Artist-Friendly Relightable and Animatable Neural Heads

Yingyan Xu[1,2] Prashanth Chandran[2] Sebastian Weiss[2]
Markus Gross[1,2] Gaspard Zoss[2] Derek Bradley[2]
[1]ETH Zürich [2]DisneyResearch|Studios
{yingyan.xu,grossm}@inf.ethz.ch
{prashanth.chandran,sebastian.weiss,gaspard.zoss,derek.bradley}@disneyresearch.com

Abstract

An increasingly common approach for creating photo-realistic digital avatars is through the use of volumetric neural fields. The original neural radiance field (NeRF) allowed for impressive novel view synthesis of static heads when trained on a set of multi-view images, and follow up methods showed that these neural representations can be extended to dynamic avatars. Recently, new variants also surpassed the usual drawback of baked-in illumination in neural representations, showing that static neural avatars can be relit in any environment. In this work we simultaneously tackle both the motion and illumination problem, proposing a new method for relightable and animatable neural heads. Our method builds on a proven dynamic avatar approach based on a mixture of volumetric primitives, combined with a recently-proposed lightweight hardware setup for relightable neural fields, and includes a novel architecture that allows relighting dynamic neural avatars performing unseen expressions in any environment, even with nearfield illumination and viewpoints.

1. Introduction

Creating realistic digital avatars of real people has many applications, for example in video games, films, VR experiences and telepresence. Original methods involved scanning and tracking geometry and appearance properties from one or more cameras and then applying a traditional graphics rendering pipeline to generate novel images. The challenge lies in the fact that avatars consist of several complex components like skin, eyes, teeth, and hair, each with complex material properties that are difficult to acquire and render realistically. As such, there has been a recent push towards *neural* representations of avatars, which forego triangles and texture maps and bypass traditional ray-tracers in exchange for neural rendering, where the avatar is represented by a neural network that can be queried at render-time, with equal handling of skin, eyes, teeth and hair in a single model.

The original Neural Radiance Field (NeRF) [29] representation laid the ground work for creating neural avatars that could be re-rendered photo-realistically from novel viewpoints. NeRFs are trained on large collections of images and represent a static scene as an MLP that is queried multiple times during volume rendering. Since the advent of NeRFs, several extended representations have emerged, which aim to increase the rendering speed [30], create NeRF avatars from smart phones [31], morph between neural avatars [47], or create generative neural heads [4].

An important component for the adoption of neural avatars is the ability to represent motion, which is not possible with the original neural field approaches. To account for this, researchers have devised extended neural representations like Neural Volumes [24], NeRFBlendShapes [9], NeRSemble [19] and a Mixture of Volumetric Primitives (MVP) [25], which aim to learn dynamic representations of human heads from video data.

A second important aspect of digital avatars is the ability to relight the head in any novel illumination. As with addressing the motion requirement, dedicated architectures have been proposed to address the relightability requirement, such as NeLFs [43], NRTFs [26] and ReNeRF [50], which can reliably relight neural representations of static scenes.

In order to have the most flexible and artist-friendly neural head avatars they should be both animatable *and* relightable. In this work we propose a new architecture for neural head avatars that can be relit to match any distant environment map or nearfield light sources, while at the same time providing full dynamic control over the facial shape. This allows both the playback of captured performances as well as the generation of novel artistically-created performances (e.g. driven by rig controls, retargeting or input video tracking), all with full control over the scene illumination and viewpoint.

Our approach is to build on a dynamic avatar representation that uses a Mixture of Volumetric Primitives (MVP) [25], which achieves efficient rendering of animatable neural heads with high visual quality. The idea of MVP is to decode the geometry and appearance of a person-specific facial

expression into a collection of geometric primitives containing color and opacity information, which can be sampled during traditional volumetric rendering. In our work, we propose to condition the appearance branch not just on the view direction but also on the scene lighting, in order to achieve controllable appearance changes based on different lighting conditions at run time. To achieve this we propose to train the model on captured dynamic performances under one-light-at-a-time (OLAT) illumination, which simplifies the lighting representation to a single light direction vector per frame. Importantly, we propose to compute per-primitive local light and view directions during the conditioning of the appearance branch, which allows us to represent both nearfield lights and distant environment map illumination, as well as nearfield viewpoints with varying focal length. We show that our method achieves high quality animatable and relightable neural avatars without the need for training data from a dense light stage, but instead using a less expensive sparse array of LED bars often used in photogrammetry setups. Once trained, the result is an artist-friendly neural representation of a complete dynamic head that can be controlled via traditional mesh deformation and scene illumination like in the familiar graphics pipeline.

2. Related Work

Realistic digital double creation requires accurate modeling of the face geometry, appearance and motion. In this section, we focus our discussion on image-based relighting, mesh-based representations and more recent work on neural volumetric avatars.

Image-based relighting techniques exploit the linearity of light transport to synthesize the scene under novel illumination conditions as a linear combination of a set of one-light-at-a-time (OLAT) images. Debevec *et al*. [6] were the first to use a light stage to acquire a dense reflectance field of a human face. Sun *et al*. [41] proposed a neural network trained on light stage OLAT data that takes as input a single portrait image and directly predicts a relit image given an environment map. Later work [42] has also explored ways to supersample the fixed basis used during capture to allow continuous high frequency relighting. These methods can only be applied to a static expression as the subject has to stay still during the capturing of OLAT images, and relighting is limited to the captured view point.

Compared to image-based techniques, mesh-based representations can handle novel views by design and offer explicit control of motion if temporally consistent tracking is provided [57]. To enable photorealistic rendering of digital humans, reflectance acquisition is also important in addition to geometry. Traditional facial appearance capture systems [11, 12, 27, 36] obtain parameters of predefined BRDFs as texture maps via inverse rendering, but are often limited to the skin regions. In recent years, mesh-based [16, 35, 44] or point-based [1] neural rendering approaches have gained popularity because they do not assume simplified reflectance models and can deal with imperfect geometry. Zhang *et al*. [53] proposed to model non-diffuse and global illumination as residuals added to a physically-based diffuse base rendering in texture space. But they have shown free viewpoint relighting of only a static expression. Meka *et al*. [28] used spherical gradient illumination to allow dynamic capture. However, their method can only be applied to performance playback due to the lack of correspondence between frames. Bi *et al*. [2] proposed a deep relightable appearance model (DRAM) as a VAE that takes as input a track mesh and an average texture and outputs the mesh vertices and view-dependent OLAT textures. All of these methods share the disadvantages of mesh-based representations, such as thin structures (*e.g*., hair), semi-transparent shiny materials (*e.g*., eyes), and large occlusions (*e.g*., teeth and tongues), which can be difficult to be tracked, reconstructed, or rendered.

More recently, there has been a rise of interest in volumetric representations since Neural Radiance Fields (NeRFs) [29, 39] were proposed. NeRFs represent the scene using a coordinate-based network that outputs color and density for each point observed from any view direction, trained on multi-view image input along with camera parameters. However, the original NeRFs are limited to static scenes under a fixed lighting condition. Therefore, motions or scene lighting cannot be modeled or controlled.

Followup work has enabled NeRFs to perform relighting [3, 26, 43, 45, 50, 52, 54]. ReNeRFs [50] take the idea of image-based relighting and extend NeRFs with an OLAT MLP and a spherical codebook to allow smooth lighting interpolation without a dense light stage. However, ReNeRFs can only handle static scenes.

Many methods have extended NeRFs to dynamic scenes. Some commonly used schemes are: (1) use a deformation field for motion and model appearance in a canonical space [22, 31, 33, 40, 46]; (2) learn a time-conditioned radiance field [21, 32]; (3) learn a radiance field for each time step often with spatial decomposition [7, 15, 24, 37, 38]. While being generic to model any dynamic scenes, these methods are restricted to performance playback, or only with limited motion manipulation capability without semantic control. Various methods [8, 10, 56] have also explored combining 3D morphable models, *e.g*., FLAME [20] with NeRFs, or include skeletal animations [13, 48]. However, these methods have only limited fidelity for re-animation and do not support relighting.

Different from explicit mesh representations or fully volumetric representations like NeRFs, the Mixture of Volumetric Primitives (MVP) representation [25] is a hybrid representation that inherits the strengths of both. MVP models the scene as a collection of geometric primitives with spatially varying color and opacity driven by a coarse guide

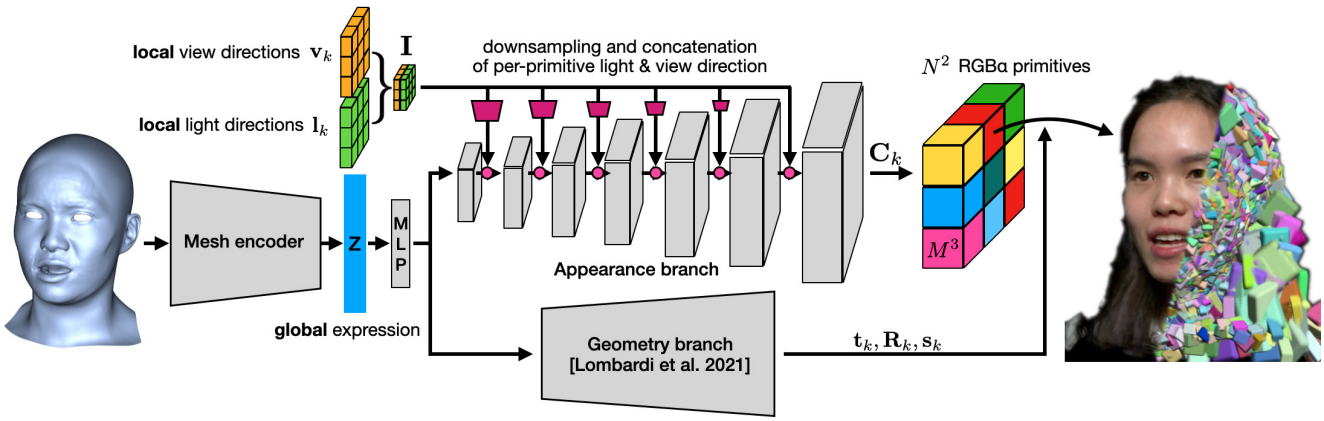

Figure 1. Overview of our pipeline, based on the Mixture of Volumetric Primitives (MVP) architecture [25]. A global expression code $\mathbf{z}$ obtained from a mesh encoder is fed into appearance and geometry decoder branches. The convolutional appearance branch predicts color and opacity of N^2 primitives with a grid resolution of M^3 each, which are placed in the scene as 3D primitives using the geometry branch (unchanged from Lombardi et al. [25]). To support relighting and especially near-field lighting, we augment the appearance branch and introduce per-primitive local view and light directions $\mathbf{v}_k, \mathbf{l}_k$ that are concatenated at every layer of the appearance branch network.

mesh. Follow-up work has extended MVP to articulated human bodies [34]. Iwase *et al*. [14] combined MVP and the student-teacher relighting framework in DRAM [2] and demonstrated the result on articulated hand models. Concurrent with our work, TRAvatar [51] extends MVP with a linear lighting branch designed to explicitly follow the linear nature of lighting. However, this method assumes a fixed basis, which often needs to be very dense for high fidelity relighting, and cannot interpolate/extrapolate novel lighting directions or model nearfield effects. Our work addresses all of these shortcomings and does not require an expensive light stage.

3. Relightable and Animatable Neural Heads

We now describe our method to create relightable and animatable neural heads. Our approach is to start with the baseline animatable head model of Lombardi et al. [25], which uses a Mixture of Volumetric Primitives (MVP) to describe a person-specific deformable neural head model, and extend the architecture and training data to allow for relighting. An overview of our method is given in Fig. 1, and the details are described in the following sub-sections. We first provide a brief overview of MVP for background information (Section 3.1), and then describe our extensions starting with the data requirements (Section 3.2), the main architecture (Section 3.3), and implementation details (Section 3.4).

3.1. Mixture of Volumetric Primitives

MVP is a state-of-the-art neural representation for human heads [25]. The key idea is that a collection of simple geometric primitives can be used collectively to render the complex geometry of a human face with high fidelity. The inputs to MVP include a person-specific latent expression vector $\mathbf{z}$, combined with the desired view vector $\mathbf{v}$ for rendering. The output is a collection of 3D primitives $\{\mathcal{V}_k\}$, which cover the occupied regions of the scene; each primitive containing volumetric RGBα information that can be used in traditional volumetric rendering. Formally,

$$\{\mathcal{V}_k\} = \text{MVP}(\mathbf{z}, \mathbf{v}), \tag{1}$$

where each of the N^2 primitives is defined by

$$\mathcal{V}_k = (\mathbf{t}_k, \mathbf{R}_k, \mathbf{s}_k, \mathbf{C}_k). \tag{2}$$

Here, the primitive geometry is defined by a translation $\mathbf{t}_k \in \mathbb{R}^3$, a rotation $\mathbf{R}_k \in \text{SO}(3)$, and a non-uniform scale $\mathbf{s}_k \in \mathbb{R}^3$. The appearance is defined by a dense voxel grid of color information $\mathbf{C}_k \in \mathbb{R}^{4 \times M_x \times M_y \times M_z}$, which stores the RGB$\alpha$ value per voxel (in our implementation, $N^2 = 16384$ and $M_x = M_y = M_z = M = 8$).

The MVP decoder consists of a geometry branch that depends only on $\mathbf{z}$ and an appearance branch that depends on both $\mathbf{z}$ and $\mathbf{v}$. The primitives are organized in a 2D grid of size $N \times N$ and are associated to positions on a guide mesh through a UV parametrization. The guide mesh is predicted by the geometry branch as a means to initialize the per-primitive transformations, which are further refined in the geometry branch. The convolutional appearance branch predicts $\mathbf{C}_k$ directly in the UV-space of the mesh. We refer to the original formulation [25] for more details. In our work, we use the MVP geometry branch directly, but we extend the appearance branch to provide the ability to relight the dynamic neural head, as described in Section 3.3.

3.2. Data Acquisition and Preprocessing

Before describing our architecture, we first describe important differences in the training data as compared to the

original MVP formulation, which was trained on multi-view video sequences of a performing actor, observed by ≈100 different cameras under constant full-on illumination.

Image Data. Our model is also actor-specific, requiring multi-view video data of a performing subject. However, as we aim to relight the performances, we require a diverse set of lighting conditions in the dataset. So instead of constant illumination we capture our subject under a time-varying light pattern consisting of both one-light-at-a-time (OLAT) frames and full-on illumination frames. In contrast to OLAT acquisition methods that require a dense light stage [6, 41, 42, 55], we use the camera and light setup of ReNeRF [50], which showed impressive relighting of *static* neural heads with far less expensive hardware. The setup consists of only 10 cameras and 32 individually-controllable LED light bars placed in the frontal hemisphere of the capture volume with calibrated 3D positions relative to the cameras [49]. We propose to capture dynamic performances while flashing a dedicated lighting sequence consisting of one illuminated light bar per frame (OLAT frames). We also intermix a full-on shot with all light bars illuminated once every 3 frames to help with mesh tracking (as described below). So the lighting sequence over time is $F, O_1, O_2, F, O_3, O_4, F, ..., F, O_{31}, O_{32}$, where F corresponds to a full-on frame and O_i corresponds to the i-th OLAT bar. The pattern is repeated indefinitely at 24 frames per second. Importantly, we choose an OLAT ordering such that neighboring OLAT frames are as dissimilar as possible (e.g. an OLAT from the left or top followed by an OLAT from the right or bottom, respectively). The motivation is that neighboring frames contain very similar facial expressions and so we maximize data efficiency if lighting conditions are as different as possible from one frame to the next. We do not impose any particular facial expression sequence, but in practice we obtained good results by capturing two different facial muscle workouts followed by three scripted lines of dialog and one free dialog performance. In total, we capture on average about 2700 images per camera for a single subject. A short clip from one subject is illustrated in Fig. 2, which shows a partial sequence of OLAT and full-on images from one frontal camera.

Mesh Tracking. In addition to the image data, we require a representation of the per-frame facial expressions for training. To this end, we apply the common practice of precomputing a tracked 3D mesh sequence corresponding to our input imagery. We employ a recent landmark-based 3D face tracking method [5], which optimizes for the parameters of an actor-specific local blendshape model to match detected 2D landmarks in all camera views. We build the local blendshape model from a small set of 3D face scans in a pre-process. Tracking is performed only on the full-on illumination frames, and then the model parameters are interpolated linearly across the OLAT frames. The parameters include both the expression blendweights and the head pose. As we wish to train our network in a stabilized space with respect to the skull position, we use only the expression parameters to construct the face meshes and we use the small per-frame head pose transformations to inversely offset the per-frame camera positions. The tracked mesh geometry corresponding to a small input sequence is shown alongside the input images in Fig. 2.

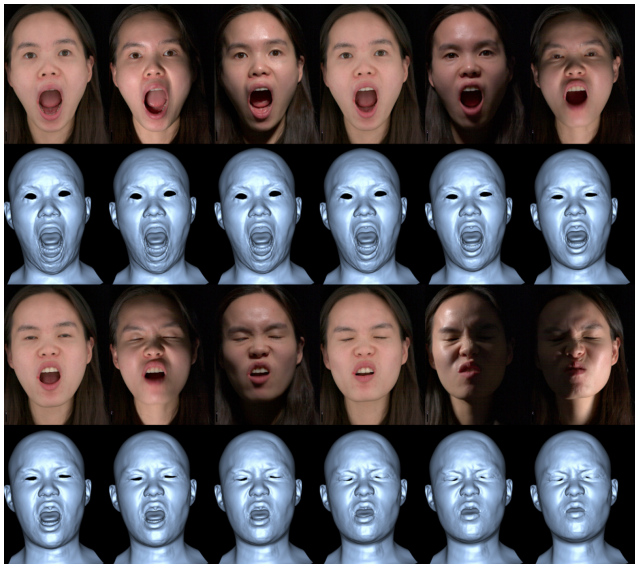

Figure 2. An overview of our training data, which includes dynamic performances illuminated by interleaved OLAT and full-on lighting conditions. We also obtain per-frame 3D geometry for the face as a representation of the expression.

To summarize the dataset, the ultimate training data is approximately 1800 frames per capture subject on average (after removing the full-on frames), each frame containing:

- 10 multi-view images from calibrated camera positions,
- one 3D light position corresponding to the center of the OLAT bar illuminating that frame,
- and a tracked 3D face mesh.

3.3. Relightable MVP

Our new architecture for creating animatable and relightable neural heads can be considered as a relightable version of MVP [25]. As shown in Fig. 1, there are three main trainable components (shown in gray): a mesh encoder network to project the input facial expression mesh to a latent global expression parameter $\mathbf{z}$, and parallel geometry and appearance branches similar to MVP as described above. A small 1-layer MLP designed to prepare $\mathbf{z}$ for the downstream branches is also learned. It maps the 256-dimensional vector $\mathbf{z}$ to a 16384-dimensional vector that is then reshaped to an 8×8 feature map with 256 channels and sent to the convolutional geometry and appearance branches. For the geometry branch

we use exactly the MVP architecture, but we make important changes to the appearance branch to support relighting. The mesh encoder and the illumination-modulated appearance branches are described in the following.

Mesh Encoder. Similar to MVP, we require a latent expression vector $\mathbf{z}$ to drive our neural head decoder. When tracked geometry is available, the original MVP formulation used the encoder architecture of a Deep Appearance Model [23], which takes the tracked geometry and a color texture as input. In contrast to MVP, however, we do not bake appearance information into $\mathbf{z}$ and instead drive our model purely from expressions, in the form of tracked geometry alone. This will allow us to artistically control the neural head at inference time with novel unseen expressions. We therefore construct our mesh encoder from a modified Deep Appearance Model encoder, specifically omitting the texture branch. Our encoder is trained end-to-end along with the geometry and appearance decoders.

Illumination-Modulated Appearance. The most significant contribution of our work is the proposed illumination-modulated appearance branch. Rather than conditioning the appearance only on the view vector as in MVP, we additionally consider the per-frame OLAT lighting condition during training. Notably, we compute per-primitive *local* light directions by first evaluating the geometry branch to get the world-space transformations for each primitive, and then computing the grid of local light directions $\mathbf{l}_k$ as the difference between the 3D OLAT light position $\mathbf{p}_{olat}$ and the center of each transformed primitive. Specifically,

$$\mathbf{l}_k = \mathbf{p}_{olat} - \mathbf{R}_k \cdot \mathbf{t}_k. \tag{3}$$

Conditioning the appearance branch on per-primitive light directions rather than a single global *distant* light allows our model to support relighting with nearfield illumination, as we demonstrate in Section 4. Analogously, the camera view for rendering is also at a discrete 3D location in the scene, which we denote as $\mathbf{p}_{cam}$, and so we can similarly compute per-primitive *local* view directions $\mathbf{v}_k$ rather than a single global view vector $\mathbf{v}$ for conditioning the appearance branch. Specifically,

$$\mathbf{v}_k = \mathbf{p}_{cam} - \mathbf{R}_k \cdot \mathbf{t}_k. \tag{4}$$

Just as local light directions enable rendering with nearfield illumination, local view directions enable rendering with nearfield camera views, which we demonstrate by synthesizing a dolly-zoom effect in Section 4. Both nearfield illumination and nearfield viewpoints are not possible with the original MVP architecture.

Our proposed appearance branch generates the voxel color grids as

$$\{\mathbf{C}_k\} = \text{RelMVP}(\mathbf{z}, \{\mathbf{v}_k\}, \{\mathbf{l}_k\}), \tag{5}$$

where we denote RelMVP() as our relightable version of the MVP appearance branch.

The relightable appearance branch is implemented as a convolutional architecture, which takes the reshaped expression vector as input and gradually produces the primitives' RGBα tiles in UV space, modulated by the local view and light directions. The local view and light directions per primitive are combined and stored as a single 6-channel image in UV space at the full network output resolution (view and light vectors are copied for every voxel within a primitive). We denote $\mathbf{I} \in \mathbb{R}^{6\times(N\cdot M)\times(N\cdot M)}$ as the concatenated set of $\{\mathbf{v}_k\}$ and $\{\mathbf{l}_k\}$. At each convolutional level, $\mathbf{I}$ is bilinearly downsampled and concatenated to the intermediate feature layers before proceeding to the next layer. This downsampling operation has the effect of averaging local view and light directions across neighboring primitives, which is acceptable since neighboring primitives are located close to each other in 3D space. At early layers, the averaged view and light directions resemble global view and light vectors, but then at deeper layers the per-primitive view and light directions can specialize the appearance of each primitive individually, allowing us to achieve nearfield lighting and viewpoints. Note that the local view and light conditioning is only applied to the RGB component of the output, as opacity α is independent of illumination and view direction.

The result is a set of primitive volumes $\{\mathbf{C}_k\}$ that are transformed by the output of the geometry branch and rendered with a differentiable raytracer [25].

3.4. Implementation Details

We employ a fully-convolutional network for the appearance branch RelMVP(). The input is the local light and view vectors, reshaped to a feature map $\mathbf{I} \in \mathbb{R}^{6\times(N\cdot M)\times(N\cdot M)}$, as well as the expression code $\mathbf{z}$ transformed and reshaped to a feature map $\mathbf{z}' \in \mathbb{R}^{256\times8\times8}$ (channels $\times$ height $\times$ width). The appearance branch then consists of seven transpose-convolution layers with a kernel size of 4, stride 2 and padding 1, which increase the feature map resolution from 8×8 by a factor of two at every step until the final resolution of 1024×1024 (i.e. $N \cdot M = 1024$) is reached. The inputs to the convolutional layers are the previous feature maps, starting with $\mathbf{z}'$ and the six channels from $\mathbf{I}$ bilinearly downsampled to match the current spatial resolution. The output features have channels $256, 128, 128, 64, 64, 32, 48$ where the final 48 channels are interpreted as $rgb \times M_z$.

Since opacity α does not depend on the local view or light direction, we follow the pracitce of the original MVP architecture [25] and estimate opacity in a separate branch. This branch is identical to the above architecture, with the difference of predicting M_z output channels and not depending on $\mathbf{I}$. We apply ReLU activation on the RGB output and all intermediate layers are followed by LeakyReLU activations.

Our neural rendering pipeline is trained end-to-end on multi-view OLAT sequences, see Section 3.2. We drop the background model that estimates foreground and background objects in the original MVP architecture since our data is recorded in front of a pure black backdrop. Instead we add a matting loss to the MVP loss functions that compares a target matting mask $\mathcal{M}$ to the accumulated density per ray $\tilde{\alpha}(\Theta)$, which avoids floating primitives at the background,

$$\mathcal{L}_{\text{mat}} := \text{MAE}(\mathcal{M}, \tilde{\alpha}(\Theta)). \tag{6}$$

We extract the matting masks $\mathcal{M}$ from the input images using MODNet [17]. The final loss function is then

$$\begin{aligned}\mathcal{L} :=& \lambda_{\text{pho}}\mathcal{L}_{\text{pho}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{vol}}\mathcal{L}_{\text{vol}} \\ &+ \lambda_{\text{kld}}\mathcal{L}_{\text{kld}} + \lambda_{\text{mat}}\mathcal{L}_{\text{mat}}\end{aligned} \tag{7}$$

where all loss terms other than $\mathcal{L}_{\text{mat}}$ are the same as in the MVP implementation, and the weights are defined as

$$\begin{aligned}&\lambda_{\text{pho}} = 1.0,\ \lambda_{\text{geo}} = 10.0,\ \lambda_{\text{vol}} = 0.01, \\ &\lambda_{\text{kld}} = 0.001,\ \lambda_{\text{mat}} = 0.1.\end{aligned} \tag{8}$$

We employ ADAM [18] as the optimizer with a learning rate of $lr = 0.0001$. For training and evaluation, we downsample the input images to a resolution of 1024×768 to reduce the training time. Each subject is trained for 200,000 iterations with a batch size of 12, taking around two days to train on an A6000 GPU.

4. Experiments

In this section we perform several experiments to validate our method. We start with a number of qualitative results, including relighting our dynamic neural heads with novel viewpoints, expressions and illumination conditions (Section 4.1). We then perform a quantitative evaluation to show the performance of our model on held-out validation data, including a comparison to related work (Section 4.2). For all results, we recommend to also view the accompanying supplemental video, in order to see the animations in motion.

4.1. Qualitative Results

We begin by showing the ability of our model to re-render our captured neural heads from novel viewpoints with artist-controllable lighting and expressions. Fig. 3 depicts one of our subjects in the studio capture environment, but relit with novel point lights (top row) and novel interpolated expressions (bottom row). All renders are generated from unseen viewpoints outside of the training views.

In Fig. 4 we highlight the application of re-rendering dynamic performances under arbitrary environment map illumination. Here we also show the performances under a turntable of novel viewpoints. Even though our training data consisted only of OLAT lighting frames, we can sample multiple individual light directions from environment maps and combine the rendered results into final high quality and consistent renders.

Figure 3. Novel view, light, and expressions: Our method allows for artistically-generated renders under novel point lights (top) and novel expressions (bottom), all rendered from novel viewpoints.

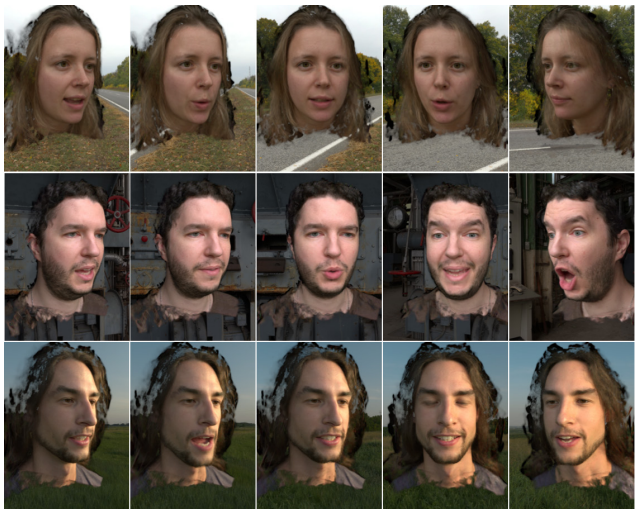

Figure 4. Novel environment maps: Since our method generalizes to novel light directions, we can densely sample light directions over the hemisphere to render performances under any environment map. Three examples are shown, rendered from novel viewpoints.

We further push the capabilities of our model by rendering several performances under a number of challenging light conditions in Fig. 5. Here we use six different environment maps from both indoor and outdoor scenes, during daytime and night, and we render the dynamic neural heads of 3 subjects from a fixed camera view while rotating the environment maps around the subjects. The two rows belonging to the same subject show the same expression but rendered with different lighting conditions, highlighting the versatility of our method for relighting animated performances.

An important aspect of our architecture is that the light and view directions are computed locally per primitive, allowing nearfield lighting and viewpoint effects. We demonstrate nearfield lighting in Fig. 6, which shows a static ex-

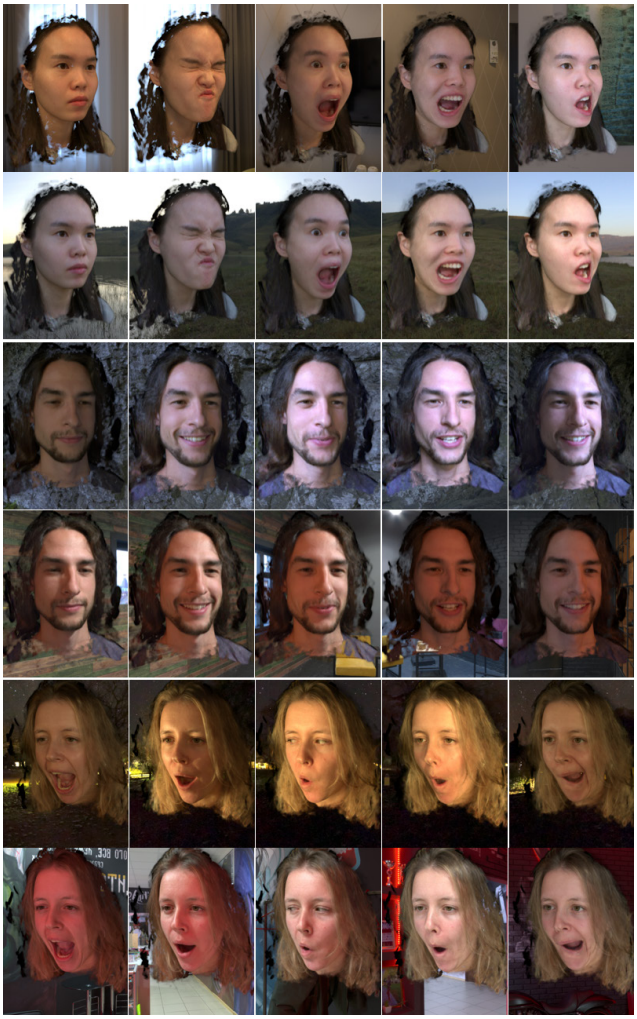

Figure 5. Here we demonstrate the quality our method achieves on further subjects: we can visualize performances under novel views and arbitrary, temporally rotated environment maps that produce complex lighting conditions.

pression of 2 captured subjects relit by a moving point light source (and varying unseen viewpoints). In the top row of each subject the point light is farther away from the subject than the corresponding frame in the bottom row, where the light is near to the face. Our method supports natural relighting under these conditions. In a similar spirit, having per-primitive local view directions allows us to synthesize complex camera motions like a dolly-zoom effect, where the camera pushes in close on a subject's face while decreasing the focal length of the lens (i.e. increasing the FOV). We demonstrate this effect in Fig. 7, which shows a realistic simulation of this commonly-used practical camera move.

The presented results show a non-exhaustive sample set of applications that are enabled by our artist-friendly method for relightable and animatable neural heads.

Figure 6. Near-field lighting: By specifying a per-primitive light direction in the appearance branch, we can render both far-field lights, as well as near-field lights.

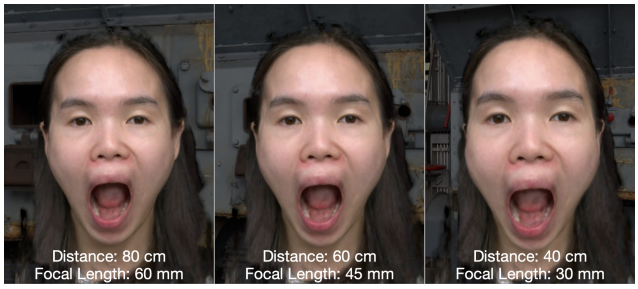

Figure 7. Near-field view: By conditioning the appearance branch also on per-primitive view directions, we can change the focal length of the camera and realize effects like a dolly zoom.

4.2. Quantitative Evaluation

To evaluate our method quantitatively we trained a version with some held-out data for validation. Specifically, we constructed three validation sets: one with held-out OLAT directions, one with held-out performances, and one combined one that contains both held-out light directions and performances. Our validation data consists of a variety of frames from three different captured subjects.

As we evaluate our model's performance, we simultaneously perform a comparison to related work. Unfortunately there are very few existing methods for both relightable and animatable neural avatars, in particular with code available for testing. To conduct a comparison with a baseline method,

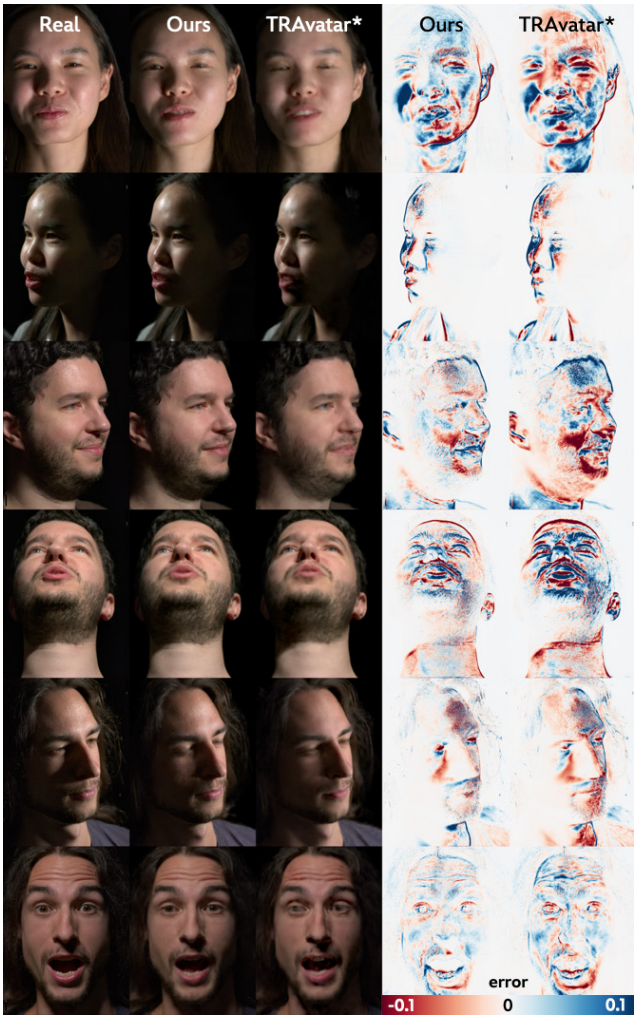

Figure 8. Example frames of the quantitative comparison against TRAvatar* on held-out data, see Section 4.2. The comparisons were conducted on three subjects, called S1 to S3 from top to bottom. The heatmaps to the right depict the per-pixel errors of our method and TRAvatar* against the ground truth images.

we re-implemented the lighting branch of TRAvatar [51], denoted by TRAvatar* below.

Table 1 shows the comparisons between our method and TRAvatar* on the held-out data. Example frames from the comparison with a heatmap visualizing the per-pixel errors are shown in Fig. 8. Note that given a novel directional light, TRAvatar can only evaluate it using barycentric coordinates within their fixed basis used in training. Our method clearly outperforms TRAvatar* in terms of both qualitative render quality and quantitative analysis in all cases. Furthermore, as mentioned earlier, TRAvatar* is unable to interpolate novel lighting directions or model nearfield effects, which our method can easily achieve.

Table 1. Quantitative evaluation of our method against TRAvatar*. Results are shown for 3 subjects (see Fig. 8), evaluated for novel light directions, novel performances, and both novel light directions and performances. Our method consistently outperforms TRAvatar.

			PSNR ↑	MAE ↓	SSIM ↓	LPIPS ↓
novel light	S1	ours	**32.19**	**2.88**	**0.899**	**0.278**
		TRAvatar*	29.68	4.03	0.870	0.326
	S2	ours	**32.20**	**3.11**	**0.864**	**0.296**
		TRAvatar*	30.13	4.06	0.832	0.355
	S3	ours	**33.73**	**2.65**	**0.873**	**0.343**
		TRAvatar*	32.37	3.24	0.854	0.390
novel perf.	S1	ours	**39.75**	**3.70**	**0.881**	**0.282**
		TRAvatar*	28.86	4.12	0.869	0.323
	S2	ours	**29.52**	**4.09**	**0.835**	**0.300**
		TRAvatar*	28.88	4.46	0.829	0.350
	S3	ours	**31.28**	**3.27**	**0.863**	**0.338**
		TRAvatar*	31.21	3.36	0.861	0.373
light & perf.	S1	ours	**28.68**	**4.30**	**0.865**	**0.300**
		TRAvatar*	27.30	5.27	0.844	0.342
	S2	ours	**29.11**	**4.34**	**0.826**	**0.316**
		TRAvatar*	28.13	4.94	0.812	0.366
	S3	ours	**30.76**	**3.55**	**0.855**	**0.354**
		TRAvatar*	30.65	3.80	0.848	0.392

5. Conclusion

We present a novel architecture for relightable, animatable neural avatars, building on top of the Mixture of Volumetric Primitives architecture. Our architecture extends the appearance branch with per-primitive light and view directions, allowing for nearfield lighting and viewpoint effects. The networks are trained end-to-end on OLAT sequences obtained by flashing 32 LED bars during a dynamic performance of a subject, captured with inexpensive hardware. Our model is capable of novel-view and novel-light estimation and also generalizes well to novel expressions and performances.

We note a few practical limitations of our method. So far, we have focused primarily on the frontal part of the head and do not recover a complete 360-degree neural avatar. This can result in artifacts at the boundary (*e.g.*, the hair and neck regions). That said, we do not believe that our model is fundamentally restricted here, and our results are only limited by the physical capture setup we used. Further, during very fast motions, if motion-blur were to occur in the training data then this will lead to blurry reconstructions. We also found that extrapolation to extreme facial expressions outside our training data was not possible, but we note that our results were generated from an order of magnitude fewer training frames than the original MVP algorithm, and therefore simply adding additional expression variation during training would help alleviate this problem. Finally, as our input to the network is a mesh, gaze changes and hair motions are currently not controllable.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural Point-Based graphics. In *Computer Vision – ECCV 2020*, pages 696–712. Springer International Publishing, 2020. 2

[2] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 2, 3

[3] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34:10691–10704, 2021. 2

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 1

[5] P. Chandran, G. Zoss, P. Gotardo, and D. Bradley. Continuous landmark detection with 3d queries. In *CVPR*, pages 16858–16867, 2023. 4, 1

[6] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 2, 4

[7] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[8] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 2

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

[10] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 2

[11] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)*, 30(6):1–10, 2011. 2

[12] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. 2018. 2

[13] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. HDHumans: A hybrid approach for high-fidelity digital humans. *Proc. ACM Comput. Graph. Interact. Tech.*, 6(3):1–23, 2023. 2

[14] Shun Iwase, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Timur Bagautdinov, Rohan Joshi, Fabian Prada, Takaaki Shiratori, Yaser Sheikh, and Jason Saragih. Relightablehands: Efficient neural relighting of articulated hand models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16663–16673, 2023. 3

[15] Hankyu Jang and Daeyoung Kim. D-TensoRF: Tensorial radiance fields for dynamic scenes. 2022. 2

[16] Shubhendu Jena, Franck Multon, and Adnane Boukhayma. Neural Mesh-Based graphics. In *Computer Vision – ECCV 2022 Workshops*, pages 739–757. Springer Nature Switzerland, 2023. 2

[17] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W H Lau. MODNet: Real-Time Trimap-Free portrait matting via objective decomposition. *AAAI*, 36(1):1140–1147, 2022. 6

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. 6

[19] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM TOG*, 42(4), 2023. 1

[20] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[21] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 2

[22] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[23] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM TOG*, 37(4):68:1–68:13, 2018. 5

[24] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM TOG*, 38(4), 2019. 1, 2

[25] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM TOG*, 40(4), 2021. 1, 2, 3, 4, 5

[26] Linjie Lyu, Ayush Tewari, Thomas Leimkuehler, Marc Habermann, and Christian Theobalt. Neural radiance transfer fields for relightable novel-view synthesis with global illumination. In *ECCV*, 2022. 1, 2

[27] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized

spherical gradient illumination. *Rendering Techniques*, 2007 (9):10, 2007. 2

[28] 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. Deep relightable textures - volumetric performance capture with neural rendering. In *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*, 2020. 2

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4), 2022. 1

[31] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 1, 2

[32] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[33] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 2

[34] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[35] Gernot Riegler and Vladlen Koltun. Stable view synthesis. pages 12216–12225, 2020. 2

[36] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. 2020. 2

[37] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 2

[38] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 2

[39] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 2

[40] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2

[41] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 4

[42] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 2, 4

[43] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. 2021. 1, 2

[44] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[45] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20762–20772, 2023. 2

[46] Edith Tretschk, Vladislav Golyanik, Michael Zollhoefer, Aljaz Bozic, Christoph Lassner, and Christian Theobalt. SceNeRFlow: Time-Consistent reconstruction of general dynamic scenes. 2023. 2

[47] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH Conf.*, 2022. 1

[48] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[49] Yingyan Xu, Jérémy Riviere, Gaspard Zoss, Prashanth Chandran, Derek Bradley, and Paulo Gotardo. Improved lighting models for facial appearance capture. In *Proc. Eurographics (short paper)*, 2022. 4

[50] Yingyan Xu, Gaspard Zoss, Prashanth Chandran, Markus Gross, Derek Bradley, and Paulo Gotardo. Renerf: Relightable neural radiance fields with nearfield lighting. In *ICCV*, 2023. 1, 2, 4

[51] Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. Towards practical capture of high-fidelity relightable avatars. *arXiv preprint arXiv:2309.04247*, 2023. 3, 8, 1, 2

[52] Chong Zeng, Guojun Chen, Yue Dong, Pieter Peers, Hongzhi Wu, and Xin Tong. Relighting neural radiance fields with shadow and highlight hints. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[53] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip

Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021. 2

[54] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 2

[55] Shizhan Zhu, Shunsuke Saito, Aljaz Bozic, Carlos Aliaga, Trevor Darrell, and Christop Lassner. Neural relighting with subsurface scattering by learning the radiance transfer gradient. 2023. 4

[56] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars, 2022. 2

[57] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Comput. Graph. Forum*, 37(2):523–550, 2018. 2