This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Comprehensive Relighting: Generalizable and Consistent Monocular Human Relighting and Harmonization

Junying Wang[1†] Jingyuan Liu[2] Xin Sun[2] Krishna Kumar Singh[2] Zhixin Shu[2]
He Zhang[2] Jimei Yang[3] Nanxuan Zhao[2] Tuanfeng Y. Wang[2] Simon S. Chen[2]
Ulrich Neumann[1] Jae Shin Yoon[2]

[1]University of Southern California [2]Adobe Research [3]Runway



Figure 1. We introduce Comprehensive Relighting, a generalizable and consistent model for relighting and harmonization, which controls the lighting property from a single image or video of humans with arbitrary body parts. Given target lighting coefficients, *e.g.*, Spherical harmonics (second), background scenes (third), or their combination (fourth), our model performs consistent and harmonized relighting.

Abstract

This paper introduces Comprehensive Relighting, the first all-in-one approach that can both control and harmonize the lighting from an image or video of humans with arbitrary body parts from any scene. Building such a generalizable model is extremely challenging due to the lack of dataset, restricting existing image-based relighting models to a specific scenario (e.g., face or static human). To address this challenge, we repurpose a pre-trained diffusion model as a general image prior and jointly model the human relighting and background harmonization in the coarse-to-fine framework. To further enhance the temporal coherence of the relighting, we introduce an unsupervised temporal lighting model that learns the lighting cycle consistency from many real-world videos without any ground truth. In inference time, our temporal lighting module is combined with the diffusion models through the spatio-temporal feature blending algorithms without extra training; and we apply a new guided refinement as a post-processing to preserve the high-frequency details from the input image. In the experiments, Comprehensive Relighting shows a strong generalizability and lighting temporal coherence, outperforming existing image-based human relighting and harmonization methods.

1. Introduction

Light is the key component that determines how a person looks, which is often defined by a specific time and space, where revisiting such a unique moment gives us the opportunity to experience authentic telepresence sensations. In this paper, we introduce a generalizable human relighting model that can control the lighting from an image or video of humans with arbitrary body parts (Fig. 3), which are well-harmonized with a conditioning space (*i.e.,* background image) as shown in Fig. 1.

As shown in Fig. 2, existing image-based relighting methods face two main problems, lack of generalizability and controllability. First, they are designed for a specific scenario, *e.g.*, only for face illumination or static humans [33, 61, 68] mainly due to the scarcity of large-scale

[1†]This work is partially done during an internship at Adobe Research.

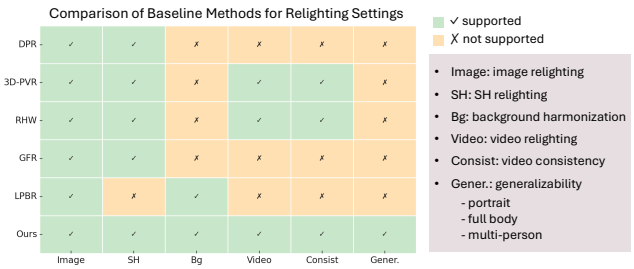


Figure 2. Comparison of various baseline methods for relighting settings and functionalities.

relighting datasets: Precise acquisition of the appearance of a static person under assorted lighting conditions requires specialized setups such as LightStage [22, 33] or expensive graphics simulation [26], which are not scalable, particularly for video contents. Learning from such limited data induces weak generalization of the model to diverse scenes. Second, most relighting algorithms struggle to effectively model multiple light sources. Typically, these algorithms are restricted to a single lighting control from either background image (*e.g.* high dynamic range lighting environment map [22]) or target lighting parameters (*e.g.* Spherical harmonics [24]). These problems inhibit the production-level application that requires general use cases.

To overcome these challenges, we propose an effective approach to achieve all-in-one relighting by utilizing a diffusion model—a general image prior that learns massive visual data with diverse lighting conditions; and repurpose this prior specialized for a human relighting and harmonization in a coarse-to-fine framework: A pretrained latent diffusion model [41] learns from limited datasets of [1]static humans to jointly perform the fine-grained relighting and harmonization from two multi-modal lighting variables: the coarse shading estimate and conditioning background scene. The coarse shading estimate can be "computed" from conditioning lighting parameters (*i.e.,* Spherical harmonics) without a neural network, and therefore, it is generalizable. The diffusion model is required to learn only the residual portion (*e.g.,* fine self-occluded shadow), which is more generalizable than direct relighting (*e.g.,* [68]). The pre-trained image prior helps with understanding the properties of the general background scenes, enabling the generalization of the background harmonization.

Our coarse-to-fine framework, however, still introduces problems: the diffusion model that learns to generate an image without temporal context produces significant temporal artifacts (*e.g.*, sudden changes of lighting distribution even for consecutive frames). We address this problem by introducing an unsupervised temporal lighting model that learns from many real videos without any ground-truth data to enforce the temporal lighting consistency over frames. This temporal module learns the unsupervised cycle consistency between the relit and many real videos to predict the future lighting distribution from the past, which can be directly combined with our coarse-to-fine relighting components without extra training.

[1]To the best of our knowledge, there exists no public ground truth data for video relighting of dynamic humans, and therefore, fine-tuning an existing video diffusion model (e.g., [17]) is not a readily available option.



Figure 3. Our model generalizes to various body parts (portrait, half-body, full-body, multiperson) for relighting and harmonization, with lighting control variables shown in the insets.

In inference time, we further push the temporal coherence of the relighting and harmonization by formulating a recurrent prediction pipeline. The generation at the current time step is conditioned onto the temporal module for the one at the next time. The features from the lighting and temporal control modules are spatially and temporally blended to enforce temporal coherence while improving the structure of the lighting distribution. Finally, our guided refinement module enhances the quality of the generated image in a way that preserves the original high-frequency details of the input image.

In the experiments, Comprehensive Relighting demonstrates high-quality relighting and background harmonization results with strong temporal coherence across the lighting, background, and pose changes. It also shows strong generalizability to any unconstrained scenes regardless of body parts, views, and poses, outperforming existing relighting and background harmonization methods.

Our contributions include: (1) To the best of our knowledge, the first approach for joint modeling of relighting and background harmonization; (2) a novel coarse-to-fine framework that enables comprehensive generalization by only learning from limited synthetic and lab-controlled data; (3) unsupervised temporal modeling with lighting cycle consistency from many unconstrained real videos; (4) an effective inference algorithm with spatio-temporal feature blending and guided refinement.

2. Related Work

Image-based Human Relighting While high-quality relighting often requires resource-intensive setups, recent efforts in mobile device relighting aim to address single-

image scenarios. Total Relighting [33] achieves photorealistic effects by leveraging detailed normal maps and albedo as priors. However, it relies on an HDR environment map that accurately matches the scene's real-world lighting, which is not always readily available, particularly in flexible and personalized relighting scenarios. [37, 46] estimate and adjust the Spherical harmonics parameters from images, and others use one-light-at-a-time (OLAT) captures to generate relighting data [49] or estimate reflectance fields [46].

Recent works have focused on diverse relighting scenarios including portrait [10, 19, 20, 25, 30, 33, 46, 58, 67, 68], full-body [9, 22, 26, 47], and object relighting [1, 55, 56]. Most of these works rely on decomposing an image into its intrinsic components, *i.e.,* albedo, normal, and lighting, and therefore, the accuracy of this decomposition directly impacts the quality of the final relighting. While some works [22, 66, 69] use single-image geometry reconstruction for traditional and neural rendering, reconstruction errors are often propagated to relighting results. Tajima et al. [47] tackled domain adaptation with a two-step relighting framework, yet noticeable texture distortion remains due to limited model generalizability. Zhang *et al.* [59] trained a 2D latent-diffusion model, allowing users to manipulate and construct face NeRFs in a zero-shot learning framework without the need for explicit 3D data. DiFaReli [36] utilizes DDIM (Denoising Diffusion Implicit Models [44]) for high-fidelity face relighting. While promising, their methods are constrained by a focus on either specific body parts, limiting their applicability for generalizable.

Background Harmonization Background harmonization seeks to harmonize color, contrast, and style discrepancies between the foreground and background, ensuring composite images appear natural and cohesive. Many existing background harmonization methods [4, 7, 8, 13–15, 23, 51, 52, 70] formulate this problem as image-to-image translation work where a neural network translate an unharmonized foreground image to the harmonized one in the context of the conditioning background image. Recently, Relightful Harmonization [38] introduced methods that harmonize both image style and lighting to match the background scene, and IC-Light [63] further enables flexible illumination control via image diffusion models, guided by text descriptions or background images. While promising, these methods lack explicit lighting control, restricting general applicability and consistent video relighting. Additionally, some ([38]) are limited to specific body parts (*e.g.*, portraits), with constrained support for full-body and multi-person scenarios.

Video Relighting Neural Radiance Fields (NeRF) [32] based methods enable novel view synthesis under varying lighting conditions in videos [43, 50, 53, 69]. Zhang *et al.* [61] achieve portrait video relighting under dynamic illuminations, while Choi *et al.* [6] ensure temporally consistent relit videos. 3D-PVR [3] present a 3D-aware, real-time method to relight videos of talking faces. Relighting4D [5] decomposes the time-varied human body as a set of neural fields of normal, occlusion, diffuse, and specular maps. However, their rendering quality are highly reliant on geometry accuracy. Some works [36] adopt a temporal modeling scheme real-time neural video portrait relighting. ST-NeRF [60] controls dynamic scenes using a neural layered radiance representation that maintains spatial and temporal coherence. Li *et al.* [28] employed multi-view reconstruction for free-viewpoint relighting videos under general illumination. Richardt *et al.* [39] use RGBZ video cameras for video effects, including relighting, but rely on multi-view reconstruction from specialized hardware, making the process highly complex and costly.

3. Method

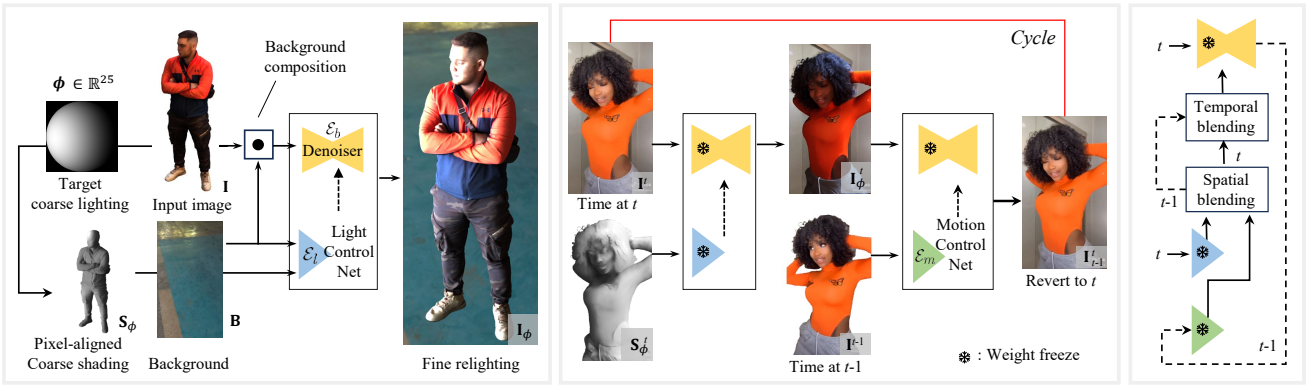
We develop a generalizable and consistent human relighting and harmonization framework using a diffusion model. Fig. 4 illustrates the overview of our framework. Given an input image of humans and control lighting variables, including coarse shading and background image, a diffusion model learns to generate a fine-grained image of the humans under the target lighting, also harmonized with the background (Sec. 3.2). An external temporal lighting module, trained on real videos using unsupervised temporal cycle consistency, is integrated into the diffusion model to enhance temporal lighting coherence (Sec. 3.3). In inference time, we blend the features between the lighting and temporal control modules over time to ensure the relighting results are accurate and temporally coherent, and we apply a guided refinement to prevent the loss of high-frequency details during the denoising process (Sec. 3.4).

3.1. Background: Image-based Relighting

Image-based relighting function can be compactly modeled with a small number of approximated basis of Spherical harmonics (SH) [9, 26, 42] which describes only essential features of illumination on the surface of a 3D sphere based on the following formulation:

$$I(\mathbf{x}) = \rho(\mathbf{x}) \cdot \sum_{l=0}^{k} \sum_{m=-l}^{l} \phi Y_{lm}(\mathbf{n}(\mathbf{x})) \tag{1}$$

where $\phi \in \mathbb{R}^{(k+1)^2}$ denotes the spherical harmonics (SH) coefficients; $Y_{lm}(\mathbf{n})$ are the SH basis functions evaluated at the surface normal $\mathbf{n}$; indices l and m represent the band and order within each band, respectively; and k is the maximum order of spherical harmonics used. While SH is computationally efficient and highly generalizable, it only captures low-frequency lighting, limiting its ability to model fine-grained, high-frequency illumination details. Addition-



(a) Coarse-to-fine relighting and harmonization (b) Temporal modeling with cycle consistency (c) Inference: Spatio-temporal feature blending

Figure 4. **System overview.** (a) Given an input image of humans with coarse lighting and background image, our diffusion model generates the relit images harmonized with background scenes (Sec. 3.2). (b) The external temporal modules learn the temporal cycle consistency from many real-world videos to construct temporal lighting features (Sec. 3.3). (c) In inference time, we blend the features from lighting and temporal modules spatially and temporally to enable coherent and generalizable human relighting (Sec. 3.4).

ally, SH-based global illumination neglects the context provided by background images, often resulting in relit images appearing unnatural when composited with different backgrounds. To overcome these limitations, we propose a coarse-to-fine human relighting and harmonization framework that leverages the strong image prior available from a pre-trained text-to-image diffusion model.

3.2. Coarse-to-Fine Relighting and Harmonization

We generate a fine-grained relit image of a person conditioned on a coarse lighting representation:

$$\mathcal{E}(\mathbf{I}; \mathbf{S}_\phi) = \boldsymbol{z}, \quad \mathcal{D}(\boldsymbol{z}) = \mathbf{I}_\phi \tag{2}$$

where $\mathcal{E}$ is an encoder that generates the latent lighting features, $\boldsymbol{z}$ as a function of the input image $\mathbf{I} \in \mathbb{R}^{w \times h \times 3}$ and a small number of global lighting parameters $\boldsymbol{\phi} \in \mathbb{R}^n$ (*i.e.*, Spherical harmonics where $n = 25$); and $\mathcal{D}$ is the decoder that generates fine-grained relit image $\mathbf{I}_\phi \in \mathbb{R}^{w \times h \times 3}$ from $\boldsymbol{z}$. $\mathbf{S}_\phi \in \mathbb{R}^{w \times h}$ is the pixel-aligned coarse shading map converted from the coarse lighting parameters ϕ as shown in Fig. 4-(a). One approach to obtaining $\mathbf{S}_\phi$ is directly computing it as demonstrated in Eq. 1 where we can compute lighting intensity along with the detected surface normal map from $\mathbf{I}$ given spherical harmonics coefficients $\boldsymbol{\phi}$. Another way, is to use a neural network to convert the surface normal to $\mathbf{S}_\phi$ as a condition of spherical harmonics coefficients $\boldsymbol{\phi}$, *i.e.*, $\mathbf{S}_\phi \leftarrow f(\mathbf{N}; \boldsymbol{\phi})$ where f is the neural shading function which maps the surface normal $\mathbf{N}$ and $\boldsymbol{\phi}$ to the coarse shading. While both approaches are highly generalizable, our experiments indicate that the latter method achieves improved accuracy and better noise correction. Please refer to the Supple. documents for more details about the coarse shading estimation.

To jointly model the relighting and background harmonization, we further condition a target background image $\mathbf{B} \in \mathbb{R}^{w \times h \times 3}$ as the additional lighting sources:

$$\mathcal{E}(\mathbf{I}; \{\mathbf{S}_\phi, \mathbf{B}\}) = \boldsymbol{z}, \quad \mathcal{D}(\boldsymbol{z}) = \mathbf{I}_\phi \tag{3}$$

where the lighting encoder $\mathcal{E}$ learns to capture the intensity and direction of light from $\mathbf{S}_\phi$ while capturing ambient environment lighting and color distribution from $\mathbf{B}$. This enables $\mathcal{D}$ to achieve complete relighting in scenarios involving new target lighting, background, or both. In the subsection, we enable $(\mathcal{D} \circ \mathcal{E})$ using a latent diffusion model.

3.2.1. Fine-grained Relighting Diffusion Model

We enable the fine-grained image relighting using a diffusion model as illustrated in Fig. 4-(a). Our encoder $\mathcal{E}$ in Eq. 3, in practice, is formulated as the composition of two encoders:

$$\mathcal{E} \rightarrow \mathcal{E}_\text{b}(\mathbf{I}; \mathcal{E}_\text{l}(\{\mathbf{S}_\phi, \mathbf{B}\})) = \boldsymbol{z}, \quad \mathcal{D}(\boldsymbol{z}) = \mathbf{I}_\phi \tag{4}$$

where $\mathcal{E}_\text{l}$ encodes lighting control variables, $\{\mathbf{S}_\phi, \mathbf{B}\}$, and $\mathcal{E}_\text{b}$ encodes the conditional variable $\mathbf{I}$ whose visual properties, *e.g.*, semantics and identity, should be preserved in the output along with the controls from $\mathcal{E}_\text{l}$. For $\mathcal{E}_\text{b}$, we use the base latent diffusion model [2, 40] pre-trained for text-to-image generation task, and for $\mathcal{E}_\text{l}$, we use ControlNet [62] (termed as Light ControlNet in Fig. 4-(a)). While noise, texts, and timestep variables are also conditioned on $\mathcal{E}_\text{b}$ to fit the modality of the latent diffusion model, they are not described in the equations and figures for conciseness. To impose the foreground attention on the lighting control, a portrait mask is also conditioned to $\mathcal{E}_l$:

$$\mathcal{E}_\text{b}(\mathbf{I}; \mathcal{E}_\text{l}(\{\mathbf{S}_\phi, \mathbf{B}\}; \mathbf{M})) = \boldsymbol{z}, \quad \mathcal{D}(\boldsymbol{z}) = \mathbf{I}_\phi \tag{5}$$

where $\mathbf{M} \in \{0, 1\}^{w \times h}$ is the foreground binary mask. In training time, the diffusion model learns to directly change the lighting distribution of the input images (without inverse

rendering techniques) as the condition of coarse lighting estimate by minimizing the latent distance of the noisy relit image with the clean one from the ground truth in the overall forward and background denoising steps. Following the findings from an existing harmonization work [38], we use the composite image between the input image and conditioning background image as $\mathbf{I}$. We randomly drop background $\mathbf{B}$ or randomly set the coarse shading $\mathbf{S}_\phi$ as binary mask such that the diffusion model learns the control the lighting from $\mathbf{B}$, $\mathbf{S}_\phi$, or both $\{\mathbf{B}, \mathbf{S}_\phi\}$.

3.3. Unsupervised Add-on Temporal Modeling

Our coarse-to-fine relighting framework ($\mathcal{D} \circ \mathcal{E}_\text{b} \circ \mathcal{E}_\text{l}$) is trained only on individual images, inherently missing temporal context, *e.g.*, how a point on a human's surface will radiate from a specific viewpoint under the continuous pose, view, and illumination changes, leading to temporal artifacts such as flickering. We model such temporal context by designing an external add-on temporal lighting module $\mathcal{E}_m$ that can be combined, in inference time, with the relighting framework without extra training:

$$\mathcal{D} \circ \mathcal{E}_\text{b} \circ \mathcal{E}_\text{l} \rightarrow \mathcal{D} \circ \mathcal{E}_\text{b} \circ (\mathcal{E}_\text{l} \times \mathcal{E}_\text{m}) \tag{6}$$

To enable this, our temporal module is designed to regress the relit image from the previous time instance, *i.e.,* $\mathbf{I}_\phi^{t-1}$, to the latent lighting distribution whose space is shared with our base relighting models. Therefore, our decoder can generate temporally coherent relit images in the current time in an auto-regressive way:

$$\mathcal{D}(\mathcal{E}_\text{b}(\mathbf{I}^t; \mathcal{E}_\text{m}(\mathbf{I}_\phi^{t-1}))) = \mathbf{I}_{t-1}^t. \tag{7}$$

However, training $\mathcal{E}_\text{m}$ with a conventional L2 loss is not possible since there exists no ground-truth video relighting data for dynamic humans. Therefore, we propose to learn $\mathcal{E}_\text{m}$ in an unsupervised way using many real videos with lighting cycle consistency as follows:

$$\mathcal{D}^*(\mathcal{E}_\text{b}^*(\mathbf{I}^t; \mathcal{E}_\text{l}^*(\{\mathbf{S}_\phi^t, \mathbf{B}^t\}; \mathbf{M}^t))) = \mathbf{I}_\phi^t \ , \tag{8}$$

$$\mathcal{D}^*(\mathcal{E}_\text{b}^*(\mathbf{I}_\phi^t; \mathcal{E}_\text{m}(\mathbf{I}^{t-1}, \mathbf{M}^{t-1}))) = \tilde{\mathbf{I}}_{t-1}^t \tag{9}$$

$$\therefore \quad \mathbf{I}^t = \tilde{\mathbf{I}}_{t-1}^t.$$

We make the hypothesis: a video sequence inherently contains temporal lighting properties whose flow can be implicitly modeled by learning to predict the lighting distribution of the future frame based on the hint from the previous frame. This involves forward and backward processes, as a cycle-training. Eq. 8 represents the forward image relighting, *i.e.,* $\mathbf{I}^t \rightarrow \mathbf{I}_\phi^t$, where our relighting diffusion model with lighting ControlNet $\mathcal{E}_\text{l}$ generates the relit image at time t as a condition of a novel lighting condition (where we pick random Spherical harmonics). Eq. 9 reverts the relighting, *i.e.*, $\tilde{\mathbf{I}}_{t-1}^t \leftarrow \mathbf{I}_\phi^t$ to the original input image conditioned by the original frame in the previous time step through our temporal lighting module where the mask $\mathbf{M}$ is used for foreground awareness. Finally, $\mathcal{E}_m$ learns the lighting cycle consistency in the diffusion process by minimizing the latent distance between the reverted and the original image. We freeze the learnable weights for the functions with $*$ during training, and $\mathcal{E}_\text{m}$ is enabled with another ControlNet [62] termed as Motion ControlNet in Fig. 4-(b).

3.4. Inference with Spatio-Temporal Blending and Guided Refinement

Given a video or image, we perform comprehensive relighting with a spatio-temporal feature blending framework. For $t = 0$, we generate the relit image without our temporal lighting module $\mathcal{E}_\text{m}$. For $t > 0$ (note that even for the static image, $t > 0$ is possible since the lighting is time-variant), the generated relit images in the previous time step are conditioned on $\mathcal{E}_\text{m}$, and therefore, the relighting is controlled by dual control modules, *i.e.*, $\mathcal{E}_\text{l}$ and $\mathcal{E}_\text{m}$ by blending their features with a (0.85:0.15) ratio as described as spatial blending block in Fig. 4-(c). The blended lighting features are recurrently combined with the one from the previous time step through the temporal blending block. For this, we adopt optical flow from the original input video as a temporal prior to spatially align the features from consecutive frames similar to [27]; and we blend the aligned temporal features with a (0.5:0.5) ratio as in Fig. 4-(c) to improve the temporal consistency.

In the denoising process of the latent diffusion model, the generation often suffers from the loss of high-frequency details. Inspired by existing image restoration techniques [45, 65], we address this problem by applying a guided refinement to each generated image. We cast the problem of guided refinement as a guided residual prediction:

$$\mathbf{I}_\phi^\text{refine} = \mathbf{I} + g(\mathbf{I}_\phi, \mathbf{I}; \mathbf{M}), \tag{10}$$

where g is the function that predicts the guided lighting residual. This residual learns to map the lighting distribution from $\mathbf{I}$ to $\mathbf{I}_\phi$. Here, $\mathbf{I}_\phi^\text{refine}$ can effectively preserve the high-frequency details of the input image $\mathbf{I}$ due to the nature of residual learning [57, 65] that is designed to preserve the visual properties from the observation space, *i.e.*, $\mathbf{I}$. We enable g with a residual network [16] by learning from our relighting data with general losses for low-level vision processing, *i.e,* L2 and VGG [11].

4. Experiments

Dataset. To train our coarse-to-fine relighting model, we use 100K ground-truth relighting samples, including 50K synthetic human renderings and 50K OLAT-captured images from LightStage, with random cropping augmentation. Ground-truth albedo, images, background, masks, and lighting coefficients are precomputed. Our dataset is categorized by gender, skin tone, and body part (full-body and

Method	Scenario 1		Scenario 2		Scenario 3	
DPR [68]	18.62/0.86/0.103	32.00 /0.94/0.030	18.14/0.89/0.089	35.29 /0.98/0.032	21.20/0.89/0.072	37.11/0.95/0.038
RHW [47]	19.78/0.87/0.113	30.74/0.95/ 0.027	20.12/0.88/0.078	36.64 /0.98/ 0.028	23.33/0.90/0.060	37.53 /0.98/0.033
GFR [22]	25.59/0.91/0.089	30.33/0.95/0.033	22.76/0.91/0.072	32.36/0.98/0.036	25.49/0.93/0.050	35.89/0.98/ 0.028
LPBR [38]	18.19/0.86/0.090	31.62/0.91/0.035	19.96/0.88/0.084	27.94/0.94/0.041	21.42/0.88/0.053	33.39/0.95/0.038
Ours	25.95/0.95/0.066	33.58/0.96/0.024	23.99/0.93/0.048	35.18/ 0.99 / 0.031	26.61/0.94/0.033	38.32/0.98/0.026

Table 1. Comparison with existing image-based human relighting methods on synthetic videos for fidelity and temporal consistency. Each column shows image fidelity (left) and video temporal consistency (right). Metrics are PSNR↑ / SSIM↑ / LPIPS↓ for accuracy, and tPSNR↑ / tSSIM↑ / tLPIPS↓ for temporal consistency. While green is used for the best values, yellow highlights the second-best values.



Figure 5. Qualitative comparison of synthetic video frames (corresponding to Tab. 1). From left to right: composite input with target lighting parameters (inset), our relit result, baseline methods, and normalized L2↓ photometric error map (inset).

multi-body), with each subject captured from 32 viewpoints under varying lighting conditions. For a detailed dataset breakdown, refer to the Supple. To train our temporal control module using real data, we used 50K frames of videos from existing works [21, 29, 34, 54] and customized videos, where, for cyclic relighting, we randomly sample spherical harmonics lighting parameters from ground-truth data. For validation on static humans with dynamic lighting, we generated a new testing set (*i.e.,* the ground-truth relit images and associated lighting parameters) using an internal simulation and LightStage. This set includes the mixture of half-body, full-body, and multi-person scenarios, with each scenario comprising 100 frames. For the validation on dynamic humans (*i.e.*, video), we newly create synthetic video sequences for three scenarios: Scenario 1 involves static lighting (environment map) across frames with a moving human; Scenario 2 features dynamic lighting (rotated environment map) with a static human; Scenario 3 combines dynamic lighting (rotated environment map) with a moving human. Please refer to the Supple. and demo video for more details about the testing results.

Metrics. To measure the relighting quality, we use L1 distance, reconstruction fidelity (PSNR [18]), local structure similarity (SSIM [18]), and the latent perceptual distance (LPIPS [64]), between the relighted and ground-truth. For measuring the temporal coherence of the relighting results, we follow the same logics as TokenFlow [12]. This involves warping the relit image from the previous time step to the next using optical flow [48], and then comparing this to the



Figure 6. Comparison with LPBR [38] on DeepFashion [31] real images for background harmonization testing. The first row shows the relit output, and the second shows the magnified results.



Figure 7. Comparison with RHW [47] on Pexels [35] real images. The lighting control variables are shown as insets. While RHW produces reasonable relighting for full-body images, its quality degrades on half-body and multi-person cases.



Figure 8. Comparison with GFR [22] on Pexels [35] real images. The lighting control variables are shown as insets. Limited generalizability of GFR results in reduced output quality for half-body and multi-person cases.

current frame using the aforementioned metrics, termed tL_1 error, tPSNR, tSSIM, and tLPIPS.

Baseline. We compare our method with existing monocular human relighting works where we chose the baselines that are applicable to general scenes (*e.g.,* any part-specific information such as a 3D face model is not a requirement): DPR [67], RHW [47], and GFR [22]. For the harmonization baseline, we compare our model with LPBR [38], a diffusion-based light-aware harmonization method. For relighting, aside from GFR, baselines use Spherical harmonics for lighting control without modeling background illumination. DPR and RHW are evaluated using their released pre-trained models. Since GFR lacks available code, we re-implemented it using our dataset, replacing HDR lighting with Spherical harmonics and background modeling to enable background harmonization in our experiments. Harmonization is directly compared with LPBR by replacing the background.

Results. For testing static humans under dynamic lighting, we present the quantitative comparison in Tab. 2. We show average numerical evaluations on our synthetic testing dataset, categorized by portrait, full-body, and multi-person. Further validation by gender, and skin color is detailed in the Supple. Methods such as RHW and DPR have

Method	$L_1\downarrow$	PSNR↑	SSIM↑
Ours-diffusion	0.05837	17.099	0.833
Ours end-to-end	0.01432	26.418	0.918
Ours-background	0.01239	27.346	0.945
Ours	0.01035	28.419	0.948
Ours+refine	0.01012	28.778	0.949

Table 3. Ablation study on coarse-to-fine relighting models.

limited generalizability for both full-body and portrait relighting. They are difficult to extend to other scenarios and tend to show reduced performance when tested in different settings. While GFR performs well in our evaluation, it struggles with significant domain gaps, leading to noticeable quality degradation on real data, including distortions and color shifts (Fig. 8). In contrast, our model exhibits strong generalizability across validation sets and real-world tests (Figs. 7, 8). Compared to the state-of-the-art background harmonization method (LPBR), our method shows the strong generalization to different body parts, and notably, LPBR often includes significant distortion when applied to the full-body images and it does not support the lighting control function as shown in Fig. 6.

Category	**DPR [68]**	**RHW [47]**	**GFR [22]**	**Ours**
Portrait	17.74 / 0.87	15.75 / 0.82	17.71 / 0.86	23.04 / 0.90
Full-body	27.62 / 0.96	27.73 / 0.95	29.51 / 0.95	30.81 / 0.97
Multi-person	25.70 / 0.95	25.69 / 0.95	29.35 / 0.97	31.49 / 0.96

Table 2. Comparison on our synthetic static testing data sorted by body-part. We compute average PSNR↑ / SSIM↑.

For video testing, in Tab. 1, we evaluate our model on scenarios 1, 2, and 3 for both fidelity and temporal consistency. Other approaches face challenges in achieving both relit fidelity and temporal consistency at the same time as also shown in Fig. 5. In contrast, our temporal module ensures our comprehensive relighting model to produce videos with strong temporal consistency. For more results on relighting results and comparisons including user study, please refer to the Supple.

Ablation Study. We conduct two ablation studies on our coarse-to-fine model using static human data and our temporal modules using dynamic human data. As shown in Tab. 3, we perform the ablation study on our coarse-to-fine approach: 1) *Ours-diffusion*: Relighting only with pixel-aligned coarse shading without diffusion model, *i.e.*, $\mathbf{S}_\phi \times \mathbf{I}$. 2) *Ours end-to-end*: Relighting in an end-to-end manner by applying target lighting parameters ϕ as a condition, trained with a diffusion model, without the coarse stage. 3) *Ours-background*: Relighting without the control of background $\mathbf{B}$. 4) *Ours+refine* (full): Applying our guided refinement to the relit image.

Tab. 3 shows the summary of the ablation study: Directly apply the coarse stage without a diffusion model introduces significant errors in the final relit result due to the detection noises. Instead of applying target lighting and end-to-end training with a diffusion model, our coarse-to-fine approach shows better performance, indicating that our coarse stage serves as a strong control prior. This control prior is both neat and effective for extending our model to diverse identities, various body parts. Based on the comparison with "Ours-background" we notice that encoding information from the background image aids in enhancing natural illumination during background harmonization. Lastly, leveraging a guided refinement enables the preservation of high-frequency information alongside robust generation capabilities.

Method	t$L_1\downarrow$	tPSNR↑	tSSIM↑
Ours	6.552	31.028	0.956
Ours+temporal	5.638	32.266	0.957
Ours+temporal+blend	4.019	33.588	0.957

Table 4. Ablation study on our temporal modules evaluated on synthetic sequences: t$L1$ error ($\times 10^{-3}$)

For our temporal module, we study three ablation studies: 1) Ours: We eliminate all temporal consistency components, which is a single-frame-based generation method. 2) Ours+temporal: We only apply temporal module $\mathcal{E}_m$ without recurrent feature blending during the test-time. 3) Ours+temporal+blend: We perform our video relighting with temporal module $\mathcal{E}_m$ and recurrent feature blending.

Tab. 4 summarizes the performance of each of our temporal modules. The temporal lighting module in "Ours+temporal" primarily enforces temporal coherence by imposing a temporal constraint on the lighting control between the current and previous frames during testing. Additionally, the recurrent blending feature further enhances temporal consistency by blending the lighting control feature between previous and current frames, thereby reinforcing the temporal context.

Limitation. Our relighting diffusion model requires heavy computational time. Significant noise on the detection (*e.g.*, mask and surface normal) affects the temporal coherence.

5. Conclusion

We introduce a method for Comprehensive Relighting that is generalizable and consistent for monocular human relighting and harmonization. We address a core dataset challenge by utilizing a large and general image prior from a pre-trained diffusion model; and repurposing the model specialized for temporally consistent image relighting. For coherent control of the lighting, we introduce a coarse-to-fine relighting framework; and combine it with an external temporal lighting module that learns many real videos. Our guided refinement network enhances the visual to preserve the fine details of an original image. In the experiments, our method outperforms other image-based relighting and harmonization models in terms of quality and temporal coherence.

6. Acknowledgement

We sincerely thank Mengwei Ren for the insightful discussions regarding the framework design, and we are grateful to Jianming Zhang for kindly providing the human normal map estimator.

References

[1] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5960–5969, 2020. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 4

[3] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6221–6231, 2024. 3

[4] Jianqi Chen, Yilan Zhang, Zhengxia Zou, Keyan Chen, and Zhenwei Shi. Dense pixel-to-pixel harmonization via continuous image representation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 3

[5] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. 3

[6] Jun Myeong Choi, Max Christman, and Roni Sengupta. Personalized video relighting with an at-home light stage. In *European Conference on Computer Vision*, pages 394–410. Springer, 2024. 3

[7] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8394–8403, 2020. 3

[8] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18470–18479, 2022. 3

[9] Yuki Endo Daichi Tajima, Yoshihiro Kanamori. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. *Computer Graphics Forum (Proc. of Pacific Graphics 2021)*, 40(7):205–216, 2021. 3

[10] David Futschik, Kelvin Ritland, James Vecore, Sean Fanello, Sergio Orts-Escolano, Brian Curless, Daniel Sỳkora, and Rohit Pandey. Controllable light diffusion for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8412–8421, 2023. 3

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 5

[12] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 6

[13] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5917–5926, 2023. 3

[14] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14870–14879, 2021.

[15] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16367–16376, 2021. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[18] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6

[19] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14719–14728, 2021. 3

[20] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4217–4226, 2022. 3

[21] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 6

[22] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *European Conference on Computer Vision*, pages 388–405. Springer, 2022. 2, 3, 6, 7, 8

[23] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021. 3

[24] Yoshihiro Kanamori and Yuki Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *arXiv preprint arXiv:1908.02714*, 2019. 2

[25] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human

portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25096–25106, 2024. 3

[26] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. Single-image full-body human relighting. *arXiv preprint arXiv:2107.07259*, 2021. 2, 3

[27] Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. Blind video deflickering by neural filtering with a flawed atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[28] Guannan Li, Yebin Liu, and Qionghai Dai. Free-viewpoint video relighting from multi-view sequence under general illumination. *Machine vision and applications*, 25:1737–1746, 2014. 3

[29] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 6

[30] Yang Liu, Alexandros Neophytou, Sunando Sengupta, and Eric Sommerlade. 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*, pages 32–40, 2021. 3

[31] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[33] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 1, 2, 3

[34] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 6

[35] Pexels: https://www.pexels.com. 7

[36] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. *arXiv preprint arXiv:2304.09479*, 2023. 3

[37] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA A*, 18 (10):2448–2459, 2001. 3

[38] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. *CVPR*, 2024. 3, 5, 6, 7

[39] Christian Richardt, Carsten Stoll, Neil A Dodgson, Hans-Peter Seidel, and Christian Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. In *Computer graphics forum*, pages 247–256. Wiley Online Library, 2012. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[42] Volker Schönefeld. Spherical harmonics. *Computer Graphics and Multimedia Group, Technical Note. RWTH Aachen University, Germany*, page 18, 2005. 3

[43] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[45] Shuangbing Song, Fan Zhong, Tianju Wang, Xueying Qin, and Changhe Tu. Guided linear upsampling. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 5

[46] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[47] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, pages 205–216. Wiley Online Library, 2021. 3, 6, 7, 8

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6

[49] Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, Christian Theobalt, et al. Monocular reconstruction of neural face reflectance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4791–4800, 2021. 3

[50] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20762–20772, 2023. 3

[51] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 3

[52] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. Semi-supervised parametric real-world image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5927–5936, 2023. 3

[53] Yuxin Wang, Wayne Wu, and Dan Xu. Learning unified decompositional and compositional nerf for editable novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18247–18256, 2023. 3

[54] Liu Wen, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, , and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 6

[55] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37 (4):1–13, 2018. 3

[56] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019. 3

[57] Wenhan Yang, Jiashi Feng, Jianchao Yang, Fang Zhao, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep edge guided recurrent residual learning for image super-resolution. *IEEE Transactions on Image Processing*, 26(12): 5895–5907, 2017. 5

[58] Mona Zehni, Shaona Ghosh, Krishna Sridhar, and Sethu Raman. Joint learning of portrait intrinsic decomposition and relighting. *arXiv preprint arXiv:2106.15305*, 2021. 3

[59] Hao Zhang, Yanbo Xu, Tianyuan Dai, Tai Chi-Keung Tang, et al. Fdnerf: Semantics-driven face reconstruction, prompt editing and relighting with diffusion models. *arXiv preprint arXiv:2306.00783*, 2023. 3

[60] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 3

[61] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 802–812, 2021. 1, 3

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 4, 5

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[65] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 5

[66] Ruichen Zheng, Peng Li, Haoqian Wang, and Tao Yu. Learning visibility field for detailed 3d human reconstruction and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 216–226, 2023. 3

[67] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 7

[68] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 1, 2, 3, 6, 8

[69] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4327, 2023. 3

[70] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3943–3951, 2015. 3