Relightable Neural Video Portrait

Simulantaneous Relighting and Reenactment using Deep Facial Reflectance Field

Youjia Wang
China
wangyj2@shanghaitech.edu.cn

Taotao Zhou
China
zhoutt@shanghaitech.edu.cn

Minzhang Li
China
limzh@shanghaitech.edu.cn

Teng Xu
China
xuteng@shanghaitech.edu.cn

Minye Wu
China
wumy@shanghaitech.edu.cn

Lan Xu
China
xulan1@shanghaitech.edu.cn

Jingyi Yu
China
yujingyi@shanghaitech.edu.cn

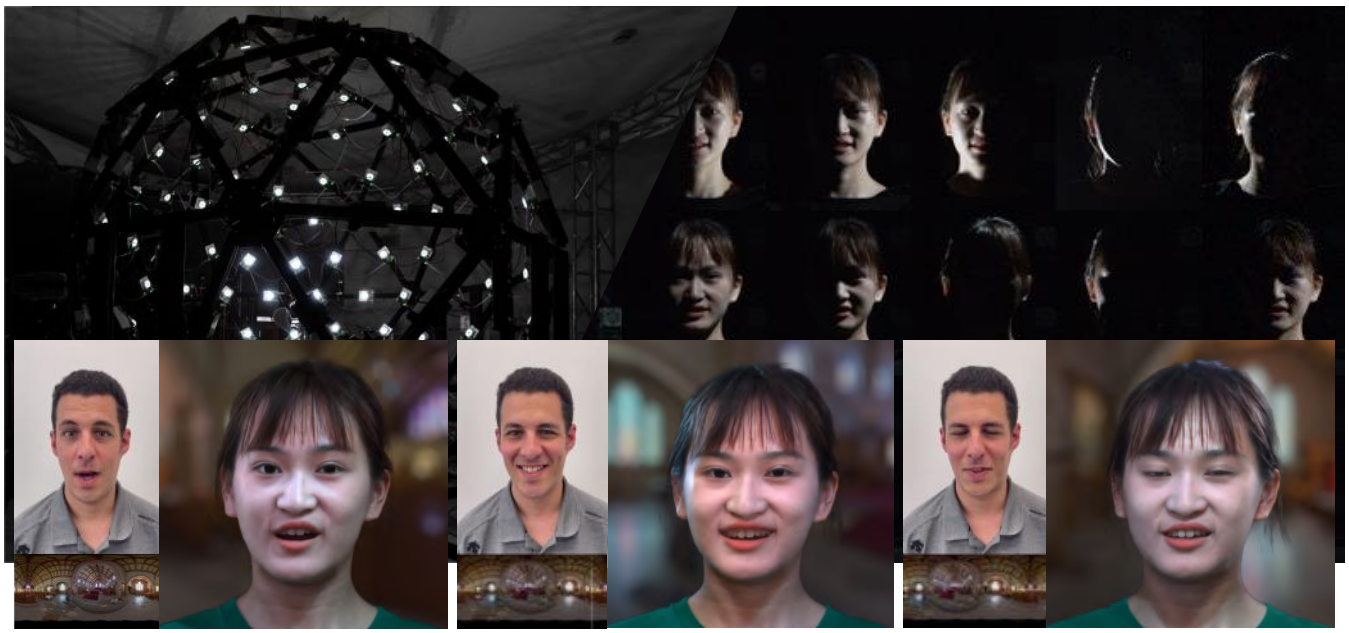

Figure 1. We introduce a relightable neural video portrait scheme for simultaneous relighting and reenactment that transfers the head pose and facial expressions from a source actor to a portrait video of a target actor with arbitrary new backgrounds and lighting conditions. It enables automatically and conveniently control of facial expression, lighting condition and background of the video portrait with high photo-realism for fancy visual effects.

Abstract

Photo-realistic facial video portrait reenactment benefits virtual production and numerous VR/AR experiences. The task remains challenging as the portrait should maintain high realism and consistency with the target environment. In this paper, we present relightable neural video portrait, a simultaneous relighting and reenactment scheme that transfers the head pose and facial expressions from a source actor to a portrait video of a target actor with arbitrary new backgrounds and lighting conditions. Our approach combines 4D reflectance field learning, model-based facial performance capture and target-aware neural rendering. Specifically, we

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

adopt a rendering-to-video translation network to first synthesize high-quality OLAT imagesets and alpha mattes from hybrid facial performance capture results. We then design a semantic-aware facial normalization scheme to enable reliable explicit control as well as a multi-frame multi-task learning strategy to encode content, segmentation and temporal information simultaneously for high-quality reflectance field inference. After training, our approach further enables photo-realistic and controllable video portrait editing of the target performer. Reliable face poses and expression editing is obtained by applying the same hybrid facial capture and normalization scheme to the source video input, while our explicit alpha and OLAT output enables high-quality relit and background editing. With the ability to achieve simultaneous relighting and reenactment, we are able to improve the realism in a variety of virtual production and video rewrite applications For instance, we can match the transferred facial expressions with the new environment and lighting condition at high fidelity for virtual conferencing, or enable highly realistic face-swapping into arbitrary desired lighting conditions for virtual cinematography. Our approach can also generate high frame-rate OLAT images (i.e, 1000 fps) that are regarded as extremely challenging to be captured by existing physical devices, which enables slow-motion relighting. Extensive experiments demonstrate the effectiveness of our approach to achieve photo-realistic relightable reenactment of facial video portraits.

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

Keywords: video editing, neural rendering, relighting, facial reenactment, neural representation, rendering-to-video translation

ACM Reference Format:
Youjia Wang, Taotao Zhou, Minzhang Li, Teng Xu, Minye Wu, Lan Xu, and Jingyi Yu. 2018. Relightable Neural Video Portrait: Simulantaneous Relighting and Reenactment using Deep Facial Reflectance Field. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

The popularity of mobile cameras has witnessed a rapid development of digital facial portrait photography. Further synthesizing and editing the video portraits suggests different content, i.e, modifying the performer's expression, the background and lighting, which enables numerous applications in virtual cinematography, movie post-production, visual effects and telepresence, among others. Generating such video portraits conveniently still remains challenging and has attracted substantive attention of both the computer vision and graphics communities.

In this paper, we focus on automatically and conveniently synthesizing a photo-realistic video portrait of a *target* actor where the head pose, facial expression, the lighting condition and background are fully disentangled and controllable through various *source* actors. It is particularly challenging since humans are highly sensitive to facial inconsistency and the video portraits are influenced by the complex interaction of environment, lighting patterns and various facial attributes like material and specular reflection. The most common practice for manipulating video portraits in the film industry still relies on tedious manual labor and sophisticated devices for reflectance field acquisition [12] or chroma-keying [77]. The recent neural techniques [31, 61] bring huge potential for convenient and highly-realistic manipulations, enabling compelling visual quality in various applications, such as visual dubbing [25, 30], telepresence [36], video post-editing [17, 80] or audio-driven virtual assistant [65], even in real-time [68, 70]. However, they still suffer from pre-baked shading of the specific training scenes without manipulating the illumination, which is critical for relighting applications in portrait photography like virtual cinematography. On the other hand, various neural schemes with controllable portrait's lighting conditions have been proposed using high-quality pore-level 4D facial reflectance fields [12] acquired under multi-view and multi-lit configurations. However, they are limited to image-based relighting [42, 52], a subset of the reflectance field learning [38, 53] or relightable model and appearance reconstruction [3, 22, 39]. Researchers pay less attention to combine the background and lighting manipulation with face-interior reenactment into a universal data-driven framework.

In this paper, we attack the above challenges and present the first approach to generate relightable neural video portraits of a specific performer. As illustrated in Fig. 1, with the aid of immediate and high-quality facial reflectance field inference, our approach endows the entire photo-realistic video portraits synthesis with the ability of fully disentanglement and control of various head pose, facial expressions, backgrounds and lighting conditions. This unique capability also enables numerous photo-realistic visual editing effects.

Generating such photo-realistically editable neural video portraits in a data-driven manner is non-trivial. Our key idea is to combine high-quality 4D reflectance field learning, model-based facial performance capture and target-aware neural rendering into a consistent framework, so as to support convenient, realistic and controllable editing of neural video portraits. To this end, for the temporal ground-truth supervision, we first build up a light stage setup to collect the dynamic facial reflectance fields of the target performer with one-light-at-a-time (OLAT) images under various sequential poses and expressions. Our capture device consists of 96 LED light sources and a stationary 4K ultra-high-speed camera at 1000 frames per second (fps), resulting in a temporal OLAT imageset at 25 fps, so as to provide fine-grained facial perception. Then, with such target-specific OLAT data, the core of our approach is a novel rendering-to-video translation

network design with the explicit head pose, facial expression, lighting and background disentanglement for highly controllable neural video portrait synthesis. Specifically, our network takes the hybrid 2D and 3D model-based facial capture results as input, and generates corresponding high-quality OLAT imagesets and alpha mattes as output. To enable reliable explicit control of the pose and expression, we carefully design a semantic-aware facial normalization scheme with a hierarchical landmark structure. For high-quality reflectance field inference, we adopt a multi-task multi-frame learning framework, which not only jointly models the content and segmentation information using a hybrid loss design, but also encodes the rich temporal information in the captured dynamic OLAT dataset in a sliding window manner. After training, reliable head pose and facial expression editing are obtained by applying the same facial capture and normalization scheme to the source video input, while our explicit alpha and OLAT output enables high-quality relit and background editing. An encoder-decoder architecture is further adopted to encode the illumination weighting parameters of OLAT images from the corresponding relit images, enabling automatic lighting editing from natural source images. Thus, our approach enables automatically and conveniently control of facial expression, lighting condition and background of the performer's video portrait with high photo-realism for various visual editing effects. More impressively, our approach can generate high frame-rate OLAT images (i.e, 1000 fps) for slow-motion relighting by interpolating the facial captured input of our translation network, which are regarded as extremely challenging using existing light stage devices due to the physical frame-rate constraint. To summarize, our main contributions include:

- We demonstrate the new capability of simultaneous relighting and reenactment, which enables photo-realistic video portraits synthesis for various visual editing effects, with full disentanglement of head pose, facial expressions, backgrounds and lighting conditions.
- We introduce a rendering-to-video translation network to transfer model-based input into high-fidelity facial reflectance fields, with a multi-frame multi-task learning strategy to encode content, segmentation and temporal information.
- We propose to utilize hybrid model-based facial capture with a carefully designed semantic-aware facial normalization scheme for reliable disentanglement.

2 Related Work

Face Capture and Reconstruction. Face reconstruction methods aim to reconstruct the geometry and appearance of 3D face models from visual data and we refer to two recent reports on monocular 3D face reconstruction and applications [15, 85] for a comprehensive overview. Early solutions adopt optimization schemes by fitting a 3D template model of only the face regions into various visual input, such as a single image [4, 5], a temporal consistent video [8, 18, 21, 23, 54, 78] or even unstructured image sets [28, 29, 46]. The recent deep learning techniques bring huge potential for face modeling, which learn to predict the 3D face shape, geometry or appearance [7, 27, 44, 45, 57, 63, 64, 79, 84] Some recent work [6, 33] learn a parametric face model from a large scale of 3D facial scans, while Feng *et al.* [16] furthers model a parametric detail layer in a data-driven manner. Some recent work also illustrates that a face parametric model can be trained from only image or video input [37] or combining the implicit representation [81]. However, these method above still cannot create a photo-realistic model in a controllable manner, especially for those fine-grained facial regions like hair, mouth interior or eye gaze. Recent target-specific neural rendering techniques [3, 36, 76] further enable more realistic geometry and appearance reconstruction using the data captured under a multi-view and multi-lit configuration. However, these neural models still heavily rely on per-frame facial geometry reconstruction during training and the final rendering results still suffer from reconstruction limitation and artifacts near the hair regions. The recent work [19] enables controllable implicit facial modeling and rendering using neural radiance field, but it still relies on per-scene training and suffers from pre-baked illumination of the training scenes. In contrast, our approach enables more disentanglement and control of background, lighting condition as well as facial capture parameters for photo-realistic video portrait generation.

Face Reenactment and Replacement. Facial reenactment re-generates the face content of a target actor in a portrait video by transferring facial expression and pose from a source actor. Many recent reenactment approaches model-based expression capturing and the expressions are transferred via dense motion fields [2, 34, 55] or facial parameters [20, 32, 67–69, 71]. The recent neural approaches [31, 61] replace components of the standard graphics pipeline by learned components, which bring huge potential for convenient and highly-realistic manipulations, enabling compelling visual quality in various applications, such as visual dubbing [25, 30], telepresence [36], video post-editing [17, 80] or audio-driven virtual assistant [65], even in real-time [68, 70]. Inspired by pix2pix [24], Deep Video Portraits [31] proposes an rendering-to-image translation approach that converts synthetic rendering input to a scene-specific realistic image output. The following work [30] present reenactment to maintain the speaking style of the target identity by modeling the audio information for visual dubbing. The recent work [25] further considers the emotion signal to generate target-specific emotional video portraits through only audio input. Besides, Deferred Neural Rendering [66] utilize neural textures to produce photo-realistic output, demonstrating high quality facial re-rendering and reenactment

with the parametric 3D Morphable Model (3DMM) as input. Recently, neural reenactment approaches [11, 17, 41, 56, 65] for video post-editing driven by audio or text have been widely explored. Specifically, Fried *et al.* [17] propose to map phonemes to the parameters of the 3DMM so as to smooth text-based reenactment. Thies *et al.* [65] present Neural Voice Puppetry, which estimates parameters of a facial blendshape expression model from the audio logits followed by neural rendering for photo-realistic results. However, these reenactment approaches above still suffer from pre-baked shading of the specific training scenes without manipulating the illumination which is critical for relighting applications like virtual cinematography. Some recent work enables lighting or facial expression control during the facial portrait image generation [9, 60] or editing [58, 59] based on styleGAN [26]. However, these methods mainly focus on the task of portrait image generation and suffers from unrealistic interpolation for neural video portrait reenactment. Differently, we take advantage of learning from the temporally consistent OLAT dataset of the target performer to allow for larger changes in facial expression reenactment, background as well as the lighting condition, while maintaining high photo-realism.

Facial Portrait Relighting. High quality facial portrait relighting requires the modeling of pore-level 4D reflectance fields. Debevec *et al.* [12] invent Light Stage to capture the reflectance field of human faces, which has enabled high-quality 3D face reconstruction and illuminations rendering, advancing the film's special effects industry. Some subsequent work has also achieved excellent results by introducing deep learning [22, 38, 39, 52, 82]. Some work follows the pipeline of color transfer to achieve the relighting effects [10, 49–51], which usually needs another portrait image as the facial color distribution reference. With the advent of deep neural networks and neural rendering, some methods [35, 48, 83] adopt Spherical Harmonics (SH) lighting model to manipulate the illumination. Several works [1, 14, 73] jointly estimate the 3D face and SH [5, 34] parameters and achieved relighting by recovering the facial geometry and modify the parameters of the SH lighting model. Explicitly modeling the shadow and specular [40, 75] achieve excellent results in directional light source relighting. Mallikarjunr *et al.* [62] take a single image portrait as input to predict OLAT(one-light-at-a-time) as Reflectance Fields, which can be relit to other lighting via image-based rendering. Sun *et al.* [52] choose environment map as lighting model and use light stage captured OLAT data to generate realistic training data and train relighting networks in an end-to-end fashion. Similarly, we use the target-aware temporal OLAT images to generate training data for high-quality lighting disentanglement. Differently, we further combine the background and lighting manipulation with face-interior reenactment into a universal data-driven framework.

3 Relightable Neural Video Portrait

In this paper, we present relightable neural video portrait, a simultaneous relighting and reenactment scheme that transfers the head pose and facial expressions from a source actor to a portrait video of a target actor with arbitrary new backgrounds and lighting conditions. Fig. 2 illustrates our technical pipeline, which combines 4D reflectance field learning, model-based facial performance capture and target-aware neural rendering. We first describe our dynamic OLAT dataset capture scheme in Sec. 3.1, which provides temporal ground-truth supervision for the target performer. We further introduce a portrait parsing scheme in Sec. 3.2 to generate conditioning feature maps for further fully controllable portrait video synthesis. Specifically, our approach parses the input video of the source actor for the explicit rigid head pose, facial expression, and semantic landmarks to support explicitly editing and disentanglement. A semantic-aware facial normalization is applied to enable reliable explicit control of the pose and expression. In sec. 3.3, we introduce our rendering-to-video translation network to synthesize high-quality OLAT imagesets and alpha mattes from our hybrid sequential conditioning feature maps. We adopt a multi-frame multi-task learning strategy to encode content, segmentation and temporal information simultaneously for high-quality reflectance field inference. Finally, various novel visual editing applications are as described in Sec. 3.4 using our high-quality relightable neural video portrait scheme.

3.1 Data Acquisition

To recover the 4D reflectance fields of dynamic portrait, we build up a light stage device with 96 LED light sources and a stationary 4K ultra-high-speed camera at 1000 fps, resulting in temporal OLAT imageset at 25 fps, so as to provide fine-grained facial perception.

Under such dynamic capture setting, the target performer can freely speak, translate and rotate in a certain range to provide a continuous and dynamic OLAT imagesets sequence. However, one of the most challenging issue is that the motion of the captured target along with data acquisition will cause mis-alignments, leading to blurriness in the OLAT imagesets, making the data post-processing more challenging. We conquer such limitations using an optical flow algorithm and further retrieve results at higher frame rate using densely-acquired homogeneous full-lit frames.

3.1.1 Hardware Architecture. Our hardware architecture is demonstrated in Fig. 3, which is a spherical dome of a radius of 1.3 meters with 96 fully programmable LEDs and a 4K ultra-high-speed camera. The LEDs are independently controlled via a synchronization control system and evenly localized on the doom to provide smooth lighting conditions. One can simulate any desired environmental map by linearly combining the intensity of 8-bit RGB channels. Single PCC (Phantom Capture Camera) is leveraged with a

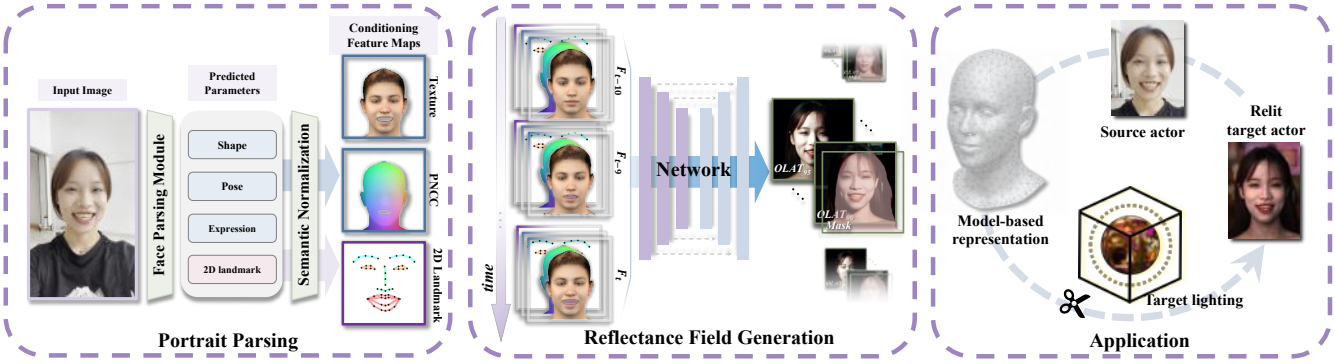

Figure 2. The algorithm pipeline of our relightable neural video portrait. Our pipeline mainly consists of three stages: Portrait parsing disentangles portrait parameters explicitly and normalizes them into the target actor's distribution for conditioning feature map generation; Then the proposed network synthesizes 4D reflectance fields via sequential conditioning feature maps; Our approach can achieve simultaneous relighting and reenactment effects on the portrait video of the target actor.

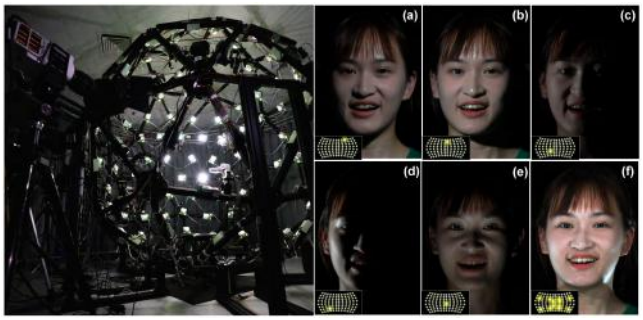

Figure 3. The demonstration of our capturing system and samples of captured data. Left: cameras and lights are arranged on a spherical structure; (a)-(e): samples of OLAT images; (f) a full-lit frame image. Subfigures on the bottom-left of (a)-(f) illustrates the corresponding lighting conditions on our system.

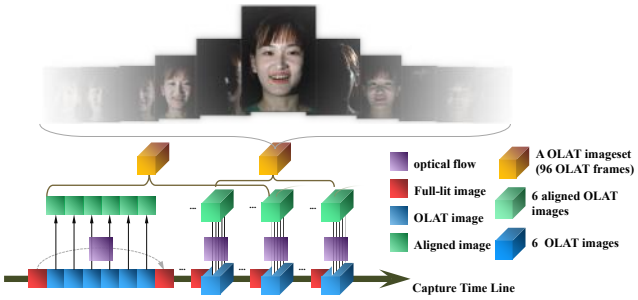

Figure 4. The illustration of capture timing and frame alignment. our method computes optical flows between full-lit images and warps OLAT images accordingly. The "overlapping" strategy allow us to reuse a same OLAT image in different OLAT imageset so that we can achieve higher capture frame rate.

global shutter setting, generating roughly 6 TB of data in total. For each data session, we collect video sequences in 25 seconds rather than isolated frames with 700 us exposure time and 1000 EI, which allows a lower noise floor. In practice, we can simultaneously control the PCC along with the LEDs system.

Capture setting. During capture, all the 96 LEDs follow the patterns shown in Fig. 3. To intensively acquire dynamic facial information, the captured target actor is asked to make at least six groups of different natural facial expressions. For the first two groups, the captured target is asked to stay comparatively calm to recite 50 predefined sentences. We set another two groups for short natural conversation with much richer expressions and translations and rotations are added to the last two groups During the capture period, the high-speed camera synchronizes with the 96 LEDs at 1000 fps and outputs an 8-bit Bayer pattern color image stream at a resolution of 2048×1440.

Data Processing. Note that it is required at least 0.1s to acquire a complete dynamic session with 96 LEDs. Such duration causes mis-alignment, making it challenging to handle blurriness in the dataset and the low frame-rate gives rise to inconsistent dynamic facial capture result. Inspired by the approach [39], we interleave "tracking frames" into the capture sequence during every six images rather a complete session to conquer such drawbacks and cast the tracking frames as references, as shown in Fig. 4. Instead of capturing a image with homogeneous illumination for every 96 images, we capture a image for tracking purposes every 6 images. This capture strategy allows us to align the OLAT data between 14 consecutive groups of full-bright frames in any pose to the middle frame with optical flow. It is equivalent to that the image between every two homogeneous

illumination frames is multiplexed thirteen times to enhance the final optical flow result.

Besides, we extract alpha mattes $\{\mathrm{M}_t\}$ of the actor for each frame using the matting algorithm [47].

3.2 Portrait Parsing

Our Portrait parsing module aims to generate a set of conditioning feature map $\{\mathrm{F}_t\}_{t=1}^{n_t}$ as inputs to our rendering-to-video network, where n_t is the number of the frames.

Conditioning Feature Maps. A conditioning feature maps F_t consists of a diffuse color image I_t^d, a coordinate image I_t^c as well as a landmark image I_t^l. Precisely, the diffuse color image I_t^d and coordinate image I_t^c are rendered from a FLAME [33] model of the target actor. Recall that FLAME [33] is a parametric 3D head model described by a function $G(\beta, \theta, \phi)$, where the coefficient parameters β, θ, and ϕ represent head shape, head pose and facial expression, respectively. We adopt an averaged model texture H of the FLAME dataset as our texture map and render the diffuse color image I_t^d via mesh rasterization with texture mapping and adopt the projected normalized coordinate code (PNCC) [84] to the FLAME model to generate the coordinate image I_t^c. These images provide a coarse estimation of the detailed facial geometry and spatial information for result synthesis. However, FLAME model itself can not precisely model fine-grained facial geometry, such as various mouth shapes, leading to a slight misalignment of facial features. To alleviate such error, we further add a landmark image in F_t to provide more accurate cues of fine-grained facial geometry and motion. We adopt the landmark defined as in the approach [27] and further connect 2D landmark points by segments with semantic meaning and form a landmark image as illustrated in Fig. 2.

Given an image of the source actor, we detect a human face and center it in a square region where the face occupies about 65% pixels. Next, we crop and resize the region from the source image to a uniform size square image patch I. Face parameters β, θ, and ϕ are extracted from this image patch via the DECA method [16], while the 2D landmarks are estimated using a regression trees approach [27].

Semantic Normalization. Note that when the source and target actors are different, their facial geometry will be different even in similar facial expressions, which will manifest in facial characteristics, such as eye size, nose length, mouth curvature, etc., as demonstrated in Fig. 5. We expect the distribution of conditioning feature maps generated from the source actor is similar to one of the target actor, so as to preserve target-aware appearance rendering results. To this end, we introduce a semantic-aware facial normalization strategy to the extracted facial parameters and 2D landmarks.

Specifically, we assume parameters as random variables which follow the normal distributions. For each parameter,

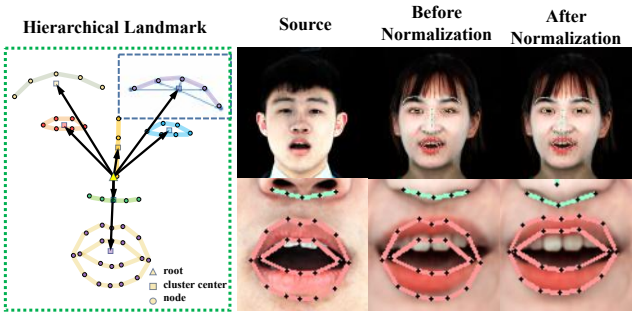

Figure 5. The illustration of our hierarchical landmark architecture and results before and after the proposed normalization scheme. Left image: black vectors represent the landmark vectors in the first level; Blue vectors in the blue rectangle on the top-right are examples of the ones in the second level. Right image: misalignment of mouth landmark is alleviated after normalization.

we normalize both the mean and variance of the source actor's distribution $\mathrm{X}_s \sim N(\mu_s, \sigma_s^2)$ to be the same as the target actor's distribution $\mathrm{X}_t \sim N(\mu_t, \sigma_t^2)$. Thus, the normalized parameter $\hat{X}_s$ can be formulated as:

$$\hat{X}_s = (\sigma_t \oslash \sigma_s) \circ (\mathrm{X}_s - \mu_s) + \mu_t \sim N(\mu_t, \sigma_t^2), \tag{1}$$

where $\oslash$ and $\circ$ are element-wise division and product respectively. In our implementation, we regard the head pose θ and facial expression ϕ of FLAME parameters as two individual random variables and use the Eqn. 1 for normalization. As shown in Fig. 5, we parameterize the 2D landmark by constructing a two-level hierarchical structure to take both global and local landmark features into account. In the first level, we group the landmark points of two eyes, two eyebrows, nose, nose bridge and mouth into seven clusters according to their categories. The root point is defined as the center of all landmark points. Then, the root point and the cluster centers form landmark vectors $\{\mathrm{l}_{1,i}\}_{i=1}^{7}$ of the first level, which represent the global spatial information of the face. In the second level, each cluster center with landmark points inside the cluster forms landmark vectors $\{\mathrm{l}_{2,i}^j\}_{j=1}^{n_i}$ of the second level for i-th category, where n_i is the number of landmark points in category i. The second level landmark vectors record the local arrangement of landmark points inside clusters, which will provide more accurate descriptions of facial features compared to FLAME model. We regard each landmark vector from all the levels as a random variable and normalize them according to Eqn. 1. Then, a normalized 2D landmark image $\tilde{\mathrm{I}}_t^l$ is reconstructed using these normalized landmark vectors in such a hierarchical manner. Thus, the conditioning feature maps F_t at time t are formulated as:

$$\mathrm{F}_t = \{\tilde{\mathrm{I}}_t^d, \tilde{\mathrm{I}}_t^c, \tilde{\mathrm{I}}_t^l\}, \tag{2}$$

where $\tilde{\mathrm{I}}_t^d, \tilde{\mathrm{I}}_t^c = \Pi(G(\beta, \tilde{\theta}, \tilde{\phi}), \mathrm{H})$ and Π is the rasterization and texture mapping function; $\tilde{\theta}$ and $\tilde{\phi}$ are normalized parameters.

3.3 Reflectance Field Generation

Our multi-branch conditioning rendering-to-video translation network takes the composed sequence of conditioning feature maps as input and generates the corresponding reflectance field and foreground mask of the portrait at each frame. Here, the generated reflectance field consists of 96 portrait images in the same head pose and facial expression from an identical camera view but are lit one light at a time. Such design in our approach enables more controllable high-quality relighting than directly predicting the portrait image under specific environment lighting conditions. Our explicit relighting strategy based on OLAT imagesets owns better generalization ability than those end-to-end learning ones which are highly relied on the diversity of training data. Meanwhile, the network is trained for a specific target actor, and such target-aware neural rendering strategy can render not only photo-realistic facial output even for those regions complying with head motion such as hair. As illustrated in Fig. 6, our translation network adopts a multi-frame multi-task learning framework, which will be described in detail below.

Sequential Input. Instead of feeding conditioning feature maps of a single frame to the network, we gather the adjacent 11 frames to be a sequential input $\mathcal{F}_{t_c} = \{\mathrm{F}_t\}_{t=t_c-10}^{t_c}$ for the network inference, where t_c is the current time. This sequential input enables the network to extract the rich temporal information in our dynamic OLAT dataset and output stabilized image sequence with temporal consistency.

Translation Network. We adopt the U-Net architecture to the proposed translation network as illustrated in Fig. 6. Our network inferences both the facial reflectance field and the foreground alpha matte simultaneously to facilitate portrait video composition applications. The proposed network consists of an encoder and two decoder modules. The encoder extracts multi-scale latent representations of conditioning feature maps, while the decoder module consists of two individual branches for generating the reflectance field and alpha, respectively. Inspired by the work [72], the alpha branch has half number of feature map channels than the branch for the reflectance field so that these decoders can independently leverage the corresponding high-level and low-level features. Such a multi-task framework enforces the encoder and the backbone of all the decoders to learn contextual information of the target actor, so as to produce more detailed reflectance fields and foreground alpha mattes. Please refer to Fig. 6 for more details.

Training Details. We leverage captured dynamic OLAT sequences of the target actor to train our target-aware translation network. The OLAT dataset can not be exploited directly due to the extreme lighting conditions which will cause DECA and landmark detection failures during portrait parsing. So, during preprocessing of training data, we synthesize a fully-illuminated image via relighting the corresponding OLAT imageset Y_t for each frame in the training sequence and then obtain a fully-illuminated portrait image patch. Portrait parsing will compute $\{\beta_t\}_{t=1}^{n_t}$, $\{\theta_t\}_{t=1}^{n_t}$, $\{\phi_t\}_{t=1}^{n_t}$, and $\{\hat{\mathrm{I}}_t^l\}_{t=1}^{n_t}$ and further render and compose them into conditioning feature maps $\{\mathrm{F}_t\}_{t=1}^{n_t}$ as described in Section 3.2. We can calculate the target distributions of these parameters for semantic normalization. The head shape of the same target actor whose head shape will not change should be a constant parameter. So we compute the average head shape parameter $\bar{\beta} = \frac{\sum_{t=1}^{n_t} \beta_t}{n_t}$ representing the actor's head shape which will be used in applications.

We utilize Mean Square Error(MSE) loss and the Multi-scale Structural Similarity(MS-SSIM) Index [74] loss to penalize the differences between the synthesized reflectance field and the ground truth, which are combined together as a photometric loss $\mathcal{L}_{color}$:

$$\mathcal{L}_{color}(\alpha) = \mathbb{E}_{\mathcal{F},\mathrm{Y}}[\|\mathrm{Y} - \Psi(\mathcal{F};\alpha)\|_2^2 + f_{mss}(\mathrm{Y}, \Psi(\mathcal{F};\alpha))], \tag{3}$$

where $f_{mss}(\cdot)$ is a differentiable MS-SSIM function [74]; α is the network weights of the proposed translation network; $\Psi(\cdot)$ outputs the predicted reflectance field of the network. Similar to [72], we use a perceptual loss to preserve details in rendered alpha mattes. The alpha loss is defined as:

$$\mathcal{L}_{alpha}(\alpha) = \mathbb{E}_{\mathcal{F},\mathrm{M}}[\|\mathrm{M} - \Psi(\mathcal{F};\alpha)\|_2^2 + f_{vgg}(\mathrm{M}, \Phi(\mathcal{F};\alpha))], \tag{4}$$

where $f_{vgg}(\cdot)$ denotes the output feature of the third-layer of pretrained VGG-19; $\Phi(\cdot)$ outputs the predicted alpha matte of the network. Meanwhile, we deploy an adversarial loss $\mathcal{L}_{GAN}$ to enhance the similarity of the distribution between predicted reflectance field and alpha matte and the ground truth so that the proposed network can produce photo-realistic results. Our discriminator $D(\cdot)$ is inspired by the PatchGAN [13] classifier and has a similar architecture but difference in inputs. $D(\cdot)$ conditions on the input, the conditioning feature maps $\mathcal{F}$, and either the predicted $\Psi(\mathcal{F};\alpha)$ or the ground-truth reflectance field Y. The adversarial loss $\mathcal{L}_{GAN}$ has the following form:

$$\begin{aligned}\mathcal{L}_{GAN}(\alpha,\omega) =& \mathbb{E}_{\mathcal{F},\mathrm{Y}}[\log D(\mathcal{F},\mathrm{Y};\omega)] \\ &+\mathbb{E}_{\mathcal{F}}[\log(1 - D(\mathcal{F},\Psi(\mathcal{F};\alpha);\omega))],\end{aligned} \tag{5}$$

where ω is the network weights of the discriminator $D(\cdot)$.

The color reflectance field branch directly outputs a reflectance field with 96 portrait images. The lack of constraints on these images let the network's outputs have slight facial geometry differences, which is not expected in our

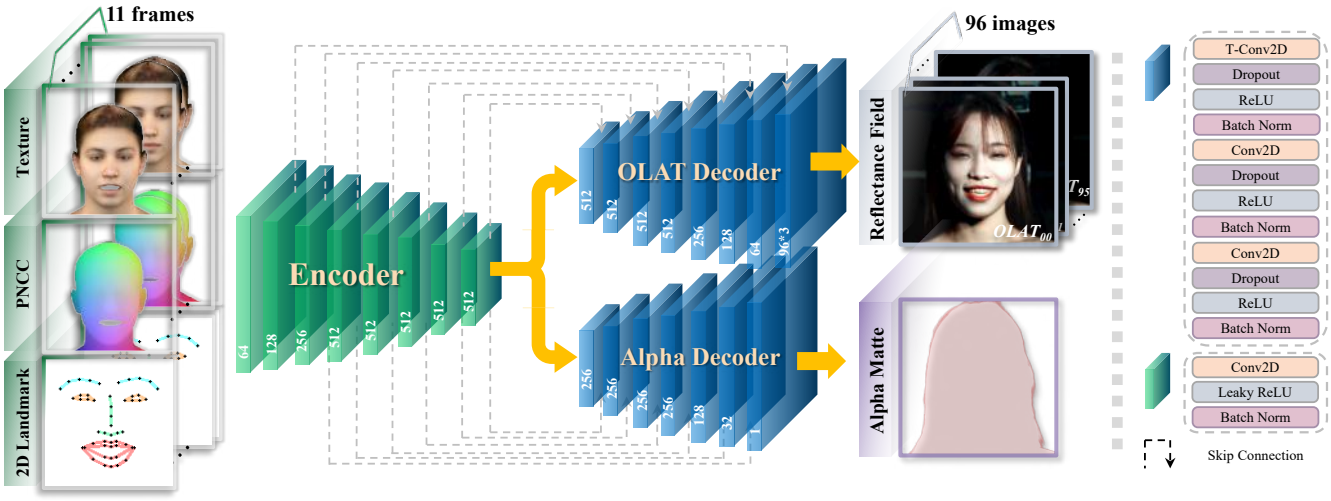

Figure 6. Our multi-frame multi-task architecture design for our rendering-to-video network. The proposed network takes a sequential conditioning feature maps as inputs and synthesizes reflectance field and corresponding alpha matte through separate decoder branches. The number on each layer indicates its number of feature channel.

task and will affect the relighting quality. We propose a context loss to alleviate this inconsistency in the predicted reflectance field. Specifically, we re-render a fully-illuminated portrait image from the predicted reflectance field which is a differentiable linear combination of output OLAT images. Next, we extract the head shape parameters $\beta(\Psi(\mathcal{F};\alpha))$ of the re-rendered portrait image and expect it to be as close to $\bar{\beta}$ as possible. The context loss $\mathcal{L}_{context}$ is formulated as:

$$\mathcal{L}_{context}(\alpha) = \mathbb{E}_{\mathcal{F}}(\|\beta(\Psi(\mathcal{F};\alpha)) - \bar{\beta}\|_1). \tag{6}$$

We linearly combine these loss function as the following objective function to find the optimal weights of translation network:

$$\begin{aligned}\alpha^* = \arg\min_{\alpha}\max_{\omega} \lambda_1 \mathcal{L}_{GAN}(\alpha,\omega) \\ +\mathcal{L}_{color}(\alpha) + \mathcal{L}_{alpha}(\alpha) + \lambda_2 \mathcal{L}_{context}(\alpha),\end{aligned} \tag{7}$$

where λ_1 and λ_2 are weights to balance the contribution of these terms. We set $\lambda_1 = \lambda_2 = 0.1$ in our implementation.

3.4 Applications

The explicit disentanglement of the head pose, facial expression and lighting condition in NeRVP that produces a photo-realistic video portrait of the target actor enables many visual effects. Here, we introduce following basic operations of our approach, whose combination empowers various applications as demonstrated in Section 4.

Relighting. Since our approach synthesizes 4D reflectance fields for the target actor, we can relight the portrait video by linearly combining the RGB channels of OLAT images according to arbitrary target environment map. Such kind of explicit approach provides more reliable and realistic outputs than learning-based methods which easily produce low-contrast results.

Face Reenactment. Given a video clip of the source actor, our approach can transfer the pose and the facial expression to the portrait video of target actor. By doing so, our method parsing the portrait images to obtain normalized parameters as described in Section 3.2. Fixed the head shape in the normalized parameters with $\bar{\beta}$, conditioning feature maps in the target actor's domain are generated. Finally, the target actor's face in the synthesized portrait video will reenact the source actor.

Frame Interpolation. Since our method parameterizes the 2D landmark in a hierarchical manner, all parameters including the head pose, the facial expression, and the head shape can be numerically interpolated between two frames. Interpolated parameters thus form a new frame for the target actor.

4 Results

In this section, we evaluate our approach in a variety of challenging scenarios. We first report the implementation details of both our approach and utilized OLAT dataset of the target performer. We further provide the comparison with previous state-of-the-art methods as well as the evaluation of our main technical components, both qualitatively and quantitatively, followed by the analysis of our results with various video editing effects. The limitations regarding our approach are provided in the last subsection.

Implementation Details. Our network model is implemented by PyTorch and we run all of our experiments with a

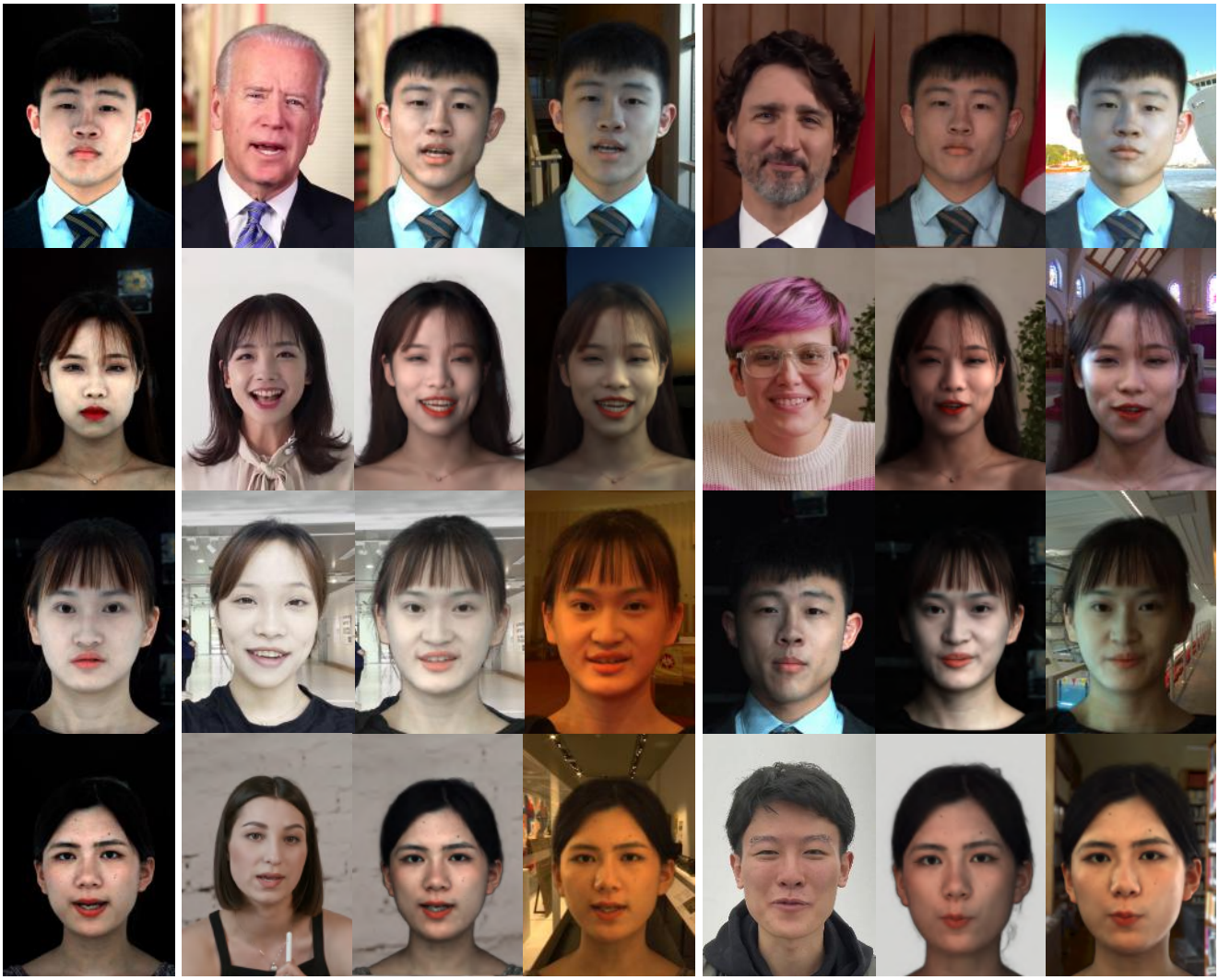

Figure 7. Our reenactment and relighting results of various neural video portraits. For each row, we obtain the expression references from two different source actors and provide the the corresponding generation results of our approach in two different lighting conditions (where the first lighting condition and background are from the source image as well).

single Nvidia GeForce RTX 3090 GPU, 24GB RAM. We adopt 500 epochs, batch size 4 to train images at the resolution of 256×256 pixels, and 200 epochs, batch size 2 for training images of 512×512 pixels. The learning rate is set to be 1e-4 for both cases. To evaluate our approach, we capture a dynamic OLAT dataset of 5 individuals with a light stage device as described in Sec. 3.1. Note that the frmae-rate of each dynamic OLAT video is 25 fps. For each individual, we obtain about 2000 OLAT imagesets covering a wide range of facial expressions where the individual is asked to provide natural conversations during the capture procedure. Note that we produce a dynamic video. The LEDs are synchronized with the PCC camera, producing a dynamic video sequence at 2048×1440 resolution with 1000 fps. In our data set, various facial expressions and motions are included to span the range of facial representative ability of the network. To handle misalignment issues, we apply an optical flow algorithm to processing raw OLAT data, followed by resizing images to 650×924 with which face bounding boxes can be detected by the DECA project [16].

4.1 Comparisons

With the aid of the immediate and high-quality dynamic facial reflectance field inference, our novel synthesized neural video portraits endow the entire capacity of combining dynamic facial expressions, various illumination conditions

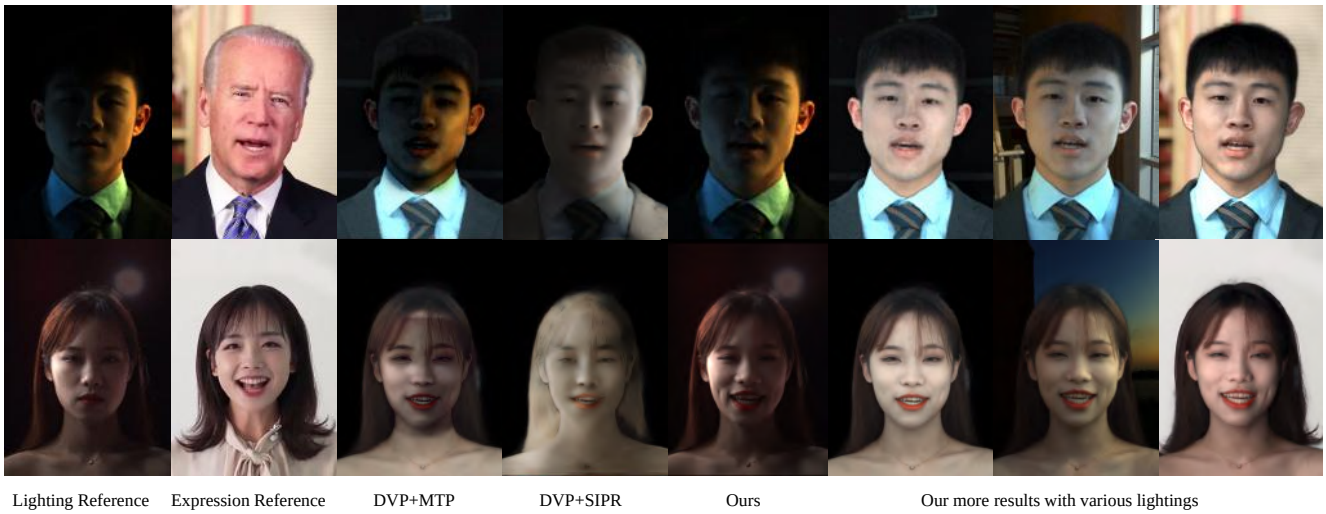

Figure 8. Comparison on in-the-wild videos: For each row, we transfer the expression and pose motions of a source person in a video to the target person. We generate the reflectance field of the target person and use an environment map to relight it. We compare our relighting result with the other two state-of-the-art methods, and the results show that our method generates a better-relighted portrait, which is closer to ground truth.

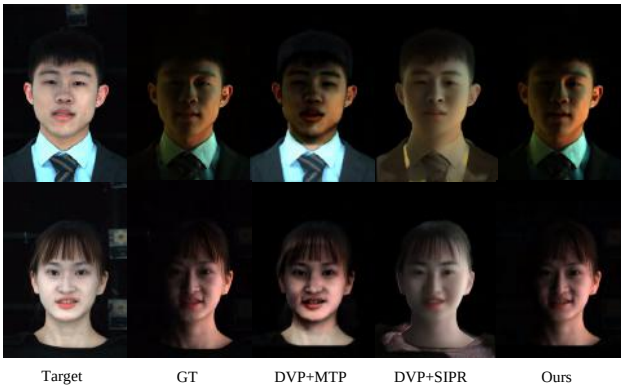

Figure 9. Comparison on OLAT test set: we divide the target's OLAT dataset into a training set and a test set. The second column is the relighting result of the real OLAT data, which is used as ground truth. The last column shows the relighting result of our neural network's output. The comparison result shows that our relighting result is more photo-realistic.

Comparison with other relighting methods			
Method	PSNR(↑)	SSIM (↑)	MAE (↓)
DVP+MTP	19.2990±1.7092	0.7015±0.0563	0.0644±0.0087
DVP+SIPR	17.6552±1.8539	0.6751±0.0431	0.0878±0.0195
Ours	**27.2709±1.6508**	**0.8618±0.0157**	**0.0291±0.0041**

Table 1. Quantitative comparision. ↑ means the larger is better, while ↓ means smaller is better. In terms of three metrics, our method outperforms the others.

and changeable environmental backgrounds. It is particularly of competitive advantages over cases that relighting portraits are interfered by the interaction of the environment. Since our method is a novel approach to generating relightable neural video portrait, to demonstrate our outstanding performance, we compare our method against the one based on DVP (Deep Video Portraits) [31], which suffers a lack of relightable capacity. Therefore, for fair comparisons, we further add another state-of-the-art methods for relighting, Single Image Portrait Relight (SIPR) and MassTransport approach (MTP), respectively.

As shown in both Fig. 8 and Fig. 9, our approach surpasses all other methods in visual effects. DVP+SIR suffers from over-smoothness, leading to insufficiency of details and unevenly distributed light on the target portrait face. Moreover, DVP+MTP shows severe flickering along with varying illumination conditions from various angles of deflection. Note that the pipeline based on MTP still relies on tedious landmark detection, which is sensitive to illumination conditions. This results in almost impossibility to render photo-realistic portrait under a backlit environment-map. In contrast, our approach achieves the perfect rendering result in terms of photo-realism without any artifact.

For further quantitative comparison, we evaluate our rendering results via three various metrics: signal-to-noise ratio (**PSNR**), structural similarity index(**SSIM**) and mean absolute error (**MAE**) to compare with existing state-of-art

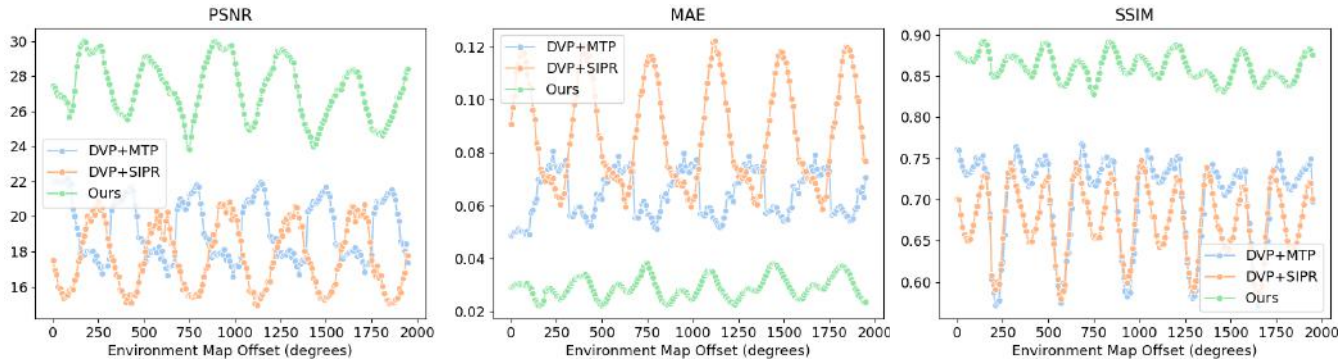

Figure 10. Quantitative comparisons on the test reference data with ground truth. Our results are more realistic and closer to ground truth.

approaches. For comparison, we generate a sequence of relighted portraits from all reference views which are evenly spaced in range of illumination angles. As shown in the Tab. 1, our method shows the outstanding performance on the photo-realistic synthesis and controllable video editing in terms of all three metrics with even less standard deviation. Moreover, as shown in Fig. 10 due to the high-quality reflectance field inference and fully disentanglement, our method enables entire photo-realistic video portraits synthesis under various illumination conditions, making the rendering quality of any temporal sequence surpassing other baselines with higher PSNR, SSIM and lower MAE. These qualitative and quantitative comparisons above reveal the effectiveness of our method in terms of full representation and control. Such qualitative comparison illustrates the effectiveness of our approach to encode the spatial and temporal information from our multi-view setting, which enables complete and realistic reconstruction of the portrait reflectance field and rendering for immersive free-view experiences.

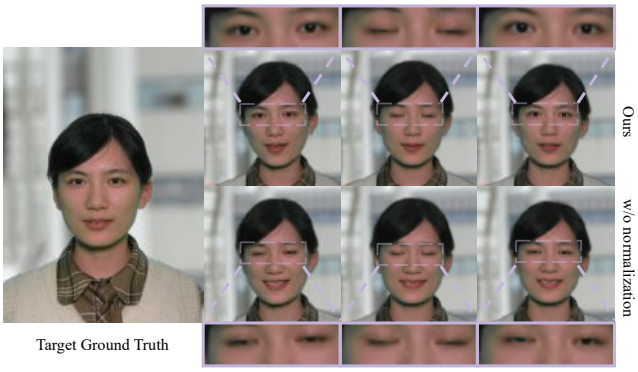

Figure 11. For the target person on the **left**, we use another person to animate. The results for our full pipeline are shown in the **first row**, and the **w/o normalization** test result is shown in the **second row**. We zoomed in on the eyes, the result shows that blinking is more natural after the normalization.

4.2 Ablation Study

In this subsection, we evaluate the performance of our approach by comparing different designs of feature maps in quality and quantity. We use **w/o normalization** to denote the variations of our approach without normalizing the parameters of the extracted 3D face model and 2D landmarks, while **w/o sparse-feature** denotes our approach without the feature map of 2D landmarks. As shown in Fig. 11, the bottom row illustrates the blinking process without normalization on the parameters of the face model and 2D landmarks. In contrast, the upper row shows the reenactment of the same eye blinking motion after normalization. We can observe that when the source person's eyes are much smaller than the target person's, the animation results look like that the target person cannot open the eyes, which is absurd and unnatural. The key problem is that the source person and target person have different facial geometries, when we simply transfer the facial motions of the source person, the size of eyes and mouth will not match the face of the target person. Normalization maps the distribution of the parameters of the source person to the target person, which ensures that we can transfer facial motions in a reasonable and photo-realistic way.

The utilization of sparse feature maps gives better results, as Fig. 12 illustrates. For the experiment in the first row, we split the OLAT dataset into 500 frames for training and 100 frames for testing. The rightmost column follows the input design proposed by Deep Video Portraits [31], which only uses rendered 3D face feature maps, without sparse features. Since the 3D face model cannot detect expressions accurately, when the test data is smiling with mouth open wide, the lips of the reconstructed 3D face model are only slightly open. In consequence, the neural network loses the information of the accurate facial motions, the generated target person cannot express the smile. By contrast, the middle column shows the results with 2D landmarks added to the neural

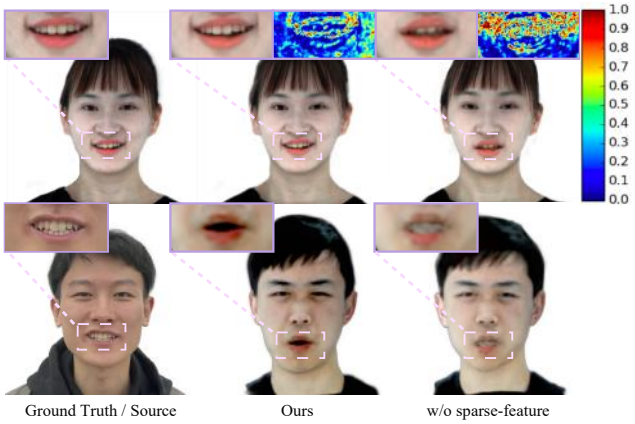

Figure 12. We divide the data into training data and test data. We compare the results of the full pipeline with **w/o sparse-feature**. For the **leftmost** ground-truth data as input, our full pipeline can reconstruct the result shown in the **middle**, and the result of **w/o sparse-feature** is shown on the right. As the figure, our results can better express the expression in the details (such as the mouth), and the error analysis shows that our results are closer to the original image.

network inputs. As we can see, the results are quite similar to the ground truth. To further evaluate the performance of our approach, we compute the error maps of our approach and DVP method [31]. We compute the MSE loss on the grayscale ground truth and grayscale generated results, and the pixel values are normalized to 0-1. The error of the result by DVP method in the upper lip region is obvious. For the experiment in the second row, we test our approach by a portrait video of another person, as shown in the first column in Fig. 12. Without the constraint of 2D landmarks, a lot of artifacts appear in regions with large facial motions. As shown in Fig. 12, there are many artifacts in the mouth of the generated target person in the rightmost column, in comparison to our approach.

4.3 Applications

Utilizing the dynamic facial reflectance field, our neural network can relight animated target person in high quality. In Fig. 7, Fig. 13 and Fig. 14, we generate OLATs of different persons in a video sequence. As we can see in the gallery Fig. 7, not only does our approach can transfer the facial expressions vividly, but also it performs pretty well with respect to relighting tasks. Our methods can handle all kinds of light conditions, including extreme cases such as indoor point lights, indoor ambient lights, the light of sunset and any linear combinations of point lights.

The leftmost column in Fig. 7 shows the re-render fully-illuminated OLAT portrait images of each target person. For each row, we use portrait videos of two different source persons as input. For each source sequence, we use two environment maps to relight the animated facial reflectance fields of the target person. We selected various environments with many kinds of light conditions.

Virtual Conference. By utilizing our generated dynamic facial reflectance field, we can achieve a relightable virtual conference. No matter whether the user is dressing casually with messy hair, or lying on the sofa with a cup of coffee, the user can appear decently in front of others in suits. The conference background can be changed arbitrarily, the most impressive effect is that the light conditions on the generated user in suits will perfectly match the environment. Furthermore, if the user is walking on the street in a rush with a shaking camera, or even he is not looking and the camera, by explicitly controlling the pose parameters, our approach can always let the user look like attending the conference naturally.

***High Frame Rate Dynamic OLAT.*.** Recall our neural video portrait technique can be alternatively viewed as a mapping from the facial parameters to the reflectance field, where the use of deep networks makes the mapping differentiable. We hence can interpolate the expression parameters to equivalently interpolate on lighting. Recall that we obtain a 96-light-sources OLAT dataset at an effective frame rate of 25 fps. In Fig. 14, via interpolating the expression parameters, we can synthesize an OLAT at 1000 fps, which in reality, would require using a camera with a frame rate of 100,000 fps, prohibitively high to even manufacture. Here, we just need the RGB OLAT dataset, and then we can generate an arbitrary number of OLAT data. In a scenario such as relighting a slow-motion shot of movies, we can directly perform the interpolation in the expression domain, which presents a photo-realistic effect, saving the time on shooting OLAT data and commissioning OLAT equipment.

Relightable Face-swapping. Popular face-swapping methods will suffer from the inconsistent light conditions between the source and the target persons. Here, we use a fully-illuminated OLAT portrait video as the source sequence. As shown in Fig. 15, We can achieve a seamless face-swapping with perfect light editing. We use the same environment map to relight the source sequence and the generated target facial reflectance field. Then, we employ the same face mask as DeepFake [43]to crop the face region of the target person. Concerning the blending part, we also use the Poisson image editing technique identical to DeepFake. The leftmost image gives a reference expression of the source person. We are able to relight the generated target sequence under any extreme light conditions.

4.4 Limitations

We have demonstrated the compelling capability of simultaneous relighting and reenactment for neural video portrait with a variety of fancy visual editing effects. Nevertheless, as

Figure 13. We demonstrated the application of our algorithm in the virtual conference. Users can participate in video conferences on formal occasions in a wider range of scenarios. Through the control of pose, we can realize that the user still presents a decent participation state in scenes such as walking and not looking at the camera.

Figure 14. High frame rate dynamic OLAT generation: by interpolating the expression parameters, our neural network can generate the 96-light-sources OLAT data with a frame rate of over 1000, which is impossible to acquire using top high-speed cameras.

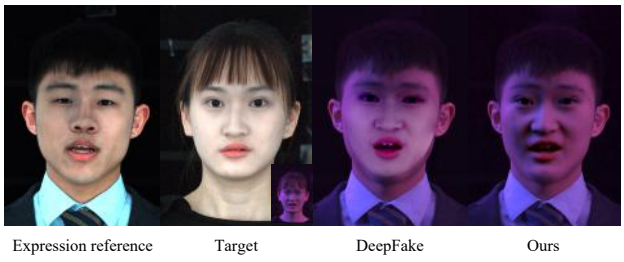

Figure 15. We demonstrate the application of our algorithm in face-swapping. Traditional face-swapping methods have discontinuities in dynamic lighting, while our method can still perform stably under dynamic lighting.

the first trial for such disentanglement of facial expression, background and lighting conditions for neural portrait generation, our approach is subject to some limitations. First, for our facial capture we do not maintain a explicit geometric relationship between the detected 2D landmarks and 3D parametric face model. As a result, when the captured subject performs extreme or fast head motions, the statistical distribution of the 2D landmarks will no longer be meaningful, leading to the failure of our normalization scheme. Furthermore, while 2D landmarks benefit the accurate facial reenactment, it also results in the heavy coupling between the head pose and facial expression, making the fine-grained separate editing infeasible. Our approach will also downgrade under those challenging or unseen poses during training since we adopt a two-stage data-driven solution. Besides, the output of our neural network is 96 OLAT imagesets, leading to a

huge GPU memory occupancy to support the training process. This is the key limitation on the spatial resolution of our generated results. In addition, to fully reenact the expression of the source person, we need to obtain a rich OLAT dataset of the target person, which is very expensive to train the person-specific neural network with a such a large-scale dataset. It's a promising direction to adopt more data-driven strategies to reduce the spatial and temporal redundancy of our OLAT dataset. Besides, our approach still requires per-target training and it's interesting to consider a general solution for all human portraits. Our approach also do not utilizes fine-grained 3D models for other regions such as hair. Thus, we can occasionally observe that the movements of these parts look unnatural and may violate the laws of physics.

5 Conclusion

We have presented the first approach to generate relightable neural video portraits of a target performer, which enables simultaneous relighting and reenactment via high-quality facial reflectance field learning. Our novel pipeline combines 4D reflectance field learning, model-based facial performance capture and target-aware neural rendering, so as to transfer the head pose and facial expressions from a source actor to a portrait video of a target actor with arbitrary new backgrounds and lighting conditions. The core of our approach is a rendering-to-video translation network to synthesize high-quality OLAT imagesets and alpha mattes from hybrid facial performance capture results, supporting explicit disentanglement. Our semantic-aware facial normalization scheme enables reliable explicit control of pose and expression with a hierarchical landmark structure. Our multi-frame multi-task learning strategy encodes the content, segmentation and temporal information simultaneously for controllable high-quality reflectance field inference. Extensive experimental results demonstrate the effectiveness of our approach for simultaneous relighting and reenactment with various visual editing effects under high realism, which compares favorably to the state-of-the-art. Our approach can even synthesize high frame-rate OLAT images for slow-motion relighting, which are challenging using existing devices. We believe that our approach is a critical step for the highly realistic synthesis of portrait video into various environments under the control of various meaningful attributes, with many potential applications for fancy visual effects in VR/AR, gaming, filming or entertainment.

References

[1] Oswald Aldrian and William AP Smith. 2012. Inverse rendering of faces with a 3D morphable model. *IEEE transactions on pattern analysis and machine intelligence* 35, 5 (2012), 1080–1093.

[2] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. 2017. Bringing portraits to life. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–13.

[3] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-EnWei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep Relightable Appearance Models for Animatable Faces. *ACM Trans. Graph. (Proc. SIGGRAPH)* 40, 4 (2021).

[4] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. 2004. Exchanging faces in images. In *Computer Graphics Forum*, Vol. 23. Wiley Online Library, 669–676.

[5] 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.

[6] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. 2018. Large scale 3d morphable models. *International Journal of Computer Vision* 126, 2 (2018), 233–254.

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

[8] Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)* 33, 4 (2014), 1–10.

[9] Anpei Chen, Ruiyang Liu, Ling Xie, and Jingyi Yu. 2020. A Free Viewpoint Portrait Generator with Dynamic Styling. *arXiv preprint arXiv:2007.03780* (2020).

[10] Xiaowu Chen, Mengmeng Chen, Xin Jin, and Qinping Zhao. 2011. Face illumination transfer through edge-preserving filters. In *CVPR 2011*. IEEE, 281–287.

[11] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? *arXiv preprint arXiv:1705.02966* (2017).

[12] 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.

[13] Ugur Demir and Gözde B. Ünal. 2018. Patch-Based Image Inpainting with Generative Adversarial Networks. *CoRR* abs/1803.07422 (2018). arXiv:1803.07422 http://arxiv.org/abs/1803.07422

[14] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. 2018. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision* 126, 12 (2018), 1269–1287.

[15] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)* 39, 5 (2020), 1–38.

[16] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2020. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *arXiv preprint arXiv:2012.04012* (2020).

[17] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[18] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. 2014. Driving high-resolution facial scans with video performance capture. *ACM Transactions on Graphics (TOG)* 34, 1 (2014), 1–14.

[19] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2020. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. *arXiv preprint arXiv:2012.03065* (2020).

[20] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4217–4224.

[21] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction

of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)* 35, 3 (2016), 1–15.

[22] 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.

[23] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–14.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[25] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. 2021. Audio-Driven Emotional Video Portraits. *arXiv preprint arXiv:2104.07452* (2021).

[26] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[27] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1867–1874.

[28] Ira Kemelmacher-Shlizerman. 2013. Internet based morphable model. In *Proceedings of the IEEE international conference on computer vision*. 3256–3263.

[29] Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M Seitz. 2010. Being john malkovich. In *European Conference on Computer Vision*. Springer, 341–353.

[30] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 2019. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

[31] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

[32] Kai Li, Qionghai Dai, Ruiping Wang, Yebin Liu, Feng Xu, and Jue Wang. 2013. A data-driven approach for facial expression retargeting in video. *IEEE Transactions on Multimedia* 16, 2 (2013), 299–310.

[33] 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 Trans. Graph.* 36, 6 (2017), 194–1.

[34] Kang Liu and Joern Ostermann. 2011. Realistic facial expression synthesis for an image-based talking head. In *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[35] Yang Liu, Alexandros Neophytou, Sunando Sengupta, and Eric Sommerlade. 2021. Relighting Images in the Wild with a Self-Supervised Siamese Auto-Encoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 32–40.

[36] 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.

[37] BR Mallikarjun, Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. 2020. Learning Complete 3D Morphable Face Models from Images and Videos. *arXiv preprint arXiv:2010.01679* (2020).

[38] 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.

[39] Abhimitra Meka, Rohit Pandey, Christian Häne, 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.

[40] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. 2020. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5124–5133.

[41] Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–14.

[42] Rohit Kumar Pandey, Sergio Orts Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. (2021).

[43] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, C. Umé, Mr. Dpfks, RP Luis, J. Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework. *ArXiv* abs/2005.05535 (2020).

[44] Elad Richardson, Matan Sela, and Ron Kimmel. 2016. 3D face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*. IEEE, 460–469.

[45] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. 2017. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1259–1268.

[46] Joseph Roth, Yiying Tong, and Xiaoming Liu. 2016. Adaptive 3D face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4197–4206.

[47] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2291–2300.

[48] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. SfSNet: Learning Shape, Reflectance and Illuminance of Facesin the Wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6296–6305.

[49] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. 2014. Style transfer for headshot portraits. (2014).

[50] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1.

[51] Yibing Song, Linchao Bao, Shengfeng He, Qingxiong Yang, and Ming-Hsuan Yang. 2017. Stylizing face images via multiple exemplars. *Computer Vision and Image Understanding* 162 (2017), 135–145.

[52] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Trans. Graph.* 38, 4 (2019), 79–1.

[53] 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.

[54] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. 2014. Total moving face reconstruction. In *European conference on computer vision*. Springer, 796–812.

[55] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2015. What makes tom hanks look like tom hanks. In *Proceedings of the IEEE International Conference on Computer Vision*.

3952–3960.

[56] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[57] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2019. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10812–10822.

[58] Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, Christian Theobalt, et al. 2021. PhotoApp: Photorealistic Appearance Editing of Head Portraits. *arXiv preprint arXiv:2103.07658* (2021).

[59] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

[60] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6142–6151.

[61] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhöfer. 2020. State of the Art on Neural Rendering. *Computer Graphics Forum* (2020). https://doi.org/10.1111/cgf.14022

[62] Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, Christian Theobalt, et al. 2020. Monocular Reconstruction of Neural Face Reflectance Fields. *arXiv preprint arXiv:2008.10247* (2020).

[63] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2549–2559.

[64] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1274–1283.

[65] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*. Springer, 716–731.

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

[67] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183–1.

[68] 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.

[69] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2018. FaceVR: Real-time gaze-aware facial reenactment in virtual reality. *ACM Transactions on Graphics (TOG)* 37, 2 (2018), 1–15.

[70] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. 2018. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

[71] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2006. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*. 24–es.

[72] Cen Wang, Minye Wu, Ziyu Wang, Liao Wang, Hao Sheng, and Jingyi Yu. 2020. Neural Opacity Point Cloud. *IEEE transactions on pattern analysis and machine intelligence* 42, 7 (2020), 1570–1581.

[73] Yang Wang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2007. Face re-lighting from a single image under harsh lighting conditions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[74] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[75] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–13.

[76] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–16.

[77] Steve Wright. 2006. *Digital compositing for film and video*. Routledge.

[78] 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.

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

[80] Xinwei Yao, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala. 2020. Iterative Text-based Editing of Talking-heads Using Neural Retargeting. arXiv:arXiv:2011.10688

[81] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2020. i3DMM: Deep Implicit 3D Morphable Model of Human Heads. *arXiv preprint arXiv:2011.14143* (2020).

[82] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. 2021. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)* 40, 1 (2021), 1–17.

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

[84] 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.

[85] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 523–550.