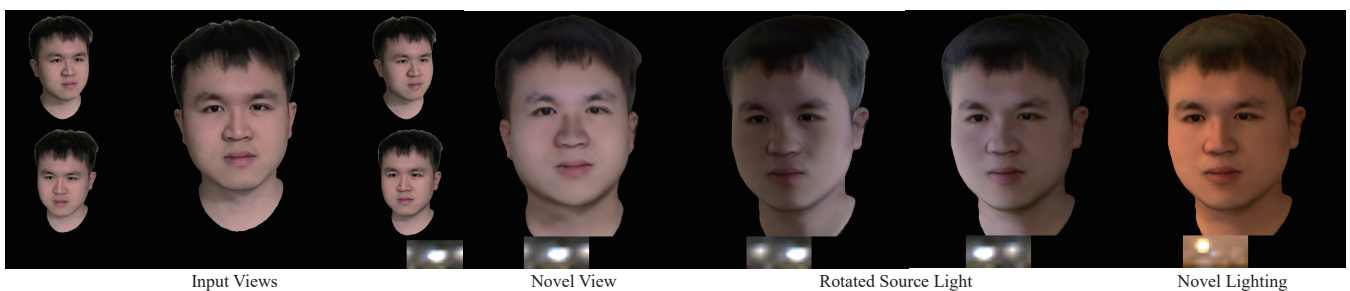# NeLF: Neural Light-transport Field for Portrait View Synthesis and Relighting

Tiancheng Sun[1*], Kai-En Lin[1*], Sai Bi[2], Zexiang Xu[2], Ravi Ramamoorthi[1]

[1]University of California, San Diego, [2]Adobe Research
*Equal contribution

| Input Views | Novel View | Rotated Source Light | Novel Lighting |

Figure 1: *We introduce neural light-transport field (NeLF), which learns to infer the light-transport and volume density from a sparse set of input views. NeLF enables joint relighting and view synthesis of real portraits from only five input images.*

Abstract

Human portraits exhibit various appearances when observed from different views under different lighting conditions. We can easily imagine how the face will look like in another setup, but computer algorithms still fail on this problem given limited observations. To this end, we present a system for portrait view synthesis and relighting: given multiple portraits, we use a neural network to predict the light-transport field in 3D space, and from the predicted Neural Light-transport Field (NeLF) produce a portrait from a new camera view under a new environmental lighting. Our system is trained on a large number of synthetic models, and can generalize to different synthetic and real portraits under various lighting conditions. Our method achieves simultaneous view synthesis and relighting given multi-view portraits as the input, and achieves state-of-the-art results.

CCS Concepts
• ***Computing methodologies*** → *Image-based rendering; Computational photography;*

1. Introduction

Digitizing human portraits from natural portrait images and re-synthesizing novel images under new lighting and viewpoints is a long-standing graphics and vision problem with many applications like computational photography and video conferencing. Recently, NeRF [MST*20] has initiated significant progress in neural rendering for photo-realistic image synthesis. Various radiance field extensions [SDZ*21; MRS*21; BXS*20a; LGL*20] have been presented and many of the concurrent works [GTZN20; LSS*21; PSB*20] aim to address problems in human portrait capture.

However, previous NeRF-based portrait capture methods usually require capturing a large number of images and an expensive

per-scene optimization process. Besides, most works consider only view synthesis tasks and do not support relighting. In general, a relightable portrait representation cannot simply be a radiance field that bakes in the original lighting condition; it instead requires explicitly disentangling and modeling lighting and portrait appearance in the image captures, which is a highly challenging problem.

In this paper, we propose a novel neural rendering approach that can jointly estimate the geometry and appearance of a human portrait and the lighting environment the human stands in, from only a sparse set of input images. We present a novel deep neural network that can generalize across scenes to regress a *neural light transport field*, i.e. NeLF, from portrait images. At an arbi-

trary 3D point, this neural light transport field outputs the volume density and light transport coefficients (unlike the view-dependent radiance in NeRF), which linearly explains the portrait appearance under distant illumination conditions represented by environment maps, similar to the classic light transport functions in image-based relighting [DHT*00] and precomputed radiance transfer [SKS02; NRH03]. With this novel representation, our approach enables high-quality portrait relighting and view synthesis done simultaneously, as shown in Fig. 1.

Our approach incorporates classic light transport functions into neural volumetric rendering. In particular, we first use a UNet-like CNN as a feature extractor to convert each input portrait image to a neural feature map that encodes per-view pixel-wise portrait geometry and appearance. For an arbitrary 3D point, we fetch multi-view neural features from its projections in the feature maps and use MLPs to regress volume density and radiance from the features to enable differentiable ray marching. As opposed to earlier works [YYTK20; WWG*21] that directly output radiance per shading point on marching rays, we propose to first regress a *light transport vector* – that can linearly compute the radiance under any novel lighting conditions – to enable relighting for neural volumetric rendering. In addition, we apply a sub-network to estimate the lighting as environment maps from the bottleneck features of the CNN feature extractor, contributing to the disentanglement of portrait appearance and light effects in the original images. Our approach jointly estimates portrait geometry (as volume density), appearance (as light transport vectors) and lighting conditions (as environment maps) from portrait images, and can regress final ray colors under arbitrary novel viewpoints and lighting via differentiable ray marching.

Acquiring a real portrait dataset under different lighting conditions is a notoriously challenging task and traditionally requires a sophisticated light stage [DHT*00] that is not easily accessible for most researchers. In order to train our network with practical data, we instead utilize rendered images of human head models with different views and environment maps. We render our training and validation sets using the reconstructed 3D models of real human heads from a public dataset, FaceScape [YZW*20], leading to realistic renderings close to real portraits. In addition, we propose an effective domain adaptation module to enhance the generalizability of our network to real captured images. In particular, we apply additional CNN layers, appending to our CNN feature extractor, to regress the original input image sent to the feature extractor. We train this regression module using a large number of real portrait images in CelebAMask-HQ [LLWL20], which effectively regularizes our feature extractor along with our full network to understand diverse real portrait appearance. This regularization module effectively improves our rendering quality on real portrait images (compared with the rightmost image in Fig. 5).

We train our entire network from end to end with a final loss that combines the lighting estimation loss, domain regularization loss, and governing rendering losses with multiple combinations of various lighting and viewing conditions. This leads to our final generalizable network that can synthesize realistic portrait images under novel viewpoints and lighting conditions from only five input images. We demonstrate that, when evaluating on the rendered

validation set, our approach can produce smooth and realistic relighting and view synthesis results that are very close to the ground truth; ours can qualitatively and quantitatively outperform baseline solutions that run state-of-the-art view synthesis [WWG*21] and relighting [SBT*19] techniques in a sequence (see Fig. 4). We also demonstrate that our approach can achieve photo-realistic renderings from captured real portrait images, significantly better than the comparison methods (see Fig. 7).

We summarize our contributions as follows:

- a novel neural representation that models scene appearance as light transport functions and enables relighting for neural volumetric rendering (Sec. 3.1, Sec. 3.2);
- a domain adaptation module to enhance the generalizability of the network trained on rendered images (Sec. 3.4);
- realistic practical rendering results of joint relighting and view synthesis of real portraits from only five captured images (Sec. 4.4, Fig. 1, Fig. 7).

2. Related Work

2.1. Portrait Appearance

There has been extensive research work in capturing and modeling human portraits [BV99; DHT*00; KB10; TL18; SKCJ18; ZTG*18]. Various morphable face models have been presented [TL18; GVR*14; WBLP11], mainly aiming to achieve facial animation and reenactment. Our focus is to capture human portraits and synthesize photo-realistic images. While generative models [GPM*14; KALL17; KLA19] can produce realistic images, they cannot be easily applied for capture and reconstruction, which aims to generate images of specific captured real human heads with given viewpoints and lighting conditions. While previous work can do so by reconstructing meshes with simple reflectance models from portrait images [BM14; SKCJ18], their rendering quality is often limited. We instead leverage the recent neural rendering techniques [TFT*20], leading to photo-realistic portrait renderings for joint view synthesis and relighting.

2.2. Relighting

Many image-based relighting papers have been presented [DHT*00; MLP04; PML*09; RDL*15]; they leverage the linearity of light transport and introduce various techniques to acquire the light transport function, whose discrete form is a light transport matrix that consists of per-pixel light transport vectors at a fixed viewpoint. Inspired by these prior works, we also model linear light transport functions; instead of predicting per-pixel coefficients, we predict light transport at a point in the 3D space, enabling both view synthesis and relighting.

Recently, deep learning techniques have been introduced to address the relighting problem [XSHR18]. Many deep learning works are specifically designed to relight human portraits [SBT*19; ZHSJ19; MHP*19; NLML20; SXZ*20]; however most works focus on relighting at a fixed viewpoint and cannot change the viewpoint. While other works can jointly do relighting and view synthesis [BXS*20b; MPH*20; GCD*20; ZFT*21], they require complex capture setups to acquire a large number images under controlled
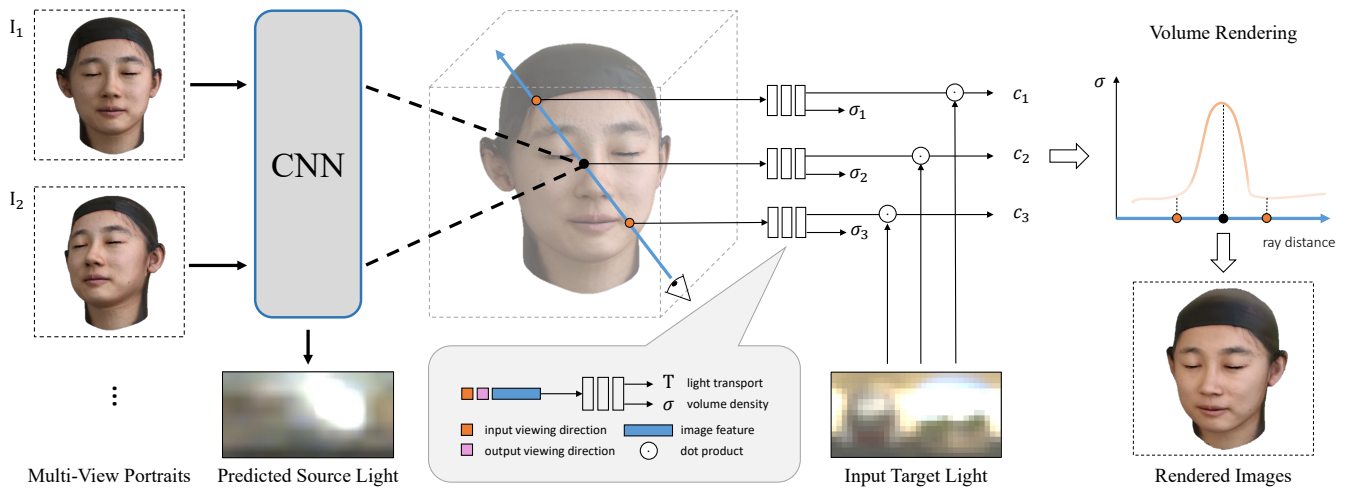
Figure 2: *Overview of our method. The proposed algorithm takes multi-view portraits as input and predicts the source environment map, light-transport and volume density at a query point. We then use the predicted light-transport and volume density to perform the joint task of view synthesis and relighting.*

lighting. In contrast, our approach enables realistic relighting and view synthesis for human portraits using only a sparse set of input images under natural illumination.

2.3. View Synthesis

View synthesis has been studied by the computer graphics and vision community for decades. Early work [CW93; MB95] utilizes view interpolation to render novel views from neighboring source views. Other classical methods, including light fields [LH96; GGSC96] and image-based rendering [DTM96; BBM*01; SSS09], have also been proposed to address view synthesis. Recently, deep learning methods [ZTF*18; MSO*19; XBS*19; FNPS16] have become dominant in this field of work. Learning-based methods have proven to be more expressive and they are able to represent various complex scenes with challenging visual effects. One recent work [XYC*20] combines morphable face models and generative techniques to reconstruct 3D portraits for rendering with changing viewpoints. However, it is non-trivial to extend these view synthesis techniques to support relighting at the same time. We propose a novel neural rendering approach that can simultaneously do relighting and view synthesis.

2.4. Neural Rendering

In addition to the aforementioned view synthesis methods, an exciting advancement is neural radiance field (NeRF) [MST*20], which encodes a 3D scene in a compact 5D continuous radiance field function represented by a multi-layer perceptron (MLP) and renders the radiance field using differentiable volume rendering. Concurrent works have extended NeRF to render human portraits [PSB*20; WBL*20; LSS*21]; however most of them rely on overfitting the network to a single target as is done in the original NeRF, which does not generalize to other unseen portraits. Other

recent works [YYTK20; WWG*21] leverage CNNs to perform per-view radiance field reasoning, leading to a generalizable neural model for view synthesis. Our approach is inspired by these CNN-based radiance field estimation techniques; we introduce novel light transport estimation modules in the per-view scene reasoning, enabling relighting in the neural volumetric rendering process.

Previous neural rendering methods have also achieved relighting. Some methods leverage controlled lighting to achieve reflectance estimation with per-scene optimization [BXS*20c; BXS*20a]. Our approach instead supports relighting under natural illumination, jointly estimating lighting and lighting transport functions. Other concurrent works [SDZ*21; MRS*21] can also change the lighting conditions but still rely on per-scene optimization with a large number of images. Our approach achieves a generalizable neural network that learns specific human portrait shape and appearance priors from large training datasets. Additionally, our approach enables efficient portrait relighting and view synthesis from only a sparse set of input images.

3. Method

In this paper, we present Neural Light-transport Field (NeLF) to solve the problem of simultaneous portrait view synthesis and relighting. Our system takes as input a small set of 5 images of an unseen human face, which are captured roughly from the frontal view of the portrait. We assume the captured portraits are lit by the same distant light, which can be modeled by an environment map.

Given the captured images as well as the corresponding camera parameters, we produce a volumetric field of *light-transport* (Sec. 3.1). Each point in the 3D scene has a volume density σ, and a light-transport T. The light-transport is a vector whose dot-product with the global environment map produces the outgoing radiance at the 3D point. The predicted neural light-transport field

(NeLF) enables us to perform view synthesis and relighting on the captured portraits. Given a new camera view and a new lighting environment, we can use the predicted light-transport to compute the outgoing radiance of each 3D point, and use a volume rendering algorithm [MST*20] to render novel views of the captured human face under the new lighting.

We show an overview of our method in Fig. 2 and a detailed illustration in Fig. 3. In Sec. 3.1, we briefly review the mathematical definition of the light-transport. Section 3.2 describes how we predict the volume density and the light-transport at each 3D point, and Sec. 3.3 explains our detailed volume rendering algorithm. We show our novel domain adaptation module in Sec. 3.4. Finally, the implementation details are included in Sec. 3.5.

3.1. Light Transport

Light transport of a 3D point describes the relationship between its radiance and the global lighting environment. The outgoing radiance L_r of each 3D point $\mathbf{x}$ can be computed according to the rendering equation [Kaj86]:

$$L_r(\mathbf{x}, \omega_o) = \int_{\omega_i} \rho(\mathbf{x}, \omega_i, \omega_o)(\mathbf{n} \cdot \omega_i) L_i(\mathbf{x}, \omega_i) d\omega_i, \qquad (1)$$

where ρ is the BSDF function, $\mathbf{n}$ is the normal direction at the 3D point, and ω_i, ω_o are the incoming and outgoing ray directions, respectively. Traditional path tracing computes the radiance L_i recursively until the ray hits the global environment L. However, after unravelling the recursion, the radiance has a linear relationship with the environment map:

$$L_i(\mathbf{x}, \omega_i) \propto L(\omega_i'). \qquad (2)$$

Here, the incoming radiance L_i of the point $\mathbf{x}$ at each incoming direction ω_i is proportional to the corresponding global environment light $L(\omega_i')$ at each direction ω_i'. Notice that this linear relation has modeled all the paths that come from the environment to this point, modeling effects caused by shadowing, inter-reflection, material properties of the ray bounces in between, etc. If we bring Eqn. 2 back to Eqn. 1, we can simplify Eqn. 1 as a linear integration:

$$L_r(\mathbf{x}, \omega_o) = \int_{\omega_i} T(\mathbf{x}, \omega_o, \omega_i) L(\omega_i) d\omega_i. \qquad (3)$$

Here, the *light transport T* has already encoded all the complexity from the environmental light L to the outgoing radiance L_r at $\mathbf{x}$.

As shown in Eqn. 3, given the light transport T, we can compute the outgoing radiance at each 3D point $\mathbf{x}$ under a novel environment map L using an integration, and then efficiently render the scene under arbitrary lighting environments. This type of method is called Precomputed Radiance Transfer (PRT) [SKS02], which has been well studied in traditional computer graphics [TS06; NRH03]. In our method, we express the global lighting simply as a 8×16 environment map $\mathbf{L}$. We predict a light-transport vector $\mathbf{T}$ at each 3D point $\mathbf{x}$ given an input viewing direction. $\mathbf{T}$ also shares the same size as the environment map $\mathbf{L}$. As a result, we can simplify Eqn. 3 to a dot product on each color channel:

$$\mathbf{L}_r = \mathbf{T} \cdot \mathbf{L}. \qquad (4)$$

This is similar to the light transport matrix in image-based relighting [DHT*00], which models per-pixel light transport. In our work, we compute the radiance at each 3D point following Eqn. 4, and later render the radiances into pixel values following the neural volumetric rendering equation in Eqn. 10.

3.2. Neural Light-transport Field

Given N different views of a human face, we use a neural network to predict a volumetric light-transport field. The key concept of our algorithm is to predict the light-transport vector $\mathbf{T}$ of each 3D point by aggregating the information from the input portraits. To this end, we utilize the ideas of the recent image-based rendering techniques [WWG*21] to predict the light-transport. Our system consists of two parts (see Fig. 3): a convolutional neural network (CNN) and several multi-layer perceptrons (MLP). The CNN operates on the captured portraits to extract the image features and predict the source lighting condition. On the other hand, the MLPs predict the light transport $\mathbf{T}$ as well as the volume density σ for each 3D point in the scene, using multiple features. Then for any given target environment map $\mathbf{L}'$, we can simple calculate the color with Equ. 4 and use it for volumetric rendering (Sec. 3.3).

3.2.1. Image Feature Extraction and Source Light Prediction

We apply an U-Net style convolutional neural network [RFB15] to N captured images. For an input portrait $\mathbf{I}_k$ at viewpoint k, the CNN extracts the image feature $\mathbf{F}_k$, as well as the source lighting feature $\mathbf{L}_k$ and confidence weights $\mathbf{W}_k^{\mathbf{L}}$ (see Fig. 3(a)). The size of image feature $\mathbf{F}_k$ is half of the original input image $\mathbf{I}_k$, and it is later used for light transport prediction.

Each lighting feature consists of two parts: a predicted source environment map $\mathbf{L}_k$, and a per-direction confidence map $\mathbf{W}_k^{\mathbf{L}}$ [SBT*19]. We predict the confidence map $\mathbf{W}_k^{\mathbf{L}}$ for all cameras, as each camera covers only parts of the portrait, providing partial lighting information. For example, it is less accurate to predict the light coming from the left by looking at the right side of the face. To resolve this issue, we can merge the lighting predictions $\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_N$ with their corresponding confidence maps $\mathbf{W}_1^{\mathbf{L}}, \mathbf{W}_2^{\mathbf{L}}, ..., \mathbf{W}_N^{\mathbf{L}}$ as weights.

Nonetheless, another issue is that the network is unaware of the camera pose when predicting the lighting environment. As a result, the network is only able to predict the lighting relative to each camera, and each predicted environment-map is defined in its corresponding camera coordinate system. In order to align multiple light predictions from different coordinate systems and merge them into a global environment map, we define a rotation operator $\mathcal{R}$ that rotates them to a canonical world coordinate, provided camera extrinsics. Finally, we compute the global lighting environment with a weighted average:
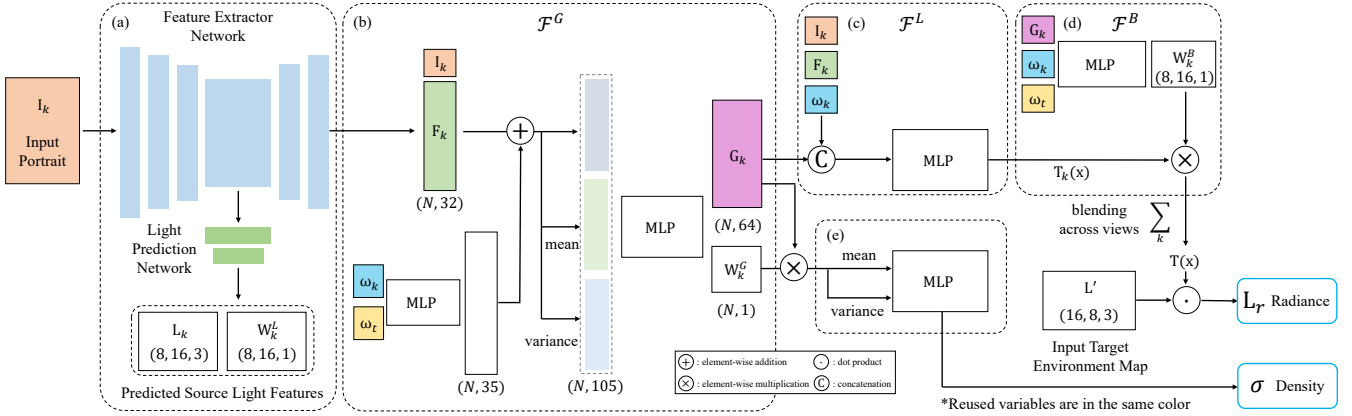
$$\mathbf{L} = \frac{\sum_k \mathcal{R}(\mathbf{L}_k \odot \mathbf{W}_k^{\mathbf{L}})}{\sum_k \mathcal{R}(\mathbf{W}_k^{\mathbf{L}})}, \qquad (5)$$

where $\odot$ means element-wise multiplication.

3.2.2. Volume Density and Light-transport Prediction

We use multiple MLPs to predict the volume density $\sigma(\mathbf{x})$ and the light transport $\mathbf{T}(\mathbf{x})$ at each point $\mathbf{x}$ observed from the target cam-

Figure 3: *Detailed network structures of our proposed algorithm. (a) For an input image $\mathbf{I}_k$, we first extract its lighting feature $\mathbf{L}_k$, weights $\mathbf{W}_k^L$ and image feature $\mathbf{F}_k$ (Sec. 3.2.1). (b) Then, as described in Sec. 3.2.2, we aggregate features from different views to extract the geometry feature, $\mathbf{G}_k$, and weights $\mathbf{W}_k^G$. From here, our network diverges into two paths: the first part (c) predicts the per-view light transports $\mathbf{T}_k$ using a per-view MLP $\mathcal{F}^L$ and then (d) aggregates them with a blending MLP $\mathcal{F}^B$ to generate the full light transport $\mathbf{T}$; the second part (e) predicts the volume density σ using an MLP from the multi-view statistics of $\mathbf{G}_k$ and $\mathbf{W}_k^G$.*

era t. We first project the point $\mathbf{x}$ to all source cameras to acquire the corresponding image features $\mathbf{F}_k(\pi_k(\mathbf{x}))$, where π_k denotes the projection to camera k. We also compute source viewing direction ω_k and target viewing direction ω_t. The image features, together with ω_k and ω_t, are fed into MLPs, $\mathcal{F}^G$, to extract the multiview-aware geometry feature $\mathbf{G}_k$ and a corresponding weight $\mathbf{W}_k^G$ for each source view (see Fig. 3):

$$\mathbf{G}_k, \mathbf{W}_k^G = \mathcal{F}^G\left(\omega_t, \{\omega_k\}_{k=1}^N, \{\mathbf{F}_k(\pi_k(\mathbf{x}))\}_{k=1}^N\right). \quad (6)$$

Shown in Fig. 3(b), the per-view geometry feature is extracted using a PointNet style MLP structure [QSMG17], using the per-element mean and variance of the feature as additional inputs (see Fig. 3). This shares the idea of traditional image-based rendering [MB95]: the projected image features from multiple cameras should be consistent around the actual object surface. In our setup, $\mathcal{F}^G$ compares the input feature vector with its per-element mean and variance, and learns to assign more weight $\mathbf{W}_k^G$ to the best matched view. This geometry feature is used to predict the volume density $\sigma(\mathbf{x})$ of the query point using another MLP as shown in Fig. 3 (e).

It is highly challenging to predict the light transport vector for the novel viewing direction. Therefore, we instead predict per-view light transports using an MLP $\mathcal{F}^L$ (Fig. 3(c)) and then blend them for the novel view (Fig. 3(d)). While we can apply $\mathcal{F}^L$ to directly regress the light transport vector, this can easily lead to the network memorizing the portrait appearance in the synthetic training set. We propose to let $\mathcal{F}^L$ regress scales relative to the pixel colors; the predicted scales have the same dimensions as the light transport, and compute the light transport vector by multiplying by the pixel colors. This design effectively retains the high-frequency information in the original input images and enables better generalizability of our network to unseen real portraits (see Fig. 5). In particular, the light-transport vector of the source view k at point $\mathbf{x}$ is computed

by:

$$\mathbf{T}_k(\mathbf{x}) = \mathbf{I}_k(\pi_k(\mathbf{x})) \cdot \mathcal{F}^L(\omega_k, \mathbf{G}_k, \mathbf{F}_k(\pi_k(\mathbf{x}))), \quad (7)$$

Note that, to ensure view-consistent light transport estimations, we leverage the multiview-aware feature $\mathbf{G}_k$ in this per-view light transport prediction.

We repeat this operation to acquire $\mathbf{T}_1(\mathbf{x}), \mathbf{T}_2(\mathbf{x}), ..., \mathbf{T}_N(\mathbf{x})$. These light-transport functions encode the information at the same 3D point, but are observed from different angles. We then calculate final target light-transport $\mathbf{T}(\mathbf{x})$ as a linear combination of $\mathbf{T}_1(\mathbf{x}), \mathbf{T}_2(\mathbf{x}), ..., \mathbf{T}_N(\mathbf{x})$. We use another MLP $\mathcal{F}^B$ to predict the blending weights

$$\mathbf{W}_k^B = \mathcal{F}^B(\omega_k, \omega_t, \mathbf{G}_k), \quad (8)$$

and linearly combine the light-transports of the source views to get the final light-transport

$$\mathbf{T}(\mathbf{x}) = \sum_k \mathbf{W}_k^B \odot \mathbf{T}_k(\mathbf{x}). \quad (9)$$

3.3. Efficient Volume Rendering

For a given target environment map $\mathbf{L}'$, we follow Equ. 4 to compute the radiance $\mathbf{L}_r(\mathbf{x})$. The pixel value $\mathbf{I}_t$ of the novel viewpoint can be computed using the volumetric rendering equation from the original NeRF formulation [MST*20]. Suppose the point at depth u is $\mathbf{x}(u) = \mathbf{x}_c + u \cdot \omega_c$, where $\mathbf{x}_c$ is the camera location and ω_c is the ray direction, we have:

$$\mathbf{I}_t = \int_{u_n}^{u_f} \exp\left(-\int_0^u \sigma(\mathbf{x}(v))dv\right)\sigma(\mathbf{x}(u))L_r(\mathbf{x}(u))du. \quad (10)$$

To be more specific, we integrate along the light ray with u_n, u_f as the near and far bounds, respectively. We additionally predict the depth and the alpha channel, which are then used for supervision

(Sec. 3.5):

$$\mathbf{D}_t = \int_{u_n}^{u_f} \exp\left(-\int_0^u \sigma(\mathbf{x}(v))dv\right)\sigma(\mathbf{x}(u))u \cdot du,$$
$$\mathbf{A}_t = \int_{u_n}^{u_f} \exp\left(-\int_0^u \sigma(\mathbf{x}(v))dv\right)\sigma(\mathbf{x}(u))du. \quad (11)$$

The original NeRF paper computes the radiance at each sample point. However, most of the sample points are 0 when the scene is spatially sparse. This creates a lot of redundancy when rendering a new image. In our setup, since portraits are usually sparse, we exploit the concept of *visual hull* to prune out the queries that are unnecessary. From the input portraits, we utilize a portrait matting algorithm [KLZ*20a] to extract their mask $\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_N$. Then for each 3D point $\mathbf{x}$ viewed by the target camera, we project its position onto all the masks, and query the MLPs if all the projections onto $\mathbf{M}_k(\pi_k(\mathbf{x}))$ are nonzero. Otherwise, we directly set the light transport $\mathbf{T}(\mathbf{x})$ and the density $\sigma(\mathbf{x})$ to be 0. This is essentially restricting the network to learn within the visual hull defined by the silhouette of the portraits. By doing this, we achieve faster convergence during training and more efficient rendering during inference.

3.4. Domain Adaptation

We train our network on synthetically rendered human face data (see Sec. 4.1). This data does not model the distribution of the real human faces very well. Thus, naively training our network using the rendered faces results in poor generalizability on real portraits. It is optimal to include real multi-view portraits in our training data. However, there is no publicly available large-scale human face dataset and it is also challenging to capture real portraits under different viewpoints and lighting without a light stage.

To this end, we propose a novel domain adaptation module that effectively enhances the generalizability of our network by augmenting the CNN feature extractor with a large number of real portrait images in the CelebAMask-HQ dataset [LLWL20]. These real images do not have groundtruth labels of their lighting and camera parameters; however we show that they can be effectively used to regularize our feature extractor to adapt to the distribution of real portrait images. In particular, for each training iteration, we additionally feed in a real portrait into the feature extractor and get an image feature. Rather than using the image feature to predict the light transport and perform volumetric rendering, we append 2 more convolution layers to the U-Net to directly recreate the original image, enhancing the expressiveness of the feature extractor. With this joint training strategy, we can equip the feature extractor with stronger capabilities to reason about real portrait images and avoid overfitting to the biased distribution of the synthetic portrait dataset.

3.5. Implementation Detail

In general, there are two kinds of relighting tasks we need to solve: changing to a completely novel environment light, or rotating the original light. In order to solve both tasks in NeLF, we implement two modes during training: novel light mode and self rotation mode. In the novel light mode, we provide a new lighting as the

target light to the MLP to compute the radiance. In the self rotation mode, we predict the lighting of the input portrait from the CNN, rotate the predicted light for a certain angle, and use that as the target light. We allocate 70% of the training for novel light, and the rest for self rotation.

Our network is supervised with multiple losses. We implement a rendering loss $\mathcal{L}_c$, a depth loss $\mathcal{L}_d$, an alpha mask loss $\mathcal{L}_a$, a lighting loss $\mathcal{L}_t$ and an image consistency loss $\mathcal{L}_p$. To begin with, the rendering loss $\mathcal{L}_c$ is simply an $\mathcal{L}_1$ loss on the predicted RGB $\mathbf{I}_t$ such that it is as close as possible to the ground truth value $\mathbf{I}_{gt}$

$$\mathcal{L}_c = ||\mathbf{I}_t - \mathbf{I}_{gt}||_1. \quad (12)$$

We additionally supervise the predicted depth and the alpha channel using the ground truth values. This is to make sure the predicted shape is meaningful and obeys multi-view constraints. These losses are formulated as:

$$\mathcal{L}_a = ||\mathbf{A}_t - \mathbf{A}_{gt}||_1, \mathcal{L}_d = \frac{1}{d}||\mathbf{D}_t - \mathbf{D}_{gt}||_1, \quad (13)$$

where $d = 200mm$ is the average size of a human head. For lighting, we use log-$\mathcal{L}_1$ loss on the predicted environment map L and the ground truth environment map $\mathbf{L}_{gt}$:

$$\mathcal{L}_t = ||\log(1+\mathbf{L}) - \log(1+\mathbf{L}_{gt})||_1 \quad (14)$$

As mentioned in Sec. 3.4, we train our network jointly with a self consistency constraint on the image encoder to ensure that the CNN can learn meaningful representations of real human faces. For each iteration, we randomly pick a in-the-wild portrait $\mathbf{I}_c$ from the CelebAMask-HQ dataset. We feed the image only into the feature extractor to get a reconstruction of the same image $\hat{\mathbf{I}}_c$. We supervise the reconstruction using an image consistency loss $\mathcal{L}_p$:

$$\mathcal{L}_p = ||\hat{\mathbf{I}}_c - \mathbf{I}_c||_1. \quad (15)$$

Notice that we only apply the image consistency loss on the random in-the-wild portrait $\mathbf{I}_c$. We don't enforce the image encoder to reproduce the calibrated multi-view images. Together, the final loss $\mathcal{L}_{total}$ is a sum of all above-mentioned losses:

$$\mathcal{L}_{total} = \mathcal{L}_c + \mathcal{L}_d + \mathcal{L}_a + \mathcal{L}_t + \mathcal{L}_p. \quad (16)$$

The detailed network structures of our CNN and MLP are shown in Fig. 3. We implement our method in PyTorch [PGM*19]. We use Adam [KB14] as our optimizer and the learning rates are set to 10^{-4} for the MLPs and 2×10^{-4} for the image encoder. Our network is trained on 4 NVIDIA 2080Ti GPUs for 300k steps, which takes around 36 hours.

4. Results

4.1. Data

In order to train our novel representation, we choose the FaceScape dataset [YZW*20] as it provides a good collection of realistic human head models with high-resolution texture and displacement maps. Due to privacy issues, some of the models have blurred textures around the eyes. Thus, we select 360 unblurred models from the whole dataset and use Blender [Ble20] with the Cycles renderer to generate the training and testing data images in 512×512 resolution. To be more specific, we generate data triplets that contain

| Inputs | Groundtruth | Ours | SIPR+IBRNet | IBRNet+SIPR |
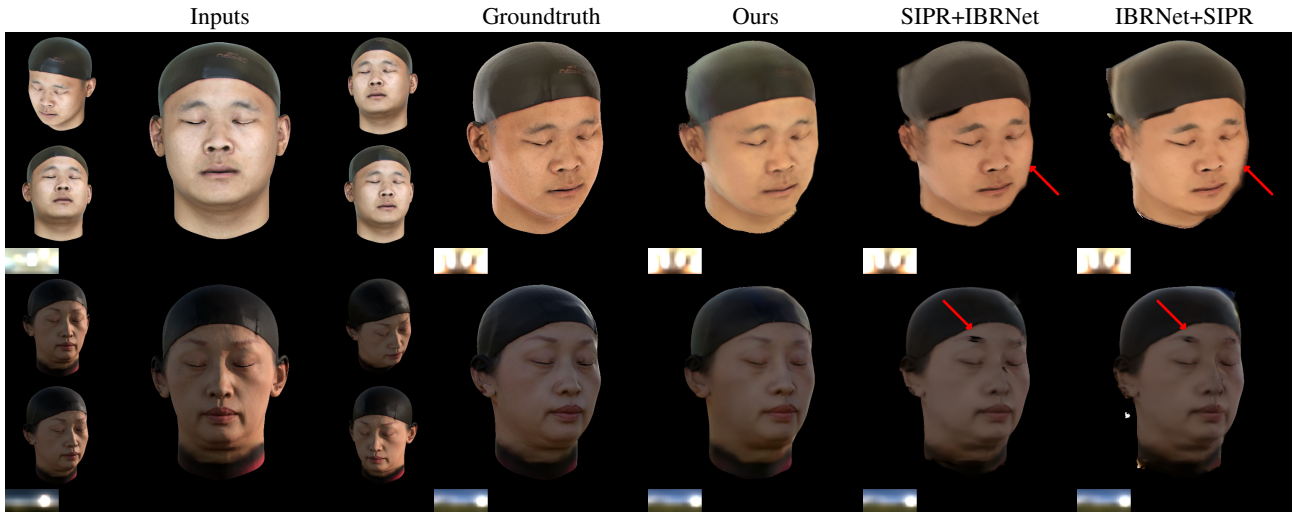
Figure 4: *We compare our results of simultaneous view synthesis and relighting with possible baseline methods on unseen samples from our synthetic evaluation dataset. Baseline methods fail to provide accurate geometry and contain random hole artifacts. For instance, SIPR + IBRNet shows some hole artifacts near the nose area and IBRNet + SIPR shows blurry results around the cheek in the top right image. Please refer to the supplementary video for more results.*

Table 1: *Quantitative comparison on synthetic evaluation dataset. Our method performs better in both PSNR and SSIM over baseline methods, providing better visual quality. For comparison against direct prediction of light transport, while the visual artifacts are disturbing in Fig. 4, they reflect only a small part of the image, and therefore do not have as much of an impact on the PSNR and SSIM scores. Please refer to supplementary video for better comparisons.*

Method	PSNR	SSIM
SIPR + IBRNet	33.55	0.8826
IBRNet + SIPR	33.18	0.8611
Ours w/o encoder augmentation	33.54	0.8916
Ours w/ direct light transport	**33.70**	**0.8928**
Ours	**33.61**	**0.8922**

(a) 5 source views with the first one being the frontal view, (b) a novel view with the same lighting but randomly rotated, and (c) a novel view with lighting randomly selected from a pool of environment maps. (a) is used as input to our network, (b) is the self rotation mode supervision, and (c) is the novel light mode supervision. Both (b) and (c) also provide novel view supervision. The novel views are uniformly sampled within $30°$ in azimuthal and $30°$ in elevational angle from the frontal face view. The distance of the cameras is chosen randomly from 100cm to 200cm to simulate real life captures. We also adjust the field-of-view accordingly so that the face would take up most space. This training triplet design allows for learning of both geometric and lighting information.

4.2. Comparison with previous methods

To the authors' best knowledge, our algorithm is the first to achieve simultaneous view synthesis and relighting on unseen subjects

from a sparse set of multi-view images under natural illumination. SIPR [SBT*19] can perform relighting on unseen portraits given a single image, while IBRNet [WWG*21] achieves view synthesis on novel scenes. We combine these previous two state-of-the-art algorithms and use the combinations as the baseline of our algorithm. There are two possible ways to combine: first do the relighting on the input views, and then synthesize the new view (SIPR+IBRNet); or first do view synthesis and then change the lighting of the synthesized image (IBRNet+SIPR). Notice that we have to correctly rotate the target light to each view in order to align the target light with the camera.

We retrain the network of IBRNet and SIPR on our synthetic dataset for fair comparisons. We choose 4 subjects out of the total 360 identities to serve as our evaluation dataset. Table 1 shows the quantitative comparison of these two baselines, as well as our method. Our method outperforms both of the baselines with both higher PSNR and SSIM. In addition, we observe that the baseline methods often lead to obvious visual artifacts in their renderings as shown in Fig. 4 (please see our supplementary video for more examples). Our approach instead can achieve much higher visual quality, consistently producing realistic renderings across different lighting and viewpoints.

These two baselines fail for different reasons. If relighting the input images first, the relighting algorithm might perform slightly differently on each input view, which will break the multi-view consistency for view synthesis. On the other hand, doing the relighting after the view synthesis also does not work well, since the relighting algorithm now has no access to the multi-view information. In contrast, our approach learns to effectively aggregate multi-view appearance features to predict the light transport of each shading point, leading to accurate relighting effects that are con-
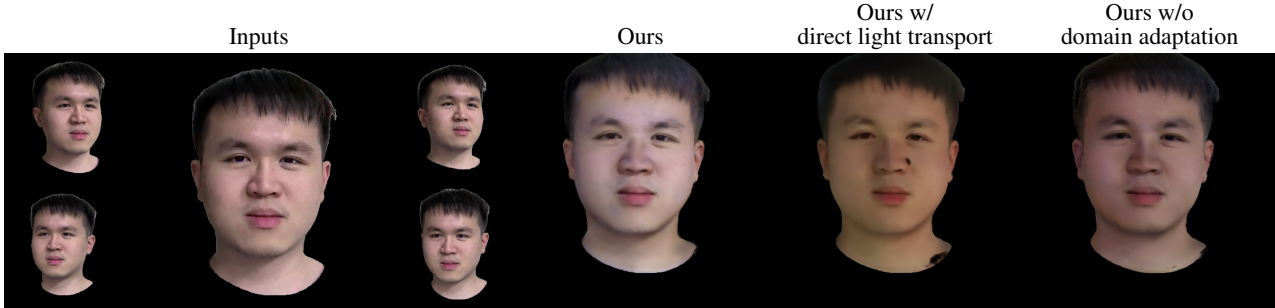
Figure 5: *We compare our method with two possible ablations on a real captured image. We show that by modulating the light transport prediction and applying our domain adaption method, our method can generalize well to real portraits with minimum color shift or artifacts. Please refer to the supplementary video for more results.*

Table 2: *Quantitative comparison on the task of relighting and view-synthesis individually on synthetic dataset. Our method performs comparably, but slightly worse than the state-of-the-art on each task.*

	Ours	SIPR	IBRNet
Relighting	0.9053	**0.9279**	—-
View-synthesis	0.8683	—-	**0.8949**

Table 3: *Quantitative comparison on different input view counts. We show that 5 views provide the best possible results in both PSNR and SSIM metrics. Visual quality starts to degrade as view count reduces.*

View Num	PSNR	SSIM
2 Views	30.91	0.6388
3 Views	32.77	0.8177
4 Views	33.40	0.8731
5 Views	**33.61**	**0.8922**

sistent across multi-view viewpoints. Our approach can simultaneously achieve high-quality relighting and view synthesis.

We have also evaluated our method on relighting and view-synthesis respectively using our synthetic dataset. We perform relighting by rendering the view at the frontal camera under new lighting, and do view synthesis by first estimating the original light, and then relighting under the predicted lighting from new views. As shown by the SSIM values in Tab. 2, our method performs comparably, but slightly worse than the baselines on each task. This is because our main goal is not to improve individual performance, but to enable the combination of these two tasks. Thus, our view-synthesis performance also includes the error from lighting estimation and relighting. Even though SIPR and IBRNet perform slightly better in their individual tasks, their combination does not work as well as ours as shown in Fig. 4. Our method aggregates multi-view features to predict the light-transport on each shading point, leading to multi-view consistent relighting effects.

4.3. Ablations

We ablate our algorithm by directly predicting the light transport rather than modulating the prediction using the pixel value in Eqn. 7. The quantitative results of synthetic validation set and qualitative results of real portraits are shown in Tab. 1 and Fig. 5 respectively. Note that, although the network with direct light transport achieves slightly higher performance on the validation dataset, our full model performs much better on the real portraits as shown in Fig. 5. In particular, directly predicting the light-transport can lead to artifacts, for instance the black hole around the nose area in the portrait. In addition, the method introduces more color shift compared to our full algorithm. Our full model learns to predicts scales

of the original pixel colors; this retains the original color signals and can generalize better to unseen real portraits.

In addition, we show the effect of our domain adaptation module in Fig. 5. We show that with the additional adaptation module, our network is able to generalize to unseen color distributions much better than without this module. As can be seen from the figure, without the adaptation, the network fails to recreate the vibrancy of the input images. Moreover, the left part of the portrait exhibits more orange color than our proposed method. Our domain adaptation module essentially regularizes the network to be able to better reproduce the appearance in the original input images. This is not only helpful when testing on real portrait images, but can also improve our performance on the synthetic validation set as shown in Tab. 1.

We also study the effects of input view numbers on the rendering quality. Our network is originally trained with 5 images as the input views. We test our network on unseen subjects from the validation dataset by feeding 2, 3, and 4 views into the network for view synthesis and relighting. Table 3 shows the quantitative evaluation and Figure 6 shows the results. Our network fails to render meaningful content when only 2 images are given. However, as we have 3 or more images as the input views, our network can render new views under new lighting with high quality.
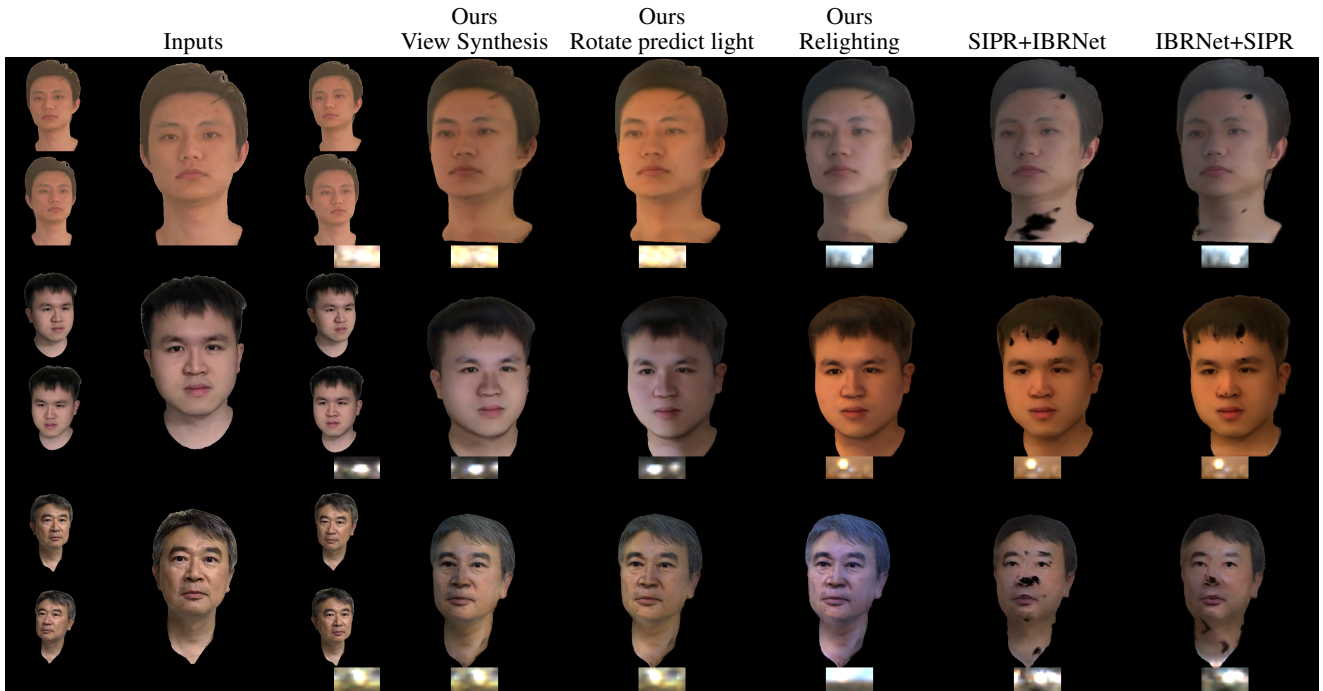
4.4. Real Portraits

In order to show the performance of our method on real data, we capture real portraits with smartphones by shooting a video
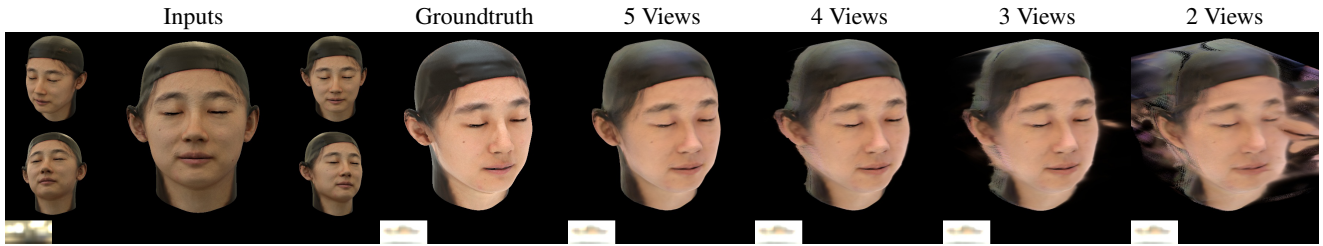
Figure 6: *Qualitative evaluation on the effects of different input view counts. We show that there is a small difference between 4 and 5 views, while 2 and 3 views result in more artifacts due to occlusions and the lack of stereo coverage.*

Figure 7: *Qualitative results on real portraits. Our method produces better visual quality than the baseline methods of SIPR and IBRNet. For example, in the forehead and neck regions, baseline methods produce hole artifacts, whereas our proposed method is able to synthesize realistic novel images under new lighting conditions. Please refer to the supplementary video for more results.*

clip of the subject. We employ COLMAP [SZPF16; SF16] to recover camera poses and rough depth statistics to determine near and far planes. We also normalize the camera poses such that the origin is approximately the head center. This is done by first extracting the center of facial landmarks with Bulat et al. [BT17], and estimating the face direction to calculate the head center. We apply MODNet [KLZ*20b] to clean up the background and use BiSeNet [YWP*18] to further remove garments and keep the face portion. As our training data does not contain any clothes, this can ensure that the rendered results are not perturbed by unrelated information.

The results are shown in Fig. 1, Fig. 7 and the supplementary video. We can observe that our method achieves state-of-the-art quality on the joint task of view synthesis and relighting. Our proposed method is able to infer source environment map and render novel view portraits with the rotated source light. Additionally,

we can input a target environment map to control the new lighting condition. For the task of rendering real portraits, the comparison baselines (both SIPR+IBRNet and IBRNet+SIPR) lead to obvious artifacts with many visible holes, more obvious than their artifacts on the synthetic validation set; this is because the baseline methods' networks overfit to the training set's data distribution and the same issue of inconsistency between separate relighting and view synthesis modules (as discussed in Sec. 4.2) becomes more significant on the real data. Thanks to our effective modeling of the light transport in the 3D space and our domain adaptation module, our model can synthesize realistic relighting and view synthesis results on the real portrait images, leading to significantly better results than the baseline methods.

4.5. Limitations

Although our proposed method generates photo-realistic rendering results, it still possesses some limitations. For example, since we train on a synthetic dataset, our model might exhibit some color shifts in certain cases when tested on real portraits. This can be ameliorated by training on a real multi-view portrait dataset. Another issue is the slight blurriness in our rendered results. This is possibly caused by the limited network capacity, which can be further increased to allow extraction of image features with higher resolution. In addition, we downsample the feature map to allow for faster training and inference. With more computational resources, it is possible to use full-scale image features. Last, because our training dataset have a global specularity coefficient instead of dedicated specularity maps, some complex view-dependent effects such as glints are not well reconstructed in the training images. As a result, our method fails in some cases where high-frequency specular highlights are presented.

5. Conclusions and Future Work

In this paper, we tackle the joint problem of portrait view synthesis and relighting, which prior works fails to handle. We solve this problem by introducing the *neural light-transport field*, which encodes the volume density and light transport vectors of each 3D point in the scene, enabling relighting with a target environment map. We demonstrate that with only 5 input views, our method is able to generalize across unseen portraits and produce better portrait renderings than previous approaches built specifically for either view synthesis or relighting.

In future work, we would like to explore the possibility of even fewer input images. We show that our method degrades for a smaller number of views. This is caused by the depth ambiguities and unseen areas, which requires prior knowledge to hallucinate reasonable renderings. Another possible direction is animated relightable avatars as our proposed method does not handle a talking head explicitly. All in all, we believe that the joint problem of view synthesis and relighting is crucial in immersive applications like mixed reality, and our work can advance the field in that direction.

References

[BBM*01] BUEHLER, CHRIS, BOSSE, MICHAEL, MCMILLAN, LEONARD, et al. "Unstructured lumigraph rendering". *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM. 2001, 425–432 3.

[Ble20] BLENDER ONLINE COMMUNITY. *Blender - a 3D modelling and rendering package*. Blender Foundation. Blender Institute, Amsterdam, 2020. URL: http://www.blender.org 6.

[BM14] BARRON, JONATHAN T and MALIK, JITENDRA. "Shape, illumination, and reflectance from shading". *IEEE transactions on pattern analysis and machine intelligence* 37.8 (2014), 1670–1687 2.

[BT17] BULAT, ADRIAN and TZIMIROPOULOS, GEORGIOS. "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)". *International Conference on Computer Vision*. 2017 9.

[BV99] BLANZ, VOLKER and VETTER, THOMAS. "A morphable model for the synthesis of 3D faces". *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 1999, 187–194 2.

[BXS*20a] BI, SAI, XU, ZEXIANG, SRINIVASAN, PRATUL, et al. "Neural reflectance fields for appearance acquisition". *arXiv preprint arXiv:2008.03824* (2020) 1, 3.

[BXS*20b] BI, SAI, XU, ZEXIANG, SUNKAVALLI, KALYAN, et al. "Deep 3d capture: Geometry and reflectance from sparse multi-view images". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 5960–5969 2.

[BXS*20c] BI, SAI, XU, ZEXIANG, SUNKAVALLI, KALYAN, et al. "Deep reflectance volumes: Relightable reconstructions from multi-view photometric images". *arXiv preprint arXiv:2007.09892* (2020) 3.

[CW93] CHEN, SHENCHANG ERIC and WILLIAMS, LANCE. "View interpolation for image synthesis". *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*. 1993, 279–288 3.

[DHT*00] DEBEVEC, PAUL, HAWKINS, TIM, TCHOU, CHRIS, et al. "Acquiring the Reflectance Field of a Human Face". *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. USA: ACM Press/Addison-Wesley Publishing Co., 2000, 145–156. ISBN: 1581132085. DOI: 10.1145/344779.344855. URL: https://doi.org/10.1145/344779.344855 2, 4.

[DTM96] DEBEVEC, PAUL E, TAYLOR, CAMILLO J, and MALIK, JITENDRA. "Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach". *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM. 1996, 11–20 3.

[FNPS16] FLYNN, JOHN, NEULANDER, IVAN, PHILBIN, JAMES, and SNAVELY, NOAH. "Deepstereo: Learning to predict new views from the world's imagery". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, 5515–5524 3.

[GCD*20] GAO, DUAN, CHEN, GUOJUN, DONG, YUE, et al. "Deferred neural lighting: free-viewpoint relighting from unstructured photographs". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–15 2.

[GGSC96] GORTLER, STEVEN J, GRZESZCZUK, RADEK, SZELISKI, RICHARD, and COHEN, MICHAEL F. "The lumigraph". *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 1996, 43–54 3.

[GPM*14] GOODFELLOW, IAN J, POUGET-ABADIE, JEAN, MIRZA, MEHDI, et al. "Generative adversarial networks". *arXiv preprint arXiv:1406.2661* (2014) 2.

[GTZN20] GAFNI, GUY, THIES, JUSTUS, ZOLLHÖFER, MICHAEL, and NIESSNER, MATTHIAS. "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction". *https://arxiv.org/abs/2012.03065* (2020) 1.

[GVR*14] GARRIDO, PABLO, VALGAERTS, LEVI, REHMSEN, OLE, et al. "Automatic face reenactment". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, 4217–4224 2.

[Kaj86] KAJIYA, JAMES T. "The rendering equation". *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*. 1986, 143–150 4.

[KALL17] KARRAS, TERO, AILA, TIMO, LAINE, SAMULI, and LEHTINEN, JAAKKO. "Progressive growing of gans for improved quality, stability, and variation". *arXiv preprint arXiv:1710.10196* (2017) 2.

[KB10] KEMELMACHER-SHLIZERMAN, IRA and BASRI, RONEN. "3D face reconstruction from a single image using a single reference face shape". *IEEE transactions on pattern analysis and machine intelligence* 33.2 (2010), 394–405 2.

[KB14] KINGMA, DIEDERIK and BA, JIMMY. "Adam: A Method for Stochastic Optimization". *International Conference on Learning Representations* (Dec. 2014) 6.

[KLA19] KARRAS, TERO, LAINE, SAMULI, and AILA, TIMO. "A style-based generator architecture for generative adversarial networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, 4401–4410 2.

[KLZ*20a] KE, ZHANGHAN, LI, KAICAN, ZHOU, YUROU, et al. "Is a Green Screen Really Necessary for Real-Time Human Matting?": *arXiv preprint arXiv:2011.11961* (2020) 6.

[KLZ*20b] KE, ZHANGHAN, LI, KAICAN, ZHOU, YUROU, et al. "Is a Green Screen Really Necessary for Real-Time Portrait Matting?": *ArXiv* abs/2011.11961 (2020) 9.

[LGL*20] LIU, LINGJIE, GU, JIATAO, LIN, KYAW ZAW, et al. "Neural sparse voxel fields". *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020 1.

[LH96] LEVOY, MARC and HANRAHAN, PAT. "Light field rendering". *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 1996, 31–42 3.

[LLWL20] LEE, CHENG-HAN, LIU, ZIWEI, WU, LINGYUN, and LUO, PING. "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 2, 6.

[LSS*21] LOMBARDI, STEPHEN, SIMON, TOMAS, SCHWARTZ, GABRIEL, et al. "Mixture of Volumetric Primitives for Efficient Neural Rendering". *arXiv preprint arXiv:2103.01954* (2021) 1, 3.

[MB95] MCMILLAN, LEONARD and BISHOP, GARY. "Plenoptic modeling: An image-based rendering system". *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 1995, 39–46 3, 5.

[MHP*19] MEKA, ABHIMITRA, HAENE, CHRISTIAN, PANDEY, ROHIT, et al. "Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination". *ACM Transactions on Graphics (TOG)* 38.4 (2019), 1–12 2.

[MLP04] MATUSIK, WOJCIECH, LOPER, MATTHEW, and PFISTER, HANSPETER. "Progressively-Refined Reflectance Functions from natural Illumination." *Rendering Techniques*. 2004, 299–308 2.

[MPH*20] MEKA, ABHIMITRA, PANDEY, ROHIT, HÄNE, CHRISTIAN, et al. "Deep relightable textures: volumetric performance capture with neural rendering". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–21 2.

[MRS*21] MARTIN-BRUALLA, RICARDO, RADWAN, NOHA, SAJJADI, MEHDI S. M., et al. "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections". *CVPR*. 2021 1, 3.

[MSO*19] MILDENHALL, BEN, SRINIVASAN, PRATUL P, ORTIZ-CAYON, RODRIGO, et al. "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines". *ACM Transactions on Graphics (TOG)* 38.4 (2019), 1–14 3.

[MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P, TANCIK, MATTHEW, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". *European Conference on Computer Vision*. Springer. 2020, 405–421 1, 3–5.

[NLML20] NESTMEYER, THOMAS, LALONDE, JEAN-FRANÇOIS, MATTHEWS, IAIN, and LEHRMANN, ANDREAS. "Learning physics-guided face relighting under directional light". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 5124–5133 2.

[NRH03] NG, REN, RAMAMOORTHI, RAVI, and HANRAHAN, PAT. "All-Frequency Shadows Using Non-Linear Wavelet Lighting Approximation". *ACM Trans. Graph.* 22.3 (July 2003), 376–381. ISSN: 0730-0301. DOI: 10.1145/882262.882280. URL: https://doi.org/10.1145/882262.882280 2, 4.

[PGM*19] PASZKE, ADAM, GROSS, SAM, MASSA, FRANCISCO, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems 32*. Ed. by WALLACH, H., LAROCHELLE, H., BEYGELZIMER, A., et al. Curran Associates, Inc., 2019, 8024–8035 6.

[PML*09] PEERS, PIETER, MAHAJAN, DHRUV K, LAMOND, BRUCE, et al. "Compressive light transport sensing". *ACM Transactions on Graphics (TOG)* 28.1 (2009), 1–18 2.

[PSB*20] PARK, KEUNHONG, SINHA, UTKARSH, BARRON, JONATHAN T., et al. "Deformable Neural Radiance Fields". *https://arxiv.org/abs/2011.12948* (2020) 1, 3.

[QSMG17] QI, CHARLES R, SU, HAO, MO, KAICHUN, and GUIBAS, LEONIDAS J. "Pointnet: Deep learning on point sets for 3d classification and segmentation". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 652–660 5.

[RDL*15] REN, PEIRAN, DONG, YUE, LIN, STEPHEN, et al. "Image based relighting using neural networks". *ACM Transactions on Graphics (ToG)* 34.4 (2015), 1–12 2.

[RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-net: Convolutional networks for biomedical image segmentation". *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, 234–241 4.

[SBT*19] SUN, TIANCHENG, BARRON, JONATHAN T, TSAI, YUN-TA, et al. "Single image portrait relighting." *ACM Trans. Graph.* 38.4 (2019), 79–1 2, 4, 7.

[SDZ*21] SRINIVASAN, PRATUL P., DENG, BOYANG, ZHANG, XIUMING, et al. "NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis". *CVPR*. 2021 1, 3.

[SF16] SCHÖNBERGER, JOHANNES LUTZ and FRAHM, JAN-MICHAEL. "Structure-from-Motion Revisited". *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 9.

[SKCJ18] SENGUPTA, SOUMYADIP, KANAZAWA, ANGJOO, CASTILLO, CARLOS D, and JACOBS, DAVID W. "SfSNet: Learning Shape, Reflectance and Illuminance of Facesin the Wild'". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 6296–6305 2.

[SKS02] SLOAN, PETER-PIKE, KAUTZ, JAN, and SNYDER, JOHN. "Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments". *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 2002, 527–536 2, 4.

[SSS09] SINHA, SUDIPTA, STEEDLY, DREW, and SZELISKI, RICK. "Piecewise planar stereo for image-based rendering". (2009) 3.

[SXZ*20] SUN, TIANCHENG, XU, ZEXIANG, ZHANG, XIUMING, et al. "Light stage super-resolution: continuous high-frequency relighting". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–12 2.

[SZPF16] SCHÖNBERGER, JOHANNES LUTZ, ZHENG, ENLIANG, POLLEFEYS, MARC, and FRAHM, JAN-MICHAEL. "Pixelwise View Selection for Unstructured Multi-View Stereo". *European Conference on Computer Vision (ECCV)*. 2016 9.

[TFT*20] TEWARI, AYUSH, FRIED, OHAD, THIES, JUSTUS, et al. "State of the art on neural rendering". *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, 701–727 2.

[TL18] TRAN, LUAN and LIU, XIAOMING. "Nonlinear 3d face morphable model". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 7346–7355 2.

[TS06] TSAI, YU-TING and SHIH, ZEN-CHUNG. "All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation". *ACM Transactions on graphics (TOG)* 25.3 (2006), 967–976 4.

[WBL*20] WANG, ZIYAN, BAGAUTDINOV, TIMUR, LOMBARDI, STEPHEN, et al. "Learning Compositional Radiance Fields of Dynamic Human Heads". *arXiv preprint arXiv:2012.09955* (2020) 3.

[WBLP11] WEISE, THIBAUT, BOUAZIZ, SOFIEN, LI, HAO, and PAULY, MARK. "Realtime performance-based facial animation". *ACM transactions on graphics (TOG)* 30.4 (2011), 1–10 2.

[WWG*21] WANG, QIANQIAN, WANG, ZHICHENG, GENOVA, KYLE, et al. "IBRNet: Learning Multi-View Image-Based Rendering". *CVPR*. 2021 2–4, 7.

[XBS*19] XU, ZEXIANG, BI, SAI, SUNKAVALLI, KALYAN, et al. "Deep view synthesis from sparse photometric images". *ACM Transactions on Graphics (TOG)* 38.4 (2019), 1–13 3.

[XSHR18] XU, ZEXIANG, SUNKAVALLI, KALYAN, HADAP, SUNIL, and RAMAMOORTHI, RAVI. "Deep image-based relighting from optimal sparse samples". *ACM Transactions on Graphics (ToG)* 37.4 (2018), 1–13 2.

[XYC*20] XU, SICHENG, YANG, JIAOLONG, CHEN, DONG, et al. "Deep 3d portrait from a single image". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 7710–7720 3.

[YWP*18] YU, CHANGQIAN, WANG, JINGBO, PENG, CHAO, et al. "BiSeNet: Bilateral segmentation network for real-time semantic segmentation". *Proceedings of the European conference on computer vision (ECCV)*. 2018, 325–341 9.

[YYTK20] YU, ALEX, YE, VICKIE, TANCIK, MATTHEW, and KANAZAWA, ANGJOO. "pixelNeRF: Neural Radiance Fields from One or Few Images". *arXiv preprint arXiv:2012.02190* (2020) 2, 3.

[YZW*20] YANG, HAOTIAN, ZHU, HAO, WANG, YANRU, et al. "FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 2, 6.

[ZFT*21] ZHANG, XIUMING, FANELLO, SEAN, TSAI, YUN-TA, et al. "Neural light transport for relighting and view synthesis". *ACM Transactions on Graphics (TOG)* 40.1 (2021), 1–17 2.

[ZHSJ19] ZHOU, HAO, HADAP, SUNIL, SUNKAVALLI, KALYAN, and JACOBS, DAVID W. "Deep single-image portrait relighting". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, 7194–7202 2.

[ZTF*18] ZHOU, TINGHUI, TUCKER, RICHARD, FLYNN, JOHN, et al. "Stereo magnification: Learning view synthesis using multiplane images". *arXiv preprint arXiv:1805.09817* (2018) 3.

[ZTG*18] ZOLLHÖFER, MICHAEL, THIES, JUSTUS, GARRIDO, PABLO, et al. "State of the art on monocular 3D face reconstruction, tracking, and applications". *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library. 2018, 523–550 2.