Deep Relightable Appearance Models for Animatable Faces

SAI BI, University of California, San Diego, USA
STEPHEN LOMBARDI* and SHUNSUKE SAITO*, Facebook Reality Labs, USA
TOMAS SIMON, SHIH-EN WEI, and KEVYN MCPHAIL, Facebook Reality Labs, USA
RAVI RAMAMOORTHI, University of California, San Diego, USA
YASER SHEIKH and JASON SARAGIH, Facebook Reality Labs, USA

Fig. 1. Our relightable facial appearance model supports renderings under novel viewpoints, expressions, and lighting conditions including nearfield lighting, directional lighting, and environment lighting. Our model is also animatable and can be driven by images captured from cameras on head-mounted displays.

We present a method for building high-fidelity animatable 3D face models that can be posed and rendered with novel lighting environments in real-time. Our main insight is that relightable models trained to produce an image lit from a single light direction can generalize to natural illumination conditions but are computationally expensive to render. On the other hand, efficient, high-fidelity face models trained with point-light data do not generalize to novel lighting conditions. We leverage the strengths of each of these two approaches. We first train an expensive but *generalizable model* on point-light illuminations, and use it to generate a training set of high-quality synthetic face images under natural illumination conditions. We then train an *efficient model* on this augmented dataset, reducing the generalization ability requirements. As the efficacy of this approach hinges on the quality of the synthetic data we can generate, we present a study of lighting pattern combinations for dynamic captures and evaluate their suitability for learning generalizable relightable models. Towards achieving the best possible quality, we present a novel approach for generating dynamic relightable faces that exceeds state-of-the-art performance. Our method is capable of capturing subtle lighting effects and can even generate compelling near-field relighting despite being trained exclusively with far-field lighting data. Finally, we motivate the utility of our model by animating it with images captured from VR-headset mounted cameras, demonstrating the first system for face-driven interactions in VR that uses a photorealistic relightable face model.

CCS Concepts: • **Computing methodologies** → **Virtual reality**; **Image-based rendering**; **Neural networks**.

Additional Key Words and Phrases: Face Rendering, Appearance Acquisition, Image-based Rendering, View Synthesis, Relighting, Neural Rendering

ACM Reference Format:
Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep Relightable Appearance Models for Animatable Faces. *ACM Trans. Graph.* 40, 4, Article 89 (August 2021), 15 pages. https://doi.org/10.1145/3450626.3459829

1 INTRODUCTION

Avatar creation has seen a notable increase in the use of learning-based techniques in recent years [Lombardi et al. 2018; Nagano et al. 2018; Schwartz et al. 2020] Traditional physically-inspired methods [Seymour et al. 2017; Weyrich et al. 2006] require precise geometry and reflectance, where costly and time-consuming manual cleanup is typically needed. In contrast, learning-based methods use general function approximators in the form of deep neural networks to faithfully model the appearance of human faces. They can achieve impressive realism with completely automated pipelines without

*Both authors contributed equally to this work.

Authors' addresses: Sai Bi, bisai@cs.ucsd.edu, University of California, San Diego, San Diego, CA, USA; Stephen Lombardi, stephen.lombardi@fb.com; Shunsuke Saito, shunsukesaito@fb.com, Facebook Reality Labs, Pittsburgh, USA; Tomas Simon, tsimon@fb.com; Shih-En Wei, swei@fb.com; Kevyn McPhail, kmcphail@fb.com, Facebook Reality Labs, Pittsburgh, PA, USA; Ravi Ramamoorthi, ravir@cs.ucsd.edu, University of California, San Diego, San Diego, CA, USA; Yaser Sheikh, yasers@fb.com; Jason Saragih, jsaragih@fb.com, Facebook Reality Labs, Pittsburgh, PA, USA.

0730-0301/2021/8-ART89
https://doi.org/10.1145/3450626.3459829

relying on precise estimates of face geometry and material properties. They can also exhibit an efficient functional form that enables real-time generation and rendering in demanding applications such as VR [Lombardi et al. 2018], where classical ray-tracing methods can be too computationally intensive [Weyrich et al. 2006].

Despite their many advantages, avatars created using learning based techniques have so far been limited to a single lighting condition [Lombardi et al. 2018; Nagano et al. 2018]. For example, Lombardi et al. build avatars that support novel viewpoints and expressions, but their model is limited to the uniform lighting condition under which the data was captured. Although there has been great progress in learning-based relighting, existing methods are limited to 2D images [Sun et al. 2019; Xu et al. 2018], static scenes [Sun et al. 2019; Xu et al. 2018; Zhang et al. 2020], or performance replay [Meka et al. 2019], which are not suitable for generating dynamic renderings under novel expressions and lighting conditions (see Table 1). This limitation has prevented the broader adoption of learning-based avatars in game and film production, where consistency between character and environment is essential.

In this work, we describe Deep Relightable Appearance Models (DRAM), a learning-based method for building relightable avatars. Our model supports rendering under novel viewpoints, novel expressions and more importantly, it can be rendered under novel lighting conditions, where we can reconstruct complex visual phenomena such as specularities, glints and subsurface scattering. We build the relightable model from light-stage captures of dynamic performances under a sparse set of space- and time-multiplexed illumination patterns. Like [Lombardi et al. 2018], we train our model using the variational auto-encoder framework [Kingma and Welling 2013], which produces a well-structured latent space of expressions that is suitable for animation. To avoid overfitting to the lighting conditions observed during capture, we leverage the additive property of light transport [Busbridge 1960] and generate expression- and view-dependent textures for each light in the scene, which are then fused with intensity-defined weights into the final lit texture. Since the lighting information is fed at a later stage of the decoder network, instead of at its bottleneck, we call this model a *late-conditioned model*. It affords generalization to completely unseen lighting environments including both distant directional lighting and real environment maps (Figure 1), and it exhibits smooth interpolation of point light sources despite the discrete set of 460 lights used during capture. Finally, it can generate compelling near-field illumination effects (Figure 8), which is particularly challenging for a learning-based approach that exclusively uses data with distant light sources.

Although late-conditioned DRAM (DRAM$_\ell$) exhibits good generalization properties, its architecture is not suitable for real-time applications, since each point light in the scene requires the generation of a light-specific texture. For natural environments, the large number of illuminating directions make it computationally prohibitive to generate. This limitation is shared by many previous works [Debevec et al. 2000; Meka et al. 2019; Zhang et al. 2020]. However, we observe that early-conditioned deep neural networks that input the desired lighting condition at the network's bottleneck can exhibit enough capacity to model the span of a single person's illuminated facial appearances while being considerably more efficient to evaluate.

	Free viewpoint	Relightable	Dynamic capture	Animatable
Wenger et al.[2005]	✗	✓	✓	✗
Lombardi et al. [2018]	✓	✗	✓	✓
Xu et al. [2018]	✗	✓	✗	✗
Meka et al. [2019]	✗	✓	✓	✗
Sun et al. [2019]	✗	✓	✗	✗
Sun et al. [2020]	✗	✓	✗	✗
Zhang et al. [2020]	✓	✓	✗	✗
Meka et al. [2020]	✓	✓	✓	✗
Ours	✓	✓	✓	✓

Table 1. Feature comparison with previous methods. Ours is the only approach that enables a relightable and animatable model, in addition to free viewpoint and dynamic expressions.

The main drawback of *early-conditioned* models is their poor extrapolation properties to unseen natural illumination conditions. Thus, we use DRAM$_\ell$ to generate renderings of the face under a large number of natural illumination conditions, which we then use to train an efficient early-conditioned model, obviating the need for it to extrapolate to those conditions during test time. We call this model early-conditioned DRAM (DRAM$_\epsilon$) and propose a hyper-network architecture for its representation. It comprises two components, one network that takes the desired lighting condition as input and predicts the weights for a second network that produces the view, expression and lighting-dependent texture. Such a design further increases the capacity of the network and results in renderings of much higher quality while maintaining a low computational cost. The result is a method for creating animatable faces that can be relit using novel illumination conditions and rendered in real time. We demonstrate a use case of our relightable model by live-driving it from a VR-headset mounted camera [Wei et al. 2019] and rendering under novel and varying illumination (Figure 1).

To summarize, the contributions of this work are:

- A method for generating high-fidelity animatable personalized face avatars from dynamic multi-view light-stage data that can be relit under novel lighting environments, including challenging natural illumination and near-field lighting that are far from what is observed during training.
- A student-teacher framework for training an efficient relighting model that achieves real-time rendering while overcoming generalization limitations typically exhibited by such models.
- A novel hyper-network architecture for early-conditioned models that achieves significantly improved reconstruction accuracy while remaining efficient to evaluate.
- The first demonstration of relightable faces driven by headset mounted cameras for VR applications.

2 RELATED WORKS

Face modeling. Traditional methods for face modeling [Alexander et al. 2009; Seymour et al. 2017] depend on precise 3D reconstruction of human faces, which require a large amount of manual effort and

are not suitable for real-time applications. Recently Lombardi et al. [2018] propose a data-driven method for face modeling. It applies a conditional variational autoencoder to learn a latent representation for facial expressions and regresses a tracked mesh and a view-dependent texture to model the appearance of human faces. Schwartz et al. [2020] build on the same framework and explicitly model human eyes for better eye contact. However, these methods are limited to a single lighting condition and do not support relighting under novel lighting conditions.

Reflectance acquisition. To relight human faces under novel lighting conditions, previous approaches have tried to estimate the reflectance properties of human faces from captured images. Such methods usually assume a simplified reflectance model based on physical priors. Some previous works develop their method based on the diffuse assumption for faces. Garrido et al. [2013] and Cal et al. [2015] assume that faces are diffuse and jointly estimate the diffuse albedo and facial geometry from monocular videos. Shu et al. [2017] applies a learning-based method to infer facial normals and albedo from a single image. Other works also model specular reflections of human faces. Both diffuse and specular albedos are estimated from captures with different acquisition setups such as spherical gradient illuminations [Guo et al. 2019; Ma et al. 2007] and multi-view captures under passive illumination [Gotardo et al. 2018]. Yamaguchi et al. [2018] applies a deep-learning based approach to infer both reflectance and high-frequency displacement maps to model mesoscopic surface details on human faces from a single RGB image under uncontrolled illuminations. More complex reflectance models that consider subsurface scatterings have also been applied. Jensen et al. [2001] introduced a bidirectional surface scattering model for human faces based on a dipole diffusion approximation and proposed a method to measure the model parameters. Ghosh et al. [2008] recover layered facial reflectance including specular reflectance and scatterings at different layers from a set of twenty photographs under environmental and projected illuminations. All these physically-based approaches can only model a portion of face appearances, and fail to faithfully reproduce the complex visual appearance of human faces, especially for dynamic animations, where different expressions will result in significant differences in appearance. In addition, rendering with such reflectance models under complex lighting conditions also requires physically-based path tracers, which is computationally expensive and not suitable for real-time applications.

Image-based relighting. Methods in this category make use of the linearity of light transport and synthesize renderings of the scene under novel lighting conditions by combining images under a set of basis lighting patterns. A category of these works focus on the relighting of static scenes. Debevec et al. [2000] capture the reflectance field of human faces by capturing images under a dense sampling of directional incident illuminations. Xu et al. [2018] propose a learning-based method to synthesize renderings of static scenes at a novel lighting direction from a sparse set of captures. Sun et al. [2019] train a network to directly regress the relighting results under novel environment lightings from a single portrait image. Their results have limited fidelity and cannot recover visual effects such as specularities and detailed glints. In a later work [Sun et al. 2020], they propose a method to increase the resolution of static light stage captures and enable relighting under an arbitrary lighting direction. Most of these methods require a static capture setup, where the subject remains still and maintains a fixed expression while a one-light-at-a-time (OLAT) capture is performed. This limits their ability to capture transient expressions of the face in motion required for building dynamic animatable face avatars.

In addition to the methods mentioned above, Wenger et al. [2005] achieve dynamic relighting with time-multiplexed lightings where the subject is illuminated with a rapid series of basis lighting patterns. They warp adjacent frames to the target frame using optical flow so as to relight the target frame, which suffers from potential misalignments due to inaccuracies in flow computation. Meka et al. [2019] applies colored gradient illumination for efficient dynamic captures, and they train a network to infer the renderings under an arbitrary lighting direction from two gradient illumination captures, which are then used for relighting under novel lighting conditions. Their method requires colored illuminations, and suffers from misalignments between the two input captures. More importantly, while these methods support dynamic relighting, they can only support playback and relighting of the captured frames, and do not support animations and novel expressions.

Free-viewpoint relighting. Some existing works leverage 3D reconstructions of the scene to enable free-viewpoint rendering of the relightable models they build. Zhang et al. [2020] achieve free-viewpoint relighting of static human captures by explicitly reconstructing the geometry of the scenes and training a network to synthesize texture-space RGB images under the desired view and lighting direction. Gao et al. [2020] learn neural textures for the coarse geometric proxies of static scenes and directly encode the lighting information into rendered radiance cues with a set of basis materials. While the usage of radiance maps enables rendering under arbitrary lighting conditions, rendering radiance maps under complex natural illuminations is time-consuming, making it not suitable for real-time applications. Meka et al. [2020] propose a network to regress dynamic textures of the underlying geometry under an arbitrary lighting direction from color gradient captures. However, they use per-frame reconstructions without correspondence between the frames, restricting its use to performance replay. We provide a detailed comparison of features between previous methods and our method in Table 1. Compared to previous works, our method is the first to support novel viewpoints, novel lighting conditions, dynamic playback and animation.

3 DATA ACQUISITION

The appearance of human faces can be modeled as a function of the facial expression, viewpoint and lighting condition. We propose to use neural networks to approximate such a function. To supervise the training of such a network, ideally we could capture image data of all possible combinations of these three factors using a light stage. Our capture system consists of 140 color cameras and 460 white LED lights. All the LEDs can be independently controlled with adjustable lighting intensity. The cameras and lights are positioned on a spherical dome with a radius of 1.1m surrounding the captured subject.

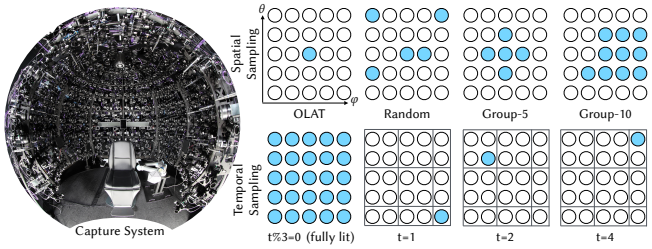

Fig. 2. Capture system with lights and cameras in a spherical dome (left) and light patterns used during capture (right). We evaluate different spatial groupings: one-light-at-a-time, 5 random lights, and spatial groups of 5 and 10 lights. Temporally, we sample lights using stratified random sampling.

To densely sample expression and viewpoint combinations, a capture-subject is asked to make a predefined set of facial expressions, recite a set of 50 phonetically balanced sentences, perform a range-of-motion sequence, and have a short natural conversation with a colleague [Lombardi et al. 2018]. During captures, all the 140 cameras synchronously capture at a frame rate of 90 frames per second, and output 8-bit Bayer-pattern color images with a resolution of 2668×4096.

The simultaneous capture of images with different lighting conditions is much more challenging in comparison. Wavelength multiplexed approaches [Gotardo et al. 2015; Hernández et al. 2007] are limited in the frequency bands that can be used, while time-multiplexed approaches [Wenger et al. 2005; Wilson et al. 2010] present challenges in capturing dynamic content with transient expressions. Our work follows the approach of Wenger et al. [2005], where time-multiplexed lighting is captured by rapidly cycling over a set of basis lighting patterns. However, instead of requiring static expressions for each cycle, we rely on amortized inference [Kingma et al. 2014] to disentangle lighting from expression in our captures of the face in motion, and evaluate the suitability of different kinds of lighting patterns for this approach. Specifically, we evaluate the efficacy of OLAT, *Random* (i.e., spatially unstructured sets of 5 lights), and two sets of *Group* patterns (i.e., spatially clustered groups of lights); one with five lights and another with ten. The rank of the basis formed by each lighting pattern ranges from 460 to 50. In all cases, a fully-lit frame is interleaved every third frame to enable face tracking [Wu et al. 2018] which produces a topologically consistent mesh, $\mathbf{M} \in \mathbb{R}^{3\times7306}$, for every frame[1]. In discussions that follow, we will use the following notation to refer to the lighting at a given frame:

$$L = \{b_1, b_2, ..., b_n\} \tag{1}$$

where b_i is the index of the i-th light that is turned on and n is the total number of lights for that frame.

The choice of lighting patterns we consider in this work is guided by a few factors that are difficult to meet simultaneously. First, it is desirable to see many different facial expressions for each lighting condition. OLAT generates the most complete set of lighting conditions with the finest spatial resolution, but has a long cycle time, minimizing the variety of facial expression seen in each lighting condition. Second, it is desirable to see many complementary lighting conditions for each facial expression. To achieve this, we temporally sample light directions using spatially stratified random sampling: lights are first stratified into 8 groups (represented as grid cells in Figure 2) with the next group chosen using furthest-group sampling across consecutive frames, and the light direction chosen randomly within a group. Third, it is preferable to have as much light as possible to overcome the *noise floor* of our cameras. Random and grouped lights trade off the spatial granularity of each lighting condition, but increase the light available to the cameras, potentially relaxing requirements on the capture system.

[1]In this work we presume the mesh between every third frame can be well approximated by linearly interpolating its adjacent tracked meshes.

4 BUILDING RELIGHTABLE AVATARS

Our goal is to build personalized expressive face avatars that can be rendered from novel viewpoints and relit to match the lighting in novel environments. We leverage the representation power of neural networks to map viewpoint, expression and lighting to highly accurate texture and geometry, which can be used to synthesize an image using standard rasterization techniques [Lombardi et al. 2018]. To overcome challenges presented by dynamic capture that are discussed in Section 3, we leverage the amortized inference properties of conditional variational auto-encoders (CVAE) [Kingma and Welling 2013] to disentangle expression from lighting in our representation. However, a naïve implementation of such an architecture generalizes poorly to novel lighting conditions that one might encounter in practice. This includes natural indoor and outdoor illumination conditions that can be quite different from the point light patterns used during data capture. An example of such a failure is illustrated in Figure 15. A key contribution of our work is a two-stage system that enables efficient relightable models that generalize to unseen lighting conditions to be learned.

The first stage of our system comprises a representation, DRAM_ℓ, that achieves generalization by leveraging the additive property of light. Although it is computationally expensive to evaluate, it allows us to synthesize high fidelity face images under lighting conditions that are far from what can be captured in our light stage. Thus, we use DRAM_ℓ to generate a large number of high-quality synthetic images to complement our real captured images, and to overcome the need for the efficient neural network architectures used in the second stage to extrapolate to those conditions.

Armed with an expanded dataset generated from the first stage, the second stage of our system involves training a novel neural network architecture, DRAM_ϵ, with high capacity but low compute. Here, we employ a hyper-network that produces lighting-specific network weights of a standard deconvolutional architecture that has previously been demonstrated to be capable of spanning the space of expressions for a single lighting condition [Lombardi et al. 2018]. The resulting model attains real-time performance of 75 frames per second on a Nvidia Tesla V100, and we demonstrate its suitability for animation by driving it from headset-mounted cameras as discussed in Section 5.

For all the models we describe in this section, we follow the data preprocessing described in [Lombardi et al. 2018]. Specifically, images, $\mathbf{I} \in \mathbb{R}^{3\times2668\times4096}$, of a specific frame and camera viewpoint,

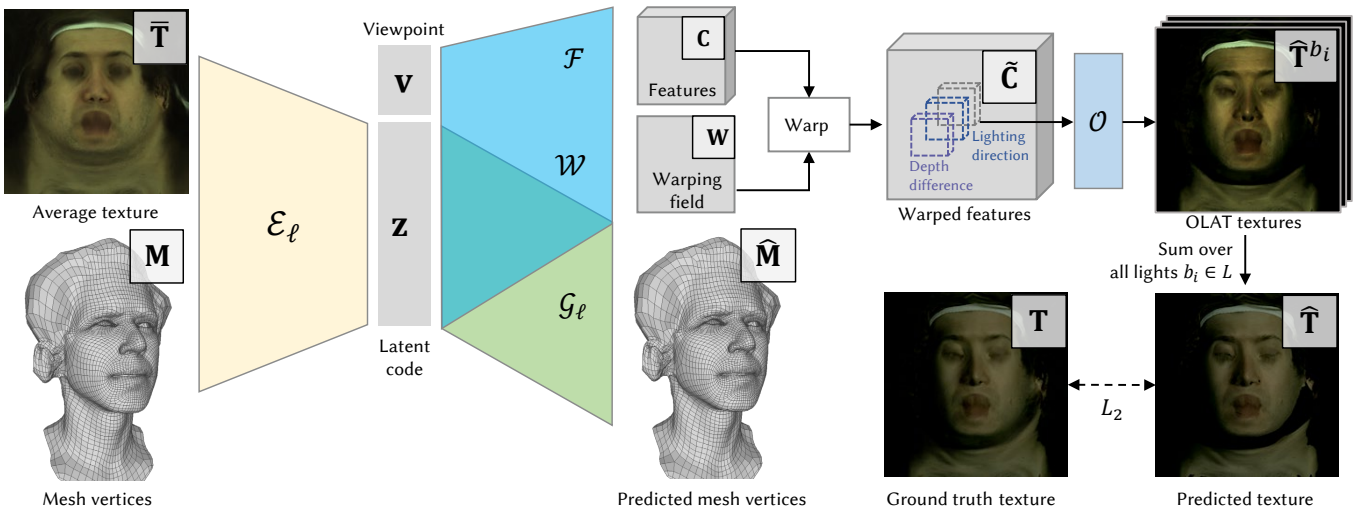

Fig. 3. Network architecture for our late-conditioned model. Expression and view-dependent features are generated with an encoder-decoder architecture, and late-conditioned with in an MLP network to produce single-light textures, which can be modulated by light intensity and summed to produce textures under more complex illuminations.

whether real or synthetic, are unwarped into a texture, $\mathbf{T} \in R^{3\times 1024^2}$, using the tracked mesh, $\mathbf{M}$, for that frame. We also calculate the average texture, $\bar{\mathbf{T}}$, for fully-lit frames by averaging the texture at each camera, which is used as input to CVAE to encourage better disentanglement between viewpoint and latent space. Representative visualizations of these elements are shown in Figure 3.

4.1 DRAM$_\ell$: A Late-conditioned Model

As shown in Figure 3, our late-conditioned model is a CVAE comprised of an encoder $\mathcal{E}_\ell$ and a decoder $\mathcal{D}_\ell$. The encoder takes the tracked mesh, $\mathbf{M}$, and the average texture, $\bar{\mathbf{T}}$, of its nearest fully-lit frame as input and outputs the parameters of its variational distribution, $\mathcal{N}$, from which the latent code $\mathbf{z} \in \mathbb{R}^{256}$ is sampled:

$$\boldsymbol{\mu}, \boldsymbol{\sigma} \leftarrow \mathcal{E}_\ell\left(\mathbf{M}, \bar{\mathbf{T}}\right) \quad , \quad \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2). \tag{2}$$

A Gaussian distribution with diagonal covariance is used for $\mathcal{N}$. The reparameterization trick [Kingma and Welling 2013] is used to ensure differentiability of the sampling process.

The input to the decoder $\mathcal{D}_\ell$ includes the latent vector $\mathbf{z}$, the view direction $\mathbf{v}$ of the camera relative to the head orientation in that frame, and the lighting condition L, transformed to the head coordinate system. The decoder outputs the reconstructed mesh $\hat{\mathbf{M}}$ and predicts the textures corresponding to each single light in L, which sum up to produce the final texture $\hat{\mathbf{T}}$. The decoder consists of two branches: the geometry branch, $\mathcal{G}_\ell$, which takes the latent vector as input and predicts the mesh, and the texture branch, $\mathcal{T}_\ell$, which additionally conditions on viewpoint and lighting to produce texture:

$$\hat{\mathbf{M}} = \mathcal{G}_\ell(\mathbf{z}) \quad , \quad \hat{\mathbf{T}} = \mathcal{T}_\ell(\mathbf{z}, \mathbf{v}, L). \tag{3}$$

Our texture branch consists of three components; a feature network $\mathcal{F}$, a warping network $\mathcal{W}$, and an OLAT prediction network $\mathcal{O}$. The feature network and the warping network output view-dependent feature maps, and the OLAT network takes per-texel features and a single lighting direction as input to predict the lighting-dependent colors at each texel. Finally we combine the colors under each light weighted by the lighting intensity to reproduce the texture. Please refer to Figure 3 for an illustration of our architecture.

Feature network. The feature network takes the latent vector, $\mathbf{z}$, and view direction, $\mathbf{v}$, as input and outputs a 64-channel feature map of size of 512×512:

$$\mathbf{C} = \mathcal{F}(\mathbf{z}, \mathbf{v}) \tag{4}$$

This feature map serves as a spatially varying encoding of expression and viewpoint across all lighting conditions.

Warping network. The warping network outputs a view-dependent warping field, $\mathbf{W} \in \mathbb{R}^{2\times 1024^2}$, which is applied to the feature map, $\mathbf{C}$, resulting in a warped feature map, $\tilde{\mathbf{C}} \in \mathbb{R}^{64\times 1024^2}$, of the same size as the texture:

$$\mathbf{W} = \mathcal{W}(\mathbf{z}, \mathbf{v}) \quad , \quad \tilde{\mathbf{C}}_t = \phi(\mathbf{C}, \mathbf{W}), \tag{5}$$

where ϕ denotes the warping operator, which performs bilinear interpolation at floating point coordinates. The warping field accounts for *texture sliding* as a result of view-dependent effects stemming from imperfect geometry, most noticeable around the mouth, eyes and hair, where accurate geometry is difficult to estimate during mesh tracking. It is also used to upscale the lower resolution feature maps, whose size is constrained by memory limitations on modern GPU hardware.

OLAT network. Given the warped feature map, $\tilde{\mathbf{C}}$, $\mathcal{O}$ is applied to each texel independently, where it predicts the color of that texel under a given lighting direction. $\mathcal{O}$ is a multi-layer perceptron (MLP) that, for a texel k and a light b_i with position $\mathbf{l}_{b_i}$, takes as input $\tilde{\mathbf{C}}_k$, the 64-dimensional feature of $\tilde{\mathbf{C}}$ at texel k, as well as the direction of light with respect to the corresponding point on the face in 3D.

Different from previous works (e.g., [Meka et al. 2019]) which assume distant lighting and where all texels share the same lighting

direction, we calculate the lighting direction of each texel using the light position and the corresponding position of the texel on the reconstructed geometry, $\hat{\mathbf{M}}$. This better models the setting in our light-stage, whose 1.1m radius results in some non-negligible foreshortening effects.

One of the most distinctive appearance change on faces is shadow by self-occlusion. While our late-conditioned model allows us to learn appearance change in a localized manner, we observe that it remains challenging for such a model to learn clear shadow boundary due to the lack of geometric information, resulting in noticeable artifacts. To alleviate this issue, we exploit the predicted geometry, $\hat{\mathbf{M}}$, to encode geometric relationship between a light source and a texel in the spirit of a shadow map [Williams 1978] as an additional input to O. Specifically, for a texel, k, and its corresponding 3D position, $\mathbf{p}_k$, we calculate the difference between the depth of $\mathbf{p}_k$ and its nearest occluder along the ray from the light to the texel in the light coordinate frame. With this, we arrive at the final form for our OLAT network:

$$\hat{\mathbf{T}}^{b_i}(k) = O\left(\tilde{\mathbf{C}}_k, \mathbf{d}_k^{b_i}, s_k^{b_i}\right) \tag{6}$$

where $\mathbf{d}_k^{b_i}$ is the lighting direction of light b_i for texel k, and $s_k^{b_i}$ is the depth difference mentioned above. Applying the OLAT network to each texel gives us the full texture $\hat{\mathbf{T}}^{b_i}$ under the current view direction and lighting, b_i.

Each frame of our training data is captured under multiple lights, and we approximate the training textures by the weighted sum of textures generated for each light independently, using weights that reflect the intensity of each light. Given the preset lighting intensity γ^{b_i} for a light b_i, our final predicted texture is constructed as follows:

$$\hat{\mathbf{T}} = \sum_{i=1}^{n} \gamma^{b_i} \hat{\mathbf{T}}^{b_i}. \tag{7}$$

Training. Our loss function consists of four terms, including a texture reconstruction loss ℓ_T, a geometry reconstruction loss ℓ_M, a regularizer loss on the warping field ℓ_W and a latent space regularizer ℓ_Z:

$$\mathcal{L}(\mathcal{E}_\ell, \mathcal{D}_\ell) = \sum_{\mathbf{v},t} \lambda_T \ell_T + \lambda_M \ell_M + \lambda_W \ell_W + \lambda_Z \ell_Z, \tag{8}$$

where $(\mathbf{v}, t)$ are the camera and frame indices over the dataset, and:

$$\ell_T = ||w \odot (\mathbf{T} - \hat{\mathbf{T}})||_2^2 \tag{9}$$

$$\ell_M = ||\mathbf{M} - \hat{\mathbf{M}}||_2^2 \tag{10}$$

$$\ell_W = ||\mathbf{W} - \mathbf{W_I}||_2^2 \tag{11}$$

$$\ell_Z = \mathrm{KL}\big(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \;||\; \mathcal{N}(0, \mathbf{I})\big) \tag{12}$$

Here, w is a weight map that avoids penalizing self-occluded texels in the current view[2]. The term $\mathbf{W_I}$ is an identity warping field, and the regularizer loss ℓ_W prevents the warped texel positions from drifting too far from their original positions. The KL-divergence loss ℓ_Z with a standard normalization encourages a smooth latent space. In all our experiments we set the weights of each loss term as $\lambda_T = 1, \lambda_M = 0.1, \lambda_W = 10, \lambda_Z = 0.001$. We use the Adam optimizer [Kingma and Ba 2014] with a learning rate of 0.0005 for training. We train the networks on 4 Nvidia Tesla V100 GPUs with a batch size of 16 for about $300k$ iterations, which takes 4-5 days on average.

Testing. Our model provides great flexibility and generalization for rendering under novel lighting conditions. We can feed in an arbitrary lighting direction for each texel as input to the OLAT network O, and predict the texture under the desired lighting conditions. Therefore, our model supports the rendering of directional lighting (Figure 7) as well as *near-field* lighting (Figure 8), which is not previously possible using existing image-based portrait relighting methods [Meka et al. 2019; Sun et al. 2019; Zhang et al. 2020][3]. For complex lighting conditions like environment maps, we can predict textures for every single pixel in the environment map, and linearly combine them to synthesize a face image in that environment. The model's runtime comprises: 24ms for shadow map calculation, 29ms for feature map generation, and 0.9ms for full texture decoding of a single lighting direction on a single Nvidia Tesla V100 GPU. Although feature map generation needs to be computed only once, the shadow map and texture decoding need to be performed for each light in the environment. So, although single light rendering using DRAM_ℓ can be relatively fast (i.e., $\sim$ 55ms), even a low-resolution (16×32)-environment map can take $\sim$ 18 seconds.

4.2 DRAM_ϵ: An Early-conditioned Model

Our late-conditioned model allows us to synthesize face images under novel expressions, viewpoints and lighting conditions. However, it is computationally expensive to evaluate for complex lighting conditions with many light sources. Unfortunately, most natural illuminations exhibit this property. Hence, they are typically modeled using an environment map, which is equivalent to having as many light sources as there are non-zero pixels in the map. Thus, this model is not suitable for interactive applications, such as VR, where real-time performance is necessary. In this section, we build on top of results from the late conditioned model described in the previous section to arrive at a formulation with similar accuracy, but that is an order of magnitude more efficient.

Data generation. We use DRAM_ℓ to generate face renderings under environment maps captured from real indoor and outdoor scenes, and use the generated textures as ground truth to supervise the training of our early-conditioned model; DRAM_ϵ. For the set of environment maps to render, we use the large-scale dataset collected by Gardner et al. [2017] and Sun et al. [2019], which contains 3094 high-resolution HDR environment illuminations including both indoor and outdoor scenes. We randomly select 2560 environment maps from the dataset for training and use the remaining 534 for testing.

We generate these synthetic lighting images for randomly sampled frames and viewpoints from our light stage capture. During

[2]We have omitted indexing the variables with $\mathbf{v}$ and t in our equations to reduce notation clutter, but they should be understood to correspond to unique values for every frame and viewpoint in the dataset.

[3]For nearfield lighting we employ a quadratic fall-off for lighting intensities used in the weighted sum in Equation 7.

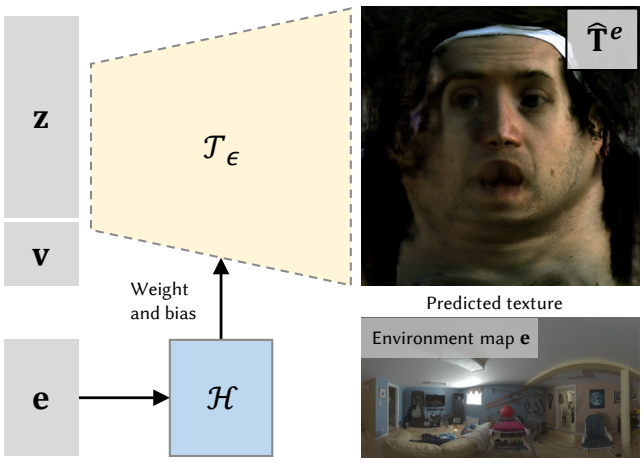

Fig. 4. We apply a hyper-network architecture for our early-conditioned model, where we use a separate network $\mathcal{H}$ that takes the current environment map as input to predict the weight and bias of the texture decoder $\mathcal{T}_\epsilon$.

rendering, we randomly select an environment map from the training dataset and apply a random rotation of $[0°, 360°]$ in longtitude and $[-30°, 30°]$ in latitude, followed by downsampling to a (16×32)-sized lat-long environment map. The resized environment map is further normalized by dividing by its sum and multiplying by a constant $\alpha \in [6, 12]$. We denote the environment map as $\mathbf{e}$. DRAM_ℓ is applied to predict the textures for each lighting direction, which correspond to individual pixels in the environment map, and perform the weighted sum in Equation 7 to produce the final texture $\mathbf{T}^e$. In total, we generate $1.2M \sim 1.8M$ textures for training each subject in our dataset. In addition to environment map renderings, we also augment our training data by rendering the captured subject under $1-5$ lights randomly selected from the 460 lights. During training, we project the selected lights onto an environment map of 16×32 and use them as input to our network to predict the corresponding textures.

Network architecture. Our early-conditioned model exhibits a similar CVAE architecture as its late-conditioned counterpart, comprising an encoder, $\mathcal{E}_\epsilon$, and a decoder, $\mathcal{D}_\epsilon$. The encoder, $\mathcal{E}_\epsilon$, shares the same architecture and input as $\mathcal{E}_\ell$, and outputs a latent vector, $\mathbf{z}$. The decoder also consists of two branches; a geometry decoder, $\mathcal{G}_\epsilon$, with the same architecture as $\mathcal{G}_\ell$, and a texture decoder, $\mathcal{T}_\epsilon$, that predicts a texture under the given environment map.

As shown in Figure 4, a naïve architecture for the texture decoder would be an extension of [Lombardi et al. 2018], where the vectorized environment map is concatenated with the latent vector, $\mathbf{z}$, and view direction, $\mathbf{v}$, and fed into to a single deconvolutional network to output the predicted texture, $\hat{\mathbf{T}}^e$. As this network architecture is designed for speed, it lacks the capacity to accurately reconstruct data that spans a large number of different environment maps. To do this, a straight-forward approach would be to increase the channel size in the hidden layers of the network. However, as we will show in Section 6.2, a considerable increase is required to achieve reasonable accuracy, which diminishes the model's efficiency, making it unsuitable for real-time applications.

Our early-conditioned model takes inspiration from recent works on hyper-networks [Ha et al. 2016], and consists of two networks: a weights network, $\mathcal{H}$, that takes the environment map as input and predicts the weights for a second network, $\mathcal{T}_\epsilon$, that takes the efficient form used in [Lombardi et al. 2018], and produces a view, lighting and expression dependent texture:

$$\Theta \leftarrow \mathcal{H}(\mathbf{e}) \quad , \quad \hat{\mathbf{T}}^e = \mathcal{T}_\epsilon(\mathbf{z}, \mathbf{v}|\Theta). \tag{13}$$

Θ denotes the weights of $\mathcal{T}_\epsilon$ that consists of 8 transposed convolution layers. For each layer, we use a small weights network that consists of 5 fully connected layers to predict the convolutional kernel weights and biases. Similar to the late-conditioned decoder, a warping field is employed on the output of the texture decoder to give us the final texture. The hyper-network architecture specializes the texture network to a specific lighting condition, which we find to be effective in improving reconstruction performance without substantially increasing computational cost, as shown in Figure 15 and Table 5.

Training. We use all the same settings for training DRAM_ϵ as we did for DRAM_ℓ. For the same number of iterations, a model can be trained within $3-4$ days on average. The trained model can synthesize face images lit by environment maps within 13ms (~75 frames per second), making it suitable for interactive applications, including demanding real-time applications such as VR.

5 ANIMATING RELIGHTABLE AVATARS

The trained early-conditioned decoder $\mathcal{D}_\epsilon$ can efficiently generate novel outputs with respect to its three inputs: expression, viewpoints, and lighting. The disentanglement of these factors in the model are important for animation, because the images coming from driving sensors can have completely unrelated viewpoints and lighting to the decoded avatar. For example, in the VR telepresence system of [Wei et al. 2019], the driving signal comes from headset-mounted IR cameras that observe facial expressions of a person wearing the headset in an arbitrary room, while being lit by headset-mounted IR lights, whereas the avatar that person is driving needs to be relit in accordance with the virtual scene, which might be arbitrarily different from where the person really is. The only factor that is desirable to match between the sensor images and the avatar, is the facial expression.

In this work, we utilize the method in [Schwartz et al. 2020], which finds correspondences between input headset images and expression codes $\mathbf{z}$ of DRAM_ℓ through analysis-by-synthesis. We similarly learn a regressor that encodes multi-view headset images into $\mathbf{z}$ and a relative pose between the headset and avatar, jointly with a style transformer that accommodates for domain differences between the headset images and the rendered avatar. An important difference here, is that we assume the lighting variation in the sensor images is small enough so that we can fix the lighting input, $\mathbf{e}$, at a constant uniform lighting. Any difference in lighting between the domains is handled by the style transformer. While this assumption holds in many cases, as shown in Section 6, an interesting future direction is to leverage our model's relighting capability and jointly optimize the model's lighting so that there is less reliance on the

Fig. 5. We show comparisons between the predicted OLAT images under novel viewpoints and expressions with our late-conditioned model and the ground truth. Our model is able to reconstruct the OLAT images accurately, even though it is trained only on group-light captures. This enables us to synthesize accurate renderings under novel lighting conditions by combining multiple OLAT predictions.

Fig. 6. Late-conditioned model: rendering under novel *directional lights*.

style transfer module, which can introduce semantic shifts during optimization.

6 RESULTS

In this section we provide qualitative and quantitative evaluations on different components of our method, including both the late-conditioned model (Section 6.1) and the early-conditioned model (Section 6.2). We perform ablation studies on each model to validate our design choices. We show relighting results with our models under novel lighting conditions, viewpoints, and expressions. We also demonstrate our relighting results animated by images captured with VR headset mounted cameras (Section 6.3).

6.1 Evaluation of Late-conditioned Models

As discussed in Section 2, none of the previous works support both free-viewpoint relighting and animations, as in our method. The work that is closest to ours is Meka et al. [2020]. However, their model is not animatable and requires color gradient illuminations as input. Therefore, in this section we focus on showing our qualitative results and validating different design choices in our system. We train the model for each subject on captures under group-light patterns, and the two subjects shown in Figure 5 are captured with the Group-5 pattern while the other subjects are captured with Group-10.

Qualitative results. We first compare our renderings to ground truth captures under novel viewpoints, expressions, and lighting conditions. To achieve this, we evaluate our model on a separate sequence of images captured using a similar acquisition setup as described in Section 3 except that each frame in this sequence is captured under a single light. We make the comparison on images captured at a set of 4 validation viewpoints that are not used in training. As shown in Figure 5, although our model is trained on images captured under group-light patterns and has never seen OLAT captures during training, our network can successfully reconstruct high-fidelity OLAT images that closely approximate the ground truth captures in terms of shadows and specularities. This demonstrates that our proposed model can not only generalize to novel expressions and viewpoints, but also effectively super-resolve the group-light captures.

Figure 6 shows renderings with our model under novel directional lights. By combining the renderings under each pixel lighting of an environment map, our model can also achieve photorealistic renderings under environment lighting. Figure 7 shows environment map renderings with our model under both outdoor and indoor

Fig. 7. Renderings under environment maps with our late-conditioned model. Our model is able to faithfully recover complex shading effects including specularities and shadows.

Fig. 8. Nearfield relighting with our late-conditioned model. Our late-conditioned model can take different lighting directions for each texel and predict their colors, which enables us to achieve efficient nearfield renderings making use of the geometry reconstructed with our geometry decoder.

environment maps. Our model can faithfully recover the glints on the forehead and the specularities on the face.

Figure 8 shows our rendering results under near-field lighting. We make use of our reconstructed geometry output by the geometry decoder to calculate the lighting direction of each texel. Since our OLAT prediction network O is applied on each individual texel, our model can predict the OLAT renderings with a single inference. In comparison, previous methods [Meka et al. 2019; Sun et al. 2019; Xu et al. 2018] do not reconstruct the geometry and therefore fail to support near-field lighting. While some other methods [Meka et al. 2020; Zhang et al. 2020] build on estimated geometry, their network can only take a single lighting direction as input at each time. To predict near-field rendering, separate evaluations of their model for the lighting direction of each individual pixel would be required, which is very time-consuming. In contrast, our model provides greater flexibility and more efficient near-field renderings. For more results, please refer to the supplementary video.

Evaluation of design choices. To validate our different design choices, we evaluate our models on testing sequences for Subject 1 and Subject 2 and compare them to the ground truth. Both subjects have two testing sequences. The testing sequences for Subject 1 consists of 18014 and 34432 frames, and the sequences for Subject 2 have 17165 and 23072 frames. There are 4 testing images

Ground truth Ours w/o depth difference

Fig. 9. We make use of the shadow mapping technique to feed self-shadowing information to the network. We can see that without the depth differences, the rendering results suffer from jagged boundaries at shadows. In contrast, our full model reproduces more accurate shadows.

	Subject 1		Subject 2	
	MSE ($\times 10^{-4}$)	SSIM	MSE ($\times 10^{-4}$)	SSIM
Our full model	**6.4377**	**0.9363**	**2.9843**	**0.9469**
w/o depth difference	6.5115	0.9344	3.0562	0.9464

Table 2. We evaluate the effectiveness of using depth differences as input to the OLAT network in our late-conditioned model on two subjects. Subject 1 and Subject 2 correspond to two subjects in Figure 5 respectively, The results on both subjects show that involving the occlusion information helps improve the accuracy of our model.

Fig. 10. Late-conditioned model: a visual comparison of different *capture lighting patterns*. The leftmost image shows a ground truth image under an "OLAT" single point-light illumination. We reconstruct this using a model trained on 5 spatially clustered lights ("Group-5"), OLAT, and 5 spatially random lights ("Random"). Both "Group-5" and "random" can use shorter camera exposures than OLAT to achieve similar camera intensities, but only "Group-5" recovers comparable details to OLAT.

for each frame captured at novel viewpoints that are not used in training. All numbers are reported on the first sequence except for those in Table 3. We consider image-space error metrics including mean-squared error (MSE) and structural similarity index (SSIM). Considering the fact that the ground truth OLAT images with our models may have different lighting intensity than the predictions, and there are potential color mismatches due to different camera calibrations, we optimize a matrix $\mathbf{Q} \in \mathbb{R}^{3\times3}$ to align our predicted image $\hat{\mathbf{I}}$ to the ground truth $\mathbf{I}$:

$$\mathbf{Q} = \arg\min_{\mathbf{Q}} ||\mathbf{Q}\hat{\mathbf{I}} - \mathbf{I}||_2^2 \tag{14}$$

Then we calculate all error metrics between $\mathbf{Q}\hat{\mathbf{I}}$ and $\mathbf{I}$.

In Table 2, we perform an ablation study to show the effectiveness of applying depth differences as input to our OLAT network. From the result we can conclude that it helps improve the accuracy of the model. We provide additional qualitative comparisons in Figure 9. As we can see from the figure, without including the depth difference information, the network predicts shadows with incorrect shapes and jagged boundaries, especially for long-range shadows on the neck. In contrast, our full model with depth differences produces more accurate shadows.

Effect of spatial light pattern. In Table 3, we make a comparison between different lighting bases for our time-multiplexed lighting. We train our late-conditioned models on captures under different lighting bases including OLAT, Random, and Group-5. We make the comparisons by predicting the OLAT images under novel expressions and viewpoints and calculate the error between the predictions and their corresponding ground-truth OLAT captures. A visual comparison is also shown in Figure 10. From the results we can see that Group-5 captures lead to slightly better reconstruction accuracy than *r*andom lighting patterns. Compared to OLAT, Group-5 achieves very similar performance despite the model never having seen OLAT images during training, and using an evaluation design

Fig. 11. We show visual comparisons between renderings on novel expressions with our late-conditioned model trained on static captures and dynamic captures. Static captures cover many fewer facial expressions than dynamic captures within the same number of frames, thus resulting in poor generalization to novel expressions.

that is favorable to OLAT captures. Compared to OLAT, grouped light captures have reduced single-light maximum power requirements (to overcome the noise floor of the cameras), or, equivalently, support capturing with shorter exposure times (which has implications for perceptual discomfort [Wenger et al. 2005]). This result

	Subject 1		Subject 2	
	MSE ($\times 10^{-4}$)	SSIM	MSE ($\times 10^{-4}$)	SSIM
OLAT	6.7205	**0.9843**	3.866	0.9931
Random	6.7588	0.9840	4.124	0.9930
Group-5	**6.5536**	0.9842	**3.676**	**0.9933**

Table 3. We compare the performance of our models trained on captures under different basis lighting patterns. We do the evaluation by comparing the predicted OLAT images under novel viewpoints and expressions to their corresponding ground-truth. From the result we can see that the performance of Group-5 capture is much better than random light patterns. The Group-5 capture is even better than the OLAT captures on 3 out of 4 metrics although it has never seen OLAT images in training. Note that this evaluation is done on a different testing sequence from that used in Table 2.

	Subject 1		Subject 2	
	MSE ($\times 10^{-4}$)	SSIM	MSE ($\times 10^{-4}$)	SSIM
Static capture	7.5862	0.9301	3.6335	0.9429
Dynamic capture	**6.4377**	**0.9363**	**2.9843**	**0.9469**

Table 4. Late-conditioned model: comparison between static captures and dynamic captures. Within the same number of frames, dynamic captures can cover more facial expressions and lead to better generalization to novel expressions, thus achieving higher accuracy.

indicates that grouping lights in spatial clusters is an attractive option for power or exposure constrained settings, with results almost indistinguishable for groups with a diameter twice the size of the single-light spatial sampling distance.

Effect of capture script content. We also compare a dynamic capture script with a static expression capture script of roughly the same total duration. For static captures, the subject is asked to remain still during each elicited expression, while a full cycle of the light patterns is captured. Each individual expression is therefore fully sampled along all spatial lighting directions. Conversely, for the dynamic capture, the subject is asked to move naturally, and the allotted capture time is used to elicit more varied expressions and poses. Instantaneous expressions are therefore sampled very sparsely along lighting directions, but a more diverse set of facial expressions is sampled, and we rely on amortized inference during model building to span the combined space of expressions and lighting directions. We capture roughly the same number of frames for these two kinds of captures. From the result in Table 4, we can see that our dynamic captures produce better results than static captures. Specifically, within the same number of frames, static captures cover a much smaller set of expressions than our dynamic captures. Therefore, as we can see in Figure 11, the model trained on static captures does not generalize well to novel expressions. In contrast, our dynamic captures provide us a more efficient way to capture the subject under a large number of expressions.

6.2 Evaluation of Early-conditioned Models

Qualitative results. In Figure 12, we show the rendering results with our early-conditioned model under novel environment maps, expressions and viewpoints. Because we are using the late-conditioned model to supervise the training of the early-conditioned model, we regard the renderings with the late-conditioned model as the ground truth. From the results, we can see that our early-conditioned model can generate photorealistic results with accurate texture details and shading effects that closely resemble its corresponding ground truth. Such results demonstrate that by extensively sampling the natural illuminations and generating renderings as training data, our early-conditioned model is able to achieve good generalization to novel lighting conditions.

Our early-conditioned model benefits from the training data augmentation with renderings under environment maps generated from projections of randomly chosen directional lights, it can therefore also generate compelling results on environment maps with high-frequency lighting, which are usually under-represented in existing environment map datasets. In Figure 13, we show our renderings under an environment map generated by projecting a single directional light onto the environment map. The results show that our early-conditioned model can generalize to high-frequency illuminations and generate accurate renderings that are comparable to the late-conditioned model.

We also compare to the state-of-the-art single image portrait relighting work of Sun et al. [2019]. To achieve this, we directly feed the ground truth fully-lit captured image as input to their method and compare their renderings to the predictions of our early-conditioned model. We use the code and model provided by Sun et al. to generate the results. From the results in Fig. 14, we can see that the method of Sun et al. fails to predict faithful shading effects such as specularities, and generates overly flat renderings. In contrast, our method produces renderings with higher fidelity.

Evaluation of design choices. To validate different design choices of our early-conditioned model, we evaluate our model and the comparison models on a separate testing sequence with novel expressions. We generate renderings with the models at a set of 4 novel viewpoints under environment maps randomly chosen from the testing dataset. The corresponding ground truth renderings are generated with our late-conditioned model. We also compare the renderings of different models on a testing sequence under group-light patterns to the ground truth group-light captures. To achieve this, we project the group-light patterns to environment maps and use them as input to the models. For both lighting conditions, the number of testing frames is 10884 for Subject 1 and 13664 for Subject 2. We apply the same error metrics as used in Section 6.1, and report the scores in Table 5. We also report the computational cost of the texture module by calculating the number of multiply-accumulate operations (MACs) for a single inference. We also show a visual comparison between our model and the baseline models in Figure 15.

From the results, we can see that the naïve decoder with fixed weights has low accuracy and generates incorrect colors on faces. In comparison, our hyper-network architecture produces more accurate renderings while maintaining a comparable computational cost. We also compare against a baseline model with twice the number of feature channels, and our hyper-network is able to achieve better performance with a much smaller computational cost.

Instead of training on environment map renderings generated with our late-conditioned model, we also compare against a model

Fig. 12. Our *early-conditioned* model is able to generate renderings under novel environment maps that have the same quality as those generated by the *late-conditioned* model. Compared to the time of around 18 seconds required by the late-conditioned model, our early-conditioned model is much more efficient and can generate environment map renderings in real time.

	Test on environment renderings				Test on group-light captures				MACs ($\times 10^9$)
	Subject 1		Subject 2		Subject 1		Subject 2		
	MSE($\times 10^{-4}$)	SSIM	MSE ($\times 10^{-4}$)	SSIM	MSE($\times 10^{-4}$)	SSIM	MSE($\times 10^{-4}$)	SSIM	
Single decoder	11.396	0.9815	3.6423	0.9885	7.5256	0.9833	3.7622	**0.9937**	1.44
Single decoder ($\times 2$ features)	8.7349	0.9862	2.6697	0.9909	7.2872	0.9838	3.5373	0.9938	5.53
Ours (trained on group-light)	108.93	0.9277	61.221	0.9417	**6.2536**	**0.9846**	**3.0830**	**0.9937**	1.50
Ours	**7.3878**	**0.9882**	**2.5309**	**0.9914**	7.4345	0.9838	3.5846	0.9938	1.50

Table 5. We compare our early-conditioned model with hyper-networks against baseline models. Compared to the naïve model that applies a single decoder network with fixed weights, our hyper-network is able to achieve much better accuracy with similar computational cost. Our model is even comparable to the naïve decoder model with twice as many feature channels at each layer, which has a much larger computational cost. In addition, from the figure we can see that while the model trained on group-light captures only can predict accurate group-light renderings, it fails to generalize to novel environment maps, which demonstrates the necessity of our two-stage student-teacher framework.

Fig. 13. Our early-conditioned model can generate photorealistic renderings under high-frequency environment lighting. Here we make a comparison between renderings with our early-conditioned and late-conditioned models under an environment map representing a single directional light. Our early-conditioned model can produce renderings of the same quality as the late-conditioned model.

that is only trained on our group-light captures by projecting the group-light onto environment maps and training the same hyper-network model on this dataset. From the table, we can see that while it can produce renderings with the highest accuracy on a testing set of group-light sequences, it produces the lowest accuracy on the environment renderings. Training only on the group-light captures makes the network overfit to the training lighting patterns and fail to generalize to novel environment lighting (see Figure 15, right). In contrast, by training on renderings with our late-conditioned model under extensively sampled natural illuminations, our hyper-network can generalize to novel environment illuminations, which demonstrates the necessity and effectiveness of our two-stage framework.

Effect of the environment map resolution. Theoretically, it is possible to generate training data under environment maps of an arbitrary resolution with our late-conditioned model. However, increasing the environment map resolution results in significantly more time for training data generation. In Figure 16, we compare rendering under environment maps with different resolutions with both the late-conditioned model and the early-conditioned model. The results show that a resolution of 16×32 has been able to well reproduce the complex lighting effects with no obvious artifacts, which motivates us to use such a resolution in our experiments to balance between the time for data generation and the rendering quality.

Fig. 14. We compare our early-conditioned model to the state-of-the-art single image portrait relighting method of Sun et al. [2019]. From the results we can see that the method of Sun et al. fails to recover accurate specularities on the face and eyes, produces softer shadows, and predicts incorrect colors. In contrast, our method can achieve much more photorealistic relighitng.

Fig. 15. We compare our early-conditioned model with hyper-networks against baseline models with a single decoder. Our hyper-network generates results of higher quality that better match the ground truth compared to baseline models. We also compare against a model that has the same architecture as ours but is trained on group-light captures only, and the result shows that such a model fails to generalize to novel lighting conditions. Instead, our models that are trained on environment map renderings with our late-conditioned model achieve better generalization, which demonstrates the effectiveness and necessity of our two-stage student-teacher framework.

6.3 Animation from Headset Mounted Cameras

The advantage of our relightable appearance model over previous works is the good disentanglement of its inputs: facial expression, viewpoints, and lighting. This makes the model animatable, and can be driven by sensors such as video captured by VR headset mounted cameras. In Figure 17, we show our early-conditioned model relighted in different environment, given the latent values $\mathbf{z}$ we extracted from the given headset images, using the method described in Section 5. The renderings of our model faithfully reproduce the facial expressions from the headset images, while photorealistically relit under novel lighting conditions. For more results, please refer to the supplementary video.

6.4 Limitations

While our models produce photorealistic relightable avatars, several limitations remain: (1) Most notably, in regions where the tracked mesh is inaccurate or lacks sufficient geometric detail, such as the hair outline and eyelashes, textured mesh rendering produces jarring border artifacts instead of blending into the environment background. For example, in Figure 10, our model fails to reconstruct the shadows caused by eyelashes. This could potentially be addressed

Fig. 16. We evaluate the quality of renderings under environment maps with different resolutions. We use the rendering under a 128×256 environment map with our late-conditioned model as ground truth, and compare the renderings under lower-resolution environment maps with both early- and late-conditioned models. Each comparison image has an inset on the bottom left to show the squared difference from the ground truth. From the result we can see that a resolution of 16×32 has been able to properly reproduce the complex shading effects on the face, while a resolution of 8×16 results in obvious artifacts. In our experiments we use environment maps with a resolution of 16×32 for our early-conditioned model to balance between the time for training data generation and the rendering quality.

Fig. 17. Our early-conditioned model under novel environment maps animated by VR headset mounted cameras. Our model is able to faithfully reproduce the expressions in the headset captures while achieving photorealistic relighting simultaneously.

via volumetric neural rendering approaches with the capacity to produce translucency. (2) Similarly, we notice some blurring in regions where the mesh geometry does not accurately track the surface, such as the mouth and eyes. Using specialized geometric models for these regions (e.g., [Bérard et al. 2019; Wu et al. 2016]) would greatly improve registration accuracy and therefore reduce the capacity required to model their appearance in texture space. (3) Very high frequency details (e.g., pores and strong specularities) are slightly blurrier in the rendered images, as shown in Figure 5. This can be potentially alleviated by increasing the network capacity and texture resolution. (4) Our models are limited by the acquisition hardware and lighting rig used to capture the training data. Due to the use of low-dynamic range 8-bit images, we notice decreased quality and color shifts in very dark regions, likely due to poor signal to noise ratio. Similarly, our model fails to reconstruct lighting directions that are very far from those that can be elicited by the light stage (e.g., lighting directly below the participant). High bit-depth HDR imaging and more complex light stage setups could improve results in these cases. (5) Finally, we have presented an efficient model for rendering animatable and relightable avatars in realtime in environmental illumination, but designing an efficient model that can render both nearfield and farfield illumination remains an open problem.

7 CONCLUSIONS

We presented Deep Relightable Appearance Models, a novel two-stage framework to achieve photo-realistic relighting of animatable face avatars. Our approach produces, for the first time, a photo-realistic face avatar that can be driven and rendered in real-time under various new illuminations. The experiments demonstrate that our late-conditioned model achieves high generalization across a wide-range of illuminations including natural indoor/outdoor illuminations, nearfield lighting, and distant directional lighting, despite being trained only with grouped point-light captures. This is possible due to the explicit modeling of the linear property of light transport and the late-stage fusion of light information in our network architecture. We further examined the effects of different light patterns and captured scripts, and show the efficacy of dynamic capture and spatial grouping of light sources. This allows us to render high-quality synthetic images under different illuminations to

generate an augmented training set for training efficient models. We also presented a hyper-network architecture for early-conditioned relightable models, which is highly efficient to run in real-time while showing comparable fidelity to a higher-capacity baseline. We believe that our two-stage framework is general and applicable to many different relighting problems and real-time applications, including volumetric rendering, and building cross-identity face models, which can be addressed in future work.

ACKNOWLEDGMENTS

This work was funded in part by a Qualcomm Innovation Fellowship, a Facebook Distinguished Faculty Award, and the Ronald L. Graham chair.

REFERENCES

Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. 2009. The digital emily project: photoreal facial modeling and animation. In *Acm siggraph 2009 courses*. 1–15.

P. Bérard, D. Bradley, M. Gross, and T. Beeler. 2019. Practical Person-Specific Eye Rigging. *Computer Graphics Forum* 38 (2019).

Ida Winifred Busbridge. 1960. *The mathematics of radiative transfer*. Number 50. University Press.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the Reflectance Field of a Human Face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 145–156. https://doi.org/10.1145/344779.344855

Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2020. Deferred Neural Lighting: Free-Viewpoint Relighting from Unstructured Photographs. *ACM Trans. Graph.* 39, 6, Article 258 (Nov. 2020), 15 pages. https://doi.org/10.1145/3414685.3417767

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017).

Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.* 32, 6 (2013), 158–1.

Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. 2008. Practical modeling and acquisition of layered facial reflectance. In *ACM SIGGRAPH Asia 2008 papers*. 1–10.

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

Paulo F. U. Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. 2015. Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. 2019. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. *ACM Trans. Graph.* 38, 6, Article 217 (Nov. 2019), 19 pages. https://doi.org/10.1145/3355089.3356571

David Ha, Andrew Dai, and Quoc V Le. 2016. Hypernetworks. *NIPS* (2016).

C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla. 2007. Non-Rigid Photometric Stereo with Colored Lights. In *ICCV* (Rio de Janeiro, Brazil).

Henrik Wann Jensen, Stephen R Marschner, Marc Levoy, and Pat Hanrahan. 2001. A practical model for subsurface light transport. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 511–518.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc., 3581–3589. https://proceedings.neurips.cc/paper/2014/file/d523773c6b194f37b938d340d5d02232-Paper.pdf

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *Proceedings ofthe 2nd International Conference on Learning Representations* (2013).

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4 (July 2018).

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul E Debevec. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. *Rendering Techniques* 2007, 9 (2007), 10.

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. 2019. Deep Reflectance Fields: High-Quality Facial Reflectance Field Inference from Color Gradient Illumination. *ACM Trans. Graph.* 38, 4, Article 77 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323027

Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. 2020. Deep Relightable Textures - Volumetric Performance Capture with Neural Rendering. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)* 39, 6. https://doi.org/10.1145/3414685.3417814

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. paGAN: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.

Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. 2020. The eyes have it: an integrated eye and face model for photorealistic facial animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020).

Mike Seymour, Chris Evans, and Kim Libreri. 2017. Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*. 1–2.

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5541–5550.

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

Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. 2020. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–12.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Trans. Graph.* 38, 4, Article 67 (July 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)* 24, 3 (2005).

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. 2006. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (TOG)* 25, 3 (2006), 1013–1024.

Lance Williams. 1978. Casting curved shadows on curved surfaces. In *Proceedings of the 5th annual conference on Computer graphics and interactive techniques*. 270–274.

Cyrus A. Wilson, Abhijeet Ghosh, Pieter Peers, Jen-Yuan Chiang, Jay Busch, and Paul Debevec. 2010. Temporal Upsampling of Performance Geometry using Photometric Alignment. *ACM Transactions on Graphics* 29, 2 (March 2010).

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

Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. 2018. Deep Incremental Learning for Efficient High-Fidelity Face Tracking. *ACM Trans. Graph.* 37, 6, Article 234 (Dec. 2018), 12 pages. https://doi.org/10.1145/3272127.3275101

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 126.

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. 2020. Neural Light Transport for Relighting and View Synthesis. *ACM Transactions on Graphics (TOG)* (2020).