Towards Practical Capture of High-Fidelity Relightable Avatars

Haotian Yang
Kuaishou Technology
China
yanghaotian03@kuaishou.com

Mingwu Zheng
Kuaishou Technology
China
zhengmingwu@kuaishou.com

Wanquan Feng
Kuaishou Technology
China
fengwanquan@kuaishou.com

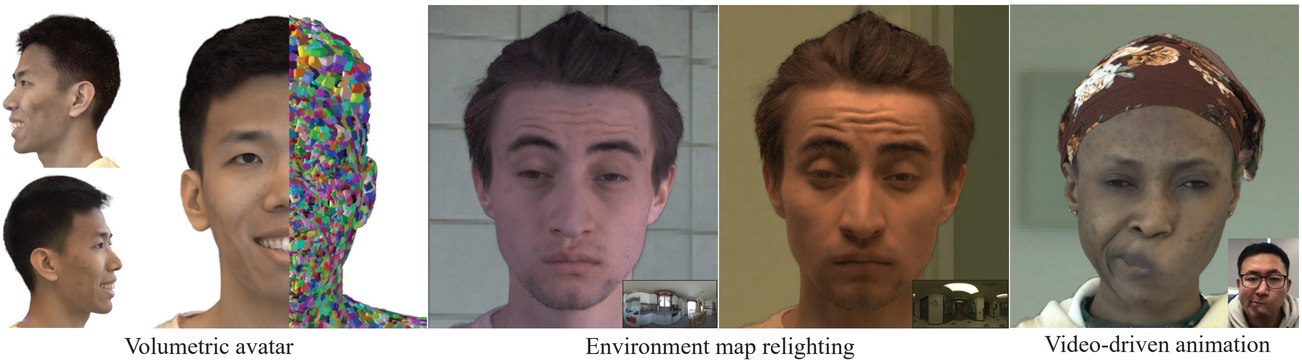
Haibin Huang
Kuaishou Technology
China
huanghaibin03@kuaishou.com

Yu-Kun Lai
Cardiff University
United Kingdom
laiy4@cardiff.ac.uk

Pengfei Wan
Kuaishou Technology
China
wanpengfei@kuaishou.com

Zhongyuan Wang
Kuaishou Technology
China
wangzhongyuan@kuaishou.com

Chongyang Ma*
Kuaishou Technology
China
chongyangma@kuaishou.com



Volumetric avatar Environment map relighting Video-driven animation

Figure 1: We present TRAvatar, a novel framework to capture and reconstruct high-fidelity volumetric avatars. Trained efficiently end-to-end on multi-view image sequences under varying illuminations, our virtual avatars can be relighted and animated in real-time of high fidelity.

ABSTRACT

In this paper, we propose a novel framework, Tracking-free Relightable Avatar (TRAvatar), for capturing and reconstructing high-fidelity 3D avatars. Compared to previous methods, TRAvatar works in a more practical and efficient setting. Specifically, TRAvatar is trained with dynamic image sequences captured in a Light Stage under varying lighting conditions, enabling realistic relighting and real-time animation for avatars in diverse scenes. Additionally, TRAvatar allows for tracking-free avatar capture and obviates the need for accurate surface tracking under varying illumination conditions. Our contributions are two-fold: First, we propose a novel network architecture that explicitly builds on and ensures the satisfaction of the linear nature of lighting. Trained on simple group light captures, TRAvatar can predict the appearance in real-time with a single forward pass, achieving high-quality relighting effects under illuminations of arbitrary environment maps. Second, we jointly optimize the facial geometry and relightable appearance from scratch based on image sequences, where the tracking is implicitly learned. This tracking-free approach brings robustness for establishing temporal correspondences between frames under different lighting conditions. Extensive qualitative and quantitative experiments demonstrate that our framework achieves superior performance for photorealistic avatar animation and relighting.

CCS CONCEPTS

• **Computing methodologies** → **Volumetric models**; **Motion capture**; **Reflectance modeling**.

* Corresponding author.

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0315-7/23/12…$15.00
https://doi.org/10.1145/3610548.3618138

KEYWORDS

Relighting; Facial animation; Neural rendering; View synthesis; Appearance acquisition.

ACM Reference Format:
Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. 2023. Towards Practical Capture of High-Fidelity Relightable Avatars. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3610548.3618138

1 INTRODUCTION

In this work, we focus on the capture and reconstruction of high-fidelity avatars in a Light Stage environment. As virtual representations of humans, avatars are crucial components in various downstream applications, such as video games, virtual reality, telepresence, and more [Bi et al. 2021; Guo et al. 2019; Lombardi et al. 2018; Moser et al. 2021; Schwartz et al. 2020].

Avatar creation has been a popular and challenging research topic in computer graphics and computer vision for decades. Despite considerable progress in this field, there are still many challenges to overcome, including expensive and sophisticated setup for avatar capture, lack of support for realistic relighting and animation, and high resource demands making training time-consuming and real-time deployment difficult to achieve. Traditional frameworks based on graphics pipeline, including geometry reconstruction [Beeler et al. 2010, 2011; Collet et al. 2015; Guo et al. 2019; Riviere et al. 2020; Wu et al. 2018] and physically-inspired reflectance capture [Debevec et al. 2000; Ghosh et al. 2011; Ma et al. 2007; Moser et al. 2021; Weyrich et al. 2006], are often difficult to set up and lack robustness, especially for dynamic subjects and non-facial parts. Recent deep learning based methods [Bi et al. 2021; Cao et al. 2022; Lombardi et al. 2018, 2021; Remelli et al. 2022] have demonstrated promising improvements for avatar representation by approximating the geometry and appearance with neural networks. However, most learning-based methods struggle to handle relighting effectively and have computationally expensive pre-processing and training steps that cannot meet the aforementioned requirements.

To this end, we propose a novel framework, Tracking-free Relightable Avatar (TRAvatar), that can circumvent the above obstacles, supporting efficient capture, high-quality reconstruction, as well as real-time animation and relighting (see Figure 1). Specifically, we improve the entire pipeline at its two primary stages, *i.e.*, both the data capture and avatar reconstruction.

For the data capture stage, we record a subject's performance under various expressions and lighting conditions. To faithfully reproduce the identity and detailed expressions of a specific subject, both dynamic geometry and reflectance should be captured. Considering the complexity of lighting conditions, it is non-trivial for the avatar network to directly learn the mapping from environment maps to the appearance. Furthermore, it is challenging to achieve satisfactory decoupling of lighting and other input conditions. To overcome this challenge, we take advantage of the prior knowledge of lighting, specifically its linear nature, to guide the network design. We design a network structure that explicitly exploits and guarantees to satisfy the linear nature of lighting, making it easy to train and enabling excellent generalization ability. Trained on dynamically captured image sequences in simple controllable group light illumination [Bi et al. 2021], our model can predict the appearance under arbitrary and complex lighting condition in a single forward pass, which facilitates real-time environment relighting. The learned disentangled representation also allows our data-driven avatar to be animated, relighted, and rendered under novel viewpoints.

For avatar reconstruction, we generate a 3D model from captured data that can be manipulated in real time. It is a challenging task to estimate temporal correspondences between captured frames with different lighting conditions. Previous learning-based methods typically rely on a pre-processing step to compute explicit tracked geometry (as a deformable base mesh), which is computationally expensive and not robust to varying light conditions. Therefore, we propose to jointly optimize the relightable appearance and latent geometry from scratch from image sequences, where the tracking is implicitly learned. Different from previous methods that separate mesh tracking and avatar creation in two stages, our tracking-free approach implicitly learns the dynamic deformation of the base mesh directly from the multi-view captured data, along with the relightable appearance in a joint optimization process. In addition to being much more efficient, this joint optimization allows our model to be directly trained on images in varying illumination, which is challenging for traditional explicit surface tracking.

Our experiments with TRAvatar show its effectiveness in creating high-quality and authentic avatars that can be animated and relighted in real-time with superior visual quality and computational efficiency compared to previous methods.

In summary, our contributions are:

- We present TRAvatar, a practical and efficient capture solution for creating high-fidelity avatars that can be animated and relighted in real time.
- We propose a novel network architecture that explicitly exploits the linear nature of lighting to improve generalizability, enabling real-time relighting with high realism for given environment maps.
- We propose to jointly optimize the relightable appearance and latent geometry of avatars from image sequences captured under varying lighting conditions, allowing more efficient and effective creation of relightable virtual avatars.
- We demonstrate that TRAvatar outperforms previous methods in terms of both visual quality and computational efficiency.

2 RELATED WORK

Creating a data-driven, relightable facial avatar of a specific subject typically involves capturing both dynamic geometry and reflectance. This is followed by constructing a parametric model from the captured data, or alternatively, employing image-based relighting techniques to synthesize the output. Below, we provide a concise overview of most relevant methods.

Geometry and reflectance acquisition. 3D face reconstruction and performance capture have been active research topics for decades. Accordingly, sophisticated 3D scanning systems have been developed for both static geometry reconstruction [Beeler et al. 2010; Ghosh et al. 2011] and dynamic performance capture [Beeler et al. 2011; Bradley et al. 2010; Collet et al. 2015; Dou et al. 2017a; Guo et al. 2019; Huang et al. 2011]. These methods utilize either multi-view stereo (MVS) or structured light for point cloud acquisition and then estimate the deforming geometry to achieve temporally consistent mesh tracking. The tracking process often involves time-consuming MVS reconstruction for thousands of frames and dense optical-flow optimization, while existing real-time face tracking algorithms cannot achieve satisfactory accuracy.

Besides, another crucial aspect of realistic relightable avatars is to estimate the way in which the light interacts with the subject, *i.e.*, the reflectance property. Previous methods usually assume physically-inspired reflectance functions modeled as bidirectional reflectance distribution function (BRDF) [Schlick 1994] and solve the parameters by observing the appearance under active or passive lighting. Active lighting methods typically require specialized setups with controllable illuminations and synchronized cameras. Debevec *et al.* [2000] pioneer in using a Light Stage for facial reflectance acquisition. One-light-at-a-time (OLAT) capture is performed to obtain the dense reflectance field. Later, polarized [Ghosh et al. 2011; Ma et al. 2007; Zhang et al. 2022] and color gradient illuminations [Fyffe and Debevec 2015; Guo et al. 2019] are used for rapid acquisition. Passive capture methods have significantly reduced the necessity for an expensive capture setup. For example, Riviere *et al.* [2020] and Zheng *et al.* [2023] propose to estimate physically-based facial textures via inverse rendering.

3D face modeling. Modeling of facial geometry and appearance has been a fundamental component of human related tasks in computer graphics and computer vision. The seminal work on 3D morphable models (3DMMs) [Blanz and Vetter 1999; Cao et al. 2013; Yang et al. 2020] employs Principal Component Analyze (PCA) to derive the shape basis from head scans. Despite its widespread use in various applications such as single-view face reconstruction and tracking [Dou et al. 2017b; Thies et al. 2016; Zhu et al. 2017], the shape space of 3DMMs is limited by its low-dimensional linear representation. Follow-up methods separate the parametric space dimensions [Jiang et al. 2019; Li et al. 2017; Vlasic et al. 2005] or use local deformation models [Wu et al. 2016] to enhance the representation power of the morphable model.

In recent years, deep learning based methods [Bagautdinov et al. 2018; Tran and Liu 2018, 2019; Zhang et al. 2022; Zheng et al. 2022] have been widely used to achieve impressive realism in face modeling. Lombardi *et al.* [2018] utilize a Variational Autoencoder (VAE) [Kingma and Welling 2013] to jointly model the mesh and dynamic texture, which is used for monocular [Yoon et al. 2019] and binocular [Cao et al. 2021] facial performance capture. Bi *et al.* [2021] propose to extend the VAE-based deep appearance model by capturing the dynamic performance under controllable group light illuminations to enable relighting.

While mesh-based methods typically require dense correspondence based on sophisticated surface tracking algorithms [Beeler et al. 2011; Wu et al. 2018] for training and degrade in non-facial regions, recent progress in neural volumetric rendering further enables photorealistic avatar creation. Lombardi *et al.* [2021] propose MVP (Mixture of Volumetric Primitives), a hybrid volumetric and primitive-based representation that produces high-fidelity rendering results with efficient runtime performance. More recently, Li *et al.* [2023] extend MVP with eyeglasses to be relightable following [Bi et al. 2021]. But it requires additional efforts for real-time relighting.

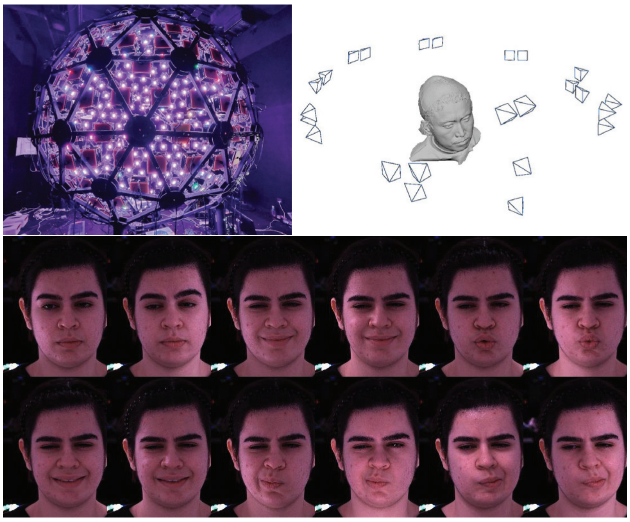


Figure 2: Illustration of our capture setup. Top left: Our customized capturing apparatus. Top right: The layout of 24 cameras. Bottom: Snapshots of captured frames from the frontal camera in a recording. Both the expression and the lighting condition change across different frames.

Some other methods have been proposed to create a facial avatar from monocular videos [Gao et al. 2022; Zielonka et al. 2023] or RGB-D input [Cao et al. 2022] without a specialized capturing apparatus. However, these approaches do not provide a relightable appearance, and their quality cannot match that of avatars built from industrial capture setups.

Image-based relighting. In contrast to model-based reflectance acquisition approaches, image-based relighting addresses the problem from an orthogonal perspective. By exploiting the linear nature of light transport, Debevec *et al.* [2000] propose to add up hundreds of images of densely sampled reflectance fields from OLAT capture to synthesize rendering results under novel lighting conditions. Subsequently, the number of sampled images is reduced by using specifically designed illumination patterns [Peers et al. 2009; Reddy et al. 2012] or employing sparse sampling [Fuchs et al. 2007; Wang et al. 2009]. Xu *et al.* [2018] propose to train a network for relighting a scene from only five input images. Meka *et al.* [2019] show that the full 4D reflectance field of human faces can be regressed from two images under color gradient light illumination. Sun *et al.* [2020] propose a learning-based method to achieve higher lighting resolution than the original Light Stage OLAT capture. Although these approaches achieve photorealistic rendering under novel lighting conditions, they only work from fixed viewpoints.

Meka *et al.* [2020] achieve relightable free viewpoint rendering of dynamic facial performance by extending Meka *et al.* [2019] with explicit 3D reconstruction and multi-view capture. However, they extract pixel-aligned features from captured raw images under

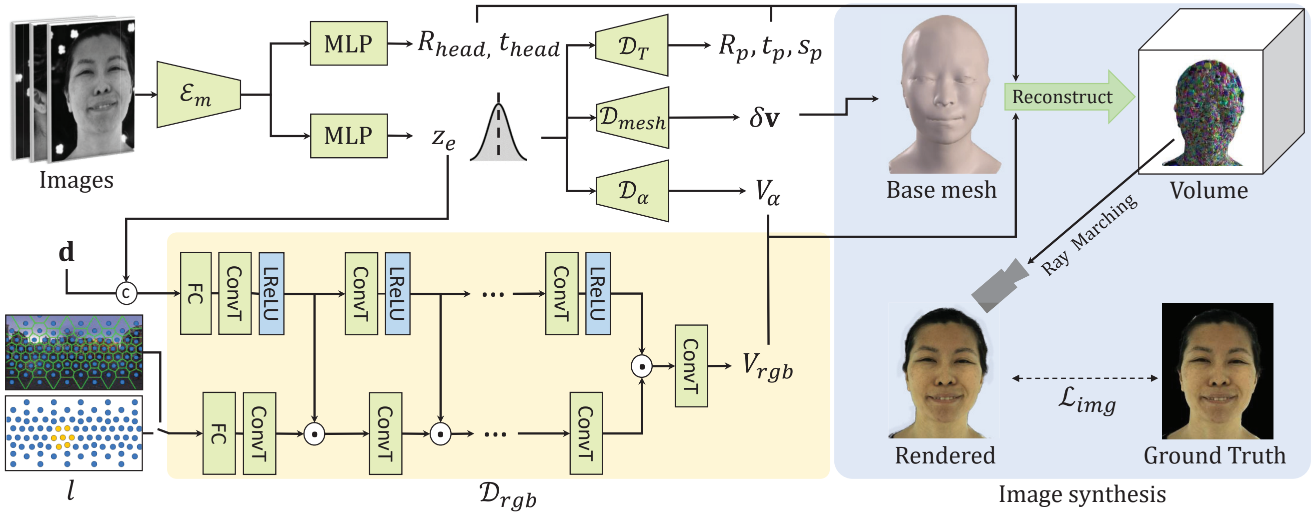


Figure 3: The pipeline of our framework. TRAvatar is a relightable volumetric avatar representation learned from multiview image sequences, including dynamic expressions and varying illuminations. For each frame, a motion encoder $\mathcal{E}_m$ forecasts the disentangled global rigid transformation $\{R_{head}, t_{head}\}$ and expression code z_e. With the given expression code, lighting condition l, and view direction d, a series of decoders subsequently predict the base mesh and the volumetric primitives mounted on it. Notably, a physically-inspired appearance decoder $\mathcal{D}_{rgb}$ (detailed in Section 4.2) is proposed to facilitate network training. Ultimately, the avatar representation is computed and then rendered, adaptable to any viewpoint and any lighting condition.

color gradient light illumination to build relightable textures, which limits its usage scenarios to performance replay. In contrast, our approach enables the creation of virtual avatars that not only allows for free viewpoint rendering with a relightable appearance but also possesses the capability of being controlled by an animation sequence of a different subject.

3 CAPTURING APPARATUS

To create an animatable and relightable avatar with ultra-high realism and specific identity, it is necessary to capture its performance under various expressions and lighting conditions. To this end, we have constructed an apparatus following the design principles of Light Stages [Debevec 2012; Guo et al. 2019]. Our customized capturing apparatus is shown in Figure 2.

Our Light Stage, installed on a spherical structure with a 3.6-meter diameter, comprises 356 lighting units and 24 machine vision cameras. We strategically place the cameras to capture the subject from multiple angles, and arrange the lighting units for precise control over illumination conditions. The Light Stage is placed in a dark room to prevent environment light interference.

Lighting units. The 356 lighting units are uniformly mounted on the sphere and are oriented towards the center. Each customized lighting unit comprises 132 high-brightness Light-Emitting Diodes (LEDs) that are controlled by a programmable embedded system. The LEDs are equipped with diffusers and lenses to ensure equal density illumination at the center.

There are five different types of LEDs on the lighting unit, namely red, green, blue, white 4500K, and white 6500K. The setup follows the latitude-longitude polarization as proposed in [Ghosh et al. 2011], and each type of LED is grouped into three categories with different polarization arrangements. The brightness of each group of lights can be adjusted independently using Pulse Width Modulation up to 100KHz. All the lighting units are connected to a central control unit and a computer via a CAN bus. The lighting pattern can be shuffled within 2ms, allowing us to capture the subject's performance under various lighting conditions quickly.

Cameras. Our apparatus includes 24 machine vision cameras installed around the sphere, with a focus on the center. The cameras consist of four 31M RGB cameras, 12 5M RGB cameras, and eight 12M monochrome cameras. The trigger ports of these cameras are linked to the central control unit, which synchronizes the cameras and lighting units to capture the subject's performance under various lighting conditions. We have disabled postprocessing features such as automatic gain adjustments in the cameras to ensure a linear response to the illuminance.

Depending on the camera types, we transmit the captured images to seven PCs via 10G Ethernet or USB ports. We calibrate the camera array with a 250mm calibration sphere similar to [Beeler et al. 2010] and undistort the images to ensure high-quality reconstruction. The mean reprojection error is less than 0.4 pixels, which facilitates high-quality creation of the target avatar.

4 METHOD

In this section, we formally introduce our novel framework, namely TRAvatar, which learns a disentangled representation for the target avatar to be animated, relighted, and rendered from novel viewpoints. As shown in Figure 3, our approach is based on a variational autoencoder (VAE) [Kingma and Welling 2013] architecture, where the latent space is designed to be disentangled with linear responses to varying lighting conditions , providing efficient and accurate modeling of dynamic geometry and reflectance fields.

We will first describe the details of our TRAvatar, including the training framework and network architecture (Section 4.1). The details of our specifically designed appearance decoder will be explained in Section 4.2. We will then describe how to use our Light Stage for data capture under various illuminations (Section 4.3). Finally, we will introduce the loss functions and regularization terms used for end-to-end network training (Section 4.4).

4.1 TRAvatar

Our volumetric avatar is built upon Mixed Volumetric Primitives (MVP) [Lombardi et al. 2021], which is a generalized hybrid representation using both a base mesh and volumetric primitives (see Figure 4). Each primitive is mounted to the base mesh and is represented as a volumetric grid with a resolution of M^3. We set $M = 8$ in our implementation.

Inspired by the success of image based relighting methods, our lighting condition is modeled as a vector $l \in \mathbb{R}_+{}^{356}$ representing the incoming light field of 356 densely sampled directions corresponding to the light positions of the Light Stage. We employ a VAE based architecture to train our relightable avatar. Different from previous methods [Bi et al. 2021; Remelli et al. 2022], we do not require tracked geometry in training. Note that the motion of a human head can be separated into global rigid motion and expression related motion. We utilize a motion encoder $\mathcal{E}_m$, to predict the disentangled motion. During training, for each frame, the convolutional motion encoder $\mathcal{E}_m$ takes a subset of the camera views as input and outputs the global head rotation $R_{head} \in SO(3)$ and translation $t_{head} \in \mathbb{R}^3$ as well as the mean $\mu \in \mathbb{R}^{256}$ and the standard deviation $\sigma \in \mathbb{R}_+{}^{256}$ of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. The expression code $z_e \in \mathbb{R}^{256}$ is sampled from this Gaussian distribution and represents expression related motion.

Taking the expression code z_e, the lighting condition l, and the view direction $\mathbf{d}$ as input, we use several decoders to predict the base mesh and volumetric primitives for output synthesis. Specifically, a mesh decoder $\mathcal{D}_{mesh} : \mathbb{R}^{256} \rightarrow \mathbb{R}^{3\times N_{mesh}}$, which is a multilayer perceptron, predicts the residual vertex positions $\delta\mathbf{v}$ based on the vertex positions $\hat{\mathbf{v}}$ of a template mesh with a fixed topology, where N_{mesh} is the number of mesh vertices. Then the resulting vertex position $\mathbf{v}$ of the base mesh is computed as $\mathbf{v} = R_{head}(\hat{\mathbf{v}}+\delta\mathbf{v})+t_{head}$.

Following [Lombardi et al. 2021], three decoders $\mathcal{D}_T$, $\mathcal{D}_\alpha$, and $\mathcal{D}_{rgb}$ with 2D convolutional architectures predict the volumetric primitives upon the base mesh. Specifically, the transformation decoder $\mathcal{D}_T : \mathbb{R}^{256} \rightarrow \mathbb{R}^{9\times N_{prim}}$ computes the rotation R_p, translation t_p, and scale s_p of N_{prim} primitives relative to the tangent space of the base mesh, which compensate for the motion that is not modeled by the mesh vertex $\mathbf{v}$. The opacity decoder $\mathcal{D}_\alpha : \mathbb{R}^{256} \rightarrow \mathbb{R}^{M^3\times N_{prim}}$ also takes the expression code z_e as input and decodes the voxel opacity V_α of the primitives. The appearance decoder $\mathcal{D}_{rgb} : \mathbb{R}^{256+356+3} \rightarrow \mathbb{R}^{3\times M^3\times N_{prim}}$ takes the expression code z_e, the lighting condition l, and the view direction $\mathbf{d}$ as input and predicts the RGB colors V_{rgb} of the primitives. The architecture of our relightable appearance decoder is designed to leverage the linear nature of lighting (see Section 4.2).

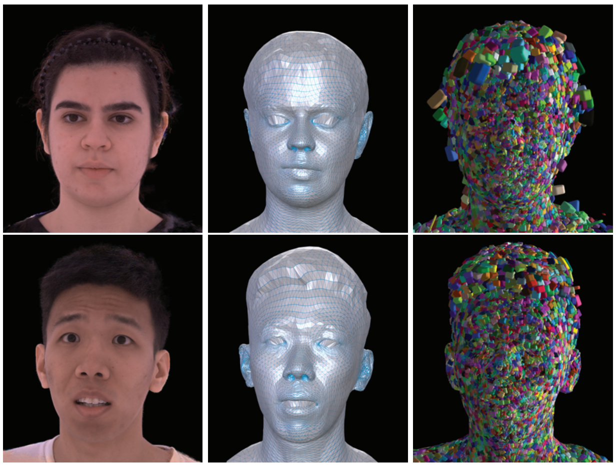


Figure 4: Illustration of our hybrid avatar representation. The base mesh and the volumetric primitives have consistent structures which provide flexible control such as video driven animation.

Output synthesis. Given the volumetric primitives, we use a differentiable accumulative ray marching algorithm [Karras and Aila 2013; Lombardi et al. 2021] to render the output images. Specifically, for a ray $\mathbf{r}_p(t) = \mathbf{o}_p + t\mathbf{d}_p$ with a direction $\mathbf{d}_p$ starting from a pixel p with a 3D position $\mathbf{o}_p$, we compute the pixel color $I_{rgb}(p)$ as:

$$I_{rgb}(p) = \int_{t_{\min}}^{t_{\max}} V_{rgb}(\mathbf{r}_p(t)) \frac{dT(p,t)}{dt}, \tag{1}$$

$$T(p,t) = \min\left(1, \int_{t_{\min}}^{t} V_\alpha(\mathbf{r}_p(t))\right), \tag{2}$$

where $t_{\min}$ and $t_{\max}$ are the predefined near and far bounds of the rendering range. The opacity of a pixel p is set as $I_\alpha(p) = T(p, t_{\max})$.

4.2 Relightable Appearance

In this section, we detail our specially designed appearance decoder $\mathcal{D}_{rgb}$ that enables high-fidelity real-time relighting using environment maps. Although the appearance changes drastically when lighting condition changes, previous methods [Basri and Jacobs 2003; Xu et al. 2018] have shown that the relighted images often lie in low-dimensional subspaces. For example, nearly all the lighting effects are linear [Chandrasekhar 2013; Debevec et al. 2000] and the full reflectance field can be predicted from a few images of the object in specific lighting conditions [Meka et al. 2019; Xu et al. 2018]. However, directly predicting all OLAT images and adding them up for environment map relighting is not feasible for real-time rendering. Our key observation is that we can design a network architecture upon the disentangled representation for our appearance decoder $\mathcal{D}_{rgb}$ to strictly satisfy the linear nature of lighting, *i.e.*:

$$\begin{aligned}\mathcal{D}_{rgb}(z_e, k_1 l_1 + k_2 l_2, \mathbf{d}) &= k_1 \mathcal{D}_{rgb}(z_e, l_1, \mathbf{d}) \\ &+ k_2 \mathcal{D}_{rgb}(z_e, l_2, \mathbf{d}), \forall k_1 \text{ and } k_2 \in \mathbb{R}.\end{aligned} \tag{3}$$

We show the architecture of $\mathcal{D}_{rgb}$ in Figure 3. Considering the spatially structured effect for each light, we use a convolutional architecture for $\mathcal{D}_{rgb}$. The expression code z_e and the view direction $\mathbf{d}$ are fed into an ordinary *non-linear* branch. The lighting condition l is injected in a separate *linear* branch, where the activation layers and the bias in the fully connected layer and transposed convolutional layers are removed. The feature maps of the linear branch $\mathcal{F}_{\text{lin}}$ is point-wise multiplied with the feature maps from the non-linear branch $\mathcal{F}_{\text{nlin}}$ at each stage:

$$\mathcal{F}_{\text{lin}}^{i+1} = ConvT\big(\mathcal{F}_{\text{lin}}^{i} \odot (\mathcal{F}_{\text{nlin}}^{i} + 1)\big), \tag{4}$$

where i is the index of the stage, $ConvT$ represents the transposed convolution operation, and $\odot$ is point-wise multiplication. The plus one term acts as a residual connection that stabilizes training (this term is omitted in Figure 3 to avoid clutter). In this way, the appearance decoder $\mathcal{D}_{rgb}$ is strictly linear to the lighting condition l while being non-linear to the expression code z_e and the view direction $\mathbf{d}$ that does not limit the representation power. We empirically find that our architecture significantly improves the generalization ability for novel lighting conditions (see Section 5.3 for some related evaluation results).

4.3 Data Acquisition

Capturing each transient facial expression under a variety of lighting conditions for relightable appearance poses a significant challenge. Instead, for each subject, we record image sequences of dynamic expressions with different lighting conditions in each frame and rely on our self-supervised training framework for disentanglement by using information across frames. Following [Bi et al. 2021; Li et al. 2023], we use group light patterns for capture, *i.e.*, for each frame seven randomly selected adjacent lights are turned to the maximum. Differently, since we do not use interleaved full-on frames for tracking, we find a large part of the face is dark in group light conditions that makes the implicit tracking in our network unstable. To provide basic illumination, we set all lights not included in the selected group to a known low brightness. Thanks to the linear nature of light and our network architecture design, the fully disentangled relightable appearance can be learned from such coalescent lighting conditions.

During the capture process, a subject is asked to perform 41 predefined expressions and read out two paragraphs. Then a freestyle performance is captured to cover extreme and complex expression combinations. We capture 10200 frames for each subject at 20fps. We show a snapshot of our captured images in Figure 2. The background without the subject is also captured.

4.4 Network Training

Our model is trained end-to-end on the multi-view image sequences under varying illuminations. The training loss $\mathcal{L}_{total}$ consists of two parts: $\mathcal{L}_{total} = \mathcal{L}_{img} + \mathcal{L}_{reg}$, where $\mathcal{L}_{img}$ is the data term and $\mathcal{L}_{reg}$ is the regularization term.

The data term $\mathcal{L}_{img}$ contains three components and measures the similarity between the captured input and the rendered output:

$$\mathcal{L}_{img} = \mathcal{L}_1 + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}, \tag{5}$$

where $\mathcal{L}_1$ is the MAE loss, $\mathcal{L}_{\text{VGG}}$ is the perceptual loss, and $\mathcal{L}_{\text{GAN}}$ is the adversarial loss that improves the visual quality. λ_{VGG} and λ_{GAN} are the balancing weights. We clip the pixel values of the rendered images I_{rgb} before calculating loss to simulate the truncation of the imaging process.

The regularization loss $\mathcal{L}_{reg}$ comprises four components:

$$\mathcal{L}_{reg} = \lambda_{\text{Lap}}\mathcal{L}_{\text{Lap}} + \lambda_{pR}\mathcal{L}_{pR} + \lambda_{vol}\mathcal{L}_{vol} + \lambda_{\text{KLD}}\mathcal{L}_{\text{KLD}}, \tag{6}$$

where $\mathcal{L}_{\text{Lap}} = ||\mathbf{L}(\mathbf{v} - \mathbf{v}_{base})||^2$ is the expression-aware Laplacian loss to encourage a smooth base mesh. $\mathbf{L}$ is the sparse Laplacian matrix. $\mathbf{v}_{base} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{v}$ is calculated in a least-squares manner based on the 51 predefined expression blendshapes $\mathbf{B} \in \mathbb{R}^{51\times 3N_{mesh}}$ from the FaceScape dataset [Yang et al. 2020]. $\mathcal{L}_{pR} = \frac{1}{N_{prim}}||(\mathcal{D}_T)_{R,t}||$ regularizes the predicted rotation and translation $(\mathcal{D}_T)_{R,t}$ to be small. We apply a predefined mask on the base mesh to assign higher weights of $\mathcal{L}_{\text{Lap}}$ and $\mathcal{L}_{pR}$ on facial regions compared to non-facial parts. $\mathcal{L}_{vol}$ and $\mathcal{L}_{\text{KLD}}$ are the volume minimization prior and KL-divergence loss as in [Lombardi et al. 2021], respectively. λ_{Lap}, λ_{pR}, λ_{vol}, and λ_{KLD} are balancing weights.

Since our training images are captured under varying illuminations, the background changes across frames. To prevent the encoding of background flashes into the avatar, the final image $\hat{I}$ in training is generated by blending the rendered foreground I_{rgb} with the captured background I_{BG} based on the computed opacity value I_α:

$$\hat{I} = I_\alpha I_{rgb} + (1 - I_\alpha)I_{BG}. \tag{7}$$

We use the Adam optimizer [Kingma and Ba 2015] to train the network with a learning rate of 10^{-4}. We choose frontal, left, and right views as input of the encoder. The input images are normalized and converted to grayscale to prevent the light from being encoded in the expression code z_e. We use the per-camera color calibration similar to [Lombardi et al. 2021]. For monochrome cameras, the rendered images are explicitly converted to grayscale before calculating loss functions. We fit a base mesh on the first frame for initialization.

The network training for each subject takes about two days on a single NVIDIA V100 graphics card. The decoding and rendering take around 22ms for a frame of a resolution 1280 × 960, enabling real-time relighting and animation. Please refer to our supplementary materials for implementation details such as network architectures and hyperparameters.

5 EXPERIMENTS

5.1 Qualitative Evaluation Results

Mesh-volume representation. Figure 4 shows two examples of our avatars based on the hybrid mesh-volume representation. Although our avatars are trained without explicit tracking, the base mesh and the volumetric primitives are roughly aligned. The inherently consistent structures enable explicit control and can be naturally used for applications such as video-driven animations and relighting.

Disentanglement of illumination and motion. Both illumination and motion are varied in our captured sequences. To evaluate the disentanglement of illumination and motion in our model, for each input frame, we keep the extracted expression code z_e fixed and change the lighting condition l extracted from environment maps to generate the relighting results. We use the appearance decoder to predict the relighted appearance of RGB channels separately for

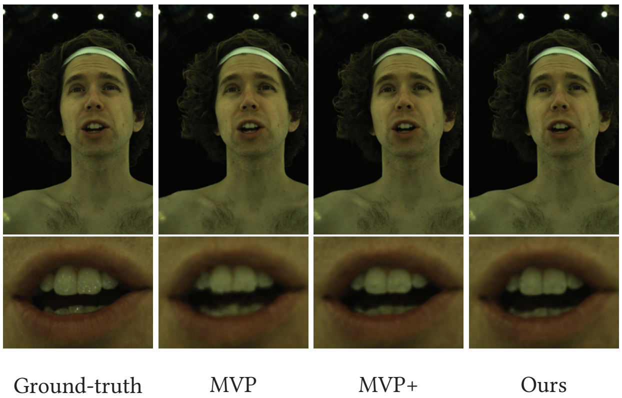


Figure 5: Comparison to MVP [Lombardi et al. 2021] on novel view synthesis. Our results are comparable to MVP and MVP+ (an improved version of MVP trained by ourselves) even without explicit tracking of the base mesh.

Table 1: Quantitative evaluation results of novel view synthesis in comparison with MVP [Lombardi et al. 2021]. The two subjects are from the Multiface Dataset [Wuu et al. 2022].

	Subject #002421669			Subject #5067077		
Method	MAE ↓	SSIM ↑	LPIPS ↓	MAE ↓	SSIM ↑	LPIPS ↓
MVP	2.08	0.910	0.273	2.21	0.923	0.232
MVP+	1.76	0.930	0.193	2.11	0.928	0.211
Ours	**1.73**	**0.932**	**0.186**	**2.01**	**0.934**	**0.208**

colorful environment map relighting. As shown in Figure 8, the lighting conditions are fully disentangled from the motion and are consistent across different subjects.

5.2 Comparisons to Prior Work

Comparison to MVP. Since existing explicit surface tracking methods [Beeler et al. 2011; Wu et al. 2018] do not generalize well under varying lighting conditions, we compare to MVP [Lombardi et al. 2021] on the publicly available Multiface Dataset [Wuu et al. 2022], which consists of high quality multi-view recordings of 13 different identities under fixed illumination. We perform qualitative and quantitative evaluations on eight held out views of two subjects. The vanilla MVP uses an L2 loss during training, which leads to blurry results. We train an improved version, namely MVP+, using the similar data term as ours for fair comparison. The other components remain identical to the vanilla MVP.

The visual comparison on Subject #002421669 from the dataset is shown in Figure 5. The Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) measurements are reported in Table 1. Both our method and MVP+ generate clearer details compared to vanilla MVP. Even without a computationally intensive tracking process, the quantitative reconstruction error of our method is slightly lower than that of MVP+. We attribute the improvement to the avoidance of information loss in the explicit surface tracking process.

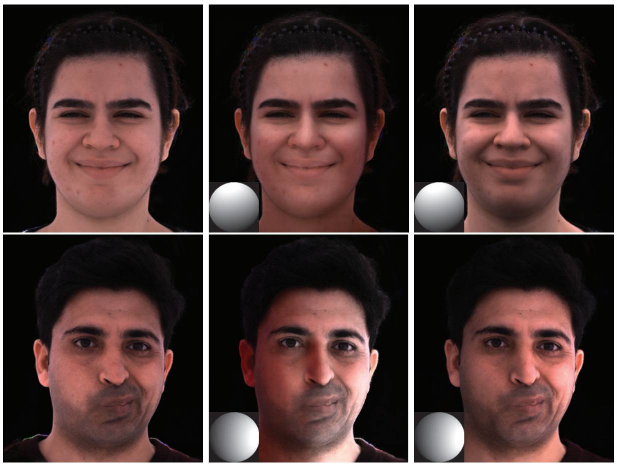


Figure 6: Comparison to DPR [Zhou et al. 2019] on single-view portrait relighting. The input illumination is shown as inset in each relighting result.

Table 2: Quantitative evaluation results of ablation study. In each column, the best number is highlighted in bold. Some corresponding visual results are shown in Figure 9.

	Subject A			Subject B		
Method	MAE ↓	SSIM ↑	LPIPS ↓	MAE ↓	SSIM ↑	LPIPS ↓
NL	10.47	0.665	0.417	13.87	0.601	0.440
NL + ENV	7.10	0.677	0.418	9.78	0.604	0.445
NL + LCL	9.74	0.661	0.428	12.19	0.601	0.449
NL + TS	8.03	0.672	0.401	9.77	0.597	0.423
Ours	**6.32**	**0.707**	**0.334**	**7.99**	**0.635**	**0.356**

Comparison to single-view portrait relighting methods. We compare our method to Deep Portrait Relighting (DPR) [Zhou et al. 2019] to evaluate the relighting results. The illumination is represented as the first three bands of Spherical Harmonics (SH) in DPR. We use their default SH coefficients and calculate the corresponding point light brightness for our model. We use a portrait in uniform illumination as the input of DPR.

As shown in Figure 6, DPR fails to predict correct relighting effects such as specularities and shadows consistent with the identity-specific geometry and skin material. As a result, the identity is shifted after relighting. In contrast, our method achieves more faithful portrait relighting results.

5.3 Ablation Study

We perform ablation studies to evaluate the effectiveness of our physically-inspired appearance decoder $\mathcal{D}_{rgb}$. Specifically, we compare our method to four alternative design options:

(1) NL: We remove the linear lighting branch of $\mathcal{D}_{rgb}$ and directly feed the concatenated lighting condition l and other latent codes to an ordinary non-linear network with the same layers as for appearance prediction.

(2) NL + ENV: We use the same network architecture as in (1) but use the Light Stage to simulate environment maps [Debevec et al. 2002] instead of group lights for training.
(3) NL + LCL: We adopt the same network architecture as in (1) and add a lighting consistency loss inspired by the recent single image portrait relighting method [Yeh et al. 2022] to enforce the linearity of lighting.
(4) NL + TS: We adopt the same network architecture as in (1) and use a two-stage training framework [Bi et al. 2021] for relighting. Specifically, we initially train an appearance decoder $\mathcal{D}_{rgb}$ for OLAT relighting, and subsequently use the trained network to synthesize data for training the environment map relighting appearance decoder.

We capture 600 frames for each subject under various preset lighting conditions in a Light Stage as ground truth for quantitative evaluation. Quantitative results are summarized in Table 2 and qualitative comparisons are shown in Figure 9. Note that not all the lighting conditions can be simulated in a Light Stage due to hardware limitations such as the maximum brightness of a lighting unit. The results demonstrate that our linear lighting branch of $\mathcal{D}_{rgb}$ significantly enhances the generalization performance for relighting.

5.4 Video-Driven Animation

Our volumetric avatar can be animated by replacing the motion encoder $\mathcal{E}_m$ with an application-specific module predicting the low-dimensional expression code z_e. Existing methods perform domain adaptation on synthetic datasets [Lombardi et al. 2018] or use triplet supervision [Zhang et al. 2022] to train the expression code predictor. In our implementation, we simply use an off-the-shelf expression regressor similar to [Weise et al. 2011] to predict the identity-independent blendshape weights of each frame from the frontal view in our captured data. Then we train a three-layer MLP to predict the expression code z_e from the blendshape weights. Our volumetric avatar can be animated by the extracted blendshape weights from monocular videos.

We find that the rigid head rotation and translation are successfully disentangled from the expression code even without explicit constraint. Thanks to the consistent structures of the base mesh and volumetric primitives, we can explicitly constrain the motion beyond face, achieving plausible animation results. Figure 7 shows some performance-driven animation results. Please refer to our accompanying video for the corresponding animations results.

6 CONCLUSION AND FUTURE WORK

In this work, we propose a novel framework, named TRAvatar, for capturing and reconstructing high-fidelity and relightable 3D avatars in a practical and efficient setting. We train the framework with dynamic image sequences captured in a Light Stage under varying lighting conditions, enabling natural relighting and video-driven animation.

Our contributions are two-fold. First, we present a novel network architecture that satisfies the linear nature of lighting, allowing for real-time appearance prediction and high-quality relighting effects. Second, we propose to jointly optimize facial geometry and relightable appearance based on image sequences, with the deformation of the base mesh implicitly learned. Our tracking-free scheme provides robustness for establishing temporal correspondences between frames under different lighting conditions. Both qualitative and quantitative experiments demonstrate that our framework achieves superior performance in photorealistic avatar animation and relighting, facilitating further advancements in content creation of 3D avatars.

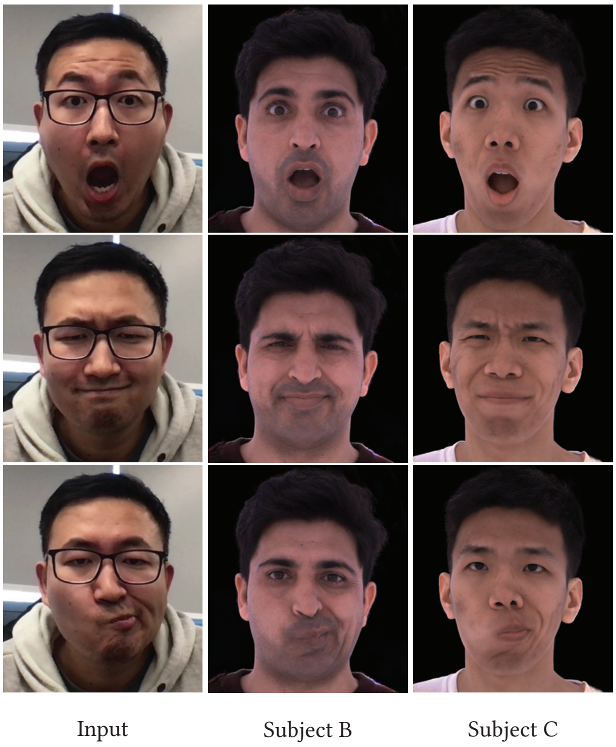


Figure 7: Video-driven animation results. Our method can faithfully generate identity-specific dynamic wrinkle details for different expressions.

Despite our promising results, there are some limitations to be addressed in future work. First, the data capturing apparatus employed in our framework is expensive, which may limit its applicability and adoption. Second, due to the lack of sufficient surface constraints, it becomes challenging to perform precise manual control on the learned implicit representation. Future work could explore methods to create relightable avatars with more affordable equipment and investigate representations that offer more flexible control. Finally, we are interested in extending our method to handle near-field and high-frequency relighting [Bi et al. 2021; Sun et al. 2020] as well as accessories such as glasses [Li et al. 2023].

ACKNOWLEDGMENTS

We would like to thank Qianfang Zou and Xuesong Niu for being our capture subjects, the authors of [Lombardi et al. 2021] and [Zhou et al. 2019] for releasing their source code, as well as anonymous reviewers for their valuable feedback.

REFERENCES

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. 2018. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3877–3886.

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

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*. 1–9.

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 papers*. 1–10.

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 Transactions on Graphics (TOG)* 40, 4 (2021), 1–15.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 papers*. 1–10.

Chen Cao, Vasu Agrawal, Fernando De La Torre, Lele Chen, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2021. Real-time 3D neural facial animation from binocular video. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–17.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. 2022. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.

Subrahmanyan Chandrasekhar. 2013. *Radiative transfer*. Courier Corporation.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–13.

Paul Debevec. 2012. *The light stages and their applications to photoreal digital actors*. Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

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*. 145–156.

Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. 2002. A lighting reproduction approach to live-action compositing. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 547–556.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017a. Motion2fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–16.

Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. 2017b. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5908–5917.

Martin Fuchs, Volker Blanz, Hendrik PA Lensch, and Hans-Peter Seidel. 2007. Adaptive sampling of reflectance fields. *ACM Transactions on Graphics (TOG)* 26, 2 (2007), 10–es.

Graham Fyffe and Paul Debevec. 2015. Single-shot reflectance measurement from polarized color gradient illumination. In *2015 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–10.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–12.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)* 30, 6 (2011), 1–10.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–19.

Haoda Huang, Jinxiang Chai, Xin Tong, and Hsiang-Tao Wu. 2011. Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition. In *ACM SIGGRAPH 2011 papers*. 1–10.

Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. 2019. Disentangled representation learning for 3D face shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11957–11966.

Tero Karras and Timo Aila. 2013. Fast parallel construction of high-quality bounding volume hierarchies. In *Proceedings of the 5th High-Performance Graphics Conference*. 89–99.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations, ICLR*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. 2023. MEGANE: Morphable Eyeglass and Avatar Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12769–12779.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 194:1–194:17.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

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

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. 2019. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. 2020. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–21.

Lucio Moser, Chinyu Chien, Mark Williams, Jose Serra, Darren Hendler, and Doug Roble. 2021. Semi-supervised video-driven facial animation transfer for production. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–18.

Pieter Peers, Dhruv K Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec. 2009. Compressive light transport sensing. *ACM Transactions on Graphics (TOG)* 28, 1 (2009), 1–18.

Dikpal Reddy, Ravi Ramamoorthi, and Brian Curless. 2012. Frequency-space decomposition and acquisition of light transport under spatially varying illumination. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 596–610.

Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. 2022. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-shot high-quality facial geometry and skin appearance capture. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 1–12.

Christophe Schlick. 1994. An inexpensive BRDF model for physically-based rendering. In *Computer graphics forum*. 233–246.

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), 91:1–91:15.

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.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2387–2395.

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3D face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7346–7355.

Luan Tran and Xiaoming Liu. 2019. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2019), 157–171.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 426–433.

Jiaping Wang, Yue Dong, Xin Tong, Zhouchen Lin, and Baining Guo. 2009. Kernel Nyström method for light transport. In *ACM SIGGRAPH 2009 papers*. 1–10.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM Transactions on Graphics (TOG)* 30, 4 (2011), 1–10.

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.

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.

Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. 2018. Deep incremental learning for efficient high-fidelity face tracking. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 234:1–234:12.

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. 2022. Multiface: A Dataset for Neural Face Rendering. *arXiv preprint arXiv:2207.11243* (2022).

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), 1–13.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. Facescape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 601–610.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–21.

Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. 2019. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4601–4609.

Longwen Zhang, Chuxiao Zeng, Qixuan Zhang, Hongyang Lin, Ruixiang Cao, Wei Yang, Lan Xu, and Jingyi Yu. 2022. Video-driven neural physically-based facial asset for production. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.

Mingwu Zheng, Zhang Haiyu, Hongyu Yang, and Di Huang. 2023. NeuFace: Realistic 3D Neural Face Rendering from Multi-view Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 16868–16877.

Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. 2022. ImFace: A Nonlinear 3D Morphable Face Model with Implicit Neural Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 20343–20352.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. 2019. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*. 7194–7202.

Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. 2017. Face alignment in full pose range: A 3D total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 78–92.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4574–4584.

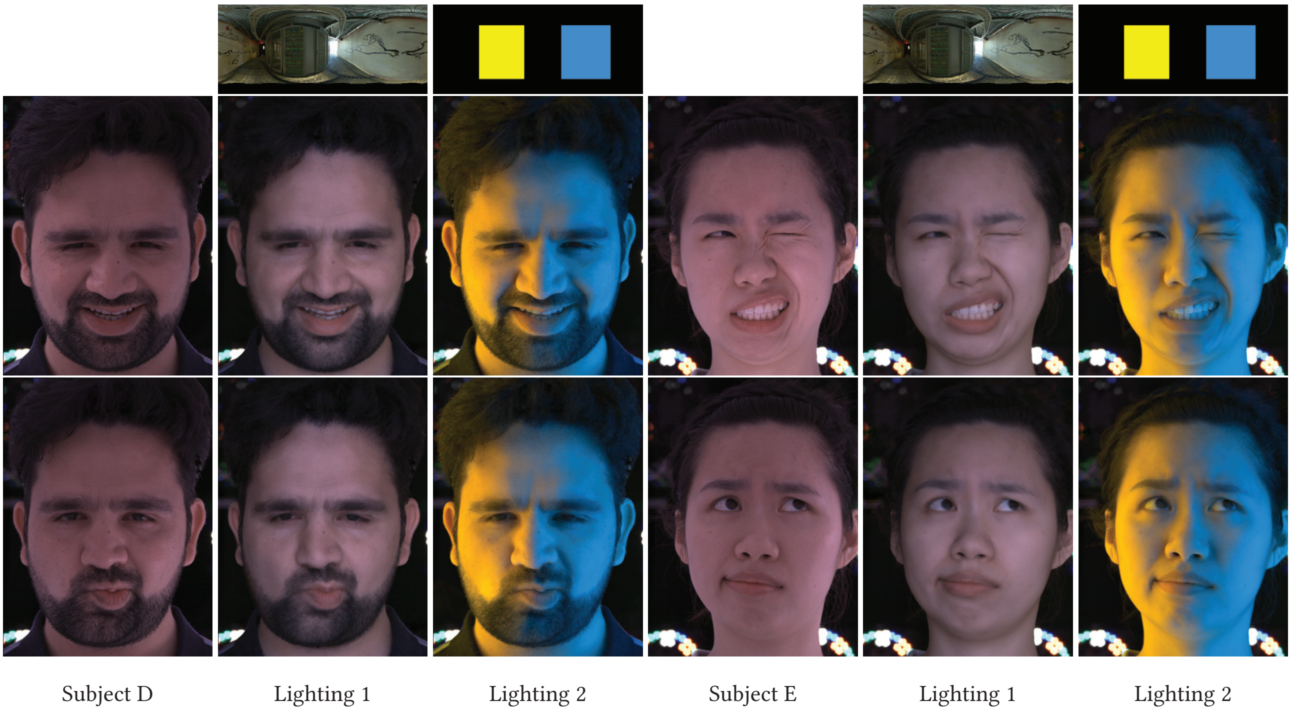


Figure 8: Evaluation results of lighting and motion disentanglement. For both subjects, we show the input frames of two different expressions on the left and the corresponding relighting results in the middle and on the right. The two input environment maps for relighting are shown on the top. The relighting effects are consistent with the dynamic expressions.

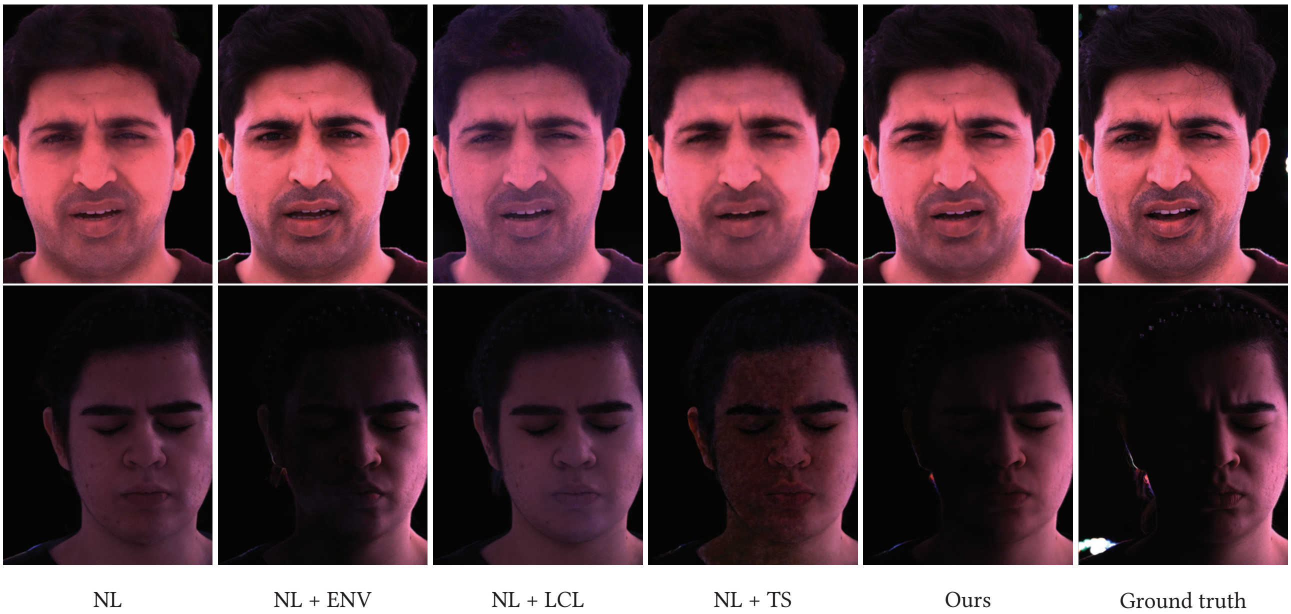


Figure 9: Ablation study results on Subjects A (top) and B (bottom) about our physically inspired linear light branch for the appearance decoder $\mathcal{D}_{rgb}$. From left to right: relighting results of four alternative baselines (see detailed explanations in Section 5.3), our results, and the ground truth. Note that here we use simulated environment map light which is similar to the lighting conditions that NL + ENV is trained on. Therefore, the results of NL + ENV are comparable to ours in this figure but downgrades significantly when using real HDR environments for testing (see more results in our supplementary materials).