DifFRelight: Diffusion-Based Facial Performance Relighting

MINGMING HE*, Netflix Eyeline Studios, United States of America
PASCAL CLAUSEN*, Netflix Eyeline Studios, Canada
AHMET LEVENT TAŞEL*, Netflix Eyeline Studios, Canada
LI MA*, Netflix Eyeline Studios, United States of America
OLIVER PILARSKI*†, Netflix Eyeline Studios, Germany
WENQI XIAN, Netflix Eyeline Studios, United States of America
LASZLO RIKKER, Netflix Eyeline Studios, United Kingdom
XUEMING YU, Netflix Eyeline Studios, United States of America
RYAN BURGERT, Netflix Eyeline Studios, United States of America
NING YU, Netflix Eyeline Studios, United States of America
PAUL DEBEVEC, Netflix Eyeline Studios, United States of America



Fig. 1. **Flat-lit input and relit results of our diffusion-based relighting method.** Top row: four flat-lit subjects transformed to colorful multi-directional lighting. The flat-lit images are novel views generated by our dynamic 3D Gaussian Splatting extension from multi-view video. Bottom row: selected frames from a dynamic flat-lit sequence alongside their relit counterparts with an animated lighting direction. Facial details show our method's high-fidelity relighting effects, including eye reflections, translucency, and self-shadowing.

*Authors contributed equally to this work.
†Corresponding author.

Authors' addresses: Mingming He, Netflix Eyeline Studios, United States of America, hmm.lillian@gmail.com; Pascal Clausen, Netflix Eyeline Studios, Canada, pascal.clausen@scanlinevfx.com; Ahmet Levent Taşel, Netflix Eyeline Studios, Canada, leventtasel@gmail.com; Li Ma, Netflix Eyeline Studios, United States of America, lmaag@connect.ust.hk; Oliver Pilarski, Netflix Eyeline Studios, Germany, oliver.pilarski@scanlinevfx.com; Wenqi Xian, Netflix Eyeline Studios, United States of America, wenqixian3@gmail.com; Laszlo Rikker, Netflix Eyeline Studios, United Kingdom, laszlo.rikker@scanlinevfx.com; Xueming Yu, Netflix Eyeline Studios, United States of America, xueming.yu@scanlinevfx.com; Ryan Burgert, Netflix Eyeline Studios, United States of America, ryancentralorg@gmail.com; Ning Yu, Netflix Eyeline Studios, United States of America, ningyu.hust@gmail.com; Paul Debevec, Netflix Eyeline Studios, United States of America, debevec@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1131-2/24/12
https://doi.org/10.1145/3680528.3687644

We present a novel framework for free-viewpoint facial performance relighting using diffusion-based image-to-image translation. Leveraging a subject-specific dataset containing diverse facial expressions captured under various lighting conditions, including flat-lit and one-light-at-a-time (OLAT) scenarios, we train a diffusion model for precise lighting control, enabling high-fidelity relit facial images from flat-lit inputs. Our framework includes spatially-aligned conditioning of flat-lit captures and random noise, along with integrated lighting information for global control, utilizing prior knowledge from the pre-trained Stable Diffusion model. This model is then applied to dynamic facial performances captured in a consistent flat-lit environment and reconstructed for novel-view synthesis using a scalable dynamic 3D Gaussian Splatting method to maintain quality and consistency in the relit results. In addition, we introduce unified lighting control by integrating a novel area lighting representation with directional lighting, allowing for joint adjustments in light size and direction. We also enable high dynamic range imaging (HDRI) composition using multiple directional lights to produce dynamic sequences under complex lighting conditions. Our evaluations demonstrate the model's efficiency in achieving precise lighting control and generalizing across various facial expressions while preserving detailed features such as skin texture and hair. The model accurately reproduces complex lighting effects like eye reflections, subsurface scattering, self-shadowing, and translucency, advancing photorealism within our framework.

CCS Concepts: • **General and reference**; • **Computing methodologies** → **Reconstruction**; **Image representations**; *Scene understanding*; **Biometrics**;

Additional Key Words and Phrases: Relighting, Stable Diffusion, image-to-image translation, facial performances, data driven, HD facial data capture, OLATs, HDRI reconstruction

ACM Reference Format:
Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, and Paul Debevec. 2024. DifFRelight: Diffusion-Based Facial Performance Relighting. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3680528.3687644

1 INTRODUCTION

Today, photorealistic digital humans often play starring roles alongside live-action actors, and appear with increasing fidelity in video games and virtual reality. However, constructing photorealistic digital human models is difficult and expensive, and animating them realistically – even with performance-capture data – is still challenging and prone to error.

Volumetric performance capture (volcap) systems use an array of inward-pointing cameras to record dynamic human performances in three dimensions, avoiding the need for traditional 3D character modeling, rigging, and animation. Volcap captures all the nuances of a performance, allowing it to be rendered into any 3D scene from any angle. However, most capture techniques record the subject under a single flat lighting condition, which hampers the achievement of cinematic lighting effects and complicates the seamless integration of characters into new scenes. To address this, some advanced systems [Guo et al. 2019] use time-multiplexed lighting to record performances under multiple lighting conditions alternating each frame. This approach enables data-driven relighting, sometimes enhanced by machine learning inference [Meka et al. 2020; Zhang et al. 2021a].

In this work, we achieve high-quality relighting of volumetrically captured facial performances as shown in Fig. 1. This is accomplished by fine-tuning a diffusion-based image-to-image translation model using subject-specific relighting examples. In addition to capturing the performance under flat lighting, we record each subject under a dense array of lighting directions across various facial expressions. We train a diffusion model for precise lighting control, enabling the generation of high-fidelity relit facial images from flat-lit inputs. This process involves spatially-aligned conditioning of flat-lit captures and random noise, integrated with lighting information for global control, leveraging prior knowledge from the pretrained stable diffusion model. Additionally, we use dynamic 3D Gaussian Splatting (3DGS) to extrapolate the performance to novel viewpoints.

Our work includes the following contributions:

(1) A novel framework for facial performance relighting that trains on a subject-specific paired dataset and generalizes to novel lighting conditions, enabling relighting from free viewpoints and unseen facial expressions.
(2) A diffusion-based relighting model that spatially conditions the flat-lit input image, utilizing lighting information as global control to generate high-quality relit results.
(3) A scalable dynamic 3DGS technique for reconstructing long sequences, ensuring temporal consistency in flat-lit inputs for coherent inference by the relighting model.
(4) A unified lighting control that combines a new area lighting representation with directional lighting, offering versatile lighting controls as well as enabling the composition of complex environment lighting.

2 RELATED WORK

2.1 Non-Diffusion-Based Relighting Methods

Parametric reflectance modeling measures a subject's geometry and reflectance properties from different views and lighting directions but struggles with complex materials and dynamic subjects [Goesele et al. 2004; Sato et al. 1997; Weyrich et al. 2006]. These techniques struggle with complex materials, such as hair, and are difficult to apply to dynamic subjects since observations under many lighting conditions are required.

Image-based relighting photographs the subject under various lighting conditions to create new lighting scenarios, which is effective but costly and challenging for dynamic subjects [Debevec et al. 2000; Einarsson et al. 2006; Nimeroff et al. 1995; Peers et al. 2007].

Intrinsic image relighting decomposes an image into intrinsic components and recomposes it with modified lighting [Barron and Malik 2014; Le and Kakadiaris 2019; Li et al. 2018; Luan et al. 2017; Mei et al. 2024; Sengupta et al. 2018; Shih et al. 2014; Shu et al. 2017a,b; Sun et al. 2019; Tewari et al. 2021a, 2020, 2021b; Wang et al. 2008; Zhou et al. 2019]. Despite advancements, these methods face challenges with complex shading effects and detail preservation. Recently, Kim et al. [2024] propose a physically based rendering model for computing diffuse and specular reflectance, inspired by [Pandey et al. 2021]. They can generalize to novel subjects but lack detail for skin and hair reflectance, although refining with a neural renderer. Rao et al. [2022] develop a model with a reflectance network trained on OLAT images to learn a light-conditioned volumetric reflectance field from a single photo, but it is limited to static faces and does not address dynamic sequences.

Neural relighting approaches either use Neural Radiance Fields (NeRFs) to reconstruct material properties and lighting effects in 3D but face challenges in novel poses and detail preservation [Boss et al. 2021; Li et al. 2023b; Liu et al. 2023; Mildenhall et al. 2020; Sarkar et al. 2023; Srinivasan et al. 2021; Wang et al. 2023a; Xu et al. 2023b; Yao et al. 2022; Zeng et al. 2023; Zhang et al. 2023b, 2021b], or introduce 3D Gaussian Splatting (3DGS) for relighting because of the fine detail reconstruction [Gao et al. 2023; Kerbl et al. 2023]. Recent research on relightable head avatars uses OLAT data and various geometry representations, such as volumetric primitives [Yang et al. 2023], neural fields [Xu et al. 2023a], and 3DGS [Saito et al. 2023], to enhance the rendering quality of facial details. The subject-specific model [Saito et al. 2023] achieves promising facial performance relighting by combining 3D Gaussians with a relightable appearance model based on learnable radiance transfer. However, it necessitates a significantly denser capture settings for expressions and requires

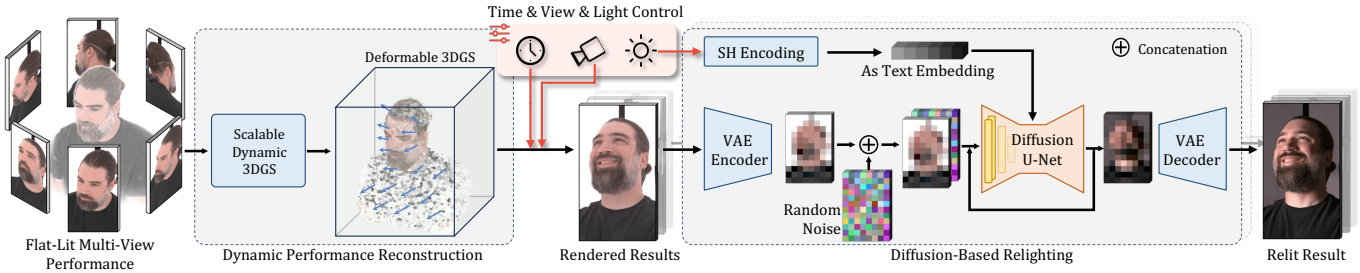


Fig. 2. **The overview of the proposed relighting pipeline including dynamic performance reconstruction and diffusion-based relighting.** Starting with multi-view performance data of a subject in a neutral environment, we train a deformable 3DGS to create novel-view renderings of the dynamic sequence. These serve as inputs for a diffusion-based relighting model, trained on paired data to translate flat-lit input images to relit results based on specified lighting. Here, we show the inference step of the diffusion model, where the latent representation of the flat-lit image is concatenated with random noise as input for the diffusion U-Net. Lighting information, encoded as SH encoding together with the text embedding, regulates the diffusion process.

precise track of facial mesh and gaze motion and explicit modeling of face and eyes, making this method difficult to generalize.

2.2 Diffusion-Based Relighting Methods

Diffusion models [Karras et al. 2022; Ke et al. 2023; Preechakul et al. 2022; Rombach et al. 2022; Song et al. 2020a,b; Wang et al. 2023b; Zhang et al. 2023a] have demonstrated remarkable capabilities in generating high-quality images by sampling from a learned distribution of images, particularly when conditioned on spatial control via image-to-image translation [Brooks et al. 2023; Ke et al. 2023; Meng et al. 2021; Parmar et al. 2023; Wang et al. 2023b].

Recently, researchers have leveraged conditional diffusion models for relighting [Ding et al. 2023; Ponglertnapakorn et al. 2023; Zeng et al. 2024]. These models leverage their ability to learn complex light interactions, drawing on a generalizable generative prior that has been pretrained on a large dataset of images under various lighting conditions. Additionally, they benefit from their controllability, which is fine-tuned using flat-relit image pairs. DiffusionRig [Ding et al. 2023] proposes to manipulate lighting and beyond by learning a mapping from computer-generated imagery (CGI) to real images using surface normals, albedo, and a diffuse shaded 3D morphable model. DiFaReli [Ponglertnapakorn et al. 2023] uses existing estimators for lighting, a 3D morphable model, the subject's identity, camera parameters, and a foreground mask, and trains a conditional diffusion network that takes a diffuse rendered model under novel lighting. Similarly, DiLightNet [Zeng et al. 2024] adopts a comparable approach, offering multiple radiance hints (diffuse and specular) without requiring a real-world captured input photograph, but is not trained on human or face datasets. Unlike one-shot methods, our approach uses multi-view, multi-pose data captured varying and controlled lighting conditions, allowing us to focus on high-fidelity identity-preserving reconstruction. We aim to generalize to arbitrary views, relighting conditions, and unseen poses, and prioritize preserving the subject identity.

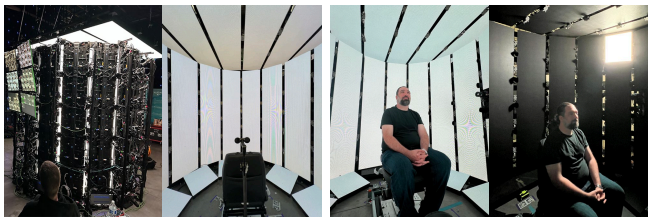
3 METHOD

Fig. 2 illustrates our dynamic facial performance relighting pipeline with two main components: dynamic 3D performance reconstruction and diffusion-based relighting. Our approach begins with the acquisition of multi-view performance sequences in a flat-lit environment. Utilizing our scalable approach of dynamic 3D Gaussian Splatting (3DGS), we first reconstruct deformable 3D Gaussians to render the sequence from novel perspectives. Then we employ the diffusion-based relighting model to generate new lighting for the rendered image sequence, based on a specified lighting direction. This model has been trained with subject-specific paired data of flat-lit and OLAT images captured using a customized LED-panel stage (Sec. 3.1), incorporating a new design of feeding flat-lit images and lighting directions as spatial and global conditions into the pretrained latent diffusion model (Sec. 3.2). To maintain temporal consistency in the dynamic relighting, we introduce a partitioning and combining scheme to train temporally-coherent 3DGS for lengthy sequences and also apply temporal blending on the dynamic relit results (Sec. 3.3). Additionally, we introduce a new variable-sized area light representation for enhanced control over different lighting types for various lighting compositions (Sec. 3.4).

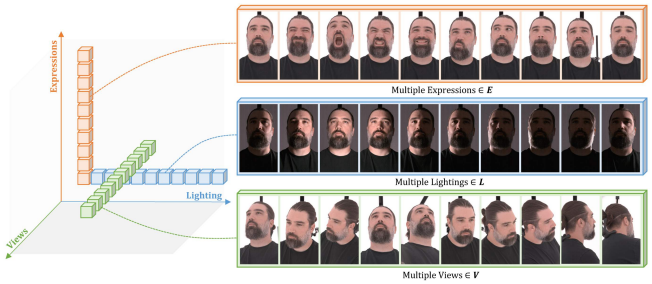
3.1 Facial Data Capture

Capture Setup. To supervise the diffusion-based relighting model and facilitate learning the translation of facial reflectance from flat lighting to arbitrary lighting, we require paired training data where lighting is the only changing variable. We use a volumetric capture stage consisting of a multi-view array camera array placed within a capped cylinder of LED panels as in Fig. 3 (a). By turning on different LED panels, we capture the same subject under varying lighting conditions. Furthermore, customizing the display patterns of LED panels enables us to record the subject under various backgrounds and image-based lighting environments. We constructed this stage as a cylindrical capture rig using off-the-shelf ROE BP2v2 LED panels and 75 synchronized 4K Z-CAM e2 cinema cameras. The rig comprises a 16×5 array of $50cm \times 50cm$ panels (16 columns, each consisting of 5 panels), forming a capture volume approximately 250 cm tall and 276 cm in diameter. The ceiling is covered with a 6×5 array of panels, and an additional 10 panels on the floor provide additional lighting angles. The cameras peer in through 5cm gaps in the walls and ceiling.

Lighting setup. Following LeGendre et al. [2020], we capture the reflectance field as OLAT images using each panel in sequence as a single light source at 24 frames per second in sync with the camera array. As Debevec and LeGendre [2022] demonstrate, LED panels are effective for image-based lighting on actors, so we use them



(a) Capture Stage (Inside & Outside *w/* and *w/o* Subject)



(b) Visualization of Captured Data

Fig. 3. **Capture stage and data.** (a) outside and inside of the LED panel volcap stage based on the LED panels, with and without the subject. (b) examples of our collected training data including variation in expression (**E**), lighting (**L**), and viewpoint (**V**).

to simulate OLATs, with one light (=OL) equivalent to one LED panel. Each panel approximates a 20-degree cone from the stage center, producing less aliased shadows than point lights while still capturing eye highlights and sharp specular reflections. We minimize the panel size while maintaining sufficient illumination and try to ensure uniformity in size and shape. Despite the large angular coverage of each panel, in practice, our OLAT-based framework works reasonably well when generalizing to environment lighting, as shown in Fig. 7 and Fig. 16. We also capture flat-lit lighting to have paired data for training the diffusion-based relighting model.

Data collection. We collected OLAT sequences for four individuals, each demonstrating 30 facial expressions and head poses. Each OLAT sequence has 123 lighting conditions with 13 interspersed flat-lit tracking frames. This same flat-lit lighting is used to capture the dynamic facial performance. Additionally, we recorded one OLAT sequence of a clean plate without a subject to facilitate background removal in 3DGS reconstruction. To compensate for the individual's slight movements during one OLAT sequence, we conduct optical flow alignment in image space with respect to the flat-lit frames for each view, following Meka et al. [2019]. When training the relighting model, we convert linear images to sRGB color space for better compatibility with the pretrained model weights.

3.2 Diffusion-Based Relighting

After data acquisition and preprocessing, we obtain a paired dataset $\{\mathbf{I}_{FlatLit}, \mathbf{I}_{OLAT}, \mathbf{d}\}$, where $\mathbf{I}_{FlatLit}$ represents the flat-lit image, and $\mathbf{I}_{OLAT}$ denotes the OLAT image under light direction $\mathbf{d}$. $\mathbf{d}$ is defined as a directional vector pointing from the stage center to the center of each panel in camera space. Our objective is to train a personalized model capable of relighting the flat-lit image of the same subject under novel views, novel lightings, and novel expressions/poses. Given the limited set of training pairs we have captured, a strong prior is essential for the model to generalize to unseen conditions. Rather than relying on physically-based lighting models, which often fail to accurately simulate complex shading effects such as anisotropic materials and subsurface scattering, we employ a purely data-driven prior to ensure the photorealism of our output. We use a diffusion-based approach that leverages the prior knowledge encapsulated in the pre-trained Stable Diffusion model [Rombach et al. 2022].

We formulate our problem as an image-to-image translation that transfers a flat-lit image $\mathbf{I}_{FlatLit}$ to the corresponding OLAT image $\mathbf{I}_{OLAT}$ under the condition $\mathbf{d}$. Inspired by recent monocular depth estimation works [Ke et al. 2023], we adapt Stable Diffusion's architecture to fine-tune it with our paired data conditioned on lighting information, achieving robust performance in relighting tasks.

The Stable Diffusion model [Rombach et al. 2022] is originally designed as a latent diffusion model for text-to-image generation. It comprises a Variational Autoencoder (VAE) that encodes an image $\mathbf{I}$ into the latent space $\mathbf{z} = \mathcal{E}(\mathbf{I})$ and decodes it back to $\mathbf{I} \approx \mathcal{D}(\mathbf{z})$. Additionally, it includes a diffusion U-Net $\hat{\epsilon}(\mathbf{z}^{(t)}; \mathbf{s}, t)$, which predicts the noise of a partially denoised latent $\mathbf{z}^{(t)}$ conditioned on some text embedding $\mathbf{s}$ and the diffusion timestep t. By iteratively removing noise from random noise, we obtain a clean image latent $\mathbf{z}^{(0)}$, which can then be decoded back into an image $\mathcal{D}(\mathbf{z}^{(0)})$.

To transform the image generation framework into an image-to-image translation model and more effectively leverage the spatial information in the flat-lit image, we modify the U-Net by concatenating the latent of the flat-lit image with the random noise map. This necessitates a minor change in the network structure: specifically, doubling the input channel count of the first convolutional layer. These minimal adjustments enable us to maximize the retention of the pre-trained weights from Stable Diffusion. Additionally, concatenating the latents with random noise improves the alignment of the spatial structure between the relit result and the input.

In addition to the flat-lit image, which offers spatial cues to the resulting output, the lighting information acts as a crucial global control signal that influences the entire image. We encode light direction $\mathbf{d}$ into higher dimensions and replace text embeddings by introducing it into the U-Net through cross-attention in diffusion models. To increase the frequency band of $\mathbf{d}$ and enhance the conditioning, we use Spherical Harmonics (SH) to encode $\mathbf{d}$, similar to techniques used by Tancik et al. [2023]. We also zero-pad the SH encoding to match the length of the text embedding:

$$\mathbf{s}_d = \mathbf{0} \oplus \mathcal{Y}(\mathbf{d}), \tag{1}$$

where $\oplus$ indicates concatenation, and $\mathcal{Y}$ indicates the SH encoding. Initially, this may appear unusual as the SH encoding is nothing like a text embedding, which is predicted by some text encoder [Schuhmann et al. 2022]. However, we find that the fine-tuning is adequate for filling the domain gap between the text embedding and SH encoding. We set the SH degree to 3 for all model training, as it showed the best performance in our experiments (see Sec. 5.1 in the supplemental material).

At training time, we freeze the VAE encoder and decoder, as both the flat-lit and OLAT-lit images lie within the manifold of the VAE's

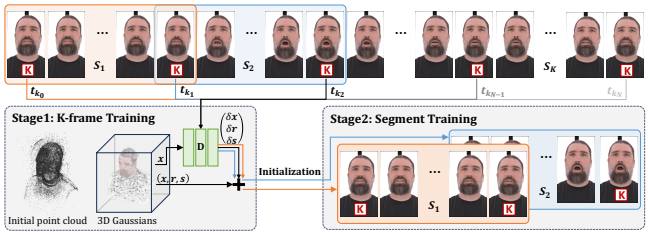


Fig. 4. **Two-stage training of the scalable dynamic 3DGS.** We first sample K frames to partition the long sequence into segments $S_1, S_2, ...S_k$. At *Stage 1*, we train the deformable 3DGS on the K frames only to generate the initialization for the training of each segment at *Stage 2*. Then we train a deformable 3DGS for each segment but conditioned on the initialization.

latent space. We randomly select training pairs from the dataset and fine-tune the diffusion U-Net using the denoising diffusion objective:

$$\mathcal{L}_{\text{diffusion}} = \|\hat{\epsilon}\left(\mathbf{z}_{OLAT}^{(t)} \oplus \mathcal{E}(\mathbf{I}_{FlatLit}); \mathbf{s}_d, t\right) - \epsilon\|_2^2, \tag{2}$$

where $\mathbf{z}_{OLAT}^{(t)}$ is a noisy ground truth latent at diffusion time step t, combining ground truth latent $\mathcal{E}(\mathbf{I}_{OLAT})$ and random noise map ϵ:

$$\mathbf{z}_{OLAT}^{(t)} = \sqrt{\alpha_t}\mathcal{E}(\mathbf{I}_{OLAT}) + \sqrt{1-\alpha_t}\epsilon. \tag{3}$$

Here α_t is a predefined value that schedules the diffusion process. We find that using naive Gaussian noise results in substantial color shift. This is likely due to the inductive bias that the network struggles to predict very dark images using zero-meaned Gaussian noise [Guttenberg 2023] for ϵ. Pyramid noise, as introduced in [Whitaker 2024], greatly improves the color consistency between the prediction and the ground truth.

We fine-tune one personalized model for each subject using the captured data. During inference, we are able to relight a flat-lit image of the same subject under arbitrary combinations of facial expression, camera pose, and light direction.

3.3 Dynamic Performance Relighting

To ensure temporally consistent relit results when applying the pre-trained diffusion-based relighting model to flat-lit facial performances, our initial goal is to minimize the domain gap between training and testing by generating photorealistic novel-view renderings. Additionally, maintaining temporal consistency in the flat-lit inputs enhances the diffusion model's performance under consistent conditions. In line with these requirements, we utilize 3D Gaussians [Kerbl et al. 2023] as the geometry representation for capturing fine details in real-time rendering. 3D Gaussian Splatting (3DGS) can be extended to the temporal domain [Jung et al. 2023; Li et al. 2023a; Luiten et al. 2023; Wu et al. 2023]. Deformable 3DGS [Jung et al. 2023] learns a deformation field for each frame within the canonical space of 3D Gaussians that is shared across all frames. The deformation field takes the positions of the 3D Gaussians $\mathbf{x}$ and the current time t as inputs, outputting the offset for position $\mathbf{x}$, rotation $\mathbf{r}$, and scaling $\mathbf{s}$ as $\delta_t\mathbf{x}$, $\delta_t\mathbf{r}$, and $\delta_t\mathbf{s}$. While preserving details in short videos, this method faces challenges in accurately representing motion and maintaining intricate details in long sequences due to its use of globally shared Gaussians, which struggle with the complexities of extended sequences.

We introduce a novel scalable method based on the deformable 3DGS [Jung et al. 2023] to optimize 3D Gaussians for lengthy performance sequences, thereby generating temporally consistent reconstruction. Instead of training all frames together with one shared set of Gaussians, we partition it into small segments with an equal number of frames, allowing varying Gaussians across segments. To minimize temporal inconsistency at the transition frame between segments and to preserve a similar level of reconstruction details across segments, we design a two-stage training strategy. As shown in Fig.4, in *Stage 1*, we evenly distribute K keyframes in the entire sequence, and then train a deformable Gaussian splatting model based on the keyframes only, generating the pretrained K-frame model. In *Stage 2*, using the K frames as transition points, we partition the entire sequence into $K-1$ segments, each containing two keyframes (one at the beginning and the other at the end). Inside every segment, we initialize the 3D Gaussians and the deformation network using the pretrained K-frame model based on the time of its first frame in the global time range. The geometry information of 3D Gaussians (such as positions $\mathbf{x}$, rotations $\mathbf{r}$ and scalings $\mathbf{s}$) is initialized as the pretained K-frame Gaussians transformed by the deformation offset of the first frame t_{k_0}. In this way, we ensure the initial states of 3D Gaussians in different segments are temporally consistent with similar level of details. To further enhance the consistency, we fix the initial Gaussians and train the deformation network only for the warm-up iterations (we use 3000 iterations as default), with the regularization term defined as L_2 loss on the deformation offset of keyframes k_0 and k_1:

$$\mathcal{L}_{reg} = \|\delta'_{t_{k_0}} - \delta_{t_{k_0}}\|_2 + \|\delta'_{t_{k_1}} - \delta_{t_{k_1}}\|_2, \tag{4}$$

where $\delta_{t_{k_0}}$ and $\delta_{t_{k_1}}$ are initial deformation offset of t_{k_0} and t_{k_1} while $\delta'_{t_{k_0}}$ and $\delta'_{t_{k_1}}$ indicates the learned offset with respect to $\mathbf{x}$, $\mathbf{r}$ and $\mathbf{s}$.

This warm-up training allows the Gaussians to freely deform to reconstruct motions while being restricted to the deformation of keyframes at transition points. After this stage, most Gaussians are in the correct position, and we relax the constraints to allow Gaussians to clone, split, and prune for detailed reconstruction for a number of iterations (defaulting to 10000 iterations). After the densification, we fix the number of Gaussians and jointly optimize the deformation network and the Gaussians. The ablation of the two-stage training method is included in the supplemental material.

Although we ensure temporal consistency in the reconstructed flat-lit sequence used as input for our relighting model, the image-based diffusion model may still produce temporally inconsistent high frequencies. To address this, we adopt the method from [Jamriška et al. 2019], applying a temporal blending approach to interpolate relit results between keyframes. We use a small step size of 4 to sample keyframes and preserve details and lighting accuracy.

3.4 Unified Lighting Control

To provide the users with flexible lighting control, we propose a novel area lighting representation, which includes both lighting direction and a variable light size. This is integrated with our directional lighting into a unified lighting control to guide the diffusion-based relighting model. Based on the single-lighting inference, we

can enable HDRI environment lighting reconstruction by compositing multiple single-lighting inferences.

Area-light training. To infer lights of various sizes – from point lighting to diffuse – we modify our training in two ways. First, we multiply the light direction by a light size factor $1 - a$, with a ranging from 0 (indicating a single OLAT) to 1 (representing the flat-lit), which represents the spread of the light and thus its sharpness. Second, we construct our ground truth target by combining OLAT images to simulate isotropic Spherical Gaussian (SG) illumination. Then the light size factor is linked to the Spherical Gaussian's sharpness by the formula $\lambda = \frac{-cos(\theta)}{(cos(\theta)^2-1)}$ with $\theta = a \cdot (\theta_{max} - \theta_{min}) + \theta_{min}$, $\frac{\pi}{2}$ being the inflection point of an SG and $\theta_{min} = 1 \cdot \frac{\pi}{180}$, $\theta_{max} = 89 \cdot \frac{\pi}{180}$.

HDRI reconstruction. To relight a performance under an HDRI environment using our model conditioned on a single lighting, we first map an HDRI latitude-longitude (lat-long) representation to the OLATs, with each OLAT's weighting coefficient being an average from its corresponding area on the HDRI lat-long. During inference, we generate an image for each OLAT direction and compute a weighted sum of these OLATs in linear space using their fitted coefficients. Additionally, the HDRI map can be reconstructed using our model conditioned on both light direction and size by first fitting Spherical Gaussians onto the HDRI lat-long representation and then inferring for each light direction and size. Similarly to the OLATs, we compute a weighted sum of the predictions using the fitted coefficients of these Spherical Gaussians. If the HDRI is animated, we adjust the conditioned directions of the OLATs or Spherical Gaussians to match the rotation of the HDRI.

4 EXPERIMENTS

We have conducted multiple experiments to evaluate the performance of our diffusion-based relighting model, particularly its generalization across various dimensions including novel lightings, expressions, and views. We compare our model with two baselines constructed using different network structures from the related methods. Additionally, we perform ablation studies to verify various technical designs within our system and conduct a simple extension to assess model generalization to novel subjects. For more results on dynamic performance relighting and detailed experiments, please refer to the accompanying video and supplemental material.

4.1 Implementation Details

We capture 30 expressions (30E), 123 OLATs (123L), and 75 views (75V) for each of the 4 subjects and downsample 4K captures to 1080×1920. For training, we use the Cartesian product of $27E \times 115L \times 69V$, with the rest reserved for validation. Capturing all OLAT data for each subject takes a few minutes. For inference, we record 10-15 seconds of performances in a multi-view capture stage under flat lighting. We implement our diffusion-based relighting system using Diffusers [Face 2024] based on Stable Diffusion v2.1. Our custom models are trained on 8 NVIDIA A100 GPUs with 40GB memory each, for 100K iterations with a batch size of 8. We use an Adam optimizer with a learning rate of $3 \cdot 10^{-5}$. We use a DDPM scheduler with 1000 steps for training and a DDIM scheduler with 30 steps for inference which takes about 5-6 seconds per frame on a single A100 GPU. Our dynamic 3DGS is also trained in full HD resolution. Given a long sequence, we divide it into segments based on the number of keyframes, typically including 20 frames per segment. We then train the deformable 3DGS for 40K iterations for each segment.

Table 1. Quantitative comparison on different relighting methods.

Novel Light	PSNR ↑	SSIM ↑	LPIPS ↓	FLIP ↓
Ours	**30.29**	**.8212**	**.1750**	**.0825**
U-Net-based	23.66	.7971	.2988	.1586
ControlNet-based	20.30	.6708	.2281	.2304
Novel Expression	PSNR ↑	SSIM ↑	LPIPS ↓	FLIP ↓
Ours	**31.55**	**.8281**	**.1585**	**.0689**
U-Net-based	23.86	.7916	.2887	.1479
ControlNet-based	20.50	.6423	.2284	.2382
Novel View	PSNR ↑	SSIM ↑	LPIPS ↓	FLIP ↓
Ours	**31.77**	**.8254**	**.1601**	**.0689**
U-Net-based	24.51	.7907	.29095	.1397
ControlNet-based	20.52	.6375	.2304	.2358
Novel LEV	PSNR ↑	SSIM ↑	LPIPS ↓	FLIP ↓
Ours	**30.98**	**.8229**	**.1729**	**.0754**
U-Net-based	24.55	.8010	.2961	.1440
ControlNet-based	20.51	.6490	.2301	.2425

4.2 Results

We showcase the effectiveness of our diffusion-based relighting method on the testing data of the four captured subjects. The visual results in Fig. 1, Fig. 10, and Fig. 11 demonstrate that our method reproduces realistic skin texture and reflectance, eye highlights, and fine hair structures while maintaining the subject-specific identity features. The novel lighting animation creates a realistic interplay of shadows and shading across the face. Fig. 11 shows our unified lighting control with both directional and area lights, and as an application to our lighting control, Fig. 7 shows the results relit by HDRI environment maps. In addition to varying lighting conditions, expression changes are illustrated in Fig. 10.

4.3 Comparisons

Since no existing works employ the same training and testing settings as our method, we establish baselines using two different network structures: a diffusion-based model built on ControlNet [Zhang et al. 2023a] and a U-Net-based model inspired by a state-of-the-art image-based relighting technique of [Pandey et al. 2021]. We adapt ControlNet to condition on a 3-channel flat-lit image and a 1-channel diffuse shading map, obtained by creating a diffuse shading map using the dot product between the lighting direction and photometric normals. Photometric normals are computed using a standard method [Ma et al. 2007] for subjects captured under gradient illumination. This model is trained under the same settings as ours at 1080×1920 resolution using 8 A100 GPUs for 100K iterations with a batch size of 8. The other baseline, a U-Net-based image-to-image translation network, is directly conditioned on the lighting direction, diverging from the use of intrinsic features (normals, albedo, specular) in the referenced architecture to minimize the negative impact

Table 2. Quantitative results for the ablations. Our method with shading map condition achieves the best result. However, it requires photometric normal as additional input, which increases the cost of capturing the data.

	PSNR ↑	SSIM ↑	LPIPS ↓	FLIP ↓
Our final model	30.04	.8141	.1719	.0859
Ours w/ shading map	**30.32**	**.8164**	**.1686**	**.0836**
Ours w/o pyramid noise	26.55	.7683	.1776	.1433
Ours w/o pre-trained	28.26	.8014	.1937	.1041



GT Ours w/o pyramid noise

Fig. 5. **Visual comparison between using and not using pyramid noise during training.** Using pyramid noise improves the color consistency between the prediction and GT. We increase the brightness of the images by 20% to exaggerate the error.

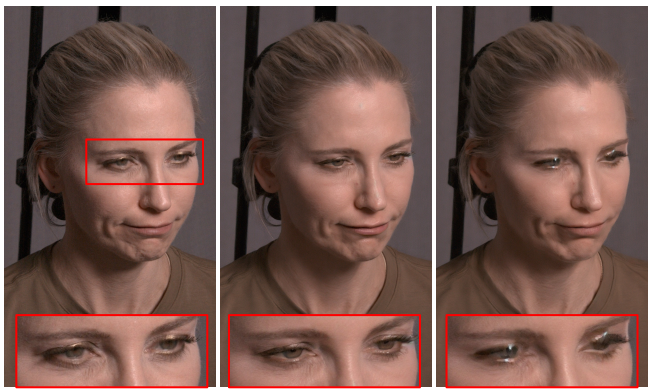
caused by inaccurate intrinsic decomposition on a small dataset. Both models are trained using the same data, and we evaluate all methods across four validation configurations: *Novel Light*, *Novel Expression*, *Novel View*, and *Novel Light+Expression+View (LEV)*, introducing complexities in one or three dimensions. We quantify differences between the relit images and ground truth by computing PSNR, SSIM, LPIPS [Zhang et al. 2018], and FLIP [Andersson et al. 2020], as detailed in Tab. 1. Visual comparisons of these methods with the ground truth are shown in Fig. 15.

The quantitative results reveal that the ControlNet-based method underperforms in preserving spatial details and color accuracy, reflected by the lowest PSNR, SSIM, and FLIP scores, because it only leverages sparse spatial information. Moreover, it is less sensitive to color shifts, leading to noticeable color differences. The U-Net-based method aligns better spatially with the flat-lit image but tends to produce blurrier results, which improves pixel-level metrics but diminishes photorealism as measured by LPIPS. Our method achieves the best quantitative outcomes, significantly enhancing lighting accuracy, color fidelity, and overall image quality.

4.4 Ablation Studies

4.4.1 Ablation studies on diffusion-based relighting. We conducted various ablation studies to evaluate the key technical designs in our diffusion-based relighting model. We first try different methods to integrate lighting control by providing the model with a rough shading estimation. We generate a shading map akin to our strategy in the ControlNet-based relighting model, i.e. dot product between light direction and photometric normal. A visualization of the normal and shading map can be seen in Fig. 14. We then concatenate the latent of the shading map with the flat-lit latent and random noise as input for the diffusion U-Net. We compare the lighting condition with shading map and SH encoding and results are shown in Fig. 14 and Tab. 2. The numerical results indicate that the shading-map-conditioned model slightly outperforms ours due to the use of its richer shading and geometric detail as spatial control. However, capturing photometric normals in multi-view dynamic settings is challenging due to hardware constraints. Notably, our method does not require time-multiplexed lighting during performance capture, allowing for more comfortable flat lighting and reducing costs by eliminating the need for high-speed global-shutter cameras. Therefore, our approach can achieve comparable performance without relying on photometric normals, making it suitable for more applications.

To determine which factor contributes the most to color fidelity and overall result quality, we disabled the training strategy that uses pyramid noise ("w/o pyramid noise") and removed the initialization that uses the pre-trained Stable Diffusion ("w/o pre-trained"). Quantitative results are detailed in Tab. 2. The model performance noticeably degrades without the pyramid noise, as reflected by a decrease in all evaluation metrics. A visual comparison in Fig. 5 further demonstrates that pyramid noise enables more accurate predictions of darker pixels with less color shifting, which likely explains why our method better preserves color fidelity. Additionally, fine-tuning the network based on pre-trained stable diffusion is crucial as it provides strong prior knowledge to generating photorealistic relit results. Without pre-trained weights, the model tends to generate more artifacts, as illustrated in Fig. 6.



GT Ours w/o pre-trained

Fig. 6. **Ablation study on using pre-trained model weights.** We ablate our method by randomly initializing the model weights instead of loading from a pre-trained model (w/o pre-trained). Results show that loading from a pre-trained stable diffusion model helps generate photorealistic results with fewer artifacts.

Table 3. Quantitative comparison on different models of dynamic 3DGS.

	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline model [Jung et al. 2023]	30.202	0.876	0.254
(a) Partition w/o K-frame	31.717	0.893	0.239
(b) Partition w/ K-frame	31.114	0.901	0.218
(c) Partition w/ K-frame & L_{reg} (Ours)	**31.979**	**0.903**	**0.216**



Fig. 7. **HDRI relighting examples**. We can relight a subject under any HDR environment by initially incorporating the HDRI into multiple OLATs or area lights. Then, we relight the subject individually using each OLAT through our diffusion-based relighting method, before compositing the final result in the image space.



Fig. 8. **Generalization capacity.** We test the generalizability of our model by evaluating on an unseen subject. Our subject-specific model best preserves identity when trained and tested on the same subject. Interestingly, the single-subject model can relight a new subject but may inherit features like skin texture and lip color from the training subject. If trained on three other subjects, the model can better preserve the new subject's unique features but still shows noticeable changes in color and fine details.

4.4.2 Ablation studies on scalable dynamic 3D Gaussian Splatting. To demonstrate the effectiveness of our scalable dynamic 3DGS, we compare it with three baselines: (1) the original deformable 3DGS model [Jung et al. 2023]; (2) a model partitioning long sequences for separate training without our K-frame initialization (K-frame) or deformation offset regularization (L_{reg}); and (3) a model using partitioning and K-frame but without L_{reg}. We evaluate by holding out 10% of the captured views from one subject's dynamic sequences, using quantitative metrics. Tab. 3 shows improvements in reconstruction quality and consistency with the proposed strategies.

5 REAL-WORLD HDRI RELIGHTING

Since our method infers OLAT images from any viewpoint that closely match the ground truth, it can reproduce real-world lighting conditions from HDRI maps, following [Debevec et al. 2000]. To evaluate this, we capture reference images of one trained subject in complex indoor and outdoor environments, using HDRI maps to reconstruct the lighting. For comparison, we render the flat-lit subject from similar viewpoints and input these into our relighting model. For each HDRI map, we generate two results: one using the OLAT-based model and the other using the area-light model. In OLAT-based relighting, we linearly combine the results of 123 OLATs, while in area-light relighting, we approximate the HDRI map using 15 Spherical Gaussians (SGs), infer for each SG, and then combine their results. To account for spectral differences between LED and real-world lighting, we capture a color chart in both environments and compute a constant 3-channel scaling factor to correct the relit image. Fig. 16 shows that our results are comparable to reference images under complex lighting conditions.

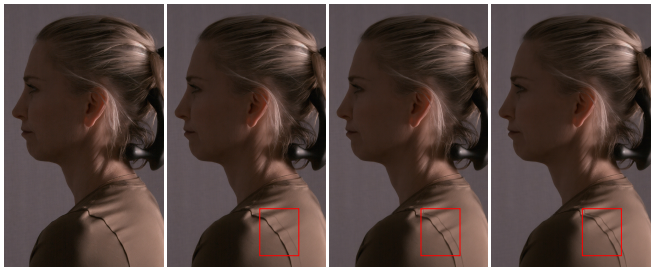


Fig. 9. **Limitation.** The post-processing keyframe interpolation may cause artifacts, such as double strips on shirts, due to inaccurate warping from fast camera motion, which affects the precision of optical flow estimation.

6 MODEL GENERALIZATION

Although our model is primarily trained on the subject-specific data, which limits its generalization capacity for novel subjects, the model learns to translate from flat to directional lighting, a process that is somewhat subject-agnostic. However, subject-specific information from the model, especially high-frequency details, may remain in the results, as in Fig. 8. To assess whether training on multiple subjects' data reduces such subject-specific information, we train a model using data from three subjects. This model shows improved performance on a novel subject, better preserving the subject's identity, though color shifts are still prominent. This suggests the potential for training a general relighting model using more subjects.

7 LIMITATIONS

Our technique has a few limitations; for one, we do not completely resolve the temporal consistency issue of image-based diffusion models due to the absence of video training data, and our optical flow post-processing occasionally introduces artifacts during rapid movements, as illustrated in Fig. 9. Emerging video diffusion models hold promise for solving this issue. Secondly, our subject-specific training is not designed to generalize to unseen subjects and can alter identity features in the relit results, as shown in our extension. We believe this limitation could be addressed by training on a diverse, multi-person dataset and/or implementing identity disentanglement techniques. Lastly, we would need to add full-body training data for our relighting technique to be applied to full-body performances.

8 CONCLUSION

The proposed method shows promising results in relighting free-viewpoint flat-lit facial performances to various lighting conditions using subject-specific OLAT training data. It demonstrates enhanced performance in replicating eye highlights, skin luster, and hair details compared to prior image-based methods, which start with a single diffuse lighting condition. Leveraging an underlying diffusion model allows us to achieve well-conditioned results while maintaining the original subject's appearance and identity by finetuning on subject-specific data. These findings suggest a promising direction for advancing relightable volumetric capture production and facilitating postproduction relighting of any flat-lit footage.

ACKNOWLEDGMENTS

We thank Stephan Trojansky and Jeffrey Shapiro for initiation and ongoing support; Gabriel Dedic for invaluable help on imaging; Jon Milward, Samuel Price for stage operation; Connie Siu, Amir Shachar, Kevin Williams, Marc Thorineaux, Alex Chun for stage coordination; Rafe Sacks, Scott Wilson for data preprocessing; Sunny Koya, Daniel Heckenberg, Pablo Delgado, Elliot Chow for technical support; the software department for stage software maintenance; Brian Tong for video editing, and our performers: Kevin Williams, Mary Carr, Zorianna Kit, and Naz Lang.



Fig. 10. **Novel facial expressions under a rotated directional lighting.** Each row shows every 12th frame of a performance sequence of 120 frames for an individual actor, having a novel facial expression every frame.



Fig. 11. **Unified lighting control combining directional and area lights.** The image sequence shows an animated area light on a performance sequence with a fixed view. The area light animates linearly back and forth between a directional light (0 degree area light) and a 360 degree area light. In order to compute the reflectance fields of an area light, weighted captured OLAT images are linearly combined.

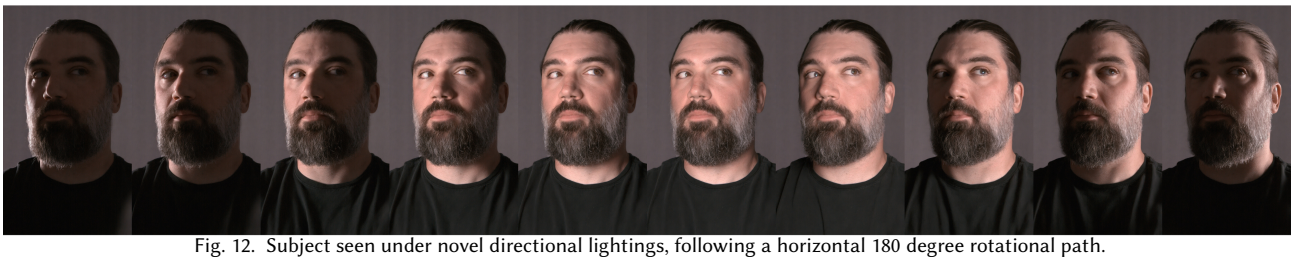


Fig. 12. Subject seen under novel directional lightings, following a horizontal 180 degree rotational path.

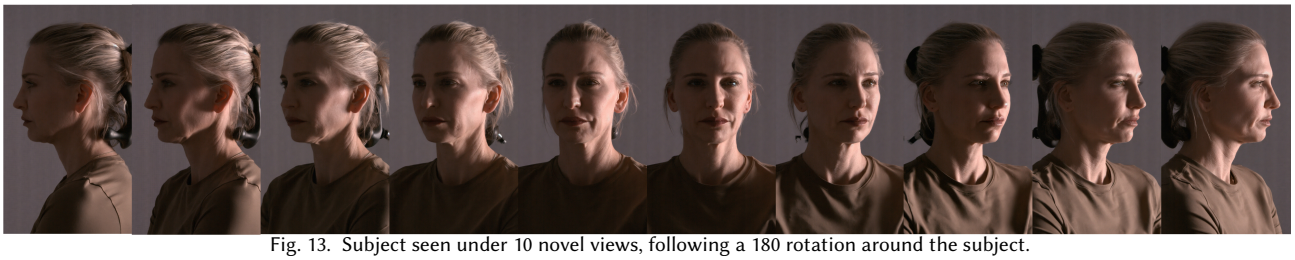


Fig. 13. Subject seen under 10 novel views, following a 180 rotation around the subject.



Input | Photometric normal | Shading map | Ours w/ shading map | Ours w/ SH encoding | GT

Fig. 14. **Comparisons between different lighting condition methods, shading map and SH encoding.** We also visualize the photometric normal and shading map, where the shading map is the dot product between photometric normal and light direction. The two methods achieve comparable visual quality, whereas the method conditioned on the shading map requires the photometric normal as additional input.

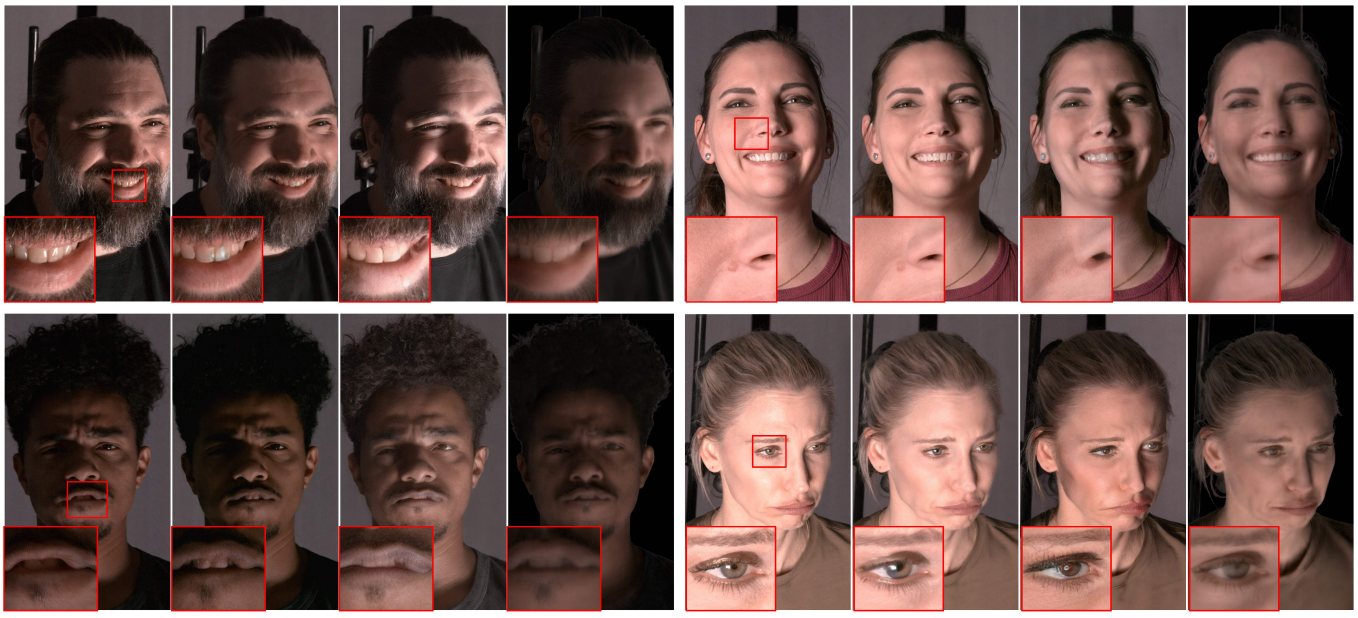


Fig. 15. **Comparison of the results generated by ControlNet-based relighting, U-Net-based relighting, and our method**. Note that the U-Net-based model is trained with foreground masks to focus on foreground relighting and generate black background while the other two methods never use masks for training; they are instead generated using 3DGS. We remove the backgrounds from the results of other methods when calculating metrics.

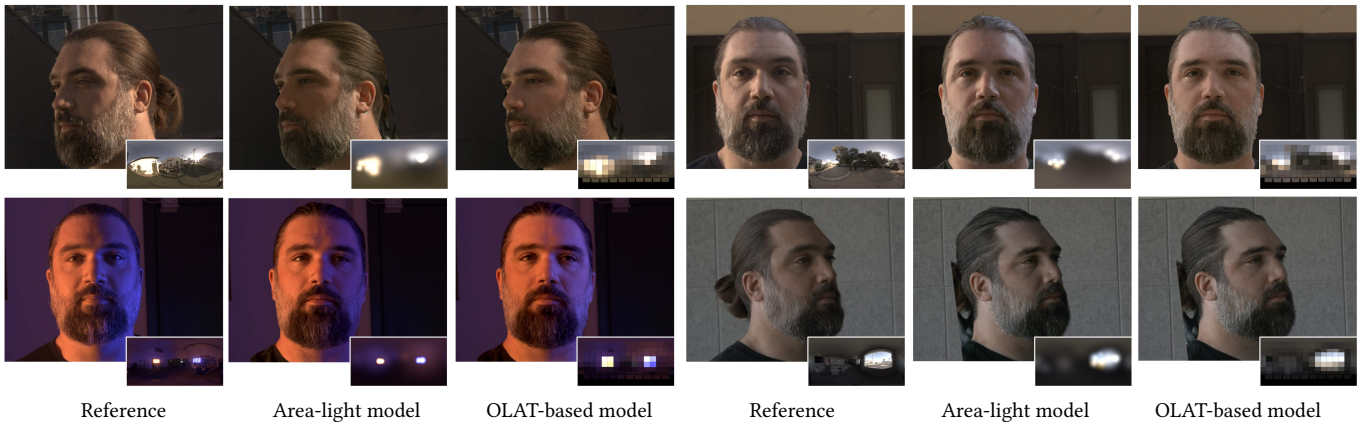


Fig. 16. **Real-world HDRI relighting**. We compare the HDRI relighting results from our method with reference images captured in four real-world environments. For each HDRI relighting example, we present the reference image, two relighting results (one using the area-light model and the other using the OLAT-based model), along with the captured HDRI map and its approximations of 15 Spherical Gaussians (SGs) and 123 OLATs overlaid on the image. Both models achieve results comparable to the reference images.

REFERENCES

Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D Fairchild. 2020. FLIP: A Difference Evaluator for Alternating Images. *Proc. ACM Comput. Graph. Interact. Tech.* 3, 2 (2020), 15–1.

Jonathan T Barron and Jitendra Malik. 2014. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence* 37, 8 (2014), 1670–1687.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P. A. Lensch. 2021. NeRD: Neural Reflectance Decomposition from Image Collections. arXiv:2012.03918 [cs.CV]

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Paul Debevec and Chloe LeGendre. 2022. HDR Lighting Dilation for Dynamic Range Reduction on Virtual Production Stages. arXiv:2205.07873 [cs.GR] https://arxiv.org/abs/2205.07873

Paul E. Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, Judith R. Brown and Kurt Akeley (Eds.). ACM, 145–156. https://doi.org/10.1145/344779.344855

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. 2023. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12736–12746.

Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. 2006. Relighting human locomotion with flowed reflectance fields. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques* (Nicosia, Cyprus) *(EGSR '06)*. Eurographics Association, Goslar, DEU, 183–194.

Hugging Face. 2024. Diffusers Documentation. https://huggingface.co/docs/diffusers/index Accessed: 2024-05-19.

Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. 2023. Relightable 3D Gaussian: Real-time Point Cloud Relighting with BRDF Decomposition and Ray Tracing. *CoRR* abs/2311.16043 (2023). https://doi.org/10.48550/ARXIV.2311.16043 arXiv:2311.16043

Michael Goesele, Hendrik P. A. Lensch, Jochen Lang, Christian Fuchs, and Hans-Peter Seidel. 2004. DISCO: acquisition of translucent objects. *ACM Trans. Graph.* 23, 3 (aug 2004), 835–844. https://doi.org/10.1145/1015706.1015807

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

Nicholas Guttenberg. 2023. Diffusion with Offset Noise. https://www.crosslabs.org/blog/diffusion-with-offset-noise. Accessed: 2024-05-15.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sỳkora. 2019. Stylizing video by example. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–11.

HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir Navab, and Benjamin Busam. 2023. Deformable 3D Gaussian Splatting for Animatable Human Avatars. arXiv:2312.15059 [cs.CV]

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems* 35 (2022), 26565–26577.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2023. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145* (2023).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023), 139:1–139:14. https://doi.org/10.1145/3592433

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. 2024. SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting. arXiv:2402.18848 [cs.CV] https://arxiv.org/abs/2402.18848

Ha A Le and Ioannis A Kakadiaris. 2019. Illumination-invariant face recognition with deep relit face images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2146–2155.

Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Ryan Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul E. Debevec. 2020. Learning Illumination from Diverse Portraits. In *SA '20: SIGGRAPH Asia 2020 Technical Communications, Virtual Event, Republic of Korea, December, 2020*. ACM, 7:1–7:4. https://doi.org/10.1145/3410700.3425432

Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. 2018. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*. 453–468.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2023a. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812* (2023).

Zhong Li, Liangchen Song, Zhang Chen, Xiangyu Du, Lele Chen, Junsong Yuan, and Yi Xu. 2023b. Relit-NeuLF: Efficient Relighting and Novel View Synthesis via Neural 4D Light Field. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 7007–7016. https://doi.org/10.1145/3581783.3612160

Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. 2023. NeRO: Neural Geometry and BRDF Reconstruction of Reflective Objects from Multiview Images. *ACM Trans. Graph.* 42, 4 (2023), 114:1–114:22. https://doi.org/10.1145/3592134

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4990–4998.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2023. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. *CoRR* abs/2308.09713 (2023). https://doi.org/10.48550/ARXIV.2308.09713 arXiv:2308.09713

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. *Rendering Techniques* 9, 10 (2007), 2.

Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M Patel. 2024. Holo-Relighting: Controllable Volumetric Portrait Relighting from a Single Image. *arXiv preprint arXiv:2403.09632* (2024).

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

Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip L. Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe LeGendre, Wan-Chun Ma, Ryan S. Overbeck, Thabo Beeler, Paul E. Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Ryan Fanello. 2020. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Trans. Graph.* 39, 6 (2020), 259:1–259:21. https://doi.org/10.1145/3414685.3417814

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934 [cs.CV]

Jeffry S Nimeroff, Eero Simoncelli, and Julie Dorsey. 1995. Efficient re-rendering of naturally illuminated environments. In *Photorealistic Rendering Techniques*. Springer, 373–388.

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul E. Debevec, and Sean Ryan Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* 40, 4, Article 43 (jul 2021), 21 pages. https://doi.org/10.1145/3450626.3459872

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul E. Debevec. 2007. Post-production facial performance relighting using reflectance transfer. *ACM Trans. Graph.* 26, 3 (2007), 52. https://doi.org/10.1145/1276377.1276442

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. 2023. DiFaReli: Diffusion face relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22646–22657.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10619–10629.

Pramod Rao, Mallikarjun B R, Gereon Fox, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Ayush Tewari, Christian Theobalt, and Mohamed Elgharib. 2022. VoRF: Volumetric Relightable Faces. *BMVC* (2022).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition. 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

Shunsuke Saito, Tomas Simon, Junxuan Li, and Giljoo Nam. 2023. Relightable Gaussian Codec Avatars. *CoRR* abs/2312.03704 (2023). https://doi.org/10.48550/ARXIV.2312.03704 arXiv:2312.03704

Kripasindhu Sarkar, Marcel C. Bühler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo F. U. Gotardo, Thabo Beeler, and Abhimitra Meka. 2023. LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces. In *SIGGRAPH Asia 2023 Conference Papers, SA 2023, Sydney, NSW, Australia, December 12-15, 2023* (<conf-loc>, <city>Sydney</city>, <state>NSW</state>, <country>Australia</country>, </conf-loc>) *(SA '23)*, June Kim, Ming C. Lin, and Bernd Bickel (Eds.). ACM, Article 42, 11 pages. https://doi.org/10.1145/3610548.3618210

Yoichi Sato, Mark D. Wheeler, and Katsushi Ikeuchi. 1997. Object shape and reflectance modeling from observation. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 379–387. https://doi.org/10.1145/258734.258885

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=M3Y74vmsMcY

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.

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

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

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. 2021. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 7495–7504. https://doi.org/10.1109/CVPR46437.2021.00741

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

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. 2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–12.

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

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.

Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, Christian Theobalt, et al. 2021b. Monocular reconstruction of neural face reflectance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4791–4800.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2023b. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015* (2023).

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2023a. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. arXiv:2106.10689 [cs.CV]

Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2008. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE transactions on pattern analysis and machine intelligence* 31, 11 (2008), 1968–1984.

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus H. Gross. 2006. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. Graph.* 25, 3 (2006), 1013–1024. https://doi.org/10.1145/1141911.1141987

John Whitaker. 2024. Multi-Resolution Noise for Diffusion Model Training. https://wandb.ai/johnowhitaker/multires_noise/reports/Multi-Resolution-Noise-for-Diffusion-Model-Training--VmlldzozNjYyOTU2. Accessed: 2024-05-15.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. *CoRR* abs/2310.08528 (2023). https://doi.org/10.48550/ARXIV.2310.08528 arXiv:2310.08528

Yingyan Xu, Prashanth Chandran, Sebastian Weiss, Markus Gross, Gaspard Zoss, and Derek Bradley. 2023a. Artist-Friendly Relightable and Animatable Neural Heads. arXiv:2312.03420 [cs.CV] https://arxiv.org/abs/2312.03420

Yingyan Xu, Gaspard Zoss, Prashanth Chandran, Markus Gross, Derek Bradley, and Paulo Gotardo. 2023b. ReNeRF: Relightable Neural Radiance Fields with Nearfield Lighting. In *International Conference on Computer Vision (ICCV)*.

Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. 2023. Towards Practical Capture of High-Fidelity Relightable Avatars. arXiv:2309.04247 [cs.CV] https://arxiv.org/abs/2309.04247

Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. 2022. NeILF: Neural Incident Light Field for Physically-based Material Estimation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI (Lecture Notes in Computer Science, Vol. 13691)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 700–716. https://doi.org/10.1007/978-3-031-19821-2_40

Chong Zeng, Guojun Chen, Yue Dong, Pieter Peers, Hongzhi Wu, and Xin Tong. 2023. Relighting Neural Radiance Fields with Shadow and Highlight Hints. In *ACM SIGGRAPH 2023 Conference Proceedings* (Los Angeles, CA, USA) *(SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, Article 73, 11 pages. https://doi.org/10.1145/3588432.3591482

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation. *arXiv preprint arXiv:2402.11929* (2024).

Jingyang Zhang, Yao Yao, Shiwei Li, Jingbo Liu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. 2023b. NeILF++: Inter-Reflectable Light Fields for Geometry and Material Estimation. *CoRR* abs/2303.17147 (2023). https://doi.org/10.48550/ARXIV.2303.17147 arXiv:2303.17147

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2024. IC-Light GitHub Page.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. 2021a. Neural Light Transport for Relighting and View Synthesis. *ACM Trans. Graph.* 40, 1, Article 9 (jan 2021), 17 pages. https://doi.org/10.1145/3446328

Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. 2021b. NeRFactor: neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics* 40, 6 (Dec. 2021), 1–18. https://doi.org/10.1145/3478513.3480496

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.

M. Zwicker, H. Pfister, J. van Baar, and M. Gross. 2002. EWA Splatting. *IEEE Transactions on Visualization and Computer Graphics* 8, 3 (07/2002-09/2002 2002), 223–238.

A APPENDIX

A.1 Method

A.1.1 3D Gaussian Splatting Reconstruction with Background Segmentation. As some post-composition like the generation of HDR environment lighting or training data of our U-Net based network necessitates foreground matting, we propose employing the clean background captured by our stage as the known background in the 3DGS reconstruction. Implementation details are provided here.

In the 3D Gaussian Splatting reconstruction, we start with a set of multi-view images, along with corresponding camera calibration and sparse point cloud generated by Metashape [1]. Following [Kerbl et al. 2023], we initialize the 3D Gaussians with the sparse point cloud and then optimize the Gaussians, which are defined by a center position $\mathbf{x}$, a covariance matrix Σ obtained from rotation $\mathbf{r}$ and scaling $\mathbf{s}$, opacity α, and color $\mathbf{c}$ represented by spherical harmonic coefficients. Given a view direction transformation matrix W, the convariance matrix Σ can be projected to camera space following [Zwicker et al. 2002] and converted to:

$$\Sigma' = JW\Sigma W^T J^T, \tag{5}$$

where J is the Jacobian of the affine approximation of the projective transformation and Σ is expressed as follows:

$$\Sigma = RSS^T R^T \tag{6}$$

with the rotation and scaling vectors $\mathbf{r}$ and $\mathbf{s}$ transformed into matrices R and S. The color of each pixel p on the image is defined as alpha blending of 3D Gaussians that covers this pixel and are sorted in depth:

$$\begin{aligned} c_p &= \sum_{i \in N} T_i \alpha_i c_i \\ \alpha_i &= \sigma e^{\frac{1}{2}(p-q_i)^T \Sigma' (p-q_i)}, \end{aligned} \tag{7}$$

where T_i is the transmittance defined as $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ and q_i is the 2D projection position of the ith Gaussian. To optimize Gaussians, the loss function is L1 combined with a D-SSIM term same as the original loss in [Kerbl et al. 2023].

To obtain precise alpha mattes for the foreground subject, we employ the captured clean plate as a known background. This facilitates the separation of foreground and background Gaussians, directing the optimization to focus only on the foreground Gaussians. Therefore, the final pixel color is computed as:

$$c'_p = \sum_{i \in N} T_i \alpha_i c_i + (1 - \sum_{i \in N} T_i \alpha_i) b_p, \tag{8}$$

where b_p is the pixel color on the clean plate. Fig. 17 shows the examples of high-quality alpha mattes of the reconstructed foreground subject by 3D GS, aided by the clean plate.

A.2 Additional Results

Fig. 18 shows complex light transport effects of our DifFRelight method. The complicated lighting effects partially originate from the stable diffusion prior, but most are learned by our model during the fine-tuning process from subject specific real-world reflectance data applying various lighting conditions. Accurately modeling these effects with physically based rendering is extremely challenging and requires complex computational models to simulate light transmittance and subsurface scattering in multi-layered materials. For example, the human skin itself consists of multiple layers—epidermis, dermis, elastic cartilage, and fatty tissue—that scatter light diffusely. Additionally, various sized blood vessels pervade the ear tissue, which absorbs light in the red spectrum producing the reddish hue from transmitting light.

[1] https://www.agisoft.com/

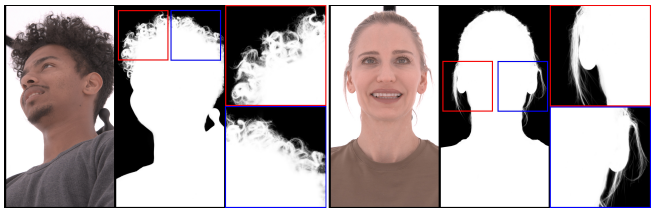


Fig. 17. Alpha mattes produced by our 3D GS reconstruction utilizing the clean plate. Each example includes the input image along with the generated alpha mattes. Zoomed-in details highlight specific elements, such as hair strands, that are distinctly segmented through the use of the clean plate.



Fig. 18. Complex light transport effects, showing scattering and transmitting light through translucent ear together with sharp specular reflection at grazing angles and subtle reflectance of hair strands, achieved with our method on unseen views and light directions.

A.3 Comparison

Next we provide the implementation details of the U-Net based relighting baseline and add two new comparisons with zero-shot relighting methods.

A.4 U-Net Training Pipeline

To train our U-Net [Ronneberger et al. 2015] based architecture for a comparison, similar to the diffusion based training pipeline, we use pairs of flat-lit captured images and captured OLAT images, along with camera space light direction as input matching the capture. In addition we also made use of the masks we acquire from the GS reconstruction to help our model skip learning the background. This helps convergence as the transform between the flatlit image and the target background is more ambigious than the foreground subject. Encoder architecture of the U-Net is based on the ResNet-50 [He et al. 2016] architecture. Matching this architecture allows weight initializations from a pre-trained model, however we elected to not use these weights as frozen. We observed that convergence is relatively faster with this initialization strategy, but a random weight initialization yields a similar convergence at longer training

times. For the decoder, we use style modulations described in StyleGAN2 [Karras et al. 2020]. Style code is used for light directions, where a multi-layer perceptron with two hidden layers with 128 neurons each hidden layer, map three dimensional light directions into the shape necessary per CNN layer of the decoder, for modulations described in the StyleGAN2 paper. Decoder weights are not initialized from a StyleGAN2 checkpoint but instead initialized randomly. We train this model per performer for about 30 epochs for convergence and it takes considerably more time to converge than the diffusion based training method. Multi GPU training is used on 4 NVIDIA A10G cards, with 2k crops from the original 4k resolution. Twenty percent of the dataset is held matching our other training configurations for comparison. Training takes multiple weeks compared to less than a day, for considerably inferior outputs. Different to the diffusion based training pipeline, we use the mask from the GS reconstruction to take out the background before the L2 norm between target OLAT image from the capture and the prediction. Adam optimizer is used for optimization with a learning rate of 3e-4. For reconstructing the eye highlights increasing the exposure by multiplying the intensity of the target image by two helped getting better results. Inference times are interactive and it was possible to build an interactive application that enables the user to pick light directions as the model shows predictions at interactive speeds of 16 to 24 frames per second on playback with commercial CUDA enabled graphics cards. However this improvement in inference time isn't justified considering the quality of results and the training times needed for convergence. We believe better results may be obtained given more variation in data and with a bigger dataset to build a prior with this method. However Stable Diffusion based method already has a prior we could employ and the diffusion based model seems to be more expressive for high frequency details. U-Net method also suffers from temporal stability where the model is not as confident about the shadow shapes or placement of some highlights. We believe this may improve with denser capture configurations of lights and cameras on our stage, which would also make the capture process more expensive. Further exploration using adversarial loss is left out, as this training method is harder to tune for reasonable convergence during training and it would not make sense for a production tool where we may need to make this work per performer in a reproducible manner without much supervision.

A.5 Comparison with Zero-Shot Relighting Methods

We first provide a comparison with SwitchLight [Kim et al. 2024], one of the most recent state-of-the-art traditional relighting methods, in Fig. 20, showing results from our method along with results generated using the same lighting information and input flat-lit image. Due to finetuning with subject specific real data and not relying on a rendering model to produce final imagery, generated skin colors with our method look more natural and detailed, as well as specular reflections on skin and eyes are a bit more distinctive. The physically based rendering approach in Switchlight emphasizes the difficulty in reproducing the color tone of the HDRI skin reflections in a natural way. Learning a general model to reconstruct geometric properties as done in Switchlight is a great challenge and comes with caveats in normal and albedo quality. Though, SwitchLight in comparison generalizes to any subject and can be used as a general relighting tool.

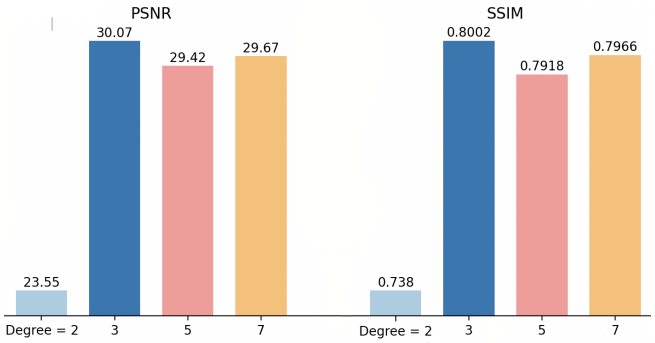


Fig. 19. We test the affect of different SH degrees to the PSNR and SSIM of the relit results. We use degree = 3 in all of other experiments.

Recently, IC-Light [Zhang et al. 2024] was released as a zero-shot relighting algorithm based on Stable Diffusion. It can harmonize the lighting of a foreground image with a given background, where the background controls the lighting of the output image.

We show qualitative results in 21 and 22. IC-Light is temporally inconsisent and is not able to produce precise lighting from a given direction. Additionally, because it only takes in a background image, it cannot be used to relight a subject given a full 360 HDRI map. Regardless of the prompt chosen, or the background chosen, IC-Light fails to add reflections to the subject's eyes - resulting in a cloudy, lifeless appearance. By contrast, our method accurately reproduces specularities in high fidelity. These results showcase the importance of our subject-specific formulation.

A.6 Additional Ablation Studies

A.6.1 Ablation on SH degree. One hyper-parameters in our method is the SH degree used for encoding the light direction into text embedding. We experiment with different SH degrees on one subject and plot the PSNR and SSIM metrics in Fig. 19 under SH degree = 2, 3, 5, 7. It can be see that when degree equals to 2, our method struggles to reason on the light direction and produce high quality results, which is likely due to the low frequency band of the light condition. When degree is larger than 2, the model achieves much better performance and achieves the best when the degree is 3, which is the final settings we choose in all of other experiments. This specific value roughly aligns with the angular sampling rate of all the possible lighting directions in the dataset. Therefore, choosing a proper SH degree is important for training a good personalized relighting model.

A.6.2 Ablation Studies on Scalable Dynamic 3DGS. We show the visual results for the ablation study on the proposed dynamic 3DGS in Fig. 23.



Fig. 20. A comparison between our method and the method Kim et al. [2024], implemented in the product named SwitchLight Studio, with the last row showing the HDRI environment lighting.

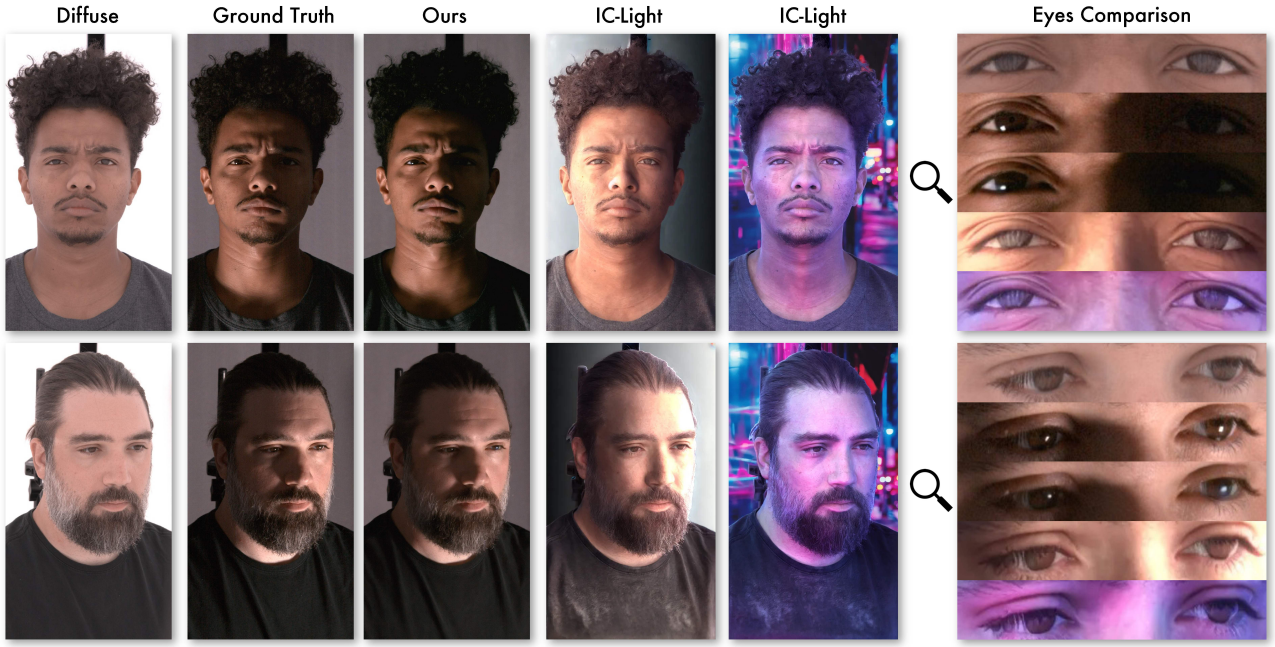


Fig. 21. A comparison between our method and IC-Light [Zhang et al. 2024]. IC-Light fails to produce highlights and reflections on eyes, and forehead wrinkles are completely absent. The official implementation's default setting for relighting from the left was used to produce the first IC-Light image, and the included cyberpunk background and prompt were used for the rightmost image.

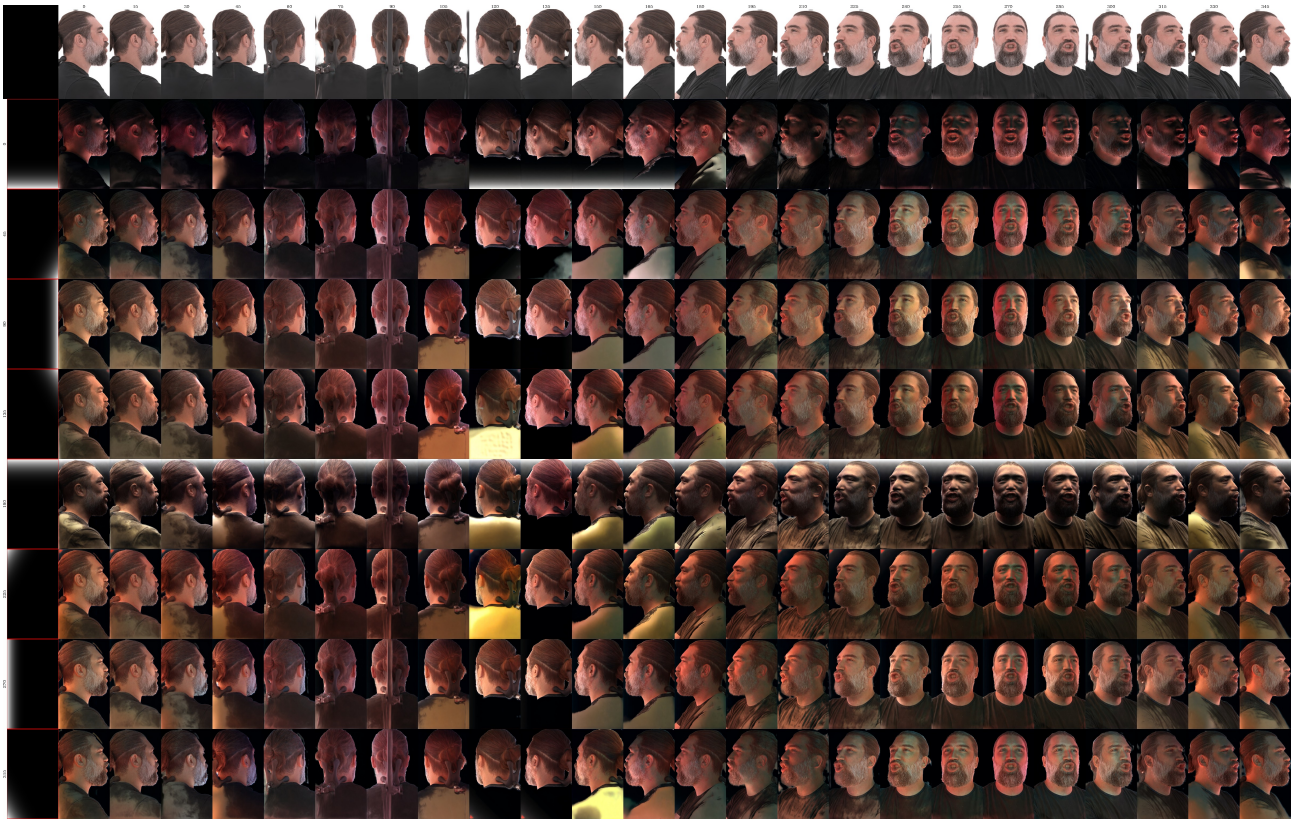


Fig. 22. IC-Light is temporally inconsistent, and not able to light the subject precisely. The y-axis is the angle of the light, and the leftmost column is the background image used. The x-axis represents the angle in which our subject is rotated in 3d.



Fig. 23. Visual comparison on different dynamic 3DGS models: (a) baseline model [Jung et al. 2023], (b) partitions long sequences into segments for separate training, (c) uses our K-frame initialization without deformation offset regularization, and (d) combines K-frame initialization with deformation offset regularization. Orange and blue frames denote transition points overlapped by two segments in (b), (c), and (d). The orange frame is learned by the previous segment model, and the blue frame comes from the next segment model, indicating better cross-segment consistency with greater similarity. In, GT and (a), they are duplicates and identical.