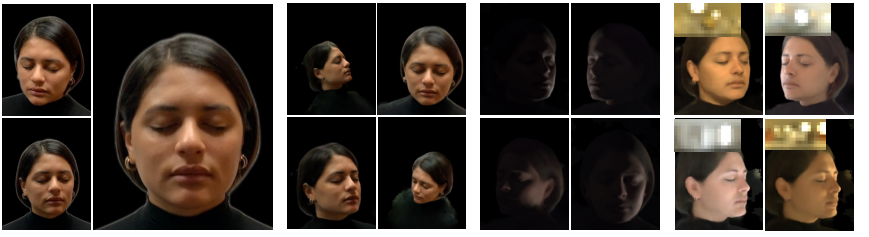
VoRF: Volumetric Relightable Faces

Pramod Rao[1]
prao@mpi-inf.mpg.de
Mallikarjun B R[1]
mbr@mpi-inf.mpg.de
Gereon Fox[1]
gfox@mpi-inf.mpg.de
Tim Weyrich[2]
tim.weyrich@fau.de
Bernd Bickel[3]
bernd.bickel@ist.ac.at
Hanspeter Pfister[4]
pfister@g.harvard.edu
Wojciech Matusik[5]
wojciech@csail.mit.edu
Ayush Tewari[5]
ayusht@mit.edu
Christian Theobalt[1]
theobalt@mpi-inf.mpg.de
Mohamed Elgharib[1]
elgharib@mpi-inf.mpg.de

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany
[3] IST-Austria, Austria
[4] Harvard University, USA
[5] MIT CSAIL, USA

Abstract

Portrait viewpoint and illumination editing is an important problem with several applications in VR/AR, movies, and photography. Comprehensive knowledge of geometry and illumination is critical for obtaining photorealistic results. Current methods are unable to explicitly model in $3D$ while handing both viewpoint and illumination editing from a single image. In this paper, we propose VoRF, a novel approach that can take even a single portrait image as input and relight human heads under novel illuminations that can be viewed from arbitrary viewpoints. VoRF represents a human head as a continuous volumetric field and learns a prior model of human heads using a coordinate-based MLP with individual latent spaces for identity and illumination. The prior model is learnt in an auto-decoder manner over a diverse class of head shapes and appearances, allowing VoRF to generalize to novel test identities from a single input image. Additionally, VoRF has a reflectance MLP that uses the intermediate features of the prior model for rendering One-Light-at-A-Time (OLAT) images under novel views. We synthesize novel illuminations by combining these OLAT images with target environment maps. Qualitative and quantitative evaluations demonstrate the effectiveness of VoRF for relighting and novel view synthesis, even when applied to unseen subjects under uncontrolled illuminations.



a) 3 input views (masked) b) Novel views c) OLAT images d) Relit novel views

Figure 1: We present VoRF, a learning framework that synthesizes novel views and relighting under any lighting conditions given a single image or a few posed images. VoRF has explicit control over the direction of a point light source and that allows the rendering of a basis of one-light-at-a-time (OLAT) images (c). Finally, given an environment map (see d, insets) VoRF can relight the input (d) by linearly combining the OLAT images.

1 Introduction

Portrait editing has a wide variety of applications in virtual reality, movies, gaming, photography, teleconferencing, etc. Synthesizing photorealistic novel illuminations and viewpoints of human heads from a monocular image or a few images is still an open challenge. While there has been a lot of work in photorealistic facial editing [2, 3, 16, 21, 30, 36, 38, 43], these methods are usually restricted by sophisticated multi-view input [3, 16], inability to edit full face region [2, 11, 38] or pure relighting capability without viewpoint editing [21, 30, 36, 43]. Some recent efforts [1, 14] have shown the ability to edit portrait lighting and viewpoint simultaneously without sophisticated input, while they still suffer from geometric distortion during multi-view synthesis as they rely on 2D representation.

Recently, NeRF [17] has proven a powerful 3D volumetric representation. and is capable of producing novel views at an unprecedented level of photorealism [17]. NeRF has been applied to tasks like human body synthesis [12, 29], scene relighting [4, 28, 41], image compositing [20, 39] and others [34]. Sun *et al*. introduced Neural Light-transport Field (NeLF) [32], a NeRF-based approach for facial relighting and viewpoint synthesis that predicts the light-transport field in 3D space and generalizes to unseen identities. However, their method struggles to learn from sparse viewpoints and requires accurate geometry for training. Besides, they need ≥ 5 views of the input face during test-time to avoid strong artifacts.

In this paper, we propose a new method that takes a single portrait image as input for synthesizing novel lighting conditions and views. We utilize a NeRF-based volumetric representation and a large-scale multi-view lightstage dataset[37] to build a space of faces (geometry and appearance) in an auto-decoder fashion, that we call the *Face Prior Network*. This network provides a suitable space to fit any test identity. In addition, our *Reflectance Network* takes a feature vector from the *Face Prior Network* as well as the direction of a point light source as input, to synthesize the corresponding "One-Light-at-A-Time" (OLAT) image. This network is supervised using the lightstage dataset [37]which captures all aspects of complex lighting effects like self-shadows, diffuse, specularity, sub-surface scattering and higher order inter-reflections. Using OLATs has been shown to improve the quality of relighting [2, 16] without assuming a BRDF model or explicit priors. After training, a test identity can be relighted by first regressing the corresponding OLAT images for the desired novel viewpoint which are then linearly combined with any target environment map to synthesize

relighted results [6]. Our comparisons to previous methods show that our approach performs novel views that are significantly better than those of SOTA methods like PhotoApp [14]. Furthermore, our approach produces results that are significantly more consistent with the input than those of NeLF [32]. It can operate directly on a monocular image and outperforms NeLF even with 3 input views.

To summarize, we make the following contributions: (1) We present a NeRF-based approach for full-head relighting that can take a single input image and produces relit results that can be observed from arbitrary viewpoints. (2) We design a dedicated *Reflectance Network* that is built over the *Face Prior Network* that allows our method to learn self-shadows, specularities, sub-surface scattering, and higher order inter-reflection through a lightstage dataset supervision. (3) VoRF is additionally able to synthesize One-Light-at-A-Time 3*D* volume for any given light direction, even though we learn from a dataset which has limited number of light sources.

2 Related Work

The literature of portrait editing is vast and here we discuss only methods which are related to relighting. OLAT images generated by a lightstage are popular for capturing the face reflectance details, as pioneered by the seminal work of Debevec *et al*. [6]. Here, it was shown that such OLAT images can be used as illumination basis to express an arbitrary environment map through a linear operation. The highly photorealistic relighting achieved by this formulation encouraged further research. This includes methods dedicated for image sequence processing [3, 40], shadow removal [42], capturing high-quality reflectance priorities from monocular images [2, 38] among others [16, 19, 21, 31, 36]. Among these, [2] is the closest in problem setting and approach. [2] can regress OLATs for any camera position given monocular image. But since they rely on 3DMM model, they can only relight face interior. The majority of these methods, can edit the face interior only [2, 36, 38], or can edit the lighting only while keeping the original camera viewpoint unchanged [16, 19, 21, 31, 40, 42]. The method proposed by Bi *et al*. [3] can edit the camera viewpoint and lighting of the full head simultaneously. But, it is person-specific.

Instead of using a lightstage OLAT data, some methods employ illumination models and/or train with synthetic data [5, 11, 26, 27, 43]. While these approaches can generalize to unseen identities, they can be limited in terms of photorealism and the overall quality [26, 27, 43] and some are constrained to editing only the face interior [11]. Recent efforts leverage the generative capabilities of the StyleGAN face model [10] to learn from in-the-wild data in a completely self-supervised manner [1, 33]. More recently, in PhotoApp B R *et al*. [14] combined the strength of both lightstage OLAT data and the generative model StyleGAN. Such formulation has two main advantages. First, it achieves strong identity generalization even when training with as few as just 3 identities. Second, it is capable of relighting the full head and editing the camera viewpoint simultaneously. However, as StyleGAN is a 2D generative model, PhotoApp suffer to generate view consistent results in 3D. In contrast, our method learns the prior space in volumetric representation, which generate significantly better view-consistent results. StyleGAN embedding can also change the original identity, leading to unacceptable results. Our method on the other hand, maintains the integrity of the original identity.

Recently, there have been multiple NeRF-based methods for general scene relighting [4, 15, 25, 28, 41]. While NeRV [28] requires the illumination of the scene as input, NeRFactor [41], NeRD [4], NeRFW [15], NeRF-OSR [25] can work with unknown input scene

illumination. The illumination space of NeRFW [15] is not based on physically meaningful semantic parameters. All of the above NeRF-based methods are scene specific and require multiple images of the scene at test time. In contrast, our method can work with images with unknown scene illumination and can even work from a single image as we build a strong face prior model in the first stage to handle such cases. And our relighting is controlled by physically-based semantically meaningful environment maps.

The closest approach to our problem setting is NeLF [32]. Based on NeRF, it has a good 3D understanding of the scene. It learns the volume density and light transport for each point in 3D space. NeLF adopts a pixelNeRF-inspired architecture where the density and color values rely heavily on localized image features. As a result, their method struggles to capture global cues and sometimes results in holes in the volume. Their method also requires high quality geometry for supervision during training, and thus fails to learn from sparse viewpoint. It also needs at least 5 viewpoints of the input face during test otherwise significant artifacts are produced. In contrast, we learn a prior geometric and appearance space of faces, which helps in capturing realistic face volumes. This allows synthesizing novel views and relightings. Our technique maintains the integrity of the facial geometry during viewpoint interpolation, relights the full head and can operate using as few as a single monocular image during test. It also provides novel generative capabilities of unseen identities and illumination during test.

3 Face Reflectance Fields

We consider scenes that contain an illuminated model of a human face. Building on previous works [6, 17] and assuming that all light sources are sufficiently far away from the face, we model the face as a *volumetric reflectance field*, which is a pair (σ, R). The *volume density function* $\sigma : \mathbb{R}^3 \to \mathbb{R}$ maps scene points to density values and the function $R(\omega, \mathbf{x}, \mathbf{d})$ indicates the fraction of $L_{\text{inc}}(\omega)$ (the radiance incident from direction ω) that is reflected from point $\mathbf{x}$ into direction $\mathbf{d}$. We assume that image formation follows a perceptive camera model. Our NeRF-based [17] models in Sec. 4 learns functions of the form $F_\Theta(\mathbf{x}, \mathbf{d}) = (L_{\text{out}}(\mathbf{x}, \mathbf{d}), \sigma(\mathbf{x}))$, based on latent codes for the facial identity and lighting conditions, where $L_{\text{out}}(\mathbf{x}, \mathbf{d})$ is the radiance emitted from point $\mathbf{x}$ into direction $\mathbf{d}$.

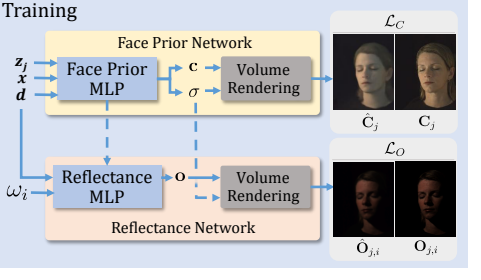


Figure 2: Our *Face Prior Network* learns to decode latent codes $\mathbf{z}_j$ to estimate radiance and volume density for each point in 3D space. Our *Reflectance Network* learns to synthesize OLAT images of the face.

"One-Light-at-A-Time" lighting means that the face is illuminated from only one single light source that is sufficiently far away. Assuming that the set I of light sources for which we can render OLAT images covers the set of all possible incident angles sufficiently densely, we can approximate any desired illumination: The radiance $L(r)$ accumulated along a ray r can be broken down as

$$L(r) \approx \sum_{i \in I} f_i \cdot L(i, r) \tag{1}$$

where $L(i, r)$ is the amount of radiance that originates from OLAT light source i and that eventually emerges from the scene along ray r. The f_i are factors with which the OLAT

light sources need to be modulated in order to implement a given lighting condition. Similar equations are known in the literature [6], showing that under the stated assumptions we can render the face under any given lighting specification $(f_i)_{i\in I}$ just as a linear combination of OLAT images. We give a more detailed derivation of Eq. 1 in our supplemental material.

To train our face prior network (see Sec. 4) and to evaluate our method, we use HDR environment maps from the Laval Outdoor dataset [8] and the Laval Indoor HDR dataset [7] to obtain coefficients f_i. This allows us to turn the OLAT basis images into depictions of faces under real world lighting conditions and we generate 600 relit images for each subject.

4 Method

We address the problem of simultaneous portrait view synthesis and relighting. Given a small set of $N \geq 1$ input images along with their camera parameters, we build a *Face Prior Network*($\mathcal{P}$) and a *Reflectance Network*($\mathcal{R}$) utilising NeRF-based representation. Firstly, the $\mathcal{P}$ is modelled in an auto-decoder fashion to learn a prior over human heads under various illumination conditions and this formulation allows VoRF to generalize to novel test identities. Furthermore, to model face reflectance that can re-illuminate a face for several viewpoints, we design a $\mathcal{R}$ that learns to predict OLAT images. Using Eq. 1, we linearly combine these OLAT images with HDR environment maps to render novel views of a given face, under new lighting conditions. An overview of our method can be found in Fig. 2.

4.1 Learning Face Prior

Neural Radiance Fields [17] learns a coordinate based representation of each scene by mapping 3D coordinates $\mathbf{x} \in \mathbb{R}^3$ and direction $\mathbf{d} \in \mathcal{S}^2$ to the densities and radiance values. However, NeRF by design is able to optimize a single scene at a time. To combat this and obtain a distribution over the entire space of faces and illumination conditions, we use an auto-decoder formulation. More specifically, we first prepare a dataset by combining a set of environment maps with OLAT images acquired from lightstage resulting in $\mathbb{J}$ combinations. For each combination $j \in \mathbb{J}$, we obtain image $\mathbf{C}_j$ and a corresponding latent code $\mathbf{z}_j$. The latent code $\mathbf{z}_j$ is partitioned into identity and illumination components as $\mathbf{z}_j^{\text{id}}$ and $\mathbf{z}_j^{\text{env}}$ respectively. We initialize the latent codes from a multivariate normal distribution and observe that separating the components individually leads to faster convergence during the training process (see supplemental for details). We design the *Face Prior Network* to take the latent code $\mathbf{z}_j$ along with $\mathbf{x}$, $\mathbf{d}$ as inputs and predict radiance $\mathbf{c}$ as well as volume density σ

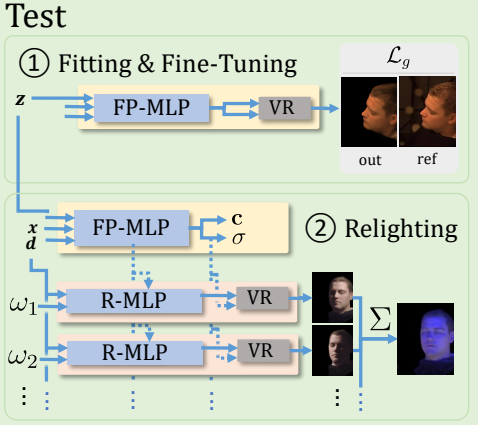


Figure 3: To reconstruct an unseen test face, we optimize latent code $\mathbf{z}$ and fine-tune the *Face Prior Network*, We can relight the reconstructed face by having the *Reflectance Network* produce a basis of OLAT images (step 2), that we linearly combine into any desired lighting condition. In this figure, MLP's with the same label share their weights.

for every point in 3D space. We represent the *Face Prior Network* as $\mathcal{P}_{\Theta_\mathcal{P}}(\mathbf{z}_j, \mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$. Following NeRF, the network weights $\Theta_\mathcal{P}$ along with the latent codes $\mathbf{z}$ are optimized jointly to regress the color values with a mean squared objective function as follows:

$$\mathcal{L}_\mathrm{C} := \sum_{j \in \mathbb{J}} \|\hat{\mathbf{C}}_j - \mathbf{C}_j\|_2^2 \tag{2}$$

where $\hat{\mathbf{C}}_j$ is the image obtained by volume rendering based on $\mathcal{P}_{\Theta_\mathcal{P}}(\mathbf{z}_j, \,.\,, \,.\,)$. To ensure smooth interpolation of the latent spaces and additionally prevent the latents $\mathbf{z}_j$ from diverging away from the origin, we introduce regularization. Inspired by [22], we regularize the distribution of the latent codes by minimizing the mean squared objective of individual latents as:

$$\mathcal{L}_\mathrm{reg} = \sum_{j \in \mathbb{J}} \|\mathbf{z}_j^\mathrm{id}\|_2^2 + \|\mathbf{z}_j^\mathrm{env}\|_2^2 \tag{3}$$

4.2 Synthesizing new OLAT images

To model a reflectance field of the faces, we propose a *Reflectance Network*($\mathcal{R}$) that learns a volumetric reflectance field by utilising the σ predictions provided by $\mathcal{P}$ (see Sec. 3). For an OLAT light source i, we consider the the incident light direction ω_i as an input to the $\mathcal{R}$. To synthesize OLAT images, we design the $\mathcal{R}$ based on NeRF and directly regress the radiance values $\mathbf{o}$. As reflectance of the face is a function of geometric and other fine details, we provide a feature vector obtained from the 9^{th} layer of $\mathcal{P}$ as an additional input to R (dotted lines in Fig. 2 and Fig. 3). We also provide the viewing direction $\mathbf{d}$ as input to capture view-dependent effects. Thus, *Reflectance Network* learns a function $\mathcal{R}_{\Theta_\mathcal{R}}$, parameterized by $\Theta_\mathcal{R}$ and is given as follows: $\mathcal{R}_{\Theta_\mathcal{R}}(\omega_i, \mathcal{F}_\mathcal{P}(\mathbf{z}_j, \mathbf{x}, \mathbf{d}), \mathbf{d}) = \mathbf{o}$. To synthesize an OLAT image $\hat{\mathbf{O}}_{j,i}$ along the light direction i for $j \in \mathbb{J}$, we combine $\mathbf{o}$ with the volume density σ predicted from $\mathcal{P}$. $\Theta_\mathcal{R}$ is optimized by minimizing HDR-based loss inspired by [18] and S is a stop gradient function:

$$\mathcal{L}_\mathrm{O} := \sum_{j \in \mathbb{J}} \left\| \frac{\hat{\mathbf{O}}_{j,i} - \mathbf{O}_{j,i}}{S(\hat{\mathbf{O}}_{j,i}) + \varepsilon} \right\|_2^2 \tag{4}$$

where $\mathbf{O}_{j,i}$ is the ground truth OLAT image from the dataset that is used in the construction of $\mathbf{C}_j$. This loss function is especially suited for handling the semi-dark lighting conditions of OLAT images. Our HDR lightstage dataset predominantly consists of dark regions and utilising an L2 loss function results in muddy artifacts in those regions [18]. In contrast, the HDR-Loss divides the absolute error by the brightness of the ground truth image giving higher weight value for darker regions. Thus, utilising this loss function helps to recover high contrast differences in dark regions.

4.3 Training

NeRF-based methods typically require dense camera views of the scene to faithfully represent the scene without cloudy artifacts. As our dataset has limited number of views, we make use of hard-loss [24] to avoid cloudy artifacts. We consider, as in previous work , the *accumulation weights* $w_{r,k}$ that are computed during volume rendering, for a given ray r (see [24]). Imposing $\mathbb{P}(w_{r,k}) \propto e^{-|w_{r,k}|} + e^{-|1-w_{r,k}|}$ for the probabilities of these weights, we minimize

$$\mathcal{L}_\mathrm{h} = \sum_{r,k} -\log(\mathbb{P}(w_{r,k})) \tag{5}$$

	View synthesis					
	NeLF		IBRNet		Ours	
Input	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
5-views	22.01	0.80	24.38	0.82	**27.45**	**0.84**
3-views	20.57	0.75	22.0	0.76	**26.67**	**0.82**
2-views	19.63	0.70	20.34	0.71	**25.44**	**0.79**
1-view	N/A	N/A	N/A	N/A	**22.49**	**0.77**

	View synthesis and relighting					
	NeLF		IBRNet+SIPR		Ours	
Input	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
5-views	21.34	0.79	19.63	0.75	**24.16**	**0.81**
3-views	19.72	0.75	18.38	0.73	**22.80**	**0.76**
2-views	19.06	0.69	17.01	0.71	**22.15**	**0.74**
1-view	N/A	N/A	N/A	N/A	**20.21**	**0.69**

Table 1: Comparing against NeLF [32] (requires at least 5 input views), IBRNet [35] and SIPR [30] in view synthesis and relighting. Our technique outperforms related methods irregardless of the number of input views (see bold).

which encourages the density functions implemented by $\mathcal{P}$ to produce hard transitions. We apply this loss during the synthesis of both $\hat{\mathbf{C}}_j$ and $\hat{\mathbf{O}}_{j,i}$, which helps to avoid cloud artifacts surrounding the face.

Our overall training loss function now is $\mathcal{L} = \alpha\mathcal{L}_\mathrm{C} + \beta\mathcal{L}_\mathrm{O} + \gamma\mathcal{L}_\mathrm{reg} + \delta\mathcal{L}_\mathrm{h}$ with hyper weights $\alpha, \beta, \gamma, \delta$.

4.4 Test

Given a small set of $N \geq 1$ input images of an unseen identity under unseen lighting conditions, we fit $\mathbf{z}$ and fine-tune $\Theta_\mathcal{P}$ by minimizing (using backpropagation)

$$\mathcal{L}_\mathrm{g} := \alpha\mathcal{L}_\mathrm{C} + \gamma\mathcal{L}_\mathrm{reg} + \delta\mathcal{L}_\mathrm{h} \tag{6}$$

where the input images now take the place of the $\mathbf{C}$ that were used during training. Note, that first, we update only $\mathbf{z}$ for 10,000 iterations (learning rate 1×10^{-3}), to make sure that it lies well within the learnt prior distribution. Then, assuming that the fitting step has converged, continue to jointly update $\mathbf{z}$ and $\Theta_\mathcal{P}$ for 3,000 iterations (learning rate 3×10^{-6}). We demonstrate the significance of this two-step approach as ablation study in the supplemental material.

With $\mathbf{z}$ and $\Theta_\mathcal{P}$ optimized in this way (part ① in Fig. 3), we can already render the face under novel views. In order to be able to also change lighting (part ② in Fig. 3), we use $\mathcal{R}$ to render an OLAT basis that by Eq. 1 we can use to synthesize any given lighting conditions.

5 Results

We evaluate our method qualitatively and quantitatively to demonstrate the efficacy of our method using our lightstage dataset, see Sec. 5.1. Additionally, we qualitatively evaluate our method on H3DS [23], a naturally lit multi-view dataset. We compare against three methods; 1) NeLF [32] 2) combination of IBRNet [35] and SIPR [30], and 3) PhotoApp [14].

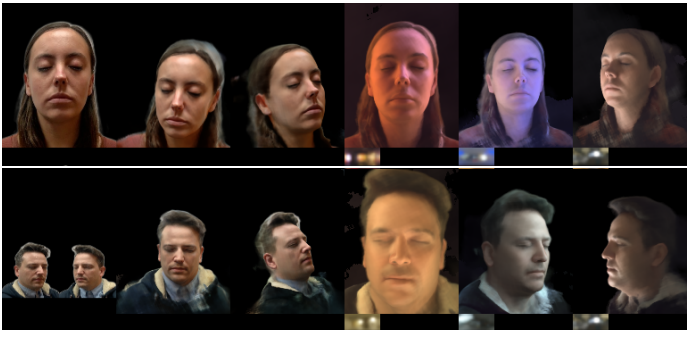


Figure 4: Novel view synthesis + relighting on unseen identities from the **H3DS** [23] dataset. We show results obtained by using a single image (top) and two images (bottom). Target environment maps are shown in the insets. Our technique performs photorealistic novel view-synthesis and relighting.

We perform ablation studies on the various design choices of our framework and discuss their significance in the supplemental material.

For a fair comparison of our method against the baselines, we retrain NeLF, IBRNet, SIPR and PhotoApp with our lightstage dataset. All the methods are retrained as suggested in the original works. We recommend referring to supplemental material for elaborate implementation details.

5.1 View synthesis and Relighting

In this section we present the results for view synthesis and relighting to demonstrate that our method can synthesize novel lighting conditions of the subject at novel viewpoints. Fig. 4 shows novel view synthesis and relighting produced by our technique. Here, we present results with single input view (top) and two input views (bottom). We observe that our method produces photorealistic renderings that are view-consistent. Our method maintains the integrity of the input identity and recovers the full head, including hair. It also maintains the integrity of the facial geometry while relighting at extreme views (third and fourth row, last column in Fig. 4).

Our *Reflectance Network* has the ability to synthesize subjects corresponding to arbitrary light directions and enable us to relight them using any HDR environment maps following Eq. 1. To achieve this, our technique predicts the 150 OLAT images as the light basis of the lightstage. In our supplemental work we show that through our rendered OLATs we are able to reproduce view-dependent effects, specular highlights and shadows. Finally, our disentangled identity and illumination latent space representation allows us to perform interesting interpolations between different subjects and illuminations. In addition to the results shown here, we recommend the readers to refer to the supplementary material for more results and evaluations.

5.2 Comparison to Related Methods

We quantitatively and qualitatively compare against the state-of-the-art view synthesis and relighting methods. All the quantitative evaluations are on unseen lightstage subjects with unseen illumination conditions (see supplemental for evaluation dataset details). We summarize

Figure 5: A sample result on the lightstage test set, with groundtruth. Our technique produces novel view synthesis and relightings that clearly outperform NeLF [32] and IBRNet [35] + SIPR [30].

our quantitative evaluations in Tab. 1 in terms of average PSNR and SSIM over all the test images. First, we compare our method for the view-synthesis task with different number of input views. Next, with same test setup we evaluate for the task of simultaneous view synthesis and relighting. For both the tasks, we observe that our method convincingly outperforms NeLF, IBRNet and IBRNet + SIPR. Unlike our approach, we observe that neither NeLF nor IBRNet can handle single input image which limits their application to multi-view setups. High evaluation scores indicate that our method recovers decent geometry and synthesizes better quality relighting. These results can be more easily understood in Fig. 5, where we clearly observe our renderings match the groundtruth more closely than the baseline methods.

We additionally compare against NeLF on **H3DS** dataset (see Fig. 6) where our approach clearly performs better. We argue this is due to NeLF's inability to recover decent geometry from sparse views. Likewise, IBRNet fails to construct multi-view consistent geometry under sparse views. Further with IBRNet+SIPR, we observe that SIPR depends on the viewpoint, which breaks down the multi-view consistent relighting. Finally, we compare against PhotoApp in Fig. 8. PhotoApp inherits the limitations of the StyleGAN space, specifically, the inversion step which modifies the input identity. Such modifications lead to highly inconsistent results limiting the application of PhotoApp. In contrast, our approach produces view-consistent results that resemble groundtruth.

Limitations: While our proposed method generates photorealistic renderings, few limitations still exist. In Fig. 7 we showcase results of our approach on FFHQ [9] and CelebA [13] datasets. Despite being trained on the lightstage dataset containing all the subjects with closed eyes and neutral expressions, we can handle novel view synthesis with opens eyes as well as natural expressions. This is attributed to the fine-tuning of the *Face Prior Network* during test. Fig. 7 further demonstrates that our method preserves the mouth and eye shape during relighting. However, it cannot synthesize their colors or texture. We argue that this is not a limitation of our approach but of the lightstage dataset. Lastly, under monocular setting our approach can sometimes generate regions that do not exist in reality. For instance, in Fig. 9 in case of single input, hair is synthesized for the bald person. Such performance is expected due to insufficient information from a single view.

6 Conclusion

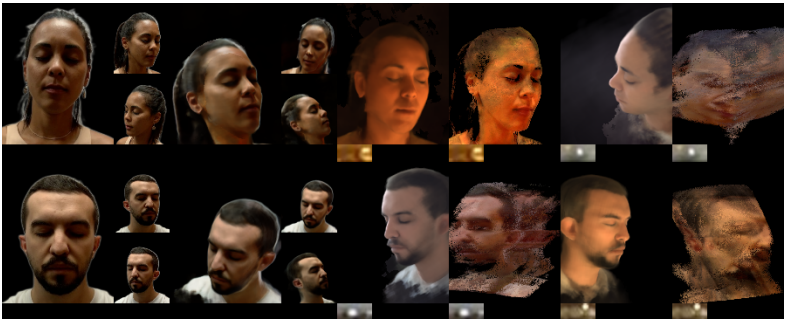


Figure 6: Novel view synthesis and relightings on the **H3DS** dataset [23]. Our technique significantly outperforms NeLF [32] especially at views far from the training set.

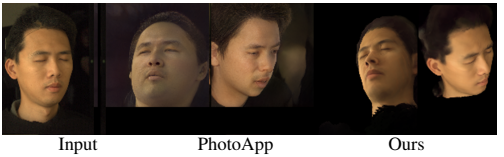


Figure 8: PhotoApp [14] view synthesis suffers from strong artifacts, including loss of the input identity and view-inconsistent results.

Figure 9: Our method produces good relightings and view synthesis using from 3, 2 or even 1 input view.

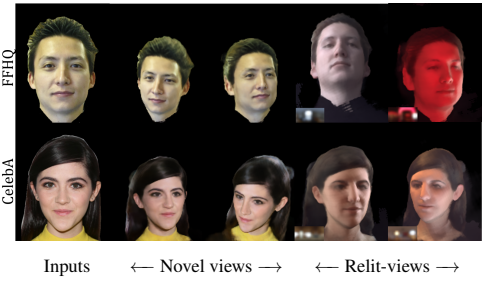


Figure 7: Given single input view from FFHQ (top) and CelebA (bottom). Although our method works well for novel view synthesis, it struggles to synthesize eyes and facial expressions during relighting.

We have presented an approach for editing light and viewpoint of human heads even with a single image as input. Based on neural radiance fields [17], our method represents human heads as a continuous volumetric field with disentangled latent spaces for identity and illumination. Our method is designed to first learn a face prior model in an auto-decoder manner over a diverse class of heads. Further, followed by training a reflectance MLP that predicts One-Light- at-A-Time (OLAT) images at every point in $3D$, parameterized by point light direction which can be combined to produce a target lighting. Quantitative and qualitative evaluations show that our results are photorealistic, view-consistent and outperforms existing state-of-the-art works.

Acknowledgments This work was supported by the ERC Consolidator Grant 4DReply (770784).

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics*, 2021.

[2] Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Monocular reconstruction of neural face reflectance fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics*, 2021.

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[5] Sreenithy Chandran, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Zhixin Shu, and Suren Jayasuriya. Temporally consistent relighting for portrait videos. In *The IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 719–728, January 2022.

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

[7] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image, 2017.

[8] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (ACM SIGGRAPH Asia)*, 2021.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[14] Mallikarjun B R, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, et al. Photoapp: Photorealistic appearance editing of head portraits. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2021.

[15] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2019.

[17] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[18] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. *arXiv*, 2021.

[19] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. Learning physics-guided face relighting under directional light. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[21] Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, 2021.

[22] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021.

[24] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision (ECCV)*, 2022.

[26] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[29] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021.

[30] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2019.

[31] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T. Barron, and Ravi Ramamoorthi. Light stage superresolution: Continuous high-frequency relighting. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 2020.

[32] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021.

[33] Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. In *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*, 2020.

[34] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in Neural Rendering. *arXiv e-prints*, 2021.

[35] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 2020.

[37] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2006.

[38] Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 2018.

[39] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2021.

[40] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[41] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics*, 2021.

[42] Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait shadow manipulation. In *ACM Transactions on Graphics (TOG)*, 2020.

[43] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single-image portrait relighting. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.