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 Francisico, 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.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8525-0/24/07
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 HDR 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

1 INTRODUCTION
Photorealistic editing and composing 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 [Bühlker 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, specularinity, 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 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 feedforward 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 volumetric 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 leveraging 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 methods 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 StyleGAN3D [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
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 $O = \{O_1, ..., O_N\} \in \mathbb{R}^{512 \times 512 \times 3}$ 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...
We adopt EG3D [Chan et al. 2022], a 3D-aware generative model. We pass \( w^* \) through a pretrained EG3D network to render the inverted image \( I_{w^*} \) and extract convolutional features \( G^k \) from \( G_{sg} \). Next, we use image residual \( \Delta I \) and \( G^k \) as inputs to the AFA module, to obtain \( F_t \). Given a target environment map \( E_t \), our relighting network \( R \) generates \( \Delta w \), which is combined with \( w^* \) to produce the relit latent code \( \hat{w}^*_t \). Subsequently, we obtain \( \hat{I}_t \) by following Eq. 7. Finally, we replace the \( k \)-th convolutional feature of \( G_{sg} \) 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: \( G_{dec} \) takes camera pose \( c \) as input.

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 \( G_{sg} \), based on StyleGAN [Karras et al. 2020], maps a latent vector \( w^* \) and camera pose \( c \) to triplane features that are further decoded to render a low-resolution image. This low-resolution image is upsampled to \( 512 \times 512 \) using a super-resolution module. We formally denote the decoder, volume rendering, and upsampling as \( G_{dec} \). The generated image is obtained as:

\[
I_{w^*} = G_{dec}(G_{sg}(w^*), c)
\]

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 \( w^* \in \mathbb{R}^{14 \times 512} \):

\[
w^* = \mathcal{E}(I_s)
\]

However, \( w^*_t \) 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 \( G_{sg} \):

\[
F_s = AFA(\Delta I, G^k) \in \mathbb{R}^{32 \times 32 \times 512}
\]

where \( G^k \in \mathbb{R}^{32 \times 32 \times 512} \) is the \( k \)-th convolutional feature of \( G_{sg} \), for some fixed \( k \). Finally, we use the remaining layers of \( G_{sg} \), starting from \( k + 1 \), to obtain the triplane features that are used to reconstruct the input image. Thus, \( w^*_t \) and \( F_t \) together serve as a robust representation of \( I_t \) 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 \( \hat{w}^*_t \in \mathbb{R}^{14 \times 512} \). 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 \( w^*_t \) and the target environment map \( E_t \) to a latent offset \( \Delta w \):

\[
\Delta w = \mathcal{R}(w^*_t, E_t) \in \mathbb{R}^{14 \times 512}
\]
To obtain $\hat{w}_t^r$, we add this offset to $w_s^r$:

$$\hat{w}_t^r = w_s^r + \Delta w$$  \hspace{1cm} (6)

A relit image can be directly obtained by giving $\hat{w}_t^r$ as the input to $G_{dec}$. 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 $\hat{w}_t^r$ is a low-dimensional representation that is insufficient to capture the subject details. But, the AFA model cannot 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 $G_{sg}$ with $\hat{w}_t^r$ as input and extract the k-th convolutional feature $G_s^k$. Finally, we combine the convolutional features of the source and target image by following [Yao et al. 2022]:

$$F_t := F_s + C_t^k - C_s^k \in \mathbb{R}^{32 \times 32 \times 512}$$  \hspace{1cm} (7)

Thus, $\hat{w}_t^r$ 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 $G_{sg}$ with $F_t$ and perform a full forward pass as follows:

$$\hat{I}_t = G_{dec}(G_{sg}(\hat{w}_t^r, F_t), c)$$  \hspace{1cm} (8)

3.4 Loss Functions

We train R, while the networks $G_{sg}$, $G_{dec}$ 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 $w_s^r$ 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:

$$L_C = \| \hat{I}_t - I_t \|_1$$  \hspace{1cm} (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 $L_{LPIPS}$ [Johnson et al. 2016] between $\hat{I}_t$ and $I_t$.

$$L_{LPIPS} = \| \Phi_{vgg}(\hat{I}_t) - \Phi_{vgg}(I_t) \|_2^2$$  \hspace{1cm} (10)

Where $\Phi_{vgg}$ 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 $L_{LPIPS}$ for relighting.

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

$$L_{lat} = \| \hat{w}_t^r - w_s^r \|_2^2$$  \hspace{1cm} (11)

The total loss is given as:

$$L_{total} = \lambda_0 L_{lat} + \lambda_1 L_C + \lambda_2 L_{LPIPS}$$  \hspace{1cm} (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 x 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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM ↑</th>
<th>LD ↓</th>
<th>PSNR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeLF (3-views)</td>
<td>0.75</td>
<td>NA</td>
<td>19.72</td>
</tr>
<tr>
<td>PhotoApp</td>
<td>0.72</td>
<td>34.08</td>
<td>29.13</td>
</tr>
<tr>
<td>VoRF</td>
<td>0.69</td>
<td>16.90</td>
<td>20.21</td>
</tr>
<tr>
<td>NeRFFaceLighting</td>
<td>0.79</td>
<td>28.31</td>
<td>13.41</td>
</tr>
<tr>
<td>Lite2Relight</td>
<td><strong>0.83</strong></td>
<td><strong>9.76</strong></td>
<td><strong>28.3</strong></td>
</tr>
</tbody>
</table>

**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 $128 \times 128$ (see blurriness in results in Fig. 4). Lite2Relight, on the other hand, based on the triplane framework, can preserve fine details at resolution $512 \times 512$. Finally, VoRF requires a two-step optimization process that takes ~ 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 visualizes 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 $F_r$ and $G^f_r$, we compare our original method to a variant that does not replace $G^f_r$ by $F_r$. The results, as depicted in Fig. 9, reveal that the incorporation of $F_r$ 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 $F_r$ 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 light-stage, where subjects are relit under novel viewpoints.

<table>
<thead>
<tr>
<th></th>
<th>SSIM ↑</th>
<th>LD ↓</th>
<th>PSNR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o F-space</td>
<td>0.83</td>
<td>10.44</td>
<td>28.29</td>
</tr>
<tr>
<td>w/o $L_{LPIPS}$</td>
<td>0.83</td>
<td>10.12</td>
<td>28.32</td>
</tr>
<tr>
<td>Full Model</td>
<td>0.83</td>
<td>9.76</td>
<td>28.33</td>
</tr>
</tbody>
</table>

Importance of Perceptual Loss. It is feasible to supervise $R$ solely using $L_{mat}$ and $L_C$, omitting $L_{LPIPS}$, 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 in-accuracies 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 $L_{LPIPS}$, as suggested before [Johnson et al. 2016], ensures a more accurate relighting, as demonstrated in the column labeled “w/ $L_{LPIPS}$”. 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 Sup-Mat.

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.
Figure 4: Qualitative Results: Comparisons to Previous Works. We compare with NFL [Jiang et al. 2023], VoRF [Rao et al. 2023, 2022] and PhotoApp [BR et al. 2021]. 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.
Input VoRF VoRF Ours Ours

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.

Input GT with F without F

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 $G^k_f$. The results demonstrate that manipulating the feature code is crucial for achieving fine-grained identity preservation.

Input GT w/ $\mathcal{L}_{\text{LPIPS}}$ w/o $\mathcal{L}_{\text{LPIPS}}$

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.

Input View 1 View 2 View 3

Figure 6: Qualitative Results: Comparisons with PhotoApp [BR 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.

Input NeLF Ours GT

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.