International Journal of Computer Vision
https://doi.org/10.1007/s11263-022-01730-5

Free-view Face Relighting Using a Hybrid Parametric Neural Model on a SMALL-OLAT Dataset

Youjia Wang[1] · Kai He[1,3] · Taotao Zhou[1] · Kaixin Yao[1] · Nianyi Li[2] · Lan Xu[1,4] · Jingyi Yu[1,4]

Received: 31 January 2022 / Accepted: 21 November 2022

Abstract
The development of neural relighting techniques has by far outpaced the rate of their corresponding training data (*e.g.,* OLAT) generation. For example, high-quality relighting from a single portrait image still requires supervision from comprehensive datasets covering broad diversities in gender, race, complexion, and facial geometry. We present a hybrid parametric neural relighting (PN-Relighting) framework for single portrait relighting, using a much smaller OLAT dataset or SMOLAT. At the core of PN-Relighting, we employ parametric 3D faces coupled with appearance inference and implicit material modelling to enrich SMOLAT for handling in-the-wild images. Specifically, we tailor an appearance inference module to generate detailed geometry and albedo on top of the parametric face and develop a neural rendering module to first construct an implicit material representation from SMOLAT and then conduct self-supervised training on in-the-wild image datasets. Comprehensive experiments show that PN-Relighting produces comparable high-quality relighting to TotalRelighting (Pandey et al., 2021), but with a smaller dataset. It further improves shape estimation and naturally supports free-viewpoint rendering and partial skin material editing. PN-Relighting also serves as a data augmenter to produce rich OLAT datasets beyond the original capture.

Keywords 3D Reconstruction · Relighting · Neural rendering

1 Introduction

There are significant demands on synthesizing high-quality 3D faces with photorealistic lighting, textures, geometry, and motions. Applications are numerous, ranging from traditional photo retouching and enhancement (Wright, 2017; Pallant, 2011; Radke, 2013) to the latest meta-human creations (Hu et al., 2017; Ichim et al., 2015) in virtual and augmented reality. The two most popular streams of approaches are physical-based modelling and image/video synthesis. The former aims to directly model the physical properties of the materials (Smolyanskiy et al., 2014; Riviere et al., 2020), lighting (Chabert et al., 2006; Kanamori & Endo, 2018), facial movements (Shin et al., 2014; Feng et al., 2021), etc., along with accurate geometry for conducting photorealistic rendering. These approaches often require exquisite skills by artists to edit on diffuse and specular normal and albedo maps, which are too expensive for a broader audience. The latter, epitomized the USC LightStage (Debevec et al., 2000), can be viewed as a special category of image-based rendering: a performer's face is first captured under varying

Communicated by Boxin Shi, Ph.D.

✉ Youjia Wang
wangyj2@shanghaitech.edu.cn

Kai He
hekai@shanghaitech.edu.cn; hekai@deemos.com

Taotao Zhou
zhoutt@shanghaitech.edu.cn

Kaixin Yao
yaokx@shanghaitech.edu.cn

Nianyi Li
nianyil@clemson.edu

Lan Xu
xulan1@shanghaitech.edu.cn

Jingyi Yu
yujingyi@shanghaitech.edu.cn

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai, China

[2] Clemson University, Clemson, USA

[3] Deemos Technology Co., Ltd, Shanghai, China

[4] Shanghai Frontiers Science Center of Human-centered Artificial Intelligence, Shanghai, China

Published online: 07 January 2023

lighting conditions to produce an OLAT (One-Light-At-a-Time) dataset that can be subsequently used to synthesize any new lighting conditions. Benefiting from comprehensive OLAT datasets, recent learning, in particular, neural rendering, techniques have further enabled single image portrait relighting (Kanamori & Endo, 2018; Nestmeyer et al., 2020; Wang et al., 2020; Pandey et al., 2021), with a quality comparable to physical-based approaches.

Despite all these advances, several fundamental challenges remain. For physical-based approaches, tremendous efforts have been focused on inferring 3D models and physical properties from real images, to reduce, if not fully eliminate, editing requirements. Latest learning-based techniques (Wang et al., 2020; Zhou et al., 2019; Nestmeyer et al., 2020; Wang et al., 2020) can produce reasonable 3D geometry but still fall short of high quality normal, reflectance, and lighting maps. Consequently, one can still easily tell real from synthesized results. An advantage there though is that the estimated models are parametric and therefore they can be adjusted in shape and movement to support free-viewpoint rendering. For neural image synthesis, the key challenge is the lack of datasets: quality OLAT data are scarce in public and the very few available ones are small in size. In contrast, to ensure quality rendering, the recent Total Relighting (Pandey et al., 2021) exploits 78 elaborately chosen subjects to cover diversities in gender, race, age, and skin complexion, with a total of over 2 million training images. Producing comparable quality results with a much small dataset is difficult but highly desirable.

In this paper, we present a hybrid parametric-neural relighting (PN-Relighting) technique for high-quality portrait relighting from a single image (Fig. 1). In a nutshell, PN-Relighting combines the benefits of the physical and image-based approaches via two core modules: appearance inference and neural relighting. It starts with an estimated parametric 3D model using well-known techniques such as 3DMM (Blanz & Vetter, 1999). The appearance inference module then adopts a learning-based scheme to infer detailed surface normals and albedo textures and to refine the parametric face. The neural relighting module constructs an implicit neural representation to reflectance (material) and combines it with the fine-scale parametric face for relighting. To eliminate the requirement of using large OLAT datasets, we use the procedures above to form a pseudo-albedo dataset to enrich the diversity of OLAT. We also adopt self-supervised training on in-the-wild image datasets to improve robustness. In particular, the implicit skin material representation accounts for variations in complexion, supporting more accurate relighting and partial material editing. Figure 2 shows the pipeline of our method.

The advantages of PN-Relighting over the state-of-the-art are multi-fold. On relighting, it uses a much smaller OLAT dataset (that we call SMOLAT), with only 30 subjects

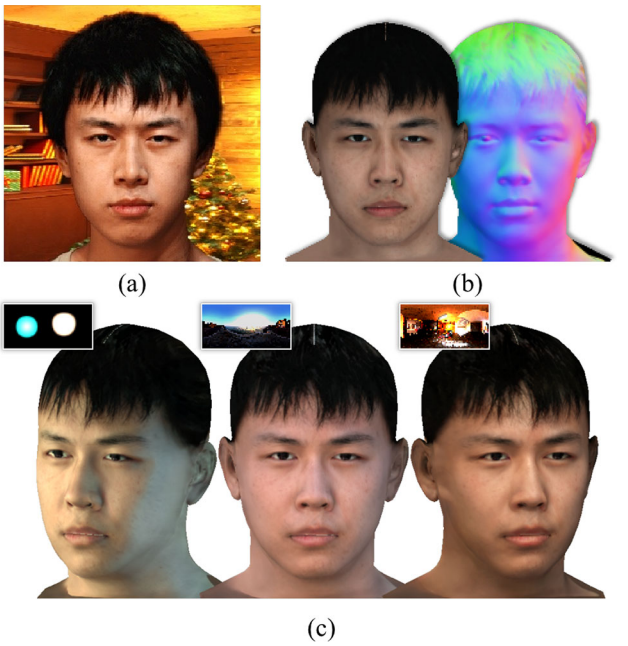

Fig. 1 We present a hybrid parametric-neural relighting (PN-Relighting) technique. Taking a single portrait image as input (**a**), we generate the surface geometry and albedo (**b**) and a free-view 3D face relightable under different illumination (**c**)

covering much fewer variations in appearance than Pandey et al. (2021). By applying the self/semi-supervised training technique to SMOLAT and our synthesized Pseudo-Albedo dataset from FFHQ (Karras et al., 2019), PN-Relighting produces realistic relighting comparable to using heavier OLAT datasets. On geometry estimation, by using a deep material model under a differentiable rendering pipeline, PN-Relighting further improves normal and albedo estimations in accuracy and robustness. On free-view rendering, PN-Relight builds upon parametric shapes, which can sustainably benefit the current technical trend of leveraging parametric models to boost portrait relighting, especially considering the limited access to high-quality lighting training data. Our approach adds another layer of sophistication to emerging 3D-aware generative models (Sela et al., 2017; Gecer et al., 2019; Lattas et al., 2021). Finally, PN-Relighting enables OLAT data augmentations, by producing strategically designed lighting patterns on in-the-wild portrait images as if they were captured in a LightStage.

To summarize, our main contributions include:

– We propose a novel neural pipeline, PN-Relighting, to produce high-quality relightable and render-ready 3D face models by only taking monocular RGB portrait images as inputs. It supports multi-scale 3D face geometry estimation, high-quality portrait relighting, and free-viewpoint rendering.

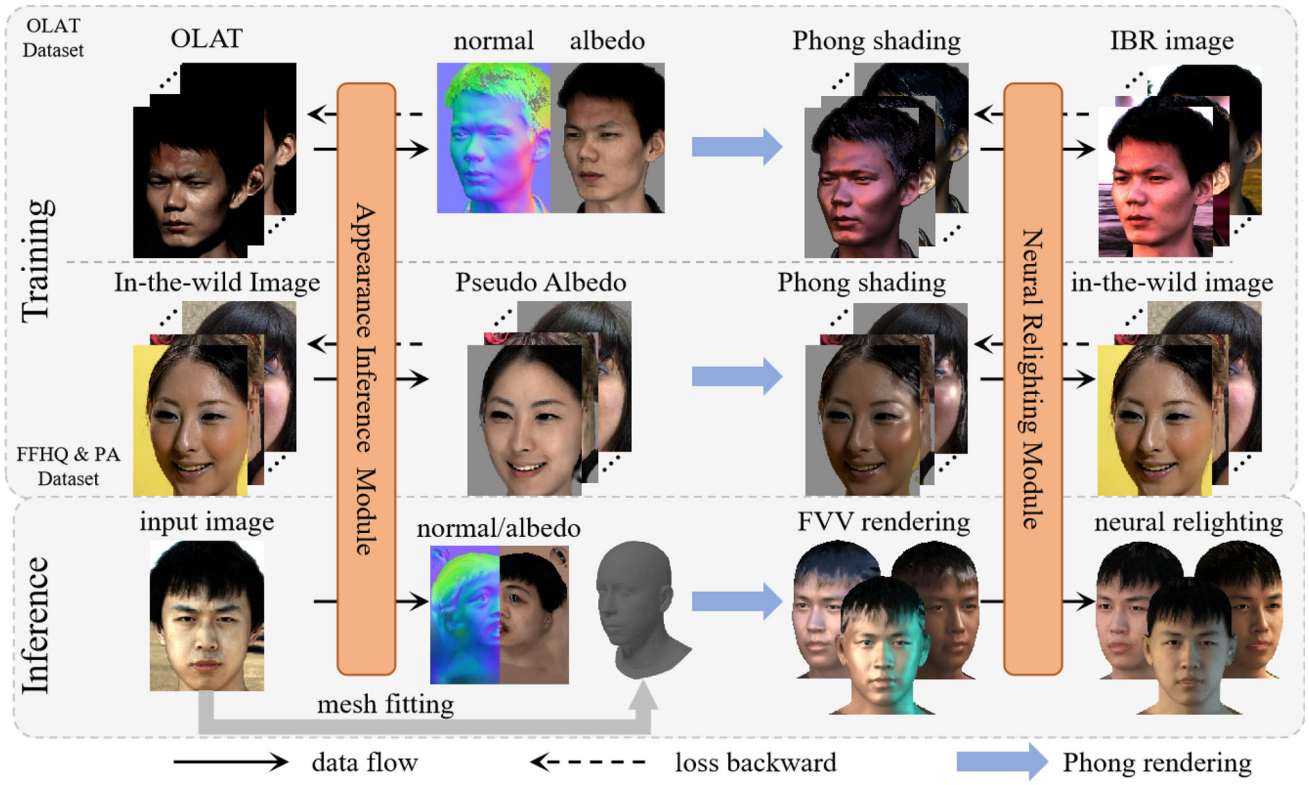

Fig. 2 The overall architecture of our PN-Relighting method. Our network is trained only using small OLAT dataset (SMOLAT) and a few hundreds of in-the-wild images (a sub-set from FFHQ) while achieving realistic relighting effect on 3D faces

– We employ a parametric-neural model to account for shape estimation, neural relighting, and implicit deep material modelling under a differentiable rendering pipeline. More importantly, we use a SMOLAT dataset of a much smaller scale than the state-of-the-art OLAT and conduct self-/semi-supervised learning to achieve comparable relighting quality on in-the-wild images.
– In addition to neural portrait relighting, PN-Relighting further enables facial material editing to support complexion adjustment. It can also be used for OLAT data augmentation.

2 Related Work

Reconstructing 3D faces from single or multiple image inputs has been thoroughly studied over the past few decades. State-of-the-art approaches aim to exploit various types of visual inputs, ranging from video frames (Garrido et al., 2016; Ichim et al., 2015; Jeni et al., 2015; Shi et al., 2014), to multi-view RGB (Cao et al., 2018; Beeler et al., 2010), and RGB-D data (Thies et al., 2015; Li et al., 2013), and to photo collections (Roth et al., 2016). In this work, we only review the most relevant ones, *i.e.,* 3D facial generation using a single RGB image as input and subsequently conducting relighting.

2.1 3D Face Reconstruction

Approaches for reconstruction from a single portrait image can be generally classified as parametric vs. non-parametric methods. Parametric methods model 3D faces by transforming the shape and texture of the facial features into a vector space, e.g., 3DMM (Blanz & Vetter, 1999), and reconstruct the 3D face geometry by fitting the learned model to the input data (Genova et al., 2018; Shang et al., 2020; Guo et al., 2020). Such morphable models can provide statistical information on physiologically sound head shapes and expression alignment (Booth et al., 2018; Dai et al., 2020; Li et al., 2017; Cao et al., 2018), and can be easily fitted into statistical linear model only using RGB data for optimization (Thies et al., 2016; Zollhöfer et al., 2018). A downside though is that parametric methods estimate the face model within a fixed linear shape space where optimization can lead to a local minimum, resulting in overly-smooth reconstructions. To overcome such limitations, (Jiang et al., 2018; Li et al., 2018; Riviere et al., 2020) extend the shape variants by fitting the parametric face model to input data, and leverage a shape from shading (SfS) method to reconstruct facial details from single RGB images. Nevertheless, these approaches have degraded performance under occlusions or viewing angle changes.

The problem is particularly severe when they are applied to the "in-the-wild" images (Yang et al., 2020) since most of these approaches rely on detailed 3D scans as training guidance. Feng et al. (2021) introduced a regression-based DECA (Detailed Expression Capture and Animation) approach to learning an animatable displacement model from in-the-wild images without 2D-to-3D supervision. This parametric face model is built on FLAME (Li et al., 2017), and can significantly restore wrinkles along with expression change. Different from parametric techniques, non-parametric methods (Jackson et al., 2017; Dou et al., 2017; Feng et al., 2018; Alp Guler et al., 2017; Wei et al., 2019; Zhu et al., 2020) directly predict 3D faces using voxels or meshes, and manage to recover fine shape details compared with the parametric models. However, they still need strong supervision from explicit 3D shapes, commonly acquired by synthesized facial data with limited shape variances. We refer readers to Zollhöfer et al. (2018) for an overview of the state-of-the-art 3D face reconstruction.

2.2 Portrait Relighting

Related to face reconstruction is the problem of portrait relighting. Many existing approaches have followed the seminal work of Debevec et al. (2000) that uses a LightStage to capture one-light-at-a-time (OLAT) faces under varying lighting conditions and subsequently conducts realistic relighting to faces under arbitrary high dynamic range (HDR) lighting environment. The LightStage-based approaches (Sagar, 2005; Chabert et al., 2006; Xu et al., 2019; Meka et al., 2019) have demonstrated robust performance on photorealistic illumination rendering (Sagar, 2005) on both static (Xu et al., 2019; Bi et al., 2020) and moving subjects (Meka et al., 2019; Chabert et al., 2006). (Zhang et al., 2021) achieves human body free-view relighting by 6D light transport function. However, it was specifically designed to relight the performer who was pre-captured within the LightStage and cannot readily extend to other subjects.

The advent of deep learning has introduced many hardware-free approaches for single image portrait relighting. Kanamori and Endo (2018) directly predicted the albedo, illumination, and an occlusion-encoded light transport map to inverse rendering the human body. However, their method downgrades quickly on specular reflectance (e.g., oily skins) or under high-frequency illumination. Zhou et al. (2019) uses synthetic data as supervision and employs a Spherical Harmonics (SH) lighting model (Basri & Jacobs, 2003) for face relighting. Their method, however, loses details due to the low-frequency nature of SH rendering. Sun et al. (2019) improved this method by estimating the illumination of input portrait, achieving plausible performance in a low-frequency lighting environment. However, it still suffers from hard shadows and specular highlight problems on human faces. Nestmeyer et al. (2020) explicitly models multiple reflectance channels of facial albedo, geometry, and lighting effects to partially account for the rendering of specularity and shadows. The technique mainly focuses on directional illumination. Wang et al. (2020) improved the method by using synthetic renderings of 3D photogrammetry scans to supervise relighting training while learning the diffuse and specular components of reflectance at the same time. They can handle non-Lambertian effects but fall short of reducing artifacts caused by errors in pixel-aligned illuminations.

The seminal work of TotalRelighting Pandey et al. (2021) produces unprecedented photorealism with the newly replaced background. It uses light maps as pixel-aligned lighting representation and demonstrates excellent performance in handling high-frequency self-shadowing effects, and specularities on faces, as well as a generalization to real-world portraits. However, similar to many data-driven approaches, it requires using heavy OLAT data. For example, in TotalRelighting, a comprehensive dataset of 78 subjects of different gender, race, skin complexion, etc, was used, accumulating over two million images in total. By far only a very small number of groups are capable of producing such comprehensive data. In this work, we also construct a mini LightStage. In contrast to using a very large dataset, we demonstrate how to use a small dataset to achieve equivalent relighting quality, by employing a hybrid parametric-neural method. In addition, our approach supports free-viewpoint viewing and partial material editing, largely missing in the prior art.

3 Overview

Given a single RGB portrait image $\mathbf{I}$, and an arbitrary HDR lighting environment $\mathbf{E}$, we set out to reconstruct a neural avatar $\mathcal{M}$ that allows for free-viewpoint rendering in arbitrary lighting environment. As shown in Fig. 2, our method consists of two consecutive modules: an Appearance Inference module and a Neural Relighting module. The appearance module inferences the intrinsic image components, *i.e.,* surface normal and albedo, from $\mathbf{I}$. Given a lighting environment $\mathbf{E}$, we project the predicted normal and albedo to its corresponding diffuse and specular reflection components of face appearance using Lambertian and Phong reflectance lobes (Phong, 1975). Next, the neural relighting module transforms $\mathbf{I}$ to an implicit neural latent vector to encode the face material, which is then used to obtain the final relighted avatar $\mathcal{M}$ by taking previously estimated diffuse and specular components as input.

With respect to datasets, we train our hybrid parametric-neural model using two datasets:

3.1 Small-OLAT Dataset

To get accurate 3D face models, we use the Dynamic OLAT dataset Zhang et al. (2021) that provides $\sim$ 600k OLAT images with 2, 810 HDR environment lighting maps for supervision. We follow the ground truth generation pipeline in Pandey et al. (2021) to acquire accurate surface normal, albedo, and the paired portraits of different subjects lit in various lighting environments with ground truth illumination from the OLAT dataset. However, due to the limited face variance of this small OLAT dataset, the trained model has degraded relighting performance on images with unconstrained lighting environments.

3.2 "In-the-wild" Dataset

To better generalize our method on images captured in an arbitrary recording conditions, we synthesize a Pseudo-albedo (PA) dataset from FFHQ dataset Karras et al. (2019) to better infer the albedo information from under-controlled lighting environments. Then, we randomly pick another non-overlapping subset from FFHQ dataset to train the Neural Relighting Module, so that it can map the estimated diffuse and specular components to the original "in-the-wild" images. To generate PA dataset, we first train the appearance inference module on FFHQ datasets in a self-supervised manner (Sun et al., 2019) to generate pseudo-surface normal and pseudo-albedo. However, due to the lack of ground truth as a strong constraint, the predicted face geometry and albedo are inaccurate in certain lighting environments. To address this issue, we selected the top-2% data of best visual correctness and manually removed noticeable highlights on the pseudo-albedo maps. The detailed data selection of PA and "in-the-wild" training is described in Sect. 5.

The rest of the paper is organized as follows: We introduce our Appearance Inference module in Sect. 4.1, and the Neural Relighting module in Sect. 4.1.3. We present our parametric 3D face model enabling free-view rendering in Sect. 4.2. Next, we show our training details and loss functions in Sect. 5. We extensively evaluate our approach on different datasets and show outperforming results compared to state-of-the-art in Sect. 6, followed by a short discussion of our limitations in Sect. 7.

4 Relightable 3D Face Generation

4.1 Appearance Inference Module

For each portrait image, we first preprocess it with the recent subject segmentation algorithm Ke et al. (2022) to mask out the background. The appearance inference module, thus, decomposes the foreground portrait image $\mathbf{I}$ to intrinsic image components, *i.e.,* the surface normal $\hat{N}$, and albedo $\hat{A}$. Specifically, we use Normal Network, resembling the structure of U-Net Ronneberger et al. (2015), to regress $\mathbf{I}$ to $\hat{N}$. Then, we feed the composited image with normal $\{\mathbf{I}, \hat{N}\}$ to Albedo Network, and generate a diffuse albedo image $\hat{A}$.

4.1.1 Normal Network

$\mathbf{\Psi}_N$. Our normal subnet takes a background-free portrait image to infer the surface geometry $\hat{N}$, which encodes the per-pixel normals. It uses the encoder-decoder structure to generate intrinsic features. The encoder consists of stacked convolutional layers with max-pooling layers. The decoder is composed of transposed convolutional layers with skip connections. We train this normal net with the normal loss functions described in Sect. 5.2.

4.1.2 Albedo Network

$\mathbf{\Psi}_A$. Our Albedo subnet predicts the diffuse albedo map $\hat{A}$ from input image $\mathbf{I}$ and the predicted surface normal $\hat{N}$. We concatenate $\hat{N}$ and $\mathbf{I}$, and feed the composited vector to another encoder-decoder network with the same architectures as Normal net. The loss functions are described in Sect. 5.2.

4.1.3 Neural Relighting Module

Our Neural Relighting Module aims to relight an image that matches an HDR lighting environment $\mathbf{E}$. In particular, this module has two subnets: $\mathbf{\Psi}_M$ generates a material latent vector from $\mathbf{I}$; and $\mathbf{\Psi}_R$ outputs the relit portrait $\hat{\mathbf{I}}_E$ or 3D avatar $\mathcal{A}$ from Phong priors $\{\mathcal{P}_n | n = 1, 2...\}$, as shown in Figs. 3 and 4.

4.1.4 Phong Priors

For time efficiency, we apply a Phong shading based method (Pandey et al., 2021) on $\mathbf{E}$ and produce a set of prefiltered environment maps with four different specular exponents $\{n = 1, 16, 32, 64\}$. Therefore, we can easily compute the diffuse and specular reflectance images, or Phong Priors, by indexing into these prefiltered light maps using the surface normal $\hat{N}$. Please refer to Phong (1975) for details about Phong Shading and the specular exponents' formulation. We also take the albedo image as a component of Phong Priors.

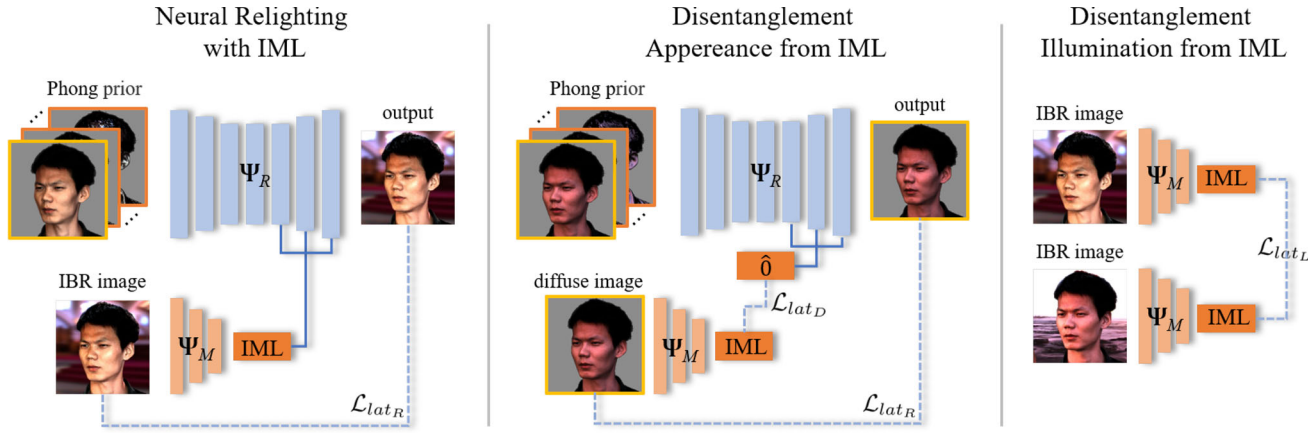

Fig. 3 The training of material network Ψ_M. We use three loss functions $\mathcal{L}_{lat_R}$, $\mathcal{L}_{lat_D}$ and $\mathcal{L}_{lat_L}$ to learn the neural implicit material latent (IML) vector, so that IMLs of the same subject are consistent on different illuminations

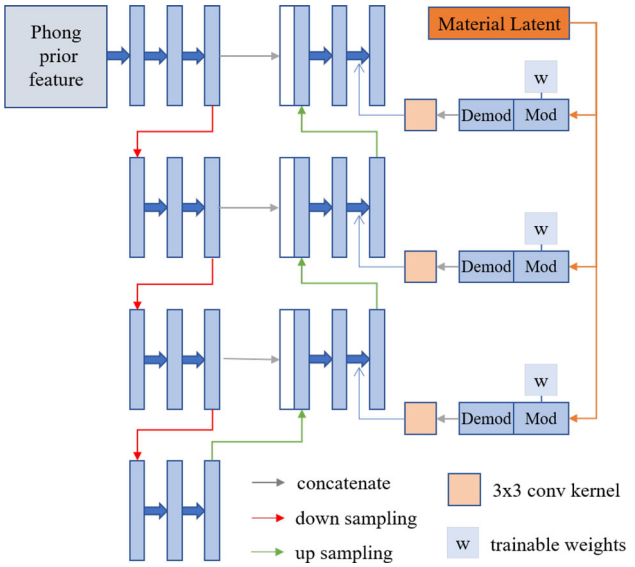

Fig. 4 The architecture of our neural relighting network Ψ_R. We use modulation (Mod) and demodulation (Demod) operators from StyleGan (Karras et al., 2020) to transform the material encoded latent to learnable weights of each layer of the decoder

4.1.5 Material Network Ψ_M

Even though these Phong priors can represent the color and lighting information of the portrait, very limited work aims to infer material properties from the data. To model variations of portrait materials, we propose Ψ_M to generate a material encoded vector $\hat{M}$ to embed the face material information into the training of neural relighting network Ψ_R. We use a common set of encoding layers to construct Ψ_M taking $\mathbf{I}$ as input, as shown in Fig. 3. This implicit skin material representation accounts for variations on complexion, supporting more accurate relighting and partial material editing. The loss functions can be found in Sect. 5.2.

4.1.6 Neural Relighting Network Ψ_R

Our relighting subnet takes the Phong priors $\{\mathcal{P}_n\}$ as input, and generate a relight image $\hat{\mathbf{I}}_E$ by incorporating the material information from Ψ_M. Figure 4 shows the structure of our relighting network. To learn details of local lighting features, we use the revised style block Karras et al. (2019) with demodulation operator Karras et al. (2020) to obtain the weight of the convolution kernel from the implicit material latent vector $\hat{M}$. We inject this block into each layer of the decoder so that the material information can be better exploited in relighting portrait data. The loss functions and training details are described in Sect. 5.2.

4.2 Free-View Relightable Facial Avatar

In the inference stage, we aim to obtain a portrait avatar that supports free-viewpoint and arbitrary lighting rendering. Specifically, we use FLAME (Li et al., 2017) to build a statistical 3D head model $\mathcal{M}$ from the surface normal $\hat{N}$ and albedo $\hat{A}$. First, we fit the parametric 3D face model from FLAME onto the input image by using a ResNet-based algorithm (Feng et al., 2021). Since the 3D face is a parametric model, we can ensure that the generated mesh is topologically consistent. Next, based on the mapping function provided by FLAME, we use grid sampling to infer the UV samples according to the 2D image-space albedo and normal. To fix the occlusions, we use the depth buffer to calculate the occlusion map. We then apply an inpainting algorithm Suvorov et al. (2022) on the occluded area, to get a complete UV-space albedo and normal map. At this point, we can fast index the albedo and normal given any viewpoints.

To enable relighting effect on the 3D face, we directly extend our relighting pipeline on 2D images to 3D, as long as we can get accurate UV-albedo and normal maps from PN-Relighting. Specifically, we use Phong Shading functions to

Table 1 We trained our network on different datasets

Network	Ψ_N	Ψ_A	Ψ_R	Ψ_M
SMALL-OLAT	✓	✓	✓	✓
Pseudo Albedo		✓	✓	
FFHQ Dataset			✓	

This table shows which datasets are used for each network separately

generate Phone priors to infer the environment illumination. Next, we use the material encoder to get the IML of the portrait. Finally, we use the relighting network with the encoded IML to produce the final free-view relight avatar (Table 1).

5 Training and Loss Functions

Our hybrid parametric-neural face is first trained on SMOLAT that only captures the faces of a small number of individuals. Brute-force training leads to degraded performance on a subject in an unconstrained environment. To generalize our algorithm on images captured in arbitrary recording conditions, we synthesize a Pseudo-albedo (PA) dataset using portrait images without the ground truth geometry and illumination to further constrain the training of Albedo Network Ψ_A. Additionally, we leverage a subset from FFHQ, called sub-FFHQ, to constrain the training of Neural Relighting Network Ψ_R to further boost our relighting performance on the "in-the-wild" images.

Table 5 shows which datasets we used to train our network. The training details of these networks will be described in the following sub-sections.

5.1 Pseudo-Albedo Generation

To further boost our relighting network using the in-the-wild data in a self-supervised manner, we introduce a novel scheme to generate pseudo albedo, normal and environment illumination for a subset of the FFHQ dataset.

Note that it's extremely ill-posed and difficult to obtain the actual ground truth per intrinsic component. Thus, we adopt the Phong model as a strong prior to mitigate the ambiguity between intrinsic components.

Specifically, we adopt the Normal and Albedo Networks trained on the OLAT dataset to obtain the initial normal and albedo for each input image. For environment illumination, we use the Spherical Harmonic (SH) Lighting coefficient and formulate an optimization problem to obtain the SH coefficients $\boldsymbol{\omega}_{sh}$ for the input image $\mathbf{I}$ from FFHQ dataset:

$$\boldsymbol{\omega}^*_{sh} = \arg\min_{\boldsymbol{\omega}_{sh}} \text{MSE}(\tilde{A} \cdot \mathcal{P}(\tilde{N}, \boldsymbol{\omega}_{sh}), \mathbf{I}) \tag{1}$$

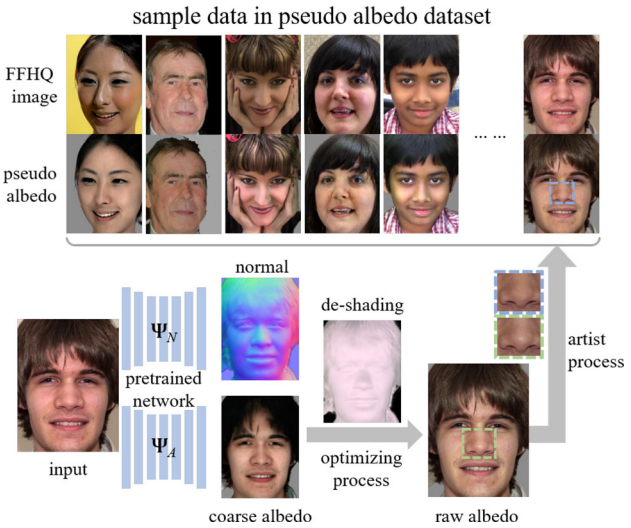

Fig. 5 For Pseudo Albedo dataset, we manually remove the specular highlights on our selected Pseudo Albedo maps

where $\mathcal{P}(\cdot)$ is the Phong Spherical Harmonic (SH) shading function Phong (1975). To solve the above optimization problem, we use the Adam optimizer and set the learning rate to 0.01.

Then, based on the optimized environment illumination $\boldsymbol{\omega}^*_{sh}$, we construct our Pseudo-Albedo dataset. Specifically, as shown in Fig. 5, for every portrait image in the FFHQ dataset, we use normal and albedo network to generate a coarse albedo $\tilde{A}$ and normal map $\tilde{N}$, and assign a score $\mathcal{S}_I$ to each image according to the normal and albedo information:

$$\mathcal{S}_I = \text{MSE}(\tilde{A} \cdot \mathcal{P}(\tilde{N}, \boldsymbol{\omega}^*_{sh}), \boldsymbol{I}). \tag{2}$$

We then ascending sort the images according to $\mathcal{S}_I$ and choose the first 2% data (∼300 images) with the lowest scores to construct the new dataset. For these selected data, we generate the pseudo-albedo A_I for each image as ground truth:

$$A_I = \boldsymbol{I} / \mathcal{P}(\tilde{N}, \boldsymbol{\omega}_{sh}) \tag{3}$$

However, due to the low-frequency property of $\mathcal{P}$, there're still a certain amount of specular highlights left on our constructed pseudo-albedo. Here, we simply manually remove the highlights on pseudo-albedo by image editing tools.

Such disentanglement and our highlight-removing operation further guarantee the effectiveness of the subsequent optimization of the SH Lighting coefficient. Note that such manual annotation to those highlight regions is accurate enough for our optimization.

5.2 Loss Functions

5.2.1 Normal Loss

We use a per-pixel normal loss to compare our predicted normal map $\hat{N}$ with the ground truth one N:

$$\mathcal{L}_N = \frac{1}{K}\sum_{p\in \boldsymbol{I}}(1 - cos(\hat{N}_p, N_p)) \tag{4}$$

where p is the pixel in $\boldsymbol{I}$, K is the total number of pixels in $\boldsymbol{I}$, and $cos(\cdot)$ is the per-pixel cosine distance.

5.2.2 Albedo Loss

We use the MSE (mean square error) loss and SSIM (structural similarity index) loss to measure the difference between our predicted albedo $\hat{A}$ and the ground truth one A:

$$\mathcal{L}_A = \text{MSE}(\hat{A}, A) + \text{SSIM}(\hat{A}, A) \tag{5}$$

Note that, the Albedo Network is trained by a mixed dataset of OLAT and PA so that the small PA dataset can compensate for the variants of people identity of OLAT data while the OLAT dataset can improve the albedo prediction's accuracy on "in-the-wild" images. For this reason, the network is capable of decomposing the albedo from light color by taking advantage of high variant light conditions from OLAT data.

5.2.3 Material Loss

Since there lacks explicit ground truth material information from the OLAT dataset, we design a material loss to enforce the consistency among the predicted material latent vectors $\hat{M}$ of the same subject. The insight is that the facial material of the same individual should be independent of the appearance features, as well as lighting environments:

$$\mathcal{L}_M = \mathcal{L}_{lat_R} + \mathcal{L}_{lat_D} + \mathcal{L}_{lat_L}, \tag{6}$$

where $\mathcal{L}_{lat_R}$ is the loss term to ensure that the input portrait $\boldsymbol{I}$ is consistent with the output of relighting network $\hat{\boldsymbol{I}}_E$ when the lighting condition $\boldsymbol{E}$ is the same:

$$\mathcal{L}_{lat_R} = \text{MSE}(\hat{\boldsymbol{I}}_E, \boldsymbol{I}). \tag{7}$$

$\mathcal{L}_{lat_D}$ is to enforce $\hat{M}$ to be a zero vector when there are no specular components in $\boldsymbol{E}$:

$$\mathcal{L}_{lat_D} = \text{MSE}(\hat{M}_D, \mathbf{0}) + \text{MSE}(\hat{\boldsymbol{I}}_D, \boldsymbol{I}_D), \tag{8}$$

where $\hat{M}_D$ is the material encoded vector taking the diffuse image $\boldsymbol{I}_D$ as input. $\hat{\boldsymbol{I}}_D$ is the relit results only with the diffuse components.

$\mathcal{L}_{lat_L}$ is to make sure the material consistency of the same subject under different lighting environments:

$$\mathcal{L}_{lat_L} = \text{MSE}(\hat{M}_{E_i}, \hat{M}_{E_j}) \tag{9}$$

where $\hat{M}_{E_i}$ and $\hat{M}_{E_j}$ is the material encoded vector in different lighting environment E_i and E_j , respectively . Figure 3 shows the training process using $\mathcal{L}_{lat_R}$, $\mathcal{L}_{lat_D}$, and $\mathcal{L}_{lat_L}$.

5.2.4 Neural Relighting Loss

For the relighting network, we use three loss terms for supervision:

$$\mathcal{L}_R = \mathcal{L}_C + \mathcal{L}_{vgg} + \max_{\boldsymbol{\Psi}_R} \mathcal{L}_{adv}. \tag{10}$$

$\mathcal{L}_C$ compares the relit results $\hat{\boldsymbol{I}}_E$ with the ground truth one $\boldsymbol{I}_E$:

$$\mathcal{L}_C = \text{MSE}(\hat{\mathbf{I}}_E, \boldsymbol{I}_E) + \text{SSIM}(\hat{\mathbf{I}}_E, \boldsymbol{I}_E). \tag{11}$$

$\mathcal{L}_{vgg}$ measures the MSE between features extracted from the relit results $\hat{\boldsymbol{I}}_E$ and the ground truth one $\boldsymbol{I}_E$ using a pre-trained VGG network on the ImageNet:

$$\mathcal{L}_{vgg} = \text{MSE}(vgg(\hat{\boldsymbol{I}}_E), vgg(\boldsymbol{I}_E)). \tag{12}$$

$\mathcal{L}_{adv}$ is an adversarial loss to encourage the relighting network $\boldsymbol{\Psi}_R$ generating photorealistic results:

$$\mathcal{L}_{adv} = \mathbb{E}[\log \mathcal{D}(\boldsymbol{I}_E)] + \mathbb{E}[1 - \log \mathcal{D}(\hat{\boldsymbol{I}}_E)] \tag{13}$$

where $\mathbb{E}[\cdot]$ is the expectation function, and $\mathcal{D}$ is a discriminator from the original GAN (Goodfellow et al., 2020).

5.2.5 Total Loss

We train all subnets in an end-to-end manner, and the total loss function for our relighting network is defined as the combinations of the above-described losses:

$$\mathcal{L}_{PNR} = \lambda_1\mathcal{L}_N + \lambda_2\mathcal{L}_A + \lambda_3\mathcal{L}_M + \lambda_4\mathcal{L}_R, \tag{14}$$

where $\lambda_{1,\ldots,4}$ are weighted factors and they are separately tuned for each subnet.

5.3 Relighting Network Training on In-the-wild Dataset

For each image $\boldsymbol{I}$ in the sub-FFHQ training set, we first estimate the normal N_I and albedo A_I using our pre-trained normal network $\boldsymbol{\Psi}_N$ and albedo network $\boldsymbol{\Psi}_A$, and generate the spherical harmonic (SH) illumination for each image in the same way as mentioned in Sect. 5.1 using Eq. 1.

Then, we utilize the lighting condition as prior and assume that the relit result from the estimated Phong priors should be equal to the input image. Due to the lack of ground-truth lighting, we only apply the self-supervising loss $\mathcal{L}_{lat_R}$ (Eq. 7) in the training of material network $\boldsymbol{\Psi}_M$. Note that we formulate the whole process in a self-supervised manner and do not require ground truth lighting when trained with sub-FFHQ dataset.

6 Experimental Results

We conduct comprehensive experiments using PN-Relighting for a number of tasks. First, we provide a detailed description of the dataset we use and provide our training details. Next, we evaluate our method qualitatively and quantitatively from three aspects: portrait appearance, portrait relighting, and novel view synthesis. We compare our method with competitive state-of-the-art methods as well as perform ablation studies to evaluate separate modules in PN-Relighting.

6.1 Training Details

We train PN-Relighting on a Linux cluster with two AMD EPYC 7742 CPUs, 16 × 64 GB RAM, and NVidia A6000 GPU with 48G memory. We set the parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ = $\{0.1, 0.1, 0.01, 1\}$ and $\{0, 0.1, 0, 1\}$ for our total loss functions on SMOLAT and in-the-wild dataset respectively. Note that there's no ground truth to guide the training of Normal and Material networks on the in-the-wild dataset. We therefore fixed the parameters of these two subnets during the training by setting the weights of corresponding loss terms to zero. We use the Adam optimizer with a learning rate of 10^{-4}. It takes around 24 hours ($\sim$ 1 day) to train our network on the SMOLAT, compared with TotalRelighting (Pandey et al., 2021) , which takes 7 days.

Once the training on SMOLAT done, we then add the Pseudo-Albedo and FFHQ dataset to the training of the networks, with probability of occurrence 0.05 and 0.1 respectively. This procedure takes around 24 hours ($\sim$ 1 day) to reach convergence. In total, we take about 48 hours to train our network.

For data augmentation, we perform regular augmentation strategies on the input, including color adjustment, image shifting, and image re-scaling.

6.2 Datasets

We train PN-Relighting on two datasets: SMOLAT from (Zhang et al., 2021), and Pseudo-Albedo Dataset from FFHQ (Karras et al., 2019).

6.2.1 SMOLAT

Zhang et al. (2021) captured this OLAT dataset for portrait relighting using video as input. They used an ultra-high speed camera to capture OLAT images of 36 subjects with 2810 HDR environment lighting maps.

For each frame of OLAT data, it contains 114 light positions, corresponding to 114 images each. The dataset contains a total of 603,288 frames of OLAT data. We split the dataset into a training set and a test set by the subject's identity. The training data contains 30 individuals and we only show results on the rest 6 subjects (unseen in training) in this section. For each frame of the OLAT data, we obtain its normal by photometric stereo method (Woodham, 1980) and use it as ground truth for training. Unlike TotalRelighting which uses a full-light image as an albedo, we generate ground truth albedo by photometric stereo to better filter out the specular effects, and thus providing higher appearance fidelity for portrait relighting. We use all the HDR environment illumination provided by SMOLAT to generate the ground truth training pair of Phong prior and image-based rendering under different illumination for our Neural Relighting networks. Figure 6 shows our relit results on SMOLAT.

6.2.2 Pseudo-Albedo Dataset

We have described the construction of Pseudo-Albedo dataset in Sect. 5. It contains 300 images in total, and each has a pseudo albedo map from Eq. 3 as training guidance. We use 250 images as training set and evaluate our methods on the rest 50 images. In Sect. 6.4 we show that this dataset improves the performance of our albedo-network on the in-the-wild data.

6.2.3 Sub-FFHQ Dataset

In addition to the PA dataset, we collect a sub-FFHQ to train the Neural Relighting Network so that our method can be generalized to "in-the-wild" images. Specifically, we collected about 50k images for training, and randomly picked about 1k images for testing. None of images from sub-FFHQ overlap with the PA dataset. Our ablation study in Sect. 6.4 shows that adding the sub-FFHQ dataset into the training procedure can help to preserve the faithful appearance realism and image sharpness.

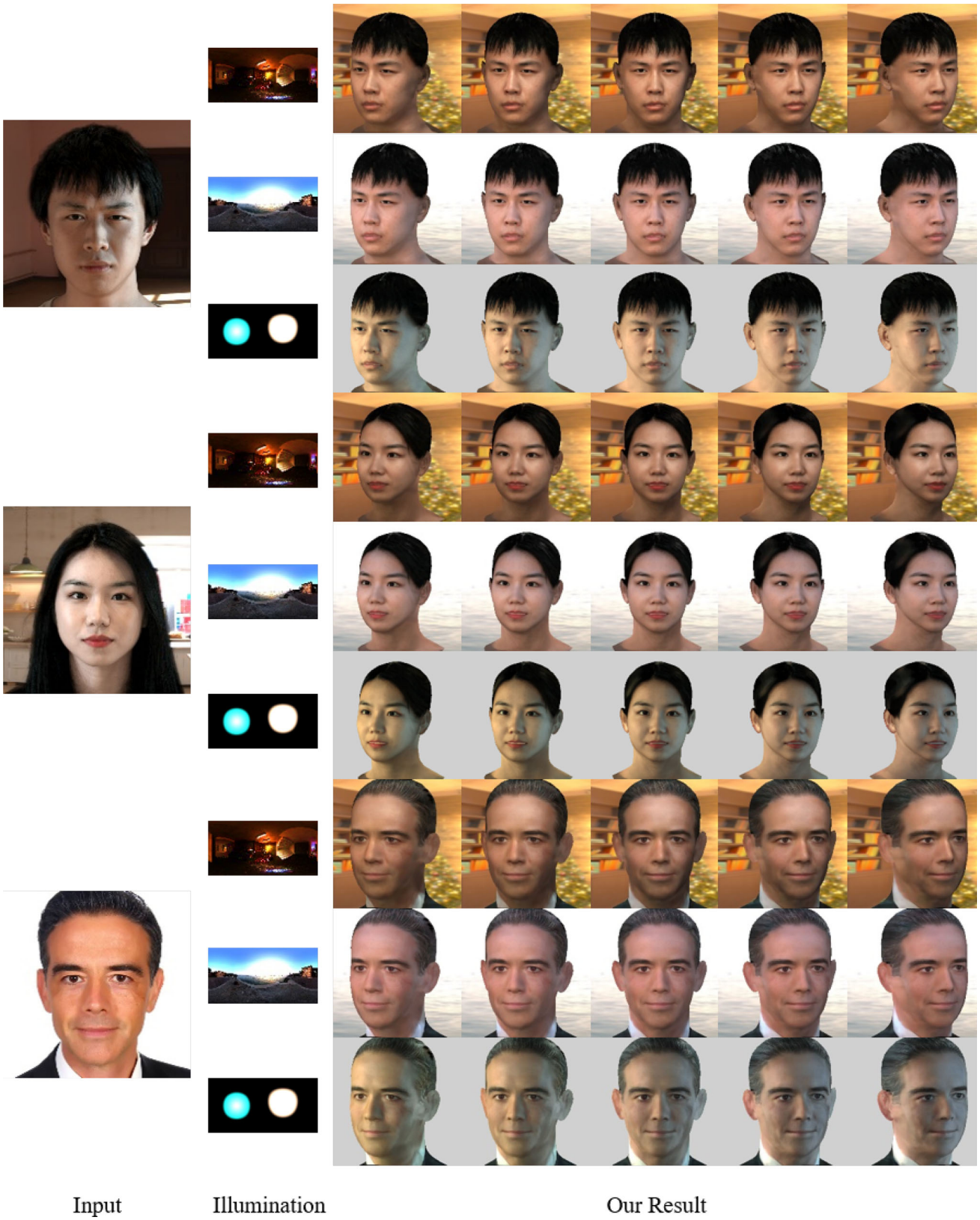

Fig. 6 Our results from a single input image and an arbitrary HDR image. We demonstrate the result in 5 different perspective of view

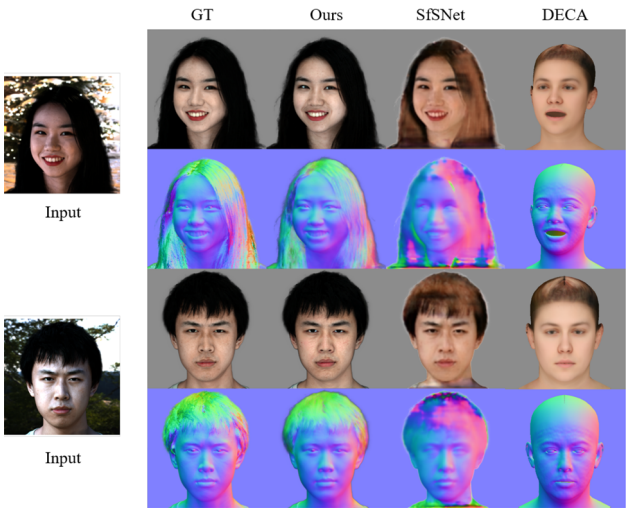

Fig. 7 Qualitative comparison of appearance reconstruction results on SMOLAT dataset, with SfSNet (Sengupta et al., 2018) and DECA (Feng et al., 2021). From left to right: Input: input image; GT: albedo and normal by photometric-stereo method on OLAT data; the result of ours, SfSNet, DECA

Table 2 Quantitative evaluation of albedo estimation on SMOLAT dataset, comparing with SfSNet(Sengupta et al., 2018) and DECA(Feng et al., 2021)

Method	SSIM ↑	PSNR ↑	RMSE ↓
SfSNet	0.899	14.499	0.190
DECA	0.860	8.395	0.390
Ours	**0.981**	**32.493**	**0.024**

We use SSIM, PSNR and RMSE to evaluate the result. ↑ / ↓ represents the higher/lower the value the better
Bolded numbers represent the best results for the same group of experiments

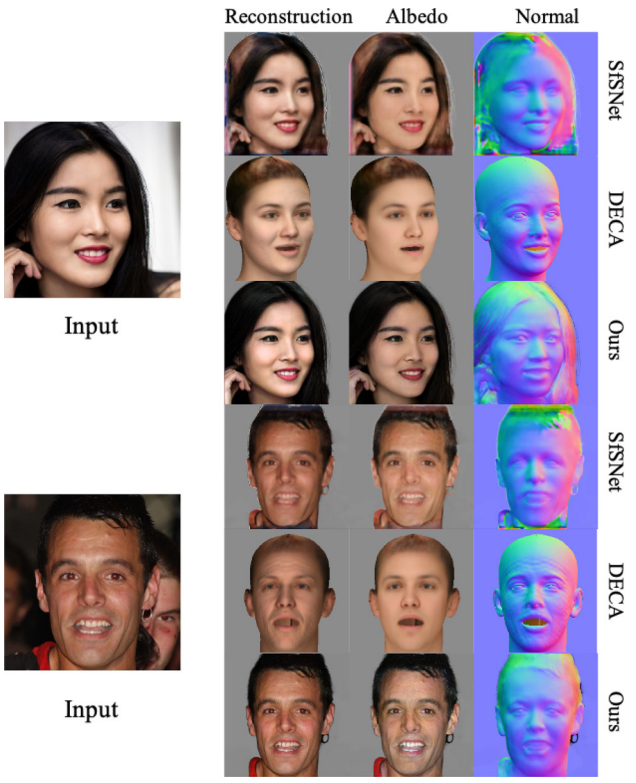

Fig. 8 Qualitative comparison of appearance reconstruction results on in-the-wild dataset, with SfSNet (Sengupta et al., 2018) and DECA (Feng et al., 2021). From left to right: Input: input image; Reconstruction: using the estimated albedo, normal, illumination and the rendering pipeline to reconstruction the input image; Albedo: estimated albedo; Normal: estimated normal

Table 3 Normal reconstruction error on SMOLAT dataset, compared with SfSNet(Sengupta et al., 2018), and DECA(Feng et al., 2021).

Algorithm	Mean	$< 5°$	$< 15°$	$< 25°$
SfSNet	11.583°	67.673%	74.872%	82.598%
DECA	8.726°	68.478%	79.071%	87.902%
Ours	**5.400°**	**73.569%**	**90.227%**	**95.082%**

The second column: mean angular error of per-pixel normal; the third to fifth columns: the percentage of correct pixels within different angular error thresholds
Bolded numbers represent the best results for the same group of experiments

6.3 Evaluation

We have evaluated PN-Relighting on three tasks: portrait appearance reconstruction, portrait relighting, and novel view synthesis under various illuminations. For each task, we compare with state-of-the-arts both qualitatively and quantitatively.

6.3.1 Facial Appearance Reconstruction

We compare our estimated surface normal $\hat{N}$ and albedo $\hat{A}$ with two state-of-the-art appearance estimation approaches: DECA (Feng et al., 2021), and SfSNet (Sengupta et al., 2018). DECA is a parametric face model that inferences $\hat{N}$ and albedo $\hat{A}$ as their intermediate results. SfSNet is more close to our method that also handles illumination change. Specifically, we measure the reconstructed albedo accuracy using PSNR, SSIM and RMSE metrics. As for normal, we use the mean error and the percentage of correct pixels at various thresholds.

On SMOLAT dataset, we conduct quantitative comparisons on the estimated normal and albedo in Tables 2 and 3. For a fair comparison, we applied the background removal and color calibration to all the other methods and only evaluate the reconstructed appearance from the original viewpoint of $\boldsymbol{I}$. Figure 7 shows the visual comparison with state-of-the-arts. Compared with other methods, normal map produced by PN-Relighting contains more details and is the closest to

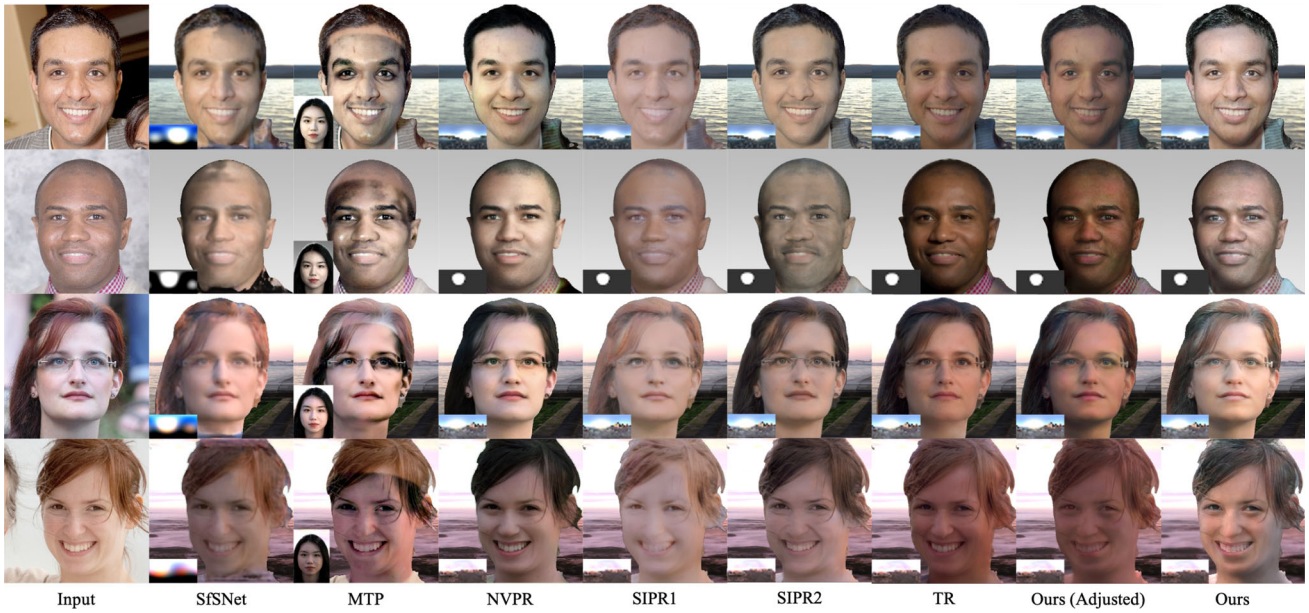

Fig. 9 Qualitative comparison of portrait relighting results on in-the-wild dataset. From left to right: Input: input image; the result of SfSNet (Sengupta et al., 2018), MTP (Shu et al., 2017) , NVPR (Zhang et al., 2021), SIPR1 (Wang et al., 2020), SIPR2 (Sun et al., 2019), TR (Pandey et al., 2021); Ours (Adjusted): Our result (adjusted exposure curve to TR), and Ours : our result

Table 4 Quantitative evaluation of reconstruction on FFHQ dataset, comparing with SfSNet(Sengupta et al., 2018) and DECA(Feng et al., 2021).

Method	SSIM↑	PSNR↑	RMSE ↓
SfSNet	0.907	24.964	0.067
DECA	0.820	20.413	0.107
Ours	**0.958**	**27.143**	**0.536**

We use SSIM, PSNR and RMSE to evaluate the result. ↑ / ↓ represents the higher/lower the value the better
Bolded numbers represent the best results for the same group of experiments

ground truth. Similarly, our albedo is of a higher resolution and presents fewer artifacts. Figure 11 shows the average normal and albedo error along lighting changes. We can tell that, our predicted intrinsic appearance parameters are stable among different illumination, which is inline with facts: surface normal and albedo are independent on lighting changes.

On the in-the-wild dataset, since there is no ground truth albedo and normal for quantitative measurement. We hence only show the qualitative comparisons of $\hat{A}$ and $\hat{N}$ in Fig. 8.

Overall PN-Relighting produces more convincing results on the in-the-wild data. Specifically, compared to SfSNet, our method reconstructs an albedo map with fewer specular and artifacts whereas our normal is sharper, preserving more high-frequency details. By introducing Phong diffuse and specular shading as a prior, our method also more faithfully recovers specular reflectance of the portrait in the reconstructed $\hat{\boldsymbol{I}}$. Table 4 shows a quantitatively comparison of the reconstructed $\hat{\boldsymbol{I}}$ on the FFHQ test set (Fig. 9) .

On SMOLAT dataset, we show the qualitative and quantitative comparison in Fig. 10, whose quantitative comparison result is in Tables. 5, and 6 respectively. Our method has achieved the best accuracy and stability under different metrics. When testing the average normal and albedo error along lighting changes, as shown in Fig. 11, our method is more robust in both geometry and albedo reconstructions, and consequently achieves better relighting compared to SfSNet. This result shows that our predicted normal and albedo keep stable when the illumination changes, thus demonstrating that our network has good decomposition ability for albedo and illumination (Table 7).

6.3.2 Portrait Relighting

We have compared our relit results $\hat{\boldsymbol{I}}_E$ with SfSNet, NVPR (Zhang et al., 2021), MTP (Shu et al., 2017), TotalRelighting (TR) (Pandey et al., 2021), SIPR1 (Wang et al., 2020) and SIPR2 (Sun et al., 2019) using SSIM, PSNR, and RMSE measurements. As we don't have access to the code and training dataset of TR, we acquired the results of our testing dataset from the authors.

As shown in Fig. 10, NVPR presents high stability in color. However, it presents deteriorated performance in high contrast environment illumination such as specular highlights due to the lacking of portrait geometry prior. To compare with MTP, we choose a reference portrait image as its input. Recall that the MTP relighting is primarily based on image color, the results exhibit artifacts when the input portrait

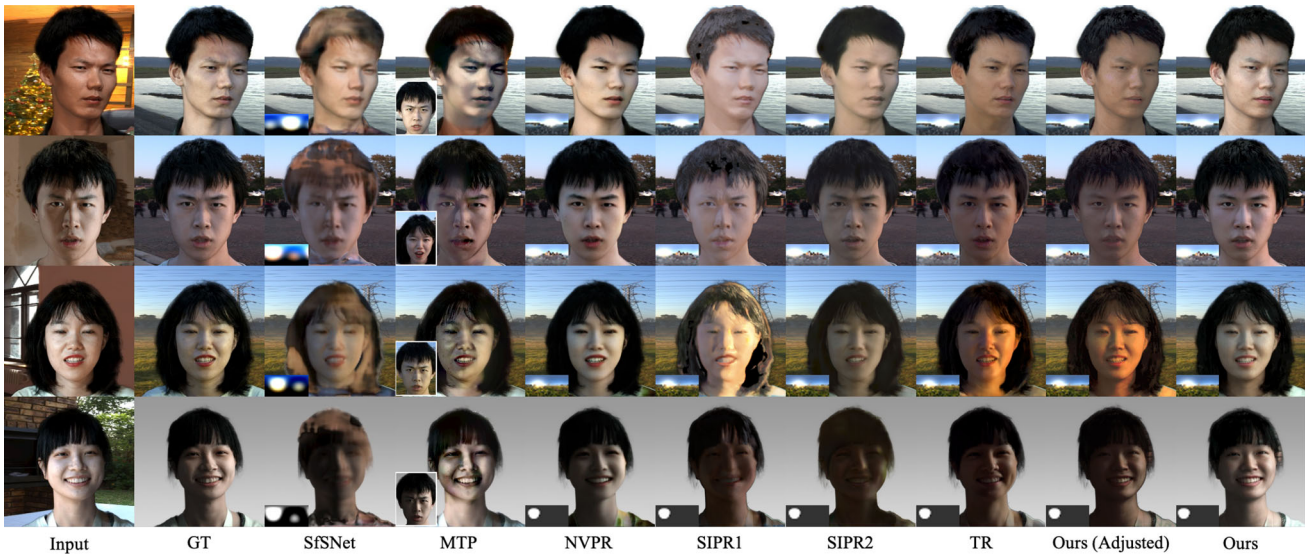

Fig. 10 Qualitative comparison of portrait relighting results on SMOLAT dataset. From left to right: Input: input image; GT: image-based rendering ground truth by OLAT data; the result of SfSNet (Sengupta et al., 2018), MTP(Shu et al., 2017) , NVPR(Zhang et al., 2021), SIPR1(Wang et al., 2020), SIPR2(Sun et al., 2019), TR(Pandey et al., 2021); Ours (Adjusted): Our result (adjusted exposure curve according to TR) and Ours : our result

Table 5 Quantitative comparison of portrait relighting on the test data in Fig.10, with SfSNet (Sengupta et al., 2018), MTP (Shu et al., 2017), NVPR (Zhang et al., 2021), SIPR1 (Wang et al., 2020), SIPR2 (Sun et al., 2019) and TR (Pandey et al., 2021)

Method	SSIM↑	PSNR↑	RMSE↓
SfSNet	0.920	24.372	0.064
MTP	0.894	21.818	0.096
NVPR	0.964	29.779	0.034
SIPR1	0.923	23.059	0.073
SIPR2	0.935	24.391	0.069
TR	0.928	23.226	0.070
Ours	**0.966**	**30.839**	**0.029**

We use SSIM, PSNR and RMSE to evaluate the result. ↑ / ↓ represents the higher/lower the value the better
Bolded numbers represent the best results for the same group of experiments

Table 6 Quantitative comparison of portrait relighting on SMOLAT dataset, with SfSNet (Sengupta et al., 2018), MTP (Shu et al., 2017) and NVPR (Zhang et al., 2021)

Method	SSIM↑	PSNR↑	RMSE↓
SfSNet	0.705	23.403	0.071
MTP	0.793	24.821	0.061
NVPR	0.854	27.256	0.046
Ours	**0.875**	**28.203**	**0.041**

We use SSIM, PSNR and RMSE to evaluate the result. ↑ / ↓ represents the higher/lower the value the better
Bolded numbers represent the best results for the same group of experiments

differs from the reference one in color and texture. Similar to NVPR, SIPR1 and SIPR2 do not contain a geometric prior, and therefore demonstrate noticeable artifacts of unnatural highlights and shadows when presented in high-contrast environment illumination. In contrast, by training with a large OLAT dataset, TR produces high-fidelity relighting results. Ours has achieved comparable performance with TR while using a much smaller training data size. Besides, due to the special post-processing, the global hue of the relit results from TR is significantly different from the other baseline methods. Thus, for a more fair visual comparison, we manually align the exposure curve of our results to TR, so that the results have a similar hue, as presented in the column *Ours(Adjusted)* of Figs. 9 and 10. We can tell that, after the adjustment, our results are of the similar, if not better, quality and preserve the same amount of details when compared with TR.

On the in-the-wild dataset, we further conduct a qualitative comparison in Fig. 9. Compared to other methods, PN-Relight produces more photo-realistic results. Moreover, we demonstrate the effect of editing the implicit material latent in Fig. 14 to change the material of the portrait. In this example, we gradually reduce the implicit material latent extracted from the original image to zero for relighting, resulting in the portrait material gradually approaching to diffuse during relighting.

6.3.3 Novel View Synthesis and Relighting

For multi-view rendering, we further compare with SfSNet, StyleFlow (Abdal et al., 2021), and StyleNerf (Gu et

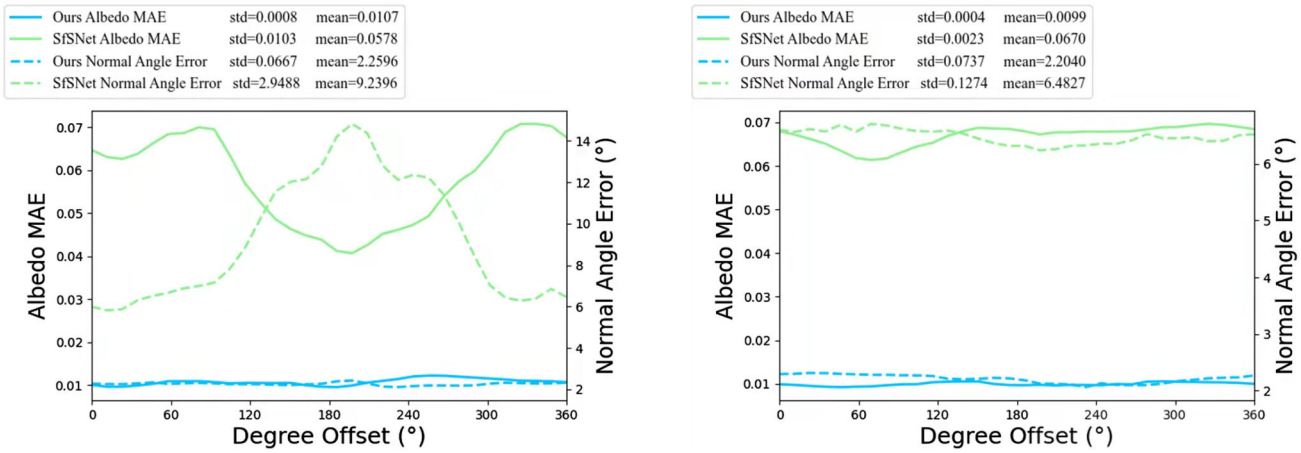

Fig. 11 We compare the estimated surface normal and albedo error with SfSNet in dynamic illuminations. A point light source (left) and a low-contrast HDR environment map (right) are used to generate novel lighting

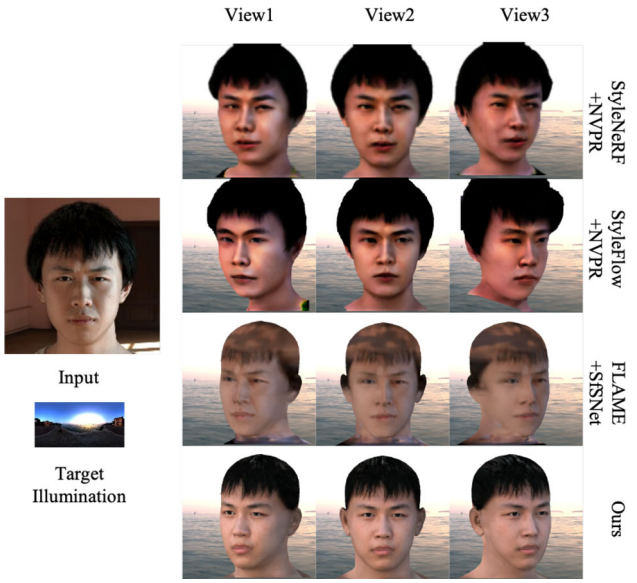

Fig. 12 Qualitative result of novel view synthesis and relighting results on SMOLAT dataset. For an input image and a target illumination, we demonstrate the three novel view results of the following method: StyleNeRF (Gu et al., 2022), StyleFlow (Abdal et al., 2021), SfSNet (Sengupta et al., 2018) and Ours

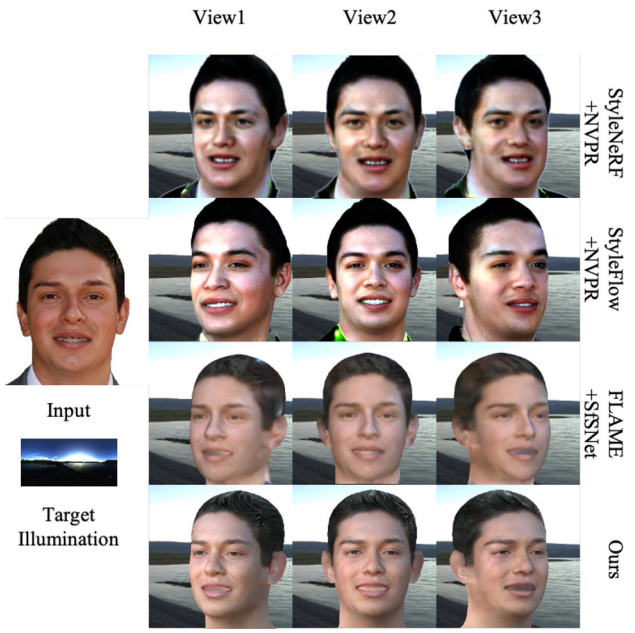

Fig. 13 Qualitative result of novel view synthesis and relighting results on in-the-wild dataset. For an input image and a target illumination, we demonstrate the three novel view results of the following method: StyleNeRF (Gu et al., 2022), StyleFlow (Abdal et al., 2021), SfSNet (Sengupta et al., 2018) and Ours

Table 7 Quantitative comparison on free-view relighting with SfSNet(Sengupta et al., 2018).

Method	SSIM↑	PSNR↑	RMSE↓
SfSNet	0.914	22.402	0.076
Ours	**0.921**	**23.622**	**0.068**

We use SSIM, PSNR and RMSE to evaluate the result. ↑ / ↓ represents the higher/lower the value the better
Bolded numbers represent the best results for the same group of experiments

al., 2022). Both StyleNerf and StyleFlow construct 3D face model based on GAN architecture, and exhibits inconsistency when the viewpoint changes. Since the two methods are not designed for relighting tasks, for fair comparison, we apply NVPR to add lighting effects to their reconstructed 3D faces. SfSNet can only generate surface normal and albedo from portrait image, and is not for 3D face generation. We therefore use the same way as described in Section. 4.2 to form 3D faces using their predicted normal and albedo maps. In Figs. 12 and 13, we show the qualitative results on SMOLAT and in-the-wild dataset, respectively. We observe that the GAN-based modeling approaches still present poor consistency on both shape and lighting under varying viewing

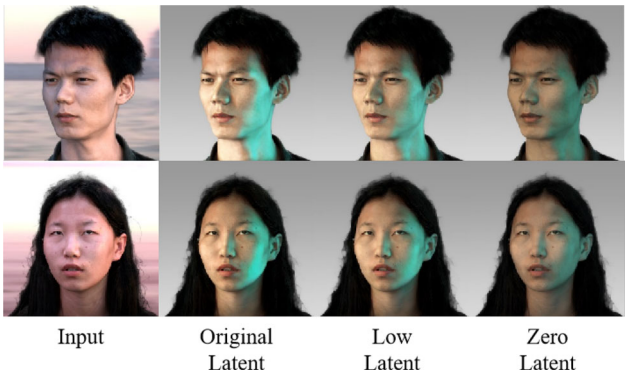

Fig. 14 The application of editing material by implicit material latent

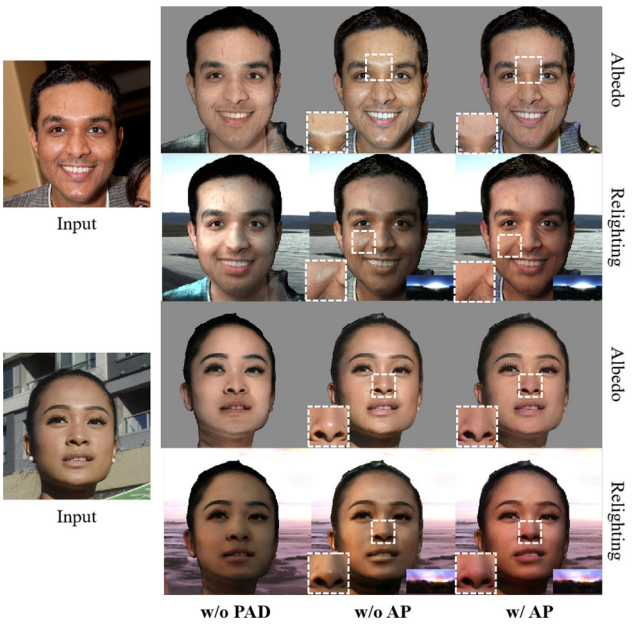

Fig. 15 We demonstrate the importance of Pseudo-Albedo dataset. Our full network are more correct in color and of less specular highlights on the estimated albedo

angles, due to their inaccurate geometry estimation. For SfS-Net, their reconstructed appearance is affected by lighting changes as we discussed above, and therefore fails to handle specular highlights as shown in Fig. 12. Our method, in contrast, demonstrates the most consistent rendering effects on both viewpoint and illumination changes.

It is important to note that there is *no* available multi-view OLAT datasets that can enable quantitative measurement of the identity and relighting consistency directly. For better evaluation, we generate reference face images under different viewing points and illuminations by using CG rendering pipeline Zhang et al. (2022). Specifically, take one view with a environment illuminations to construct the input image, and take two new views with a new environment illumination as the condition . As shown in Table 6, our results achieves the best consistency w.r.t. identity and relighting (Fig. 14).

Table 8 Quantitative ablation study on the importance of in-the-wild training of relighting module

Category	Method	PSNR↑	SSIM↑	RMSE↓
Albedo	w/o PAD	27.494	0.832	0.044
	w/o AP	31.706	0.926	0.029
	w/ AP	**34.333**	**0.970**	**0.020**
Relighting	w/o PAD	24.960	0.909	0.058
	w/o AP	26.086	0.939	0.051
	w/ AP	**28.182**	**0.958**	**0.043**

The result of our complete pipeline preserves richer detail. With in-the-wild training, we achieve better albedo prediction and relighting results. We use SSIM, PSNR and RMSE to evaluate the result. ↑ / ↓ represents the higher/lower the value the better
Bolded numbers represent the best results for the same group of experiments

6.4 Ablation Study

6.4.1 Pseudo-Albedo Dataset for Appearance Inference

To validate that our pseudo-albedo rendering pipeline can effectively improve the generalization ability of PN-Relighting on SMOLAT, we create a variation of our network: (1) **w/o PAD** that is trained only on SMOLAT; (2) **w/o AP** that is trained on SMOLAT and Pseudo-Albedo dataset without manually removing highlights on pseudo-albedo maps; (3) **w/ PAD** denoting the full pipeline. The qualitative comparison results are shown in Fig. 15, and the quantitative comparison results are shown in Table 8. We observe that our Pseudo-albedo generation pipeline enables PN-Relighting to preserve fine details in the relighting results, including the makeup, skin texture, eyebrows, pupils, hair color, etc. **w/o PAD**, in contrast, still exhibits a number of specular highlights in the predictions that should ideally be removed .

6.4.2 In-the-wild Training for Relighting

We have also verified the importance of training PN-Relighting network on the in-the-wild dataset by creating two variations: (1)**w/o Wild** represents the network using only the OLAT dataset; (2) **w/ Wild** represents our full pipeline. We show a qualitative comparison in Fig. 16. Compared to the **w/o Wild**, the network trained with FFHQ achieves superior performance in generalization, by faithfully reproducing the original portrait's in both appearance realism and image sharpness.

6.4.3 Implicit Material Editing

We compare our relighting network with the material encoder to the relighting network with the U-net structure. **w/o IML**

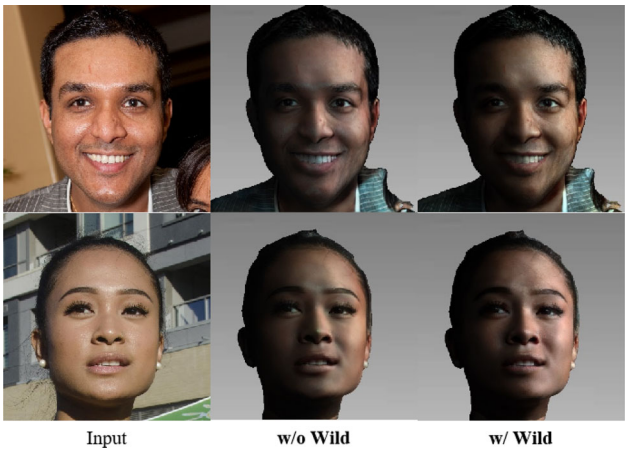

Fig. 16 We demonstrate the importance of in-the-wild training of relighting module. The result of our complete pipeline preserves richer detail

Fig. 17 The ablation study of Implicit Material Latent. Using IML will get more realistic highlights and shadows on the resulting image

Table 9 The ablation study on implicit material latent

Method	SSIM↑	PSNR↑	RMSE↓
w/o IML	0.906	28.769	0.041
w/ IML	**0.913**	**29.319**	**0.039**

By adding IML to the relighting network, we improve the relighting results
Bolded numbers represent the best results for the same group of experiments

denotes the effect without the implicit material latent. **w/ IML** denotes our complete pipeline.

Regarding the "w/o IML" variant, we adopt the *same* network structure as "w/ IML" (the full pipeline), but remove loss terms that related to material latent vector $\hat{M}$ in the training phase, *i.e.*, the $\mathcal{L}_{lat_D}$ (Eq. 8) and $\mathcal{L}_{lat_L}$ (Eq. 9).

Qualitative comparisons are shown in Fig. 17 and quantitative comparisons are illustrated in Table 9. The network

Table 10 Ablation study on OLAT dataset size vs. in-the-wild training

Method	OLAT	PSNR↑	SSIM↑	RMSE↓
w/Wild	100%	**34.333**	**0.970**	**0.020**
	50%	33.890	0.966	0.021
	25%	32.669	0.959	0.025
	12.5%	32.369	0.956	0.026
w/o Wild	100%	31.706	0.926	0.029
	50%	31.661	0.936	0.028
	25%	30.427	0.921	0.032
	12.5%	30.397	0.923	0.032

We demonstrate the importance of in-the-wild training
Bolded numbers represent the best results for the same group of experiments

with implicit material latent allows the network to produce more realistic specular highlights, more consistent with the real-life cases.

6.4.4 Training size of OLAT vs. In-the-wild Dataset

In order to verify the advantage of our in-the-wild training strategy, we conduct an ablation study on different training data partitions among OLAT and in-the-wild data. Specifically, we use 25%-OLAT, 50%-OLAT and 100%-OLAT to denote variants created by using 25%, 50% and 100% OLAT data for training, and use **w/o Wild** and **w/ Wild** to denote variants using in-the-wild training and not using in-the-wild training respectively. The quantitative comparison is presented in Table 10. We can tell that the model of 25% OLAT + **w/Wild** outperforms the model using 100% OLAT but doesn't include the "in-the-wild" training data. This further demonstrates the effectiveness of employing PA & FFHQ in the network training to boost the efficiency of OLAT data usage.

6.4.5 Normal Network with Mesh-Prior

Even though the parametric model can provide normal information, we found out that these parametric normals cannot model the pixel-aligned geometry details in those facial regions near the eyes and wrinkles. Such normal artifacts further lead to inaccurate albedo modeling and the following relighting module. Thus, we chose to rely on our OLAT dataset to provide pixel-aligned facial normal estimation and only adopt the parametric model to enable more stable free-view relighting. To verify this, we create a variation **w/ mesh-prior** by using the mesh normal from the parametric model as prior. In detail, for **w/ mesh-prior**, we first project mesh-normal onto the input's perspective of view, and attach the transformed normal to the original input of Normal Net-

Table 11 Quantitative ablation study on Normal Network

Algorithm	Mean	$< 5°$	$< 15°$	$< 25°$
w/ DECA prior	5.685 ± 1.666 °	72.159 ± 2.710 %	88.884 ± 3.768 %	94.892 ± 1.833 %
w/o DECA prior	$\mathbf{5.400 \pm 1.591}$ °	$\mathbf{73.569 \pm 2.706}$ %	$\mathbf{90.227 \pm 3.693}$ %	$\mathbf{95.082 \pm 1.749}$ %

The first row demonstrates the normal error predicted by the Normal Network initiated with DECA mesh normal prior. The second row shows the normal error of our Normal Network training on the OLAT dataset only. We can see that, since the normal provided by OLAT dataset is pixel aligned, it produces the lower error

Bolded numbers represent the best results for the same group of experiments

work Ψ_N. The rest training procedure is the same as "**w/o mesh-prior**" (our full pipeline). The quantitative comparison is shown in Table 11. We can tell that the "parametric normal" will degrade the normal estimation accuracy. For this reason, we don't use the parametric normal as prior in our optimization pipeline.

7 Discussions and Future Work

We have presented a novel hybrid parametric-neural approach for producing high-quality portrait relighting. Our approach PN-Relighting achieves comparable single image relighting performance to the latest TotalRelighting (Pandey et al., 2021). Different from prior art though, PN-Relighting uses a much smaller OLAT dataset or SMOLAT. To address the small data learning problem, we have employed parametric 3D faces and coupled them with appearance inference and implicit material modeling. The key insight here is although small, SMOLAT provides a viable implicit model to account for material variations that in return compensate for limitations in parametric models. Specifically, we have tailored a differentiable rendering pipeline that combines the benefits of parametric and neural approaches. Another major benefit of our hybrid model is that it directly supports free-viewpoint rendering, as the parametric model provides a 3D model. Further, the implicit material model from SMOLAT supports partial reflectance editing. Putting them all together, PN-Relighting not only enables single portrait relighting but potentially serves as a virtual LightStage, to supplement limited OLAT data with more varieties through rendering.

A major limitation in our current implementation is that we have relied on OLAT image based rendering and Phong rendering, neither can sufficiently render self-shadowing. The lack of shadow handling leads to artifacts in the inference stage. One possible direction is to first eliminate shadows from the image and subsequently add them back once relighting is conducted. In addition, our relighting module employs pairwise training. Consequently, the network exhibits color shifts in lighting that differ significantly from the environment illumination dataset. Emulating additional illumination patterns can potentially mitigate the problem but would require longer training time.

In addition, the regular acquisition of the OLAT dataset is not able to separate highlights from diffuse, leading to potential errors in our calculated normal and albedo in areas of strong specularity (e.g. hair, glasses, teeth). Adding polarizers to cameras may be a viable solution to separate highlights from diffuse in subsequent acquisitions.

Furthermore, our current UV-space normal and albedo generation pipeline leverage a few structurally symmetrical features, *e.g.*, nose, and mouth, to inpaint the missing part. Consequently, a large posed lateral face may fail the inpainting algorithm, and introduce strong artifacts to the novel-view results.

In addition, the parametric models are not 100% correct in the construction of the facial details, especially in the cheeks and the bridge of the nose areas. Since the face reconstruction algorithms are sensitive to features from these areas, it can cause a change of identity in the rendering results when viewing point changes are large.

Inherent to parametric models, the PN-Relighting cannot model objects adhesive to the face, e.g., hair and glasses, and produce artifacts on these objects. The missing of these regions, as important features of the human face, may bring about changes visually in identities, thus affecting the generalizability of the method. Recent works, such as EG3D Chan et al. (2022), styleNeRF Gu et al. (2022), etc., have demonstrated outperforming results in the field of 3D face generation by leveraging the 3D-aware StyleGAN framework. Compared to parametric models, these GAN-based methods can handle challenging scenes including rendering specular components, such as hair, teeth, glasses, and other portrait details. However, StyleGAN is trained to tune the input latent vectors, which are hard to extract explicit semantic facial features to match the target image. For this reason, to enable free-view portrait relighting on StyleGAN-based frameworks, a prototype solution is to adopt PTI Roich et al. (2022) to match the latent vector from StyleGAN to the target image. However, this vector-image matching procedure is extremely time-consuming and is not feasible to apply in a real-world application. In addition, the dedicated fine-tuning procedure introduced by PTI is hard to ensure facial feature

consistency among different viewpoints, as shown in Figs. 12 and 13. We can tell that the mouth shape of subjects is not consistent along with viewpoint changes. In our future work, we plan to adopt a GAN-based method to fill in the missing parts and components that FLAME cannot well handle to enable a better photo-realistic free-view relighting effect.

PN-Relighting partially addresses the issue by using a small set of subjects and allows a user to synthesize more comprehensive relighting datasets. PN-Relighting addresses multi-view rendering by using a parametric 3D face model. The latest trend in multi-view image synthesis is to combine the Neural Radiance Field Mildenhall et al. (2020) and styleGAN Karras et al. (2019), e.g., in Gu et al. (2022) and Chan et al. (2022). We hence intend to investigate how to integrate our material model in conjunction with these approaches, to form an end-to-end free-view relighting pipeline without employing explicit models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-022-01730-5.

Acknowledgements This work was supported by Shanghai YangFan Program (21YF1429500), Shanghai Local college capacity building program (22010502800), NSFC programs (61976138, 61977047), the National Key Research and Development Program (2018YFB2100500), STCSM (2015F0203-000-06), and SHMEC (2019-01-07-00-01-E00003).

References

Abdal, R., Zhu, P., Mitra, N. J., & Wonka, P. (2021). Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG), 40*(3), 1–21.

Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I. (2017) Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp 6799–6808)

Basri, R., & Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(2), 218–233.

Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M. (2010) High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, (pp 1–9)

Bi, S., Xu, Z., Sunkavalli, K., Kriegman, D., Ramamoorthi, R. (2020) Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 5960–5969

Blanz, V., Vetter, T. (1999) A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques*, pp 187–194

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., & Zafeiriou, S. (2018). Large scale 3d morphable models. *International Journal of Computer Vision, 126*(2), 233–254.

Cao, X., Chen, Z., Chen, A., Chen, X., Li, S., Yu, J. (2018) Sparse photometric 3d face reconstruction guided by morphable models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4635–4644

Chabert, CF., Einarsson, P., Jones, A., Lamond, B., Ma, WC., Sylwan, S., Hawkins, T., Debevec, P. (2006) Relighting human locomotion with flowed reflectance fields. In *ACM SIGGRAPH 2006 sketches*, pp 76–es

Chan, ER., Lin, CZ., Chan, MA., Nagano, K., Pan, B., Mello, SD., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G. (2022) Efficient geometry-aware 3D generative adversarial networks. In *CVPR*

Dai, H., Pears, N., Smith, W., & Duncan, C. (2020). Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision, 128*(2), 547–571.

Debevec, P., Hawkins, T., Tchou, C., Duiker, HP., Sarokin, W., Sagar, M. (2000) Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on computer graphics and interactive techniques*, pp 145–156

Dou, P., Shah, SK., Kakadiaris, IA. (2017) End-to-end 3d face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp 5908–5917

Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X. (2018) Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pp 534–551

Feng, Y., Feng, H., Black, M. J., & Bolkart, T. (2021). Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG), 40*(4), 1–13.

Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., & Theobalt, C. (2016). Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG), 35*(3), 1–15.

Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S. (2019) Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1155–1164

Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, WT. (2018) Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 8377–8386

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2020) Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144.

Gu, J., Liu, L., Wang, P., Theobalt, C. (2022) Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International conference on learning representations*, https://openreview.net/forum?id=iUuzzTMUw9K

Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, SZ. (2020) Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16, Springer, pp 152–168

Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y. C., & Li, H. (2017). Avatar digitization from a sin-

gle image for real-time rendering. *ACM Transactions on Graphics (ToG), 36*(6), 1–14.
Ichim, A. E., Bouaziz, S., & Pauly, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG), 34*(4), 1–14.
Jackson, AS., Bulat, A., Argyriou, V., Tzimiropoulos, G. (2017) Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pp 1031–1039
Jeni, LA., Cohn, JF., Kanade, T. (2015) Dense 3d face alignment from 2d videos in real-time. In *11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, IEEE, vol 1, pp 1–8
Jiang, L., Zhang, J., Deng, B., Li, H., & Liu, L. (2018). 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing, 27*(10), 4756–4770.
Kanamori, Y., & Endo, Y. (2018). Relighting humans: Occlusion-aware inverse rendering for full-body human images. *ACM Trans Graph, 10*(1145/3272127), 3275104.
Karras, T., Laine, S., Aila, T. (2019) A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 4401–4410
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. (2020) Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 8110–8119
Ke, Z., Sun, J., Li, K., Yan, Q., & Lau, R. W. (2022). Modnet: Real-time trimap-free portrait matting via objective decomposition. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(1), 1140–1147. https://doi.org/10.1609/aaai.v36i1.19999
Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Ghosh, A., & Zafeiriou, S. P. (2021). Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 01*, 1.
Li, H., Yu, J., Ye, Y., & Bregler, C. (2013). Realtime facial animation with on-the-fly correctives. *ACM Trans Graph, 32*(4), 42–1.
Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4d scans. *ACM Trans Graph, 36*(6), 194–1.
Li, Y., Ma, L., Fan, H., Mitchell, K. (2018) Feature-preserving detailed 3d face reconstruction from a single image. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pp 1–9
Meka, A., Haene, C., Pandey, R., Zollhöfer, M., Fanello, S., Fyffe, G., Kowdle, A., Yu, X., Busch, J., Dourgarian, J., et al. (2019). Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG), 38*(4), 1–12.
Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), 99–106.
Nestmeyer, T., Lalonde, J. F., Matthews, I., & Lehrmann, A. (2020). Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5124–5133.
Pallant, C. (2011) Demystifying Disney: a history of Disney feature animation. Bloomsbury Publishing USA
Pandey, R., Escolano, S. O., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., & Fanello, S. (2021). Total relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG), 40*(4), 1–21.
Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM, 18*(6), 311–317.
Radke, R. J. (2013). *Computer vision for visual effects*. Cambridge University Press.
Riviere, J., Gotardo, P., Bradley, D., Ghosh, A., & Beeler, T. (2020). Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans Graph, 39*(4). https://doi.org/10.1145/3386569.3392464.
Roich, D., Mokady, R., Bermano, A. H., & Cohen-Or, D. (2022). Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG), 42*(1), 1–13.
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, Springer, 234–241.
Roth, J., Tong, Y., & Liu, X. (2016). Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4197–4206.
Sagar, M. (2005) Reflectance field rendering of human faces for" spiderman 2". In *ACM SIGGRAPH 2005 Courses*, pp 14–es
Sela, M., Richardson, E., & Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, 1576–1585.
Sengupta, S., Kanazawa, A., Castillo, C. D., & Jacobs, D. W. (2018). Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6296–6305.
Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L. (2020) Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Computer Vision–ECCV 2020: 16th European conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer, pp 53–70
Shi, F., Wu, H. T., Tong, X., & Chai, J. (2014). Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG), 33*(6), 1–13.
Shin, I. K., Öztireli, A. C., Kim, H. J., Beeler, T., Gross, M., & Choi, S. M. (2014). Extraction and transfer of facial expression wrinkles for facial performance enhancement. In *PG (Short Papers)*
Shu, Z., Hadap, S., Shechtman, E., Sunkavalli, K., Paris, S., & Samaras, D. (2017). Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG), 36*(4), 1.
Smolyanskiy, N., Huitema, C., Liang, L., & Anderson, S. E. (2014). Real-time 3d face tracking based on active appearance model constrained by depth data. *Image and Vision Computing, 32*(11), 860–869.
Sun, T., Barron, J. T., Tsai, Y. T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P. E., & Ramamoorthi, R. (2019). Single image portrait relighting. *ACM Trans Graph, 38*(4), 79–1.
Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., et al. (2022). Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.
Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Trans Graph, 34*(6), 183–1.
Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.
Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., & Xu, F. (2020). Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG), 39*(6), 1–13.
Wei, H., Liang, S., Wei, Y. (2019) 3d dense face alignment via graph convolution networks. arXiv preprint arXiv:1904.05562

Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. *Optical engineering, 19*(1), 191139.

Wright, S. (2017) Digital compositing for film and vídeo: Production Workflows and Techniques. Routledge

Xu, Z., Bi, S., Sunkavalli, K., Hadap, S., Su, H., & Ramamoorthi, R. (2019). Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics (ToG), 38*(4), 1–13.

Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., & Cao, X. (2020). Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 601–610.

Zhang, L., Zhang, Q., Wu, M., Yu, J., & Xu, L. (2021). Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 802–812.

Zhang, L., Zeng, C., Zhang, Q., Lin, H., Cao, R., Yang, W., Xu, L., & Yu, J. (2022). Video-driven neural physically-based facial asset for production. *ACM Trans Graph (Proc SIGGRAPH Asia), 41*(6), 1–16.

Zhang, X., Fanello, S., Tsai, Y. T., Sun, T., Xue, T., Pandey, R., Orts-Escolano, S., Davidson, P., Rhemann, C., Debevec, P., et al. (2021). Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG), 40*(1), 1–17.

Zhou, H., Hadap, S., Sunkavalli, K., & Jacobs, D. W. (2019). Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7194–7202.

Zhu, X., Yang, F., Huang, D., Yu, C., Wang, H., Guo, J., Lei, Z., Li, SZ. (2020) Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, Springer, pp 343–358

Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., & Theobalt, C. (2018). State of the art on monocular 3d face reconstruction, tracking, and applications. *Computer Graphics Forum, Wiley Online Library, 37*, 523–550.