataset to obtain a generalizable prior.
- Demonstration of view synthesis and light editing of human faces from a single portrait image using the proposed prior at interactive frame rates.
- Extensive quantitative and qualitative evaluation of the proposed method against state-of-the-art techniques to demonstrate its enhanced efficacy.

Code and pre-trained checkpoints is available under https://vcai.mpi-inf.mpg.de/projects/Lite2Relight//.

2 RELATED WORKS

We first discuss face reflectance modeling methods that are constrained by a 2D prior and/or incomplete face modeling. We then discuss 3D neural representations and facial editing methods that utilize a 3D generative model. Finally, we discuss 3D portrait lighting methods, which are the most relevant to our work.

Face Appearance Modelling. Capturing and modeling human faces to achieve highly authentic digital faces has been an active area of research [Debevec et al. 2000; Weyrich et al. 2006; Zollhöfer et al. 2018]. Several recent learning-based methods have exploited 2D generative image models for facial relighting [Abdal et al. 2021; B R et al. 2021a; Kwak et al. 2022; Richardson et al. 2021; Tewari et al. 2020a,b]. However, they cannot consistently disentangle the underlying identity-specific geometry from the view-dependent appearance, leading to inconsistent view synthesis. Parametric face models [Blanz and Vetter 1999; Li et al. 2017] have traditionally provided 3D priors for such tasks, but suffer from their low-dimensional representations, which limit their capacity to model high-frequency details such as wrinkles, and completely fail for unstructured regions like hair. Although there exist methods that accurately capture and model face reflectance fields, rendering such digital avatars [Alexander et al. 2010; Seymour et al. 2017] requires significant manual effort. Several traditional method use hand-crafted models and target specific parts of the face, such as facial hair [Echevarria et al. 2014], skin wrinkles [Gotardo et al. 2018], eyes [Li et al. 2022] teeth [Wu et al. 2016] and lips [Garrido et al. 2016] using computationally expensive optimization routines, and often require a dense and invasive data capture mechanisms. Using a parametric face model, multiple methods [B R et al. 2021b; Yamaguchi et al. 2018] enable face reflectance editing in the face interior region for monocular inputs. Several image-based relighting methods [Meka et al. 2019b; Nestmeyer et al. 2020; Pandey et al. 2021; Sun et al. 2019; Wang et al. 2020; Zhou et al. 2019] relight entire human heads for a fixed viewpoint or identity-specific settings [Bi et al. 2021]. Due to the lack of an underlying 3D representation such methods are limited to only relighting as they cannot modify the camera viewpoints.

3D Neural Representations and GANs. Neural Radiance Fields (NeRF) [Mildenhall et al. 2020], model a 3D scene as a 5D continuous radiance field function using a multi-layer perceptron (MLP) network and differential volume rendering from multiple viewpoints. This innovative approach enables precise 3D representations without the need for explicit geometric modeling. In the realm of 3D neural rendering, NeRF-based methods [Tewari et al. 2022b] have successfully achieved realistic rendering of human avatars [Gafni et al. 2021; Teotia et al. 2023] in a consistent 3D manner. Several innovative approaches [Bühler et al. 2023; Cao et al. 2022; Hong et al. 2022; Khakhulin et al. 2022; Ramon et al. 2021] have expanded these techniques into multi-identity models, learning a facial prior and demonstrating personalization even with sparse input data.

There has been a significant effort [Chan et al. 2021, 2022; Deng et al. 2022; Gu et al. 2022; Or-El et al. 2022] to blend Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] with NeRF [Mildenhall et al. 2020], facilitating the learning of a latent facial manifold. Notably, EG3D [Chan et al. 2022] incorporates StyleGAN [Karras et al. 2020] into a 3D framework to generate a comprehensive generative 3D prior of faces. These rich generative priors have opened avenues for portrait editing, primarily utilizing GAN inversion techniques. Methods such as Richardson et al. [2021]; Roich et al. [2021]; Yao et al. [2022] have demonstrated embedding portrait images into StyleGAN's latent space. In 3D GANs, especially EG3D, optimization-based inversion methods [Ko et al. 2023; Xie et al. 2022] have been used to update inverted latent codes to minimize reconstruction loss. However, such methods can be slow and often yield subpar editing quality [Yao et al. 2022], as the optimized latent code could diverge from the original sampling space. Encoder-based inversion methods [Trevithick et al. 2023b; Yuan et al. 2023], in contrast, offer faster performance with better regularization due to the lack of an optimization loop. Live 3D Portrait [Trevithick et al. 2023b] uses a ViT-based architecture [Dosovitskiy et al. 2021] to learn a new triplane representation [Chan et al. 2022] using synthetic data from EG3D, efficiently converting 2D portraits to 3D while foregoing the rich latent manifold for semantic editing.

Drawing inspiration from E4E [Richardson et al. 2021], GOAE [Yuan et al. 2023] trains an encoder to embed subjects within

Table 1: Our approach achieves a mix of novel capabilities for monocular in-the-wild portrait image editing such as 3D consistent pose synthesis, physically accurate relighting, semantic editing, and the efficiency of a feedforward encoder-based inference pipeline that enables interactive performance without the complexity of optimization-based fitting or finetuning. Note that while some methods like LumiGAN [Deng et al. 2023] and NeRFFaceLighting [Jiang et al. 2023] *can* perform relighting, they are not physically accurate due to the low-dynamic range of in-the-wild training data.

	Monocular	3D Consistency	Physical Relighting	Optimization-free	Semantic Editing
EG3D [Chan et al. 2021]	✓	✓	✗	✗	✓
Live 3D Portrait [Trevithick et al. 2023a]	✓	✓	✗	✓	✗
PhotoApp [B R et al. 2021a]	✓	✗	✓	✓	✗
LumiGAN [Deng et al. 2023]	✓	✓	✗	✗	✓
NeRFFaceLighting [Jiang et al. 2023]	✓	✓	✗	✗	✓
NeLF [Sun et al. 2021]	✗	✓	✓	✓	✗
VoRF [Rao et al. 2023, 2022]	✓	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓

EG3D's W^+ space, also incorporating an attention-based module to recover identity-specific features. This approach allows inverted portraits to be elevated to 3D while retaining the capability for semantic editing. We adapt such a 3D-aware encoder to invert desired portraits into the latent space of EG3D and perform relighting at interactive rates.

3D Portrait Relighting. Volumetric rendering approaches have enabled simultaneous editing of viewpoints and illuminations for both general scenes and human avatars without requiring explicit 3D geometry. The work of Boss et al. [2021]; Rudnev et al. [2022]; Zhang et al. [2021] perform intrinsic decomposition of general scenes and relight under novel illumination. Using a lightstage setup, Sarkar et al. [2023]; Yang et al. [2023b,a] demonstrate person-specific relighting. Using a synthetic OLAT dataset Sun et al. [2021] adapt PixelNeRF [Yu et al. 2021] to learn a generalizable 3D portrait relighting method. Similarly, MEGANE [Li et al. 2023] trains an MVP [Lombardi et al. 2021] representation that can generalize to unseen subjects. Both methods need at least three multi-view inputs, which limits their application in many real-world scenarios. Instead, our method takes a single monocular image as input.

Using a similar lightstage dataset, VoRF [Rao et al. 2023, 2022] trains a NeRF-based autodecoder network that generalizes to unseen identities under monocular settings. However, these models, trained on data that was captured in controlled setups with limited numbers of subjects often lead to a less diverse face prior, resulting in limited generalization towards in-the-wild samples. The method we present here not only avoids these issues using the EG3D prior but also computes results at interactive rates because it does not rely on a rather costly implicit representation. Generative models like EG3D offer a rich face prior and can synthesize arbitrary numbers of faces. On this basis, recent methods [Deng et al. 2023; Ranjan et al. 2023; Tan et al. 2022] relight synthetic identities sampled from a latent space. Here, Deng et al. [2023] combine precomputed radiance transfer [Sloan et al. 2002] with adversarial learning to relight portraits, but due to the self-supervised learning paradigm this method struggles to learn physically accurate lighting. Additionally, the above methods focus on generating synthetic samples and offer very limited capability for editing a given input portrait. Our method on the other hand can be controlled very accurately by an explicit environment map as input.

3 METHOD

The primary aim of our method is to relight a portrait of a human from a single input image under any desired novel viewpoint and illumination. This is achieved without the need for time-intensive optimization processes. This task is inherently underconstrained due to depth ambiguity and the complex interplay between facial features and varying illumination. Directly modeling light transport is computationally expensive and approximations often lead to non-photorealistic outcomes. Our approach circumvents these challenges by implicitly managing light transport through neural networks.

As illustrated in Fig. 2, we leverage EG3D [Chan et al. 2022], a 3D-aware generative model known for its high-quality, generalizable representations of human faces. To adapt the relighting task to the feature space of EG3D, we first embed real images into the EG3D space Sec. 3.2. Following this, a mapping network, trained on a lightstage dataset (Sec. 3.1), is employed to transition the original feature vector into a target feature vector. This enables us to render the face under novel lighting conditions and viewpoints. Details of the relighting module are further elaborated in Sec. 3.3. Finally, the various loss functions utilized in our method are outlined in Sec. 3.4.

3.1 Dataset

We use the multi-view lightstage dataset captured by [Weyrich et al. 2006] that has 353 subjects illuminated under $N = 150$ point light sources where each subject is captured with 16 cameras. Hence, the dataset contains a set of one-light-at-a-time (OLAT) images $O = \{\mathbf{O}_1, ...\mathbf{O}_N | \mathbf{O}_i \in \mathbb{R}^{512\times512\times3}\}$ for every subject under 16 viewpoints. Normally, "in the wild" images are not captured under such OLAT conditions. Therefore, directly embedding OLAT images onto the EG3D space will result in unfaithful reconstructions. Thus, to alleviate this and obtain realistic scene illumination conditions, we

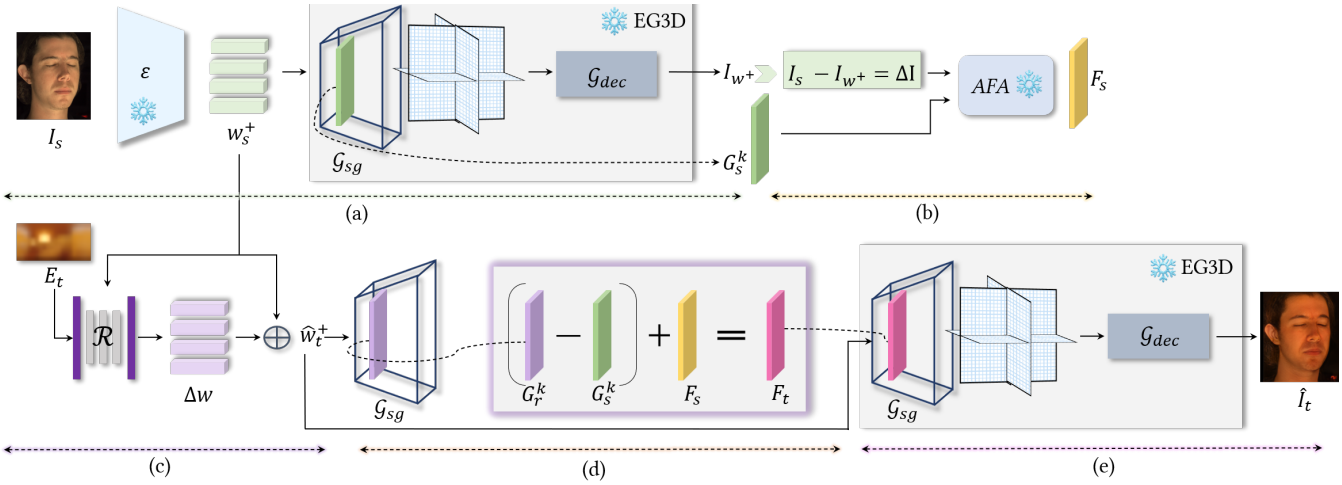

Figure 2: Method Overview. (a) Given an input image I_s, we use a pretrained encoder $\mathcal{E}$ to invert I_s and obtain the latent vector w_s^+. We pass w_s^+ through a pretrained EG3D network to render the inverted image I_{w^+} and extract convolutional features G_s^k from $\mathcal{G}_{sg}$. (b) Next, we use image residual ΔI and G_s^k as inputs to the AFA module, to obtain F_s. (c) Given a target environment map E_t, our relighting network $\mathcal{R}$ generates Δw, which is combined with w_s^+ to produce the relit latent code $\hat{w}_t^+$. (d) Subsequently, we obtain F_t by following Eq. 7. (e) Finally, we replace the k-th convolutional feature of $\mathcal{G}_{sg}$ by F_t and perform a full forward pass through the EG3D network with the latent code $\hat{w}_t^+$ to generate $\hat{I}_t$, which is relit by E_t. Note: $\mathcal{G}_{dec}$ takes camera pose c as input.

linearly combine OLAT images with environment maps by following image-based relighting [Debevec et al. 2000] as follows

$$\mathbf{I} = \sum_{i=0}^{N} E(i) \cdot \mathbf{O}_i \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{512\times512\times3}$ is the relit image and $E : \mathbb{N}_{<N} \rightarrow \mathbb{R}^3$ is the downsampled version of the input environment map.

Training Data. We relight all the subjects under 50 natural illumination conditions randomly sampled from the Laval Indoor and Outdoor datasets [Gardner et al. 2017; Hold-Geoffroy et al. 2019] by using Eq. 1. To obtain paired data for supervision, for each training step we randomly sample two naturally relit images of the *same subject*: The input i.e. the source image is referred to as I_s and is relit with a source environment map E_s. The target I_t is relit with E_t.

3.2 3D GAN Inversion

We adopt EG3D [Chan et al. 2022], a 3D-aware generative model that has demonstrated remarkable generalization results, as our backbone. The generator $\mathcal{G}_{sg}$, based on StyleGAN [Karras et al. 2020], maps a latent vector w_s^+ and camera pose c to triplane features that are further decoded to render a low-resolution image. This low-resolution image is upsampled to 512×512 using a super-resolution module. We formally denote the decoder, volume rendering, and upsampling as $\mathcal{G}_{dec}$. The generated image is obtained as:

$$I_{w^+} = \mathcal{G}_{dec}(\mathcal{G}_{sg}(w_s^+), c) \tag{2}$$

To obtain robust features in the EG3D space that are representative of the training images of our dataset, we use an encoder $\mathcal{E}$, adopting [Yuan et al. 2023]. $\mathcal{E}$ maps a given source image I_s to a latent code $w_s^+ \in \mathbb{R}^{14\times512}$:

$$w_s^+ = \mathcal{E}(I_s) \tag{3}$$

However, w_s^+ is a low-dimensional latent code that is insufficient to learn a rich representation of the portrait. Thus, we follow [Yuan et al. 2023] and further adapt their "Adaptive Feature Alignment" module which has attention-based and convolutional layers to learn an additional feature code. We denote the module as *AFA* and it takes the difference between the source image and the predicted image as input, i.e. $\Delta I = I_s - I_{w^+}$, along with features of $\mathcal{G}_{sg}$:

$$F_s = AFA(\Delta I, G_s^k) \in \mathbb{R}^{32\times32\times512} \tag{4}$$

where $G_s^k \in \mathbb{R}^{32\times32\times512}$ is the k-th convolutional feature of $\mathcal{G}_{sg}$, for some fixed k. Finally, we use the remaining layers of $\mathcal{G}_{sg}$, starting from k + 1, to obtain the triplane features that are used to reconstruct the input image. Thus, w_s^+ and F_s together serve as a robust representation of I_s in a high-dimensional latent space.

3.3 Relighting

In this section, we explain the steps involved in relighting I_s given a target environment map E_t. The latent space of EG3D contains a rich representation of faces and we leverage this advantageous latent space to perform relighting. Our aim is to embed the target images I_t to this latent space and obtain $\hat{w}_t^+ \in \mathbb{R}^{14\times512}$. However, as described in Sec. 3.1, a pair of I_s and I_t have the same subject relit under different environment maps. Thus, we have an MLP $\mathcal{R}$, that maps the source latent code w_s^+ and the target environment map E_t to a latent offset Δw:

$$\Delta w = \mathcal{R}(w_s^+, E_t) \in \mathbb{R}^{14\times512} \tag{5}$$

To obtain $\hat{w}_t^+$, we add this offset to w_s^+:

$$\hat{w}_t^+ = w_s^+ + \Delta w \tag{6}$$

A relit image can be directly obtained by giving $\hat{w}_t^+$ as the input to $\mathcal{G}_{dec}$. However, we found the obtained image to not preserve subject-specific details well, altering the perceived identity (see ablation in Sec. 4.4). This is because $\hat{w}_t^+$ is a low-dimensional representation that is insufficient to capture the subject details. But, the *AFA* model can not be used as it requires the relit image which we don't have access to. Hence, we aim to transfer fine-scale details present in F_s to the target illumination space: We perform a forward pass through $\mathcal{G}_{sg}$ with $\hat{w}_t^+$ as input and extract the k-th convolutional feature G_r^k. Finally, we combine the convolutional features of the source and target image by following [Yao et al. 2022]:

$$F_t := F_s + G_r^k - G_s^k \in \mathbb{R}^{32\times32\times512} \tag{7}$$

Thus, $\hat{w}_t^+$ and F_t serve as a robust representation of the predicted relit image in a high-dimensional latent space. Finally, we replace the k-th convolutional feature of $\mathcal{G}_{sg}$ with F_t and perform a full forward pass as follows:

$$\hat{I}_t = \mathcal{G}_{dec}(\mathcal{G}_{sg}(\hat{w}_t^+, F_t), c) \tag{8}$$

3.4 Loss Functions

We train $\mathcal{R}$, while the networks $\mathcal{G}_{sg}$, $\mathcal{G}_{dec}$ and *AFA* remain frozen. As described in Sec. 3.1, we first obtain a set of paired source and target images I_s, I_t. Given I_s as input and target illumination condition E_t, the goal is to predict a $\hat{I}_t$ that is as close to I_t as possible. We obtain target latent code w_t^+ as described in Eq 3 using I_t. As a training objective, we minimize a combination of reconstruction and latent loss:

Reconstruction Loss. : We penalize deviations of $\hat{I}_t$ from I_t by $L1$ distance:

$$\mathcal{L}_{\text{C}} = \left\|\hat{I}_t - I_t\right\|_1 \tag{9}$$

Perceptual Loss. : Supervision in the image space alone resulted in poor reconstruction of certain illumination conditions in the target space. Therefore, we employ a feature-based loss $\mathcal{L}_{\text{LPIPS}}$ [Johnson et al. 2016] between $\hat{I}_t$ and I_t.

$$\mathcal{L}_{\text{LPIPS}} = \left\|\Phi_{vgg}(\hat{I}_t) - \Phi_{vgg}(I_t)\right\|_2^2 \tag{10}$$

Where Φ_{vgg} is the extracted features from the pre-trained VGG [Simonyan and Zisserman 2015] network. We conduct an ablative study (Sec. 4.4), to demonstrate the effectiveness of $\mathcal{L}_{\text{LPIPS}}$ for relighting.

Latent Loss. : To ensure that the $\hat{w}_t^+$ predicted by $\mathcal{R}$ is in the same part of the EG3D latent space as w_s^+ (and not in a region that behaves differently under $\mathcal{G}_{sg}$), we penalize the $L2$ distance between $\hat{w}_t^+$ and w_t^+ :

$$\mathcal{L}_{\text{lat}} = \left\|\hat{w}_t^+ - w_t^+\right\|_2^2 \tag{11}$$

The total loss is given as:

$$\mathcal{L}_{\text{total}} = \lambda_0\mathcal{L}_{\text{lat}} + \lambda_1\mathcal{L}_{\text{C}} + \lambda_2\mathcal{L}_{\text{LPIPS}} \tag{12}$$

4 EVALUATION

We describe the datasets used for evaluation in Sec. 4.1. In Sec. 4.2 we discuss qualitative evaluation of simultaneous view synthesis and relighting for in-the-wild portraits. We report quantitative analysis on the lightstage dataset in Sec. 4.3 and discuss ablative experiments in Sec. 4.4.

4.1 Dataset

We create an evaluation dataset based on a lightstage dataset [Weyrich et al. 2006] with 10 unseen subjects, illuminated under 10 novel illumination conditions under 12 novel camera viewpoints. Linearly combining OLATs with downsampled HDRI environment maps (like in Sec. 3.1), gives us relit ground truth images. Additionally, we qualitatively evaluate Lite2Relight using diverse subjects captured in the wild [Caselles et al. 2023; Livingstone and Russo 2018; Ramon et al. 2021; Shih et al. 2014].

4.2 Relighting in-the-wild Portraits

In Fig. 3 and Fig. 5, we show Lite2Relight 's capability to modify viewpoint and illumination of diverse in-the-wild portraits. Our method retains 3D consistency for head poses as well as relighting. This can be specifically observed in the last two columns of both figures, where the subject is relit under the same environment map, and shows consistent relighting under two different camera views. This is because the 3D representation, coupled with accurate relighting in the latent manifold, faithfully preserves the identity and expression of subjects. From (Fig. 3 and Fig. 5) we observe strong identity preservation. Moreover, intricate details like expressions (see rows 1 of Fig. 3 and Fig. 5) and accessories such as spectacles under varying illumination scenarios (see row 4 of Fig. 5) are preserved as well. Owing to the robust generative prior, our method generalizes to a wide variety of subjects, while preserving/synthesizing complex reflectance phenomena, such as subsurface scattering and specular highlights, particularly noticeable on the nose, cheeks, and forehead (see rows 2 and 3 of Fig. 3), as well as soft-shadows (see row 1 of Fig. 3, rows 2 and 3 of Fig. 5).

4.3 Comparisons to Previous Works

We compare our method to state-of-the-art methods for simultaneous viewpoint and illumination editing: (1) VoRF [Rao et al. 2023, 2022] employs a NeRF-based auto-decoder architecture learning a volumetric reflectance field from an OLAT lightstage dataset. (2) PhotoApp [B R et al. 2021a] leverages a 2D StyleGAN prior for faces, focusing on learning latent transformations to manipulate viewpoint and illumination. (3) NeRFFaceLighting [Jiang et al. 2023] leverages EG3D design principles to learn a separate appearance and lighting triplane. (4) NeLF [Sun et al. 2021] trains pixelNeRF[Yu et al. 2021] inspired 3D representation using a synthetic lightstage dataset to derive 3D geometry and facial reflectance properties. For all the methods, we use the original implementation provided by the authors and train it on the lightstage dataset. For a fair qualitative assessment, VoRF results are upsampled to 512×512 pixels. Fig. 4 and Tab. 2 indicate that Lite2Relight not only surpasses the baseline approaches in rendering high-quality relit images under various scene illuminations but also excels in maintaining 3D consistency and identity fidelity.

Figure 3: Qualitative Results: Relighting in-the-wild Portraits. col. (column) 1: input in-the-wild image, col. 2: image relit with HDRI environment maps (inset) (E_1) under the same viewpoint. col. 3: Novel View (NV) 1 with a different environment map (E_2). col. 4 and 5: NV 2 and 3 under the same map. This figure demonstrates that Lite2Relight can generalize robustly to in-the-wild images, preserve subject-specific face semantics and perform relighting under various environment maps simultaneously. Image credits to Flickr.

Table 2: Quantitative Results: Comparisons to Previous Works. We report SSIM, landmarks distance (LD), and PSNR computed on the the test data (Sec. 4.1), for Lite2Relight and previous works, relighting subjects under novel views.

	SSIM ↑	LD ↓	PSNR ↑
NeLF (3-views)	0.75	NA	19.72
PhotoApp	0.72	34.08	**29.13**
VoRF	0.69	16.90	20.21
NeRFFaceLighting	0.79	28.31	13.41
Lite2Relight	**0.83**	**9.76**	28.3

Metrics. Apart from Peak Signal-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), we also use a landmark distance metric (LD) to evaluate the 3D consistency of facial geometry. It is calculated as the average deviation of 68 facial key points [Bulat and Tzimiropoulos 2017] over the evaluation dataset.

VoRF. Fig. 8 shows that VoRF is unable to generalize to subjects that are out of the training distribution (lightstage data): It struggles with identity preservation (row 1), as well as expression preservation and capturing eye details (rows 2 and 3). This is due to VoRF's dependence on a face prior learned from the lightstage data. For lightstage subjects Fig. 4 VoRF does show reasonable results. In contrast, however, Lite2Relight can preserve facial details, expressions, and even eye details for out-of-distribution data (columns 4 and 5 in Fig. 8), thanks to our design where we leverage information from large-scale in-the-wild data (FFHQ). Furthermore, due to VoRF's NeRF-based framework, which uses large amounts of memory, its outputs are confined to resolution 128×128 (see blurriness in results in Fig. 4). Lite2Relight , on the other hand, based on the triplane framework, can preserve fine details at resolution 512×512. Finally, VoRF requires a two-step optimization process that takes $\sim$ 10mins to relight a single image, making it unsuitable for interactive applications. While Lite2Relight , with its encoder-based architecture and an efficient triplane-based volume rendering technique, achieves interactive performance (7-31 fps. See Sec. 1 of SupMat).

PhotoApp. PhotoApp achieves high image quality (see PSNR in Tab. 2), thanks to the latent face prior from StyleGAN [Karras et al. 2020]. However, PhotoApp's inherent lack of a native 3D representation leads to noticeable inconsistencies in subject identity under novel views. This results in much lower SSIM and LD scores

compared to Lite2Relight , which uses an efficient NeRF-based representation, and excels in maintaining 3D consistency. Fig. 4 visualises this advantage: PhotoApp deviates from the original identity, while Lite2Relight maintains a close resemblance to the input subject.

Furthermore, Photoapp manipulates both viewpoint and illumination in the latent space which is challenging to control (row 2 in Fig. 4). Lite2Relights's latent transformation is confined to illumination, allowing us to control the viewpoint and illumination in a disentangled way. Further results shown in Fig. 6 demonstrate that PhotoApp struggles to preserve the identity (rows 1 and 3) and even shows inconsistent illumination across different views. Lite2Relight does preserve the identity of the subject in a 3D-consistent manner and maintains the desired lighting across all views.

NeRFFaceLighting. In our analysis of NeRFFaceLighting (NFL), as illustrated in Fig. 4, NFL clearly shows significant lighting artifacts, poor identity retention (can be noticed in row 1), and overly saturated effects (observed in row 2). These shortcomings are quantitatively evidenced in Tab. 2, where NFL consistently underperforms in comparison to Lite2Relight . We hypothesize that these issues stem from the inherent complexities in jointly optimizing illumination and identity during the inversion process. NFL employs an explicit decomposition of the input into albedo and shading components, which makes the inversion process complicated and fails to recover the true identity accurately. As a result, the input light gets baked into the albedo during relighting, as shown in row 2, where the prediction of the subject has a red-colored skin taken from the input light source. In contrast, our method mitigates the albedo-lighting ambiguity by utilizing a supervised training approach with a lightstage dataset, where each subject is relit under various lighting conditions. Moreover, our strategy includes an encoder-based inversion process that helps preserve identity integrity and achieve more accurate relighting outcomes.

NeLF. NeLF requires at least three input views to obtain reasonable results. This makes NeLF unusable for almost all in-the-wild portraits. Moreover, NeLF struggles to reconstruct a reasonable facial structure even with multiview inputs as it fails to represent the underlying geometry leading to low scores in Tab. 2. We can observe the results in Fig. 7, where NeLF's results were obtained from 3 input views producing distorted facial reconstructions. In contrast, our approach demonstrates superior generalization to novel subjects from a single image. Furthermore, our method exhibits the ability to accurately relight these subjects, maintaining both the integrity of facial features and the overall photorealism.

4.4 Ablation Study

Significance of Feature Code Manipulation. To evaluate the importance of Eq. 7, particularly focusing on the comparative roles of F_t and G_r^k, we compare our original method to a variant that does not replace G_r^k by F_t. The results, as depicted in Fig. 9, reveal that the incorporation of F_t significantly enhances the recovery of fine-scale, identity-specific details. This improvement is especially noticeable in the eye region of row 2, as well as in the jawlines and overall facial contours of all the rows. The efficacy of F_t in enhancing the fidelity of these features is further corroborated by its superior performance in quantitative evaluations, see Tab. 3.

Table 3: Quantitative Results: Ablation Study. We report SSIM, landmarks distance (LD), and PSNR on the test data of lightstage, where subjects are relit under novel viewpoints.

	SSIM ↑	LD ↓	PSNR ↑
w/o F-space	0.831	10.44	28.29
w/o $\mathcal{L}_{\text{LPIPS}}$	0.830	10.12	28.32
Full Model	**0.83**	**9.76**	**28.33**

Importance of Perceptual Loss. It is feasible to supervise $\mathcal{R}$ solely using $\mathcal{L}_{\text{lat}}$ and $\mathcal{L}_{\text{C}}$, omitting $\mathcal{L}_{\text{LPIPS}}$, thereby effectively only supervising the latent vectors to preserve identity. However, as our findings in Fig. 10 suggest, relying solely on reconstruction losses may not guarantee accurate relighting: Rows 1 and 2 show inaccuracies in the overall color of the images, while shadows and highlights are noticeably absent in row 3. We hypothesize that these discrepancies arise due to the inherent limitations of reconstruction losses, which do not necessarily account for perceptual plausibility. Adding $\mathcal{L}_{\text{LPIPS}}$, as suggested before [Johnson et al. 2016], ensures a more accurate relighting, as demonstrated in the column labeled "w/ $\mathcal{L}_{\text{LPIPS}}$". Here, the relit images not only more closely resemble the reference illumination but also achieve higher scores across quantitative metrics. Additional ablations are provided in the SupMat.

5 CONCLUSION

In conclusion, our work, Lite2Relight , represents a significant advancement in the field of 3D portrait editing and relighting, effectively addressing the complex challenges associated with interactive, photorealistic image processing for AR/VR applications. By innovatively leveraging a pre-trained 3D generative model in conjunction with a lightstage dataset and a 3D-aware encoder, we have successfully developed a method that lifts 2D images into a relightable 3D space with strong geometric and reflectance accuracy. This technique not only simplifies the process of achieving photorealism in portrait editing but also ensures practical applicability by minimizing the need for extensive computational resources and complex data inputs. Our work with Lite2Relight showcases the capability of performing view synthesis, light editing, and semantic modifications at interactive rates from single, in-the-wild portrait images, a notable development in the field. The efficiency and fidelity with which Lite2Relight accomplishes these tasks have meaningful implications for the evolution of personalized digital content creation, particularly within the context of AR/VR applications. The extensive evaluations of our method against current state-of-the-art techniques highlight its superior performance and practicality, reinforcing our contribution to the domain of computational photography and graphics. We believe that the release of our code and pretrained models will foster further research and development, potentially leading to widespread adoption and continuous improvement in this field.

ACKNOWLEDGMENTS

This work was supported by the ERC Consolidator Grant 4DReply (770784). We extend our gratitude to Shrisha Bharadwaj for providing feedback and constant support.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. *ACM Transactions on Graphics*, Article 21 (2021), 21 pages.

Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. 2010. The Digital Emily Project: Achieving a Photorealistic Digital Actor. *IEEE Computer Graphics and Applications* 30, 4 (2010), 20–31. https://doi.org/10.1109/MCG.2010.65

Mallikarjun B R, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed A Elgharib, and Christian Theobalt. 2021a. PhotoApp: Photorealistic appearance editing of head portraits. *ACM Transactions on Graphics* 40, 4 (2021).

Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. 2021b. Monocular Reconstruction of Neural Face Reflectance Fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep Relightable Appearance Models for Animatable Faces. *ACM Transactions on Graphics*, Article 89 (2021), 15 pages.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. 2021. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12684–12694.

Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. 2023. Preface: A Data-driven Volumetric Prior for Few-shot Ultra High-resolution Face Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3402–3413.

Marcel C. Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, and Abhimitra Meka. 2023. Preface: A Data-driven Volumetric Prior for Few-shot Ultra High-resolution Face Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3402–3413.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Trans. Graph.* 41, 4, Article 163 (jul 2022), 19 pages. https://doi.org/10.1145/3528223.3530143

Pol Caselles, Eduard Ramon, Jaime Garcia, Gil Triginer, and Francesc Moreno-Noguer. 2023. Implicit Shape and Appearance Priors for Few-Shot Full Head Reconstruction. *arXiv preprint arXiv:2310.08784* (2023).

Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2020. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *arXiv*.

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. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

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. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *CVPR*.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Annual conference on Computer graphics and interactive techniques*.

Boyang Deng, Yifan Wang, and Gordon Wetzstein. 2023. LumiGAN: Unconditional Generation of Relightable 3D Human Faces. In *arXiv*.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

Jose I. Echevarria, Derek Bradley, Diego Gutierrez, and Thabo Beeler. 2014. Capturing and Stylizing Hair for 3D Fabrication. *ACM Trans. Graph.* 33, 4, Article 125 (jul 2014), 11 pages. https://doi.org/10.1145/2601097.2601133

Ohad Fried, Jennifer Jacobs, Adam Finkelstein, and Maneesh Agrawala. 2020. Editing Self-Image. *Commun. ACM* 63, 3, Article 10.1145/33 (March 2020), 70–79 pages.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *Proc. SIGGRAPH Asia* (2017).

P. Garrido, M. Zollhöfer, C. Wu, D. Bradley, P. Perez, T. Beeler, and C. Theobalt. 2016. Corrective 3D Reconstruction of Lips from Monocular Video. *ACM Transactions on Graphics (TOG)* 35, 6 (2016).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical Dynamic Facial Appearance Modeling and Acquisition. *ACM Trans. Graph.* 37, 6, Article 232 (dec 2018), 13 pages. https://doi.org/10.1145/3272127.3275073

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*.

Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. 2019. Deep Sky Modeling For Single Image Outdoor Lighting Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. 2023. NeRFFaceLighting: Implicit and Disentangled Face Lighting Representation Leveraging Generative Prior in Neural Radiance Fields. *ACM Transactions on Graphics (TOG)* (2023).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 694–711.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic One-Shot Mesh-Based Head Avatars. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 345–362.

Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 2023. 3D GAN Inversion with Pose Optimization. *WACV* (2023).

Jeong-gi Kwak, Yuanming Li, Dongsik Yoon, Donghyeon Kim, David Han, and Hanseok Ko. 2022. Injecting 3D Perception of Controllable NeRF-GAN into StyleGAN for Editable Portrait Image Synthesis. In *European Conference on Computer Vision*. Springer, 236–253.

Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C. Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Trans. Graph.* 41, 4, Article 166 (jul 2022), 16 pages. https://doi.org/10.1145/3528223.3530130

Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. 2023. MEGANE: Morphable Eyeglass and Avatar Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12769–12779.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13, 5 (05 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019a. Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination. *ACM Transactions*

on Graphics (Proceedings SIGGRAPH) 38, 4. https://doi.org/10.1145/3306346.3323027

Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019b. Deep Reflectance Fields: High-Quality Facial Reflectance Field Inference from Color Gradient Illumination. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (2019).

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*.

Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. 2020. Learning Physics-guided Face Relighting under Directional Light. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13503–13513.

Xingang Pan, Ayush Tewari, Lingjie Liu, and Christian Theobalt. 2022. GAN2X: Non-Lambertian Inverse Rendering of Image GANs. In *International Conference on 3D Vision (3DV)*.

Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. 2021. A Shading-Guided Generative Implicit Model for Shape-Accurate 3D-Aware Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* (2021).

Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5620–5629.

Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, and Oncel Tuzel. 2023. FaceLit: Neural 3D Relightable Faces. In *CVPR*. https://arxiv.org/abs/2303.15437

Pramod Rao, Mallikarjun B. R, Gereon Fox, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Fangneng Zhan, Ayush Tewari, Christian Theobalt, and Elgharib Mohamed. 2023. A Deeper Analysis of Volumetric Relightable Faces. *International Journal of Computer Vision* (10 2023), 1–19. https://doi.org/10.1007/s11263-023-01899-3

Pramod Rao, Mallikarjun B R, Gereon Fox, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Ayush Tewari, Christian Theobalt, and Mohamed Elgharib. 2022. VoRF: Volumetric Relightable Faces. (2022).

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Trans. Graph.* (2021).

Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. 2022. NeRF for Outdoor Scene Relighting. In *European Conference on Computer Vision (ECCV)*.

Kripasindhu Sarkar, Marcel C. Buehler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo Gotardo, Thabo Beeler, and Abhimitra Meka. 2023. LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces. In *ACM SIGGRAPH Asia 2023 Conference Papers, December 12–15, 2023, Sydney, NSW, Australia.* https://doi.org/10.1145/3610548.3618210

Mike Seymour, Chris Evans, and Kim Libreri. 2017. Meet Mike: Epic Avatars. In *ACM SIGGRAPH 2017 VR Village* (Los Angeles, California) *(SIGGRAPH '17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 2 pages. https://doi.org/10.1145/3089269.3089276

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. 2014. Style transfer for headshot portraits. *ACM Trans. Graph.* 33, 4, Article 148 (jul 2014), 14 pages. https://doi.org/10.1145/2601097.2601137

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

Peter-Pike J. Sloan, Jan Kautz, and John M. Snyder. 2002. Precomputed Radiance Transfer for Real-Time Rendering in Dynamic, Low-Frequency Lighting Environments. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2* (2002). https://api.semanticscholar.org/CorpusID:324277

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single Image Portrait Relighting. *ACM Trans. Graph.* 38, 4, Article 79 (jul 2019), 12 pages. https://doi.org/10.1145/3306346.3323008

Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. 2021. NeLF: Neural Light-transport Field for Portrait View Synthesis and Relighting. In *Eurographics Symposium on Rendering*.

Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. 2022. VoLux-GAN: A Generative Model for 3D Face Synthesis with HDRI Relighting. arXiv:2201.04873 [cs.CV]

Kartik Teotia, Xingang Pan, Hyeongwoo Kim, Pablo Garrido, Mohamed Elgharib, Christian Theobalt, et al. 2023. HQ3DAvatar: High Quality Controllable 3D Head Avatar. *arXiv preprint arXiv:2303.14471* (2023).

Ayush Tewari, MalliKarjun B R, Xingang Pan, Ohad Fried, Maneesh Agrawala, and Christian Theobalt. 2022a. Disentangled3D: Learning a 3D Generative Model with Disentangled Geometry and Appearance from Monocular Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020a. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020b. PIE: Portrait Image Embedding for Semantic Control. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)* 39, 6 (December 2020). https://doi.org/10.1145/3414685.3417803

A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik. 2022b. Advances in Neural Rendering. *Computer Graphics Forum (EG STAR 2022)* (2022).

Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2023a. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*.

Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2023b. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*.

Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single Image Portrait Relighting via Explicit Multiple Reflectance Channel Modeling. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* (2020).

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. 2006. Analysis of Human Faces using a Measurement-Based Skin Reflectance Model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (2006).

C. Wu, D. Bradley, P. Garrido, M. Zollhöfer, C. Theobalt, M. Gross, and T. Beeler. 2016. Model-Based Teeth Reconstruction. *ACM Transactions on Graphics (TOG)* 35, 6 (2016).

Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. 2022. High-fidelity 3D GAN Inversion by Pseudo-multi-view Optimization. *arXiv preprint arXiv:2211.15662* (2022).

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-Fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image. *ACM Trans. Graph.* 37, 4, Article 162 (jul 2018), 14 pages. https://doi.org/10.1145/3197517.3201364

Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. 2023b. Towards Practical Capture of High-Fidelity Relightable Avatars. In *SIGGRAPH Asia 2023 Conference Proceedings*.

Jing Yang, Hanyuan Xiao, Wenbin Teng, Yunxuan Cai, and Yajie Zhao. 2023a. Light Sampling Field and BRDF Representation for Physically-based Neural Rendering. arXiv:2304.05472 [cs.CV]

Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. 2022. A Style-Based GAN Encoder for High Fidelity Reconstruction of Images and Videos. *European conference on computer vision* (2022).

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Transactions on Graphics (TOG)* (2022).

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*.

Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. 2023. Make Encoder Great Again in 3D GAN Inversion through Geometry and Occlusion-Aware Encoding. *arXiv preprint arXiv:2303.12326* (2023).

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. 2021. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)* 40 (2021), 1–18.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. 2019. Deep Single-Image Portrait Relighting. In *The IEEE International Conference on Computer Vision (ICCV)*.

M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum* 37, 2 (2018), 523–550. https://doi.org/10.1111/cgf.13382 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13382

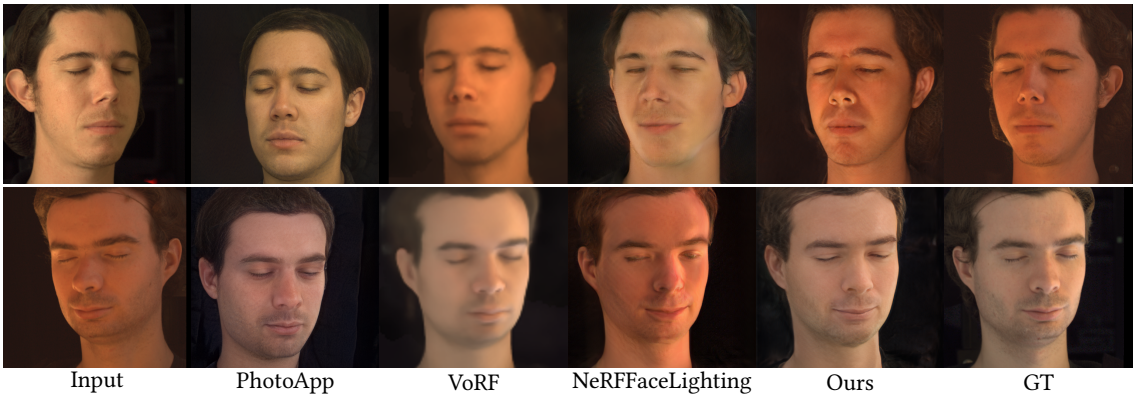

Figure 4: Qualitative Results: Comparisons to Previous Works. We compare with NFL [Jiang et al. 2023], VoRF [Rao et al. 2023, 2022] and PhotoApp [B R et al. 2021a].For each method, including ours, a single input view is utilized to generate novel views alongside relighting of the lightstage subjects. In comparison to the leading state-of-the-art techniques, Lite2Relight demonstrates superior ability in maintaining subject identity and capturing finer details.

Figure 5: Qualitative Results: Relighting in-the-wild Portraits. col. (column) 1: input in-the-wild image, col. 2: image relit with HDRI environment maps (inset) (E_1) under the same viewpoint. col. 3: Novel View (NV) 1 with a different environment map (E_2). col. 4 and 5: NV 2 and 3 under the same map. We show additional results to demonstrate generalization, 3D consistent pose of subjects, and relighting results of Lite2Relight for in-the-wild images. Image credits to Steven R. Livingstone and Flickr.

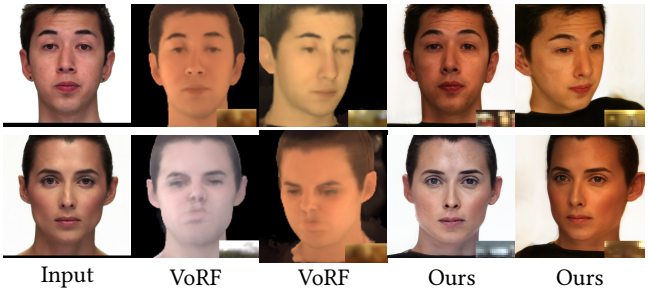

Figure 8: Qualitative Results: Comparisons with VoRF [Rao et al. 2023, 2022]. We present evaluations on the Ravdess [Livingstone and Russo 2018] dataset. The first column contains the input images, followed by columns depicting simultaneous view synthesis and relighting results under varying environment maps. VoRF struggles to generalize to subjects outside the training data distribution, exhibiting limitations in generalization. In contrast, Lite2Relight demonstrates robust generalization, preserves subject-specific details, and achieves accurate relighting.Image credits to Steven R. Livingstone.

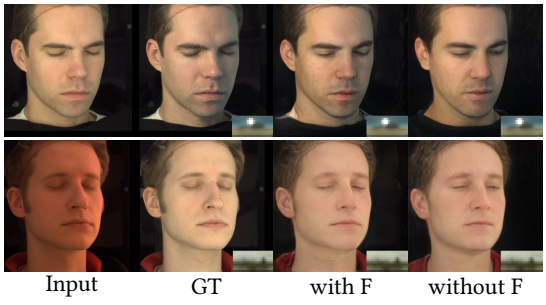

Figure 9: Ablation Study: Significance of Feature Code Manipulation. The label "With F" denotes results obtained after manipulating the feature code according to Eq. 7, whereas "Without F" implies the direct use of G_r^k. The results demonstrate that manipulating the feature code is crucial for achieving fine-grained identity preservation.

Figure 10: Ablation Study: Importance of Perceptual Loss. Qualitative comparison of results with and without $\mathcal{L}_{\text{LPIPS}}$. Results indicate that using $\mathcal{L}_{\text{LPIPS}}$ improves the quality of relighting.

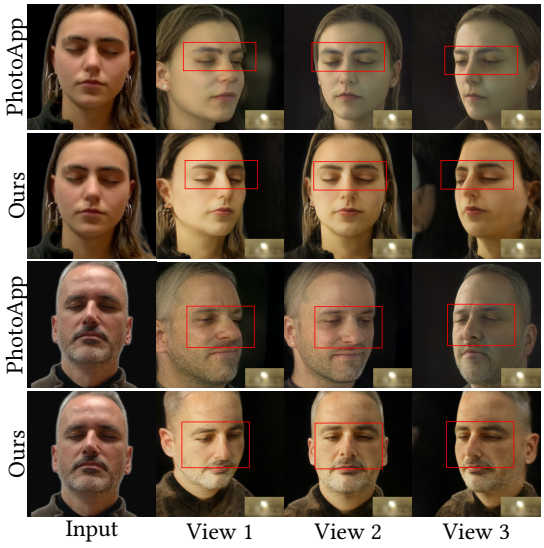

Figure 6: Qualitative Results: Comparisons with PhotoApp [B R et al. 2021a]. We conduct comparisons using the H3DS dataset [Caselles et al. 2023; Ramon et al. 2021]. The first column displays the input images, followed by three columns showcasing novel view synthesis results under the same environment map. This comparison highlights that PhotoApp fails to retain identity-specific details as effectively as our method. Notably, the subject's eyes and nose region appear altered across different views in PhotoApp, whereas it remains consistent and true to the input in Lite2Relight . Image credits to Pol Caselles.

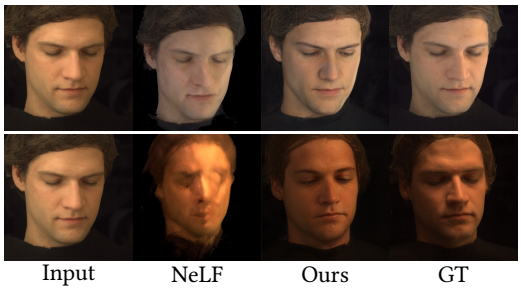

Figure 7: Qualitative Results: Comparison with NeLF [Sun et al. 2021]. We present evaluations using the lightstage dataset [Weyrich et al. 2006]. For NeLF, three input views are provided, whereas for Lite2Relight , only the first column input is used. NeLF exhibits artifacts under novel viewing conditions, whereas Lite2Relight maintains 3D consistency across different viewing angles.