Relighting4D: Neural Relightable Human from Videos

Zhaoxi Chen and Ziwei Liu(✉)

S-Lab, Nanyang Technological University, Singapore, Singapore
{zhaoxi001,ziwei.liu}@ntu.edu.sg

Abstract. Human relighting is a highly desirable yet challenging task. Existing works either require expensive one-light-at-a-time (OLAT) captured data using light stage or cannot freely change the viewpoints of the rendered body. In this work, we propose a principled framework, **Relighting4D**, that enables free-viewpoints relighting from only human videos under unknown illuminations. Our key insight is that the space-time varying geometry and reflectance of the human body can be decomposed as a set of neural fields of normal, occlusion, diffuse, and specular maps. These neural fields are further integrated into reflectance-aware physically based rendering, where each vertex in the neural field absorbs and reflects the light from the environment. The whole framework can be learned from videos in a self-supervised manner, with physically informed priors designed for regularization. Extensive experiments on both real and synthetic datasets demonstrate that our framework is capable of relighting dynamic human actors with free-viewpoints. Codes are available at https://github.com/FrozenBurning/Relighting4D.

Keywords: Neural rendering · Dynamic scenes · Inverse rendering

1 Introduction

The emergence of metaverse has fueled the demands for photorealistic rendering of human characters, which benefits applications like digital 3D human and virtual reality. Among all factors, lighting is the most crucial one for rendering quality. Recently, remarkable success in relighting humans has been achieved [4, 11, 17, 25, 26, 31, 45, 51, 60, 62]. However, the impressive quality of these methods heavily relies on the data captured by Light Stage [8]. The complicated hardware setup makes relighting systems expensive and only applicable in the constrained environment. On the other hand, a number of recent works propose to relight human images from a perspective of inverse rendering [13, 15, 21, 39, 47, 64]. They succeed in relighting 2D images, yet fail to relight with novel views. A lack of underlying 3D representations impedes their flexibility of application.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19781-9_35.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13674, pp. 606–623, 2022.
https://doi.org/10.1007/978-3-031-19781-9_35

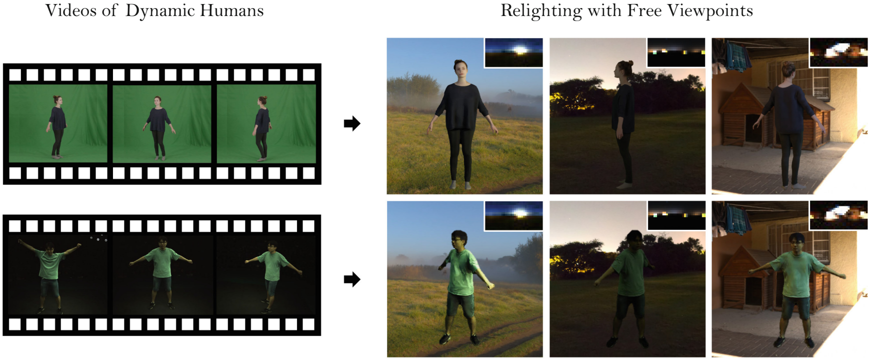

Fig. 1. Relighting of dynamic humans with free viewpoints. *Relighting4D* ;takes only videos as input, decomposing them into geometry and reflectance, which enables relighting of dynamic humans with free viewpoints by a physically based renderer.

In this paper, we focus on the problem of relighting dynamic humans from only videos, as illustrated in Fig. 1. The setting significantly reduces the cost of a flexible relighting system and broadens its scope of application. It has been proved [5,6,19,24,28,30,32,34,35,41,42,54,58,63] that a scene can be represented as neural fields to enable novel view synthesis and relighting. Among above methods, some [5,6,24,42,59,63] deal with relighting static objects but fail to model dynamic scenes. In sum, none of those methods successfully incorporate illuminations and scene dynamics simultaneously.

Different from existing methods on novel view synthesis of the human body that are either non-relightable or require expensive OLAT captured images, we seek to estimate plausible geometry and reflectance from posed human videos.

To this end, we propose ***Relighting4D***, to relight dynamic humans with free viewpoints from videos given the 4D coordinates (x, y, z, t) and the desired illumination. Specifically, our method first aggregates observations from posed human videos through space and time by a neural field conditioned on a deformable human model. Then, we decompose the neural field into geometry and reflectance counterparts, namely normal, occlusion, diffuse, and specular maps, which drive a physically based renderer to perform relighting.

We evaluate our approach on both monocular and multi-view videos. Overall, *Relighting4D* outperforms other methods on perceptual quality and physical correctness. It relights dynamic humans in high fidelity, and generalizes to novel views. Furthermore, we demonstrate our capability of relighting under novel illuminations, especially the challenging OLAT setting, by creating a synthetic dataset called BlenderHuman for quantitative evaluations.

We summarize our contributions as follows: **1)** We present a principled framework, *Relighting4D*, which is the first to relight dynamic humans with free viewpoints using only videos. **2)** We propose to disentangle reflectance and geometry from input videos under unknown illuminations by leveraging multiple physically informed priors in a physically based rendering pipeline. **3)** Extensive experiments on both synthetic and real datasets demonstrate the feasibility and significant improvements of our approach over prior arts.

2 Related Work

Neural Scene Representation. [14,18,20,28,34–36,40,41,43,46,49,56] has witnessed significant progress in representing a 3D scene with deep neural networks. NeRF [28] proposes to model the scene as a 5D radiance field. To model dynamic humans, Neural Body [34] proposes to attach a set of latent codes to a deformable human body model (i.e., SMPL [23]). However, these methods implicitly incorporate all color information in the radiance field, which impedes their application towards relighting a dynamic human.

Inverse rendering aims to disentangle the appearance from observed images into geometry, material, and lighting condition. Previous works [3,16,21,22,27,37,53,57] seek to address it by conditioning on physically based priors or synthetic data. However, they fail in novel view synthesis due to the lack of underlying 3D representations. Recently, NeRF based methods [5,6,42,59,63] propose to learn 3D reflectance fields or light transport fields from input images to enable free-viewpoint relighting. However, none of them is applicable to relight dynamic humans with space-time varying features.

Relighting of human face, avatar and body has wide-range applications [31,39,44,51,64]. As for full-body human relighting, convolutional methods [13,47] fail to relight from novel viewpoints as there is no underlying 3D representation. Other methods [11,62] heavily relies on one-light-at-a-time [8] (OLAT) images, which is neither cheap to capture nor publicly available. *Relighting4D* differentiates itself from aforementioned methods in that we achieve free-viewpoint relighting of dynamic full-body humans without the requirement on expensive capture setup.

3 Our Approach

Given a human video, *Relighting4D* can synthesize videos with free viewpoints under novel illuminations. We denote the input video as $I = \{I_1, I_2, ..., I_t\}$, where t is the time step. In general, our model learns a physically based renderer from I. During inference, it takes a 3D position $\boldsymbol{x} \in \mathbb{R}^3$, a time step t, a camera view $\boldsymbol{\omega}_o \in \mathbb{R}^3$, a desired light probes $\boldsymbol{L}_i \in \mathbb{R}^{16\times 32\times 3}$ as inputs, and outputs the corresponding outgoing radiance $\boldsymbol{L}_o \in \mathbb{R}^3$.

Framework Overview. We first give an overview of *Relighting4D* (Figure 2). It first derives latent features from the video, which is achieved by estimating a neural field. Based on the latent features, *Relighting4D* decomposes the human performer into geometry and reflectance information which drive our physically based renderer. The space-time varying geometry and reflectance of the full human body are parameterized by four multilayer perceptrons. Note that, *Relighting4D* enables relighting of dynamic humans with free viewpoints using **only** videos, without training on any captured data (e.g., OLAT or flash images).

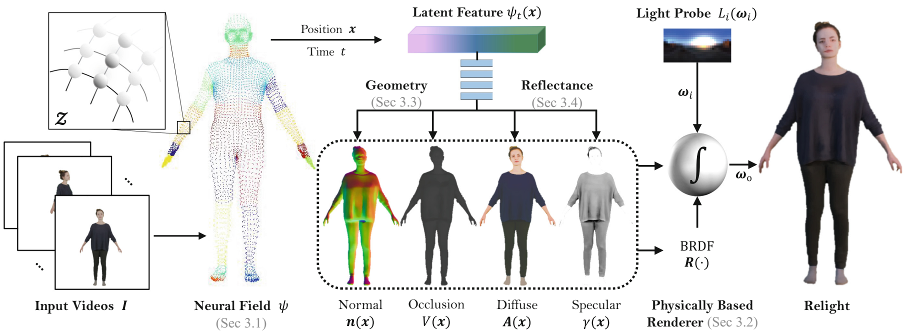

Fig. 2. Overview of *Relighting4D*. Given the input video frame at time step t, *Relighting4D* represents the human as a neural field ψ on latent vectors $\mathcal{Z}$ anchored to a deformable human model. The value of the neural field $\psi_t(\boldsymbol{x})$ at any 3D point $\boldsymbol{x}$ and time t is taken as latent feature and fed into multilayer perceptrons to obtain geometry and reflectance, which are normal, occlusion, diffuse, and specular maps respectively. Finally, a physically based renderer is raised to render the human subject to the input light probe under novel illumination.

3.1 Neural Field as Human Representation

Extracting 4D representations of dynamic human performers is a non-trivial task. Compared to the static scenes where NeRF [28] fits well, dynamic scenes in videos have factors like motion, occlusion, non-rigid deformation, and illumination that vary through space and time which hampers an accurate estimation.

Inspired by the local implicit representations [10,34], we introduce a 4D neural field ψ conditioned on a parametric human model (SMPL [23] or SMPL-X [33]) to represent a dynamic human performer, which maps the position $\boldsymbol{x}$ and time step t to the latent feature $\psi_t(\boldsymbol{x})$. Specifically, at frame I_t, we obtain the parameters of human model (i.e. locations of vertices) using this tool [12]. Then, a set of latent vectors $\mathcal{Z} \in \mathbb{R}^{N\times 16}$ is assigned to the vertices of human model, where $N = 6890$ for SMPL [23] and $N = 10475$ for SMPL-X [33]. Then we query the neural field by the 4D coordinates $(\boldsymbol{x}, t)$, extracting the latent feature $\psi_t(\boldsymbol{x}) \in \mathbb{R}^{256}$ from $\mathcal{Z}$ via trilinear interpolation of its nearby vertices.

NeuralBody [34] employs a similar strategy on human representations. But it's not relightable in the way that it fails to disentangle geometry and reflectance from the latent codes. In contrast, *Relighting4D* learns a distinct neural field that can be decomposed into geometry (Sect. 3.3) and reflectance (Sect. 3.4), which serves the physically based renderer (Sect. 3.2) for relighting.

3.2 Physically Based Rendering

While differentiable volume rendering has been used in recent works [28,34,56], these methods focus on novel view synthesis with radiance fields. In general, to enable relighting with neural representations, instead of modeling the human body as a field of vertices that *emit* light, we represent the human as a field of vertices that *reflect* the light from the environment. Specifically, we leverage a physically based renderer, which models a reflectance-aware rendering process. Mathematically, our rendering pipeline is driven by the following equation:

$$L_o(\boldsymbol{x}, \boldsymbol{\omega}_o) = \int_{\Omega} R(\boldsymbol{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o, \boldsymbol{n}(\boldsymbol{x})) L_i(\boldsymbol{x}, \boldsymbol{\omega}_i)(\boldsymbol{\omega}_i \cdot \boldsymbol{n}(\boldsymbol{x})) d\boldsymbol{\omega}_i, \tag{1}$$

where $L_o(\boldsymbol{x}, \boldsymbol{\omega}_o) \in \mathbb{R}^3$ is the outgoing radiance at point $\boldsymbol{x}$ viewed from $\boldsymbol{\omega}_o$. $L_i(\boldsymbol{x}, \boldsymbol{\omega}_i) \in \mathbb{R}^3$ is the incident radiance arriving at $\boldsymbol{x}$ from direction $\boldsymbol{\omega}_i$. Ω is an unit sphere that models all possible light directions, and $\boldsymbol{n}(\boldsymbol{x}) \in \mathbb{R}^3$ is the normal. $R(\boldsymbol{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o, \boldsymbol{n}(\boldsymbol{x}))$[2] is the Bidirectional Reflectance Distribution Function (BRDF) which defines how the incident light is reflected at the surface, and $d\boldsymbol{\omega}_i$ is the solid angle of incident light at $\boldsymbol{\omega}_i$. We use a discrete set of light samples to approximate Eq. 1 in the following way:

$$L_o(\boldsymbol{x}, \boldsymbol{\omega}_o) \approx \sum_{\boldsymbol{\omega}_i} R(\boldsymbol{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o, \boldsymbol{n}(\boldsymbol{x})) L_i(\boldsymbol{x}, \boldsymbol{\omega}_i)(\boldsymbol{\omega}_i \cdot \boldsymbol{n}(\boldsymbol{x})) \Delta\boldsymbol{\omega}_i, \tag{2}$$

where $\Delta\boldsymbol{\omega}_i$ is sampled from a light probe that depicts the distribution of light sources in space. We represent the environment light $L_i(\boldsymbol{\omega}_i)$ as a light probe image in latitude-longitude format with a resolution of $16 \times 32 \times 3$, which facilitates relighting applications by replacing the estimated light probe with an external one. Figure 3 illustrates our physically based renderer at surface $\boldsymbol{x}$.

Note that previous work [28,34] implicitly encodes $R(\cdot)$ in the radiance fields without modeling the reflectance. To enable flexible relighting applications, we leverage the microfacet model [50] to approximate a differentiable reflectance function parameterized by the surface normal $\boldsymbol{n}(\boldsymbol{x})$, the diffuse map $\boldsymbol{A}(\boldsymbol{x})$ and the specular roughness $\gamma(\boldsymbol{x})$. Due to the limited space, we introduce the implementation of $R(\cdot)$ in the supplementary.

To encode harsh shadow and occlusion, we mask the incident light $L_i(\boldsymbol{x}, \boldsymbol{\omega}_i)$ by the occlusion map $V(\boldsymbol{x}, \boldsymbol{\omega}_i)$ at $\boldsymbol{x}$:

$$L_i(\boldsymbol{x}, \boldsymbol{\omega}_i) = V(\boldsymbol{x}, \boldsymbol{\omega}_i) L_i(\boldsymbol{\omega}_i). \tag{3}$$

Physical Characteristics Disentanglement. Driven by Eq. 2, the renderer requires physical characteristics, i.e., geometry, reflectance, and light, of a given human performer, which are disentangled and estimated by *Relighting4D* from input videos. The details are introduced in the following two sections.

3.3 Volumetric Geometry

In terms of geometry, our renderer requires a normal map $\boldsymbol{n}(\boldsymbol{x}) \in \mathbb{R}^3$ and an occlusion map $V(\boldsymbol{x}, \boldsymbol{\omega}_i) \in \mathbb{R}$ as inputs. Moreover, we render on the surface to keep the computing process tractable, which requires an estimation of surface position. It can be easily obtained by querying a density field.

[2] For simplicity, we also use $R(\cdot)$ to denote BRDF when necessary.

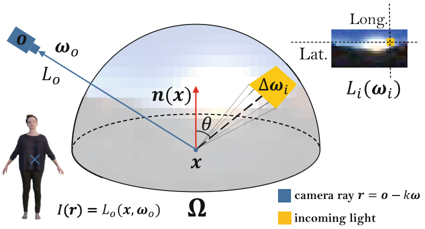
$\boldsymbol{o}$
$\boldsymbol{\omega}_o$
L_o
$\boldsymbol{n}(\boldsymbol{x})$
θ
$\Delta\boldsymbol{\omega}_i$
$\boldsymbol{x}$
Long.
Lat.
$L_i(\boldsymbol{\omega}_i)$
$I(\boldsymbol{r}) = L_o(\boldsymbol{x}, \boldsymbol{\omega}_o)$
$\boldsymbol{\Omega}$
🟦 camera ray $\boldsymbol{r} = \boldsymbol{o} - k\boldsymbol{\omega}$
🟨 incoming light

Fig. 3. Illustration of our physically based rendering pipeline. The environment light is represented as spherical coordinates in latitude-longtitude (Lat.-Long.) format. Given the surface location $\boldsymbol{x}$, the incoming light from $\boldsymbol{\omega}_i$ with the area of $\Delta\boldsymbol{\omega}_i$ is scattered by the microfacet that is parameterized by BRDF $R(\cdot)$, normal $\boldsymbol{n}(\boldsymbol{x})$, and $\cos\theta = \boldsymbol{\omega}_i \cdot \boldsymbol{n}(\boldsymbol{x})$. Then the outgoing radiance $L_o(\boldsymbol{x}, \boldsymbol{\omega}_o)$ along the ray $\boldsymbol{r} = \boldsymbol{o} - k\boldsymbol{\omega}_o$ is calculated according to Eq. 2, which equals to the corresponding pixel value.

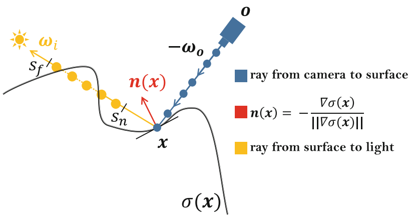
$\boldsymbol{o}$
$\boldsymbol{\omega}_i$
s_f
$-\boldsymbol{\omega}_o$
$\boldsymbol{n}(\boldsymbol{x})$
s_n
$\boldsymbol{x}$
$\sigma(\boldsymbol{x})$
🟦 ray from camera to surface
🟥 $\boldsymbol{n}(\boldsymbol{x}) = -\frac{\nabla\sigma(\boldsymbol{x})}{\|\nabla\sigma(\boldsymbol{x})\|}$
🟨 ray from surface to light

Fig. 4. The process of baking geometry. Note that we perform two different types of ray marching during training. The one is marching along the camera ray $\boldsymbol{r} = \boldsymbol{o} - k\boldsymbol{\omega}_o$ to the expected depth of termination k to get the geometry surface $\boldsymbol{x}$, while the other is marching from the surface $\boldsymbol{x}$ to the light coming from direction $\boldsymbol{\omega}_i$ to calculate the accumulated transmittance (occlusion map).

We first reconstruct the geometry of the given scene using an auxiliary density field $f_\sigma : (\boldsymbol{x}, \psi_t(\boldsymbol{x})) \rightarrow \sigma(\boldsymbol{x})$. It's derived from the latent feature $\psi_t(\boldsymbol{x})$ using an MLP. As shown in Fig. 4, *Relighting4D* leverages the auxiliary density field by baking it into surface maps, normal maps and occlusion maps.

Surface map is the 3D coordinates of points at the expected termination of depth given the camera view $\boldsymbol{\omega}_o$. We march the camera ray $\boldsymbol{r}$ from its origin $\boldsymbol{o}$ along the direction $-\boldsymbol{\omega}_o$ to the expected termination of depth k to get the surface $\boldsymbol{x} = \boldsymbol{o} - k\boldsymbol{\omega}_o$.

Normal map is computed on the surface as the normalized negative gradient of the density field: $\tilde{\boldsymbol{n}}(\boldsymbol{x}) = -\nabla\sigma(\boldsymbol{x})/||\nabla\sigma(\boldsymbol{x})||$.

Occlusion map denotes the transmittance of surface points from a specific direction. We compute the occlusion map by marching the ray $\boldsymbol{r}(s, \boldsymbol{x}, \boldsymbol{\omega}_i) = \boldsymbol{x} + s\boldsymbol{\omega}_i$ from the surface of the human body to the corresponding light at $\boldsymbol{\omega}_i$: $\tilde{V}(\boldsymbol{x}, \boldsymbol{\omega}_i) = 1 - exp(-\int_{s_n}^{s_f} \sigma(\boldsymbol{r}(s, \boldsymbol{x}, \boldsymbol{\omega}_i))ds)$, where s_n and s_f is the near and far bounds along the direction of the light. We set $s_n = 0, s_f = 0.5$ for all scenes. In other words, occlusion map considers the visibility at the given surface $\boldsymbol{x}$ by querying the density fields from s_n to s_f along the incident light direction $\boldsymbol{\omega}_i$.

Unfortunately, directly using the baked geometry causes numerous queries of f_σ(e.g., for occlusion map, we should trace $16 \times 32 = 512$ rays from all possible lighting directions for one 3D point), which is not tractable during training and rendering. Thus, we use an MLP $f_n : (\boldsymbol{x}, \psi_t(\boldsymbol{x})) \rightarrow \boldsymbol{n}(\boldsymbol{x})$ to reparameterize the surface and latent features to the normal map, and another MLP $f_V : (\boldsymbol{x}, \omega_i, \psi_t(\boldsymbol{x})) \rightarrow V(\boldsymbol{x})$ to map the surface, light direction and features to the occlusion map V. The weights of f_V, f_n are trained with the geometry reconstruction loss, intending to recover the baked geometry:

$$\mathcal{L}_{geo} = ||V(\boldsymbol{x}) - \tilde{V}(\boldsymbol{x})||_2^2 + ||\boldsymbol{n}(\boldsymbol{x}) - \tilde{\boldsymbol{n}}(\boldsymbol{x})||_2^2. \tag{4}$$

Smoothness Regularization. We regularize f_V, f_n by L1 penalty to keep the smoothness of their outputs:

$$\tau_V = |V(\boldsymbol{x}) - V(\boldsymbol{x} + \boldsymbol{\epsilon})|_1 \quad \tau_n = |\boldsymbol{n}(\boldsymbol{x}) - \boldsymbol{n}(\boldsymbol{x} + \boldsymbol{\epsilon})|_1, \tag{5}$$

where we measure the local smoothness by adding 3D perturbation $\boldsymbol{\epsilon}$ to $\boldsymbol{x}$ which is sampled from a Gaussian distribution with zero mean and standard deviation 0.01. Several works [29,63] have validated the use of similar smoothness losses for the aim of shape reconstruction.

Temporal Coherence Regularization. It is crucial for a 4D representation to incorporate temporal coherence. Otherwise, the rendered sequence will contain jitter appearance. Moreover, an accurate geometry is also important for artifact-free physically based rendering. Therefore, we add the following regularization term to encourage a temporally smooth geometry:

$$\mathcal{L}_{temp} = \frac{1}{N} \sum_{i=1}^{N} |\sigma_t(\hat{\boldsymbol{x}}_i) - \sigma_{t+1}(\hat{\boldsymbol{x}}_i)|_1, \tag{6}$$

where $\hat{\boldsymbol{x}}_i$ is the 3D position of i-th vertex of SMPL model. Eqn. 6 explicitly constrains the temporal coherence of the geometry, and also implicitly regularize the latent feature $\psi_t(\boldsymbol{x})$ which benefits the following reflectance estimation.

3.4 Reflectance

In terms of reflectance, our physically based renderer requires the BRDF $R(\cdot)$ and the light probe $L_i(\boldsymbol{\omega}_i)$ as inputs. As presented in Sect. 3.2, our BRDF estimation consists of a Lambertian RGB diffuse component $\boldsymbol{A}(\boldsymbol{x}) \in \mathbb{R}^3$ and a specular component $\gamma(\boldsymbol{x}) \in \mathbb{R}$. We parameterize the diffuse map at $\boldsymbol{x}$ with latent features $\psi_t(\boldsymbol{x})$ as an MLP $f_A : (\boldsymbol{x}, \psi_t(\boldsymbol{x})) \rightarrow \boldsymbol{A}(\boldsymbol{x})$, and parameterize the specular map as another MLP $f_\gamma : (\boldsymbol{x}, \psi_t(\boldsymbol{x})) \rightarrow \gamma(\boldsymbol{x})$.

Local Smoothness Prior. The problem that decomposes BRDF from video frames under unknown illumination is highly ill-posed. As the color information is entangled, and there is no off-the-shelf supervision on the reflectance. Inspired by work [3,6,37,59,63] on intrinsic decomposition which leverages piece-wise smoothness prior on albedo, we regularize the optimization of f_A by L1 penalty:

$$\tau_A = |\boldsymbol{A}(\boldsymbol{x}) - \boldsymbol{A}(\boldsymbol{x} + \boldsymbol{\epsilon})|_1, \tag{7}$$

where $\boldsymbol{\epsilon}$ is the same type of perturbation as Eq. 5.

Global Sparsity Prior. However, given this under-constrained problem, the local smoothness regularization in Eq. 7 is not sufficient for a plausible estimation of the diffuse map, as shown in Fig. 7. Thus, we further leverage global minimum-entropy sparsity prior on diffuse map which has been previously explored [2,3,9,38] on shadow removal. From a perspective of physically based rendering, the diffuse map represents the base color, indicating that the palette should be sparse enough. Intuitively, the diffuse map of clothes should contains a small number of colors. Thus, we minimize the Shannon entropy of diffuse map, denoted as H_A, to impose this prior on our model. Since the diffuse map $\boldsymbol{A}(\boldsymbol{x})$ is a continuous variable whose probability density function (PDF) is unknown, a naive way to estimate its entropy is using histogram to get PDF. But, it's not differentiable. Instead, it's always possible to use a soft and differentiable generalization of Shannon entropy (i.e. quadratic entropy [55]). However, it's quadratically expensive to the number of sampled camera rays.

This motivates our novel approximation of minimizing H_A in a both differentiable and efficient way. The key insight is that the PDF of $\boldsymbol{A}(\boldsymbol{x})$, $p(\boldsymbol{A}(\boldsymbol{x}))$, can be estimated by a Gaussian KDE (Kernel Density Estimator). Given a diffuse map $\boldsymbol{A}(\boldsymbol{x})$, we leverage a KDE as its PDF approximation:

$$\tilde{p}(\boldsymbol{A}(\boldsymbol{x})) = \frac{1}{n}\sum_{i=1}^{n} K_G(\boldsymbol{A}(\boldsymbol{x}) - \boldsymbol{A}_i(\boldsymbol{x})), \tag{8}$$

where K_G is the standard normal density function, n is the number of sampled rays during training, and $\boldsymbol{A}_i(\boldsymbol{x})$ is the value of diffuse map at the i-th camera ray. Then the entropy of $\boldsymbol{A}(\boldsymbol{x})$ is computed as an expectation:

$$H_A = \mathbb{E}[-\log(\tilde{p}(\boldsymbol{A}(\boldsymbol{x})))]. \tag{9}$$

In addition, as the input video is captured under unknown illuminations, we randomly initialize the light probe $L_i(\boldsymbol{\omega}_i)$ as a trainable parameter, optimizing it during the training phase to estimate a plausible ambient light of the scene. It can be replaced by a new HDR map for relighting after training.

3.5 Progressive End-to-End Learning

In the training phase, we randomly sample 1024 camera rays for each input frame. Besides, we employ a progressive training strategy which allows the resolution of video to gradually increase. In specific, before ray sampling, the input video is scaled to the resolution of $\alpha H \times \alpha W$ where $\alpha \in (0, 1]$ is a monotonically increasing function of the number of iterations.

Furthermore, we embed the surface position $\boldsymbol{x}$ and the light direction $\boldsymbol{\omega}_i$ using the positional encoding [28,48] before concatenating them with latent features $\psi_t(\boldsymbol{x})$. The maximum frequency of set to 2^{10} and 2^4, respectively. We use four fully-connected ReLU layers with 256 channels for each MLP.

Our full loss function is a summation:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{temp}\mathcal{L}_{temp} + \lambda_V \tau_V + \lambda_n \tau_n + \lambda_A \tau_A + \lambda_H H_A, \tag{10}$$

where $\mathcal{L}_{rgb}$ is the reconstruction loss against the ground-truth pixel color value. We train each model for 260k iterations with a Tesla V100 GPU. Details of training hyperparameters are deferred to the supplementary.

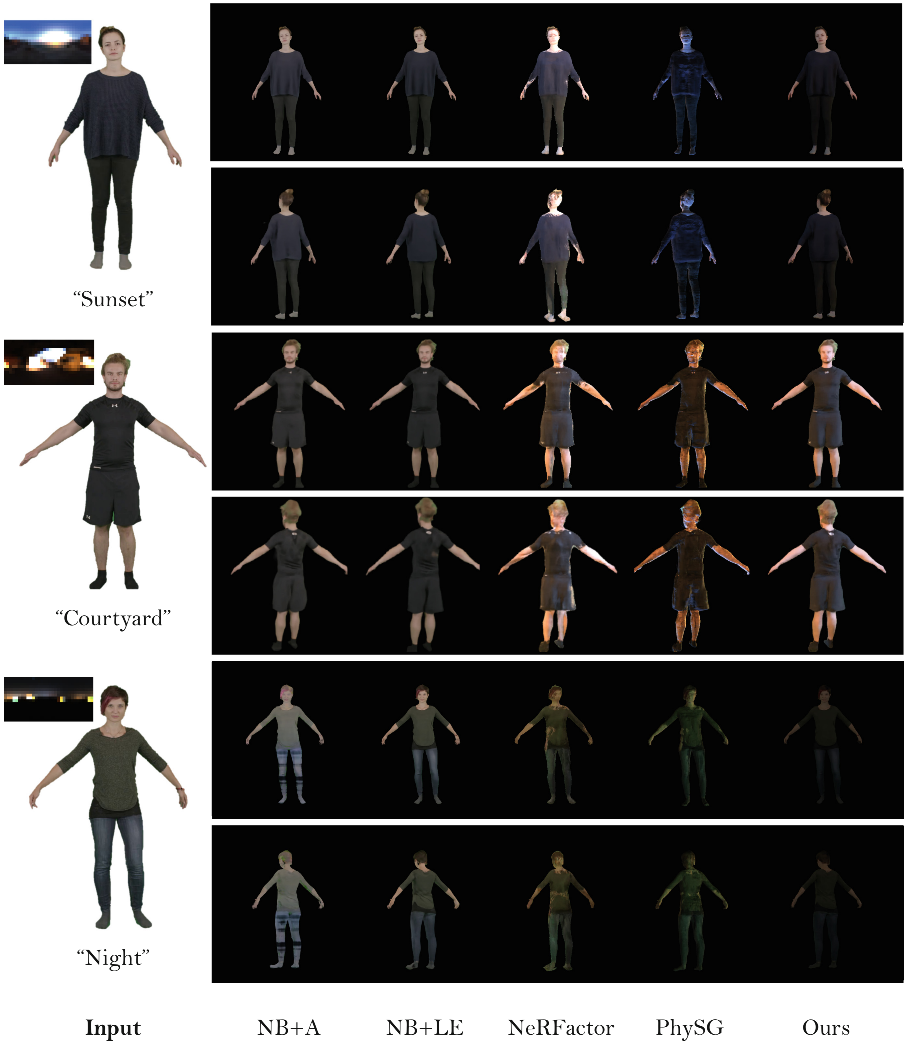

Fig. 5. Free-Viewpoint Relighting on the People-Snapshot Dataset. Two variants of NeuralBody [34] (NB+A and NB+LE) fail to incorporate the lighting in a physical way, thus are unable to reasonably relight the human actor. NB+A learns the wrong mapping between the target light and the appearance. And NB+LE reconstruct the input video well yet fails to generalize to novel lightings. NeRFactor [63] and PhySG [59] seem to model physically correct illuminations but gives blurry results due to the incapability of modeling dynamics. *Relighting4D* significantly outperforms comparison methods. **We show more results in both ambient lighting and OLAT setting in the supplementary videos.**

4 Experiments

Rendering Settings. We render humans in both the ambient lighting and the OLAT setting. For ambient lighting, we use publicly available[2] HDRi maps as light probes. Furthermore, for the OLAT setting, we simulate point lights by generating one-hot light probes given the incoming light directions.

Real Datasets. We validate our method on the People-Snapshot [1] dataset and ZJU-Mocap [34] dataset qualitatively. People-Snapshot [1] captures monocular videos with dynamic performers that keep rotating. And ZJU-Mocap [34] captures dynamic humans with complex motions using a multi-camera system.

Synthetic Dataset. To further demonstrate the effectiveness of *Relighting4D*, we create a dataset, **BlenderHuman**, using the Blender engine [7] for quantitative evaluation. Details will be deferred to the supplementary.

Comparison Methods. We compare *Relighting4D* with several competitive methods. **NeRFactor** [63] requires a pretrained NeRF as a geometry proxy and learns a data-driven BRDF to perform relighting, but it fails to represent dynamic scenes. **PhySG** [59] adopts a spherical Gaussians reflectance model which cannot handle high-frequency lights, and its geometry representation cannot model dynamic scenes. Moreover, to demonstrate the importance of physically based rendering, we implement two variants on top of NeuralBody (NB) [34] which succeeds in novel view synthesis of dynamic humans but fails to incorporate lighting and reflectance. **NB+Ambient Light** (NB+A) uses a flattened light probe as the latent code which contributes to the prediction of its color model, while **NB+Learnable Embedding** (NB+LE) maps the light probe into a latent code using an MLP with two layers.

Evaluation Metrics. For quantitative analysis, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure [52] (SSIM), and Learned Perceptual Image Patch Similarity [61] (LPIPS) as metrics. In addition, we use the error of degree(°) to measure the normal map estimations.

4.1 Results on Real Datasets

Performance on Relighting with Novel Views. Figure 5 shows qualitative results on People-Snapshot dataset. All methods train a separate model for each human performer and re-render the human given the input light probes. Two variants of NeuralBody [34], NB+A and NB+LE, are good at reconstructing appearance but fail to incorporate novel illuminations in a perceptually salient way. They fail to learn the underlying physics of rendering. For example, NB+A maps the input light probe to artifacts of texture while NB+LE even seems to discard the features from lightings. NeRFactor [63] and PhySG [59] give blurry results, which show that they cannot aggregate space-time varying geometry and reflectance of dynamic humans, leading to degraded rendering results. In contrast, our method generates photorealistic relit novel views.

[2] https://polyhaven.com/.

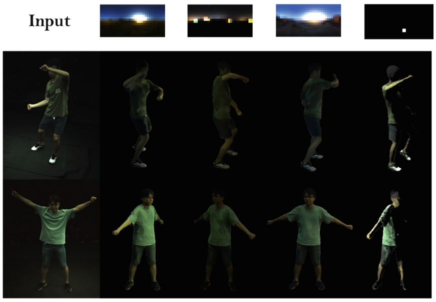

Fig. 6. Relighting of dynamic humans with complex motions on ZJU-Mocap dataset. *Relighting4D* renders high-fidelity human actors with time-varying poses under novel illuminations. **Please check the supplementary videos for more results.**

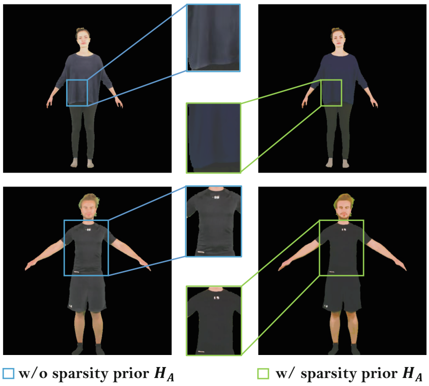

Fig. 7. Here we visualize how our model benefits from incorporating minimum-entropy sparsity prior by minimizing H_A. Without this prior, the estimation of diffuse map would suffer from shadow residuals as shown on the left side.

We also present our qualitative results in the challenging OLAT setting. Since the point light comes from only one direction, these OLAT illuminations induce hard cast shadows, effectively revealing rendering artifacts due to inaccurate geometry and materials. *Relighting4D* synthesizes shadows cast by limbs and clothes in a physically correct way. Please refer to the supplementary for details.

We demonstrate that our method is capable of relighting dynamic humans with complex motions from multi-views videos on the ZJU-Mocap dataset [34]. Figure 6 shows our qualitative results on the "Twirl" and "Swing" scenes.

Decomposition of Geometry and Reflectance. We demonstrate that our method is able to extract geometry and reflectance representations from the input videos and disentangle them into surface normals, diffuse maps and occlusion maps, which may facilitate downstream graphics tasks. The visualizations on People-Snapshot videos are presented in Fig. 10, and quantitative results on the BlenderHuman dataset are shown in Table 1. Note that, we directly take the albedo generated by NeRFactor [63] and PhySG [59] as their diffuse maps for comparisons. However, NeRFactor [63] estimates the diffuse map of the dynamic human with incorrect base color and facial details, and fails to capture the accurate geometry of dynamic humans. Though PhySG [59] captures the correct base color of clothes, due to its incapability of handling dynamic scenes, the facial details of diffuse map remains artifacts when the viewpoint changes. With the latent representation of human body, *Relighting4D* can integrate geometry information through space and time, successfully capturing the fine-grained details of the normal map and the correct color of diffuse map.

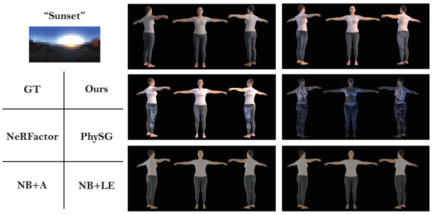

Fig. 8. Visualization of the "Sunset" scene in the BlenderHuman dataset. We customize a synthetic dataset, BlenderHuman, to provide ground truths of relit videos for quantitative evaluations. *Relighting4D* outperform other methods, producing promising results of relighting dynamic human under novel illuminations.

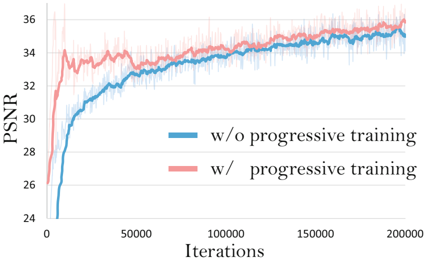

Fig. 9. The PSNR v.s. training iterations on People-Snapshot dataset. Progressive training helps the model reconstruct the scene faster and better. When trained with a constant spatial resolution, the reconstruction error falls into sub-optimal at the end (PSNR drops from 36.65 to 34.56).

Table 1. Results on the BlenderHuman Dataset. The top two techniques for each metric are highlighted in red and orange respectively. The reported numbers are the arithmetic averages of 16 different scenes. We relight the human actor with 8 HDR ambient light probes and 8 OLAT conditions. *Relighting4D* achieves the best overall performance across all metrics.

Method	Relighting			Normal Map	Diffuse Map		
	PSNR ↑	SSIM ↑	LPIPS ↓	Degree° ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NB [34]+A	20.9348	0.8559	0.2368	-	-	-	-
NB [34]+LE	22.7957	0.8721	0.2145	-	-	-	-
NeRFactor [63]	22.8037	0.8830	0.2045	43.7012	27.0585	0.9202	0.1929
PhySG [59]	23.8810	0.8427	0.2959	50.5721	28.0852	0.9350	0.1810
Ours	26.1475	0.9118	0.1639	32.1803	28.9517	0.9279	0.1502

4.2 Results on Synthetic Dataset

We quantitatively evaluate comparison methods on the BlenderHuman dataset. We show results in this section and defer the visualizations in the supplementary.

Table 1 shows the results on the BlenderHuman dataset. Overall, our model achieves the best performance. The results indicate that *Relighting4D* better handles the dynamics across video frames while feasibly modeling the light transport to relight dynamic humans.

4.3 Ablation Studies

We conduct ablation studies on BlenderHuman dataset, as presented in Table 2.

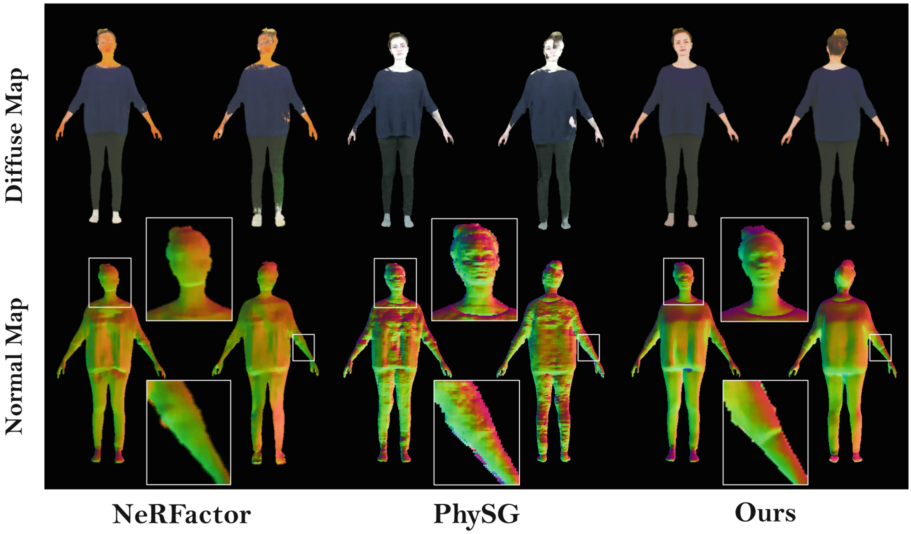

Fig. 10. Comparisons of geometry and reflectance decomposition. *Relighting4D* is able to estimate fine-grained details of geometry and physically correct reflectance. We defer the visualization of the occlusion map in the supplementary.

Impact of the Human Representation. We train our model without $\psi_t(\boldsymbol{x})$ that is proposed in Sect. 3.1. In other words, the geometry and reflectance MLPs take only the surface coordinates of the human body as inputs. The result indicates that incorporating our latent feature $\psi_t(\boldsymbol{x})$ is crucial for relighting dynamic human videos. The incapability of modeling scene dynamics leads to significant performance drops in terms of relighting quality and inverse rendering quality, which explains why NeRFactor [63] fails so badly.

Effectiveness of the smoothness regularization is validated in Table 2. We train our model without τ_V, τ_n, which leads to the decreased rendering quality.

Impact of the Baked Geometry. We train *Relighting4D* without the supervision of the baked geometry. The results indicate that the baked geometry improves the relighting performance. It reveals that poor renderings are caused by the inaccurate geometry as the error of normals drops by a large margin.

Effectiveness of the global sparsity prior on diffuse map is validated in Table 2. Without minimizing the entropy of the diffuse map, the relighting quality is perceptually decreased due to the degraded inverse rendering, which induces shadows in the estimation of diffuse map, as shown in Fig. 7.

Impact of Progressive Training. Without progressive spatial resolutions during training, the relighting quality decreases as shown in the last row of Table 2. We believe the progressive strategy helps the model quickly learn coarse geometry in the early training phase, which is even validated on real datasets (Fig. 9).

We plot the reconstruction error (PSNR) versus iterations on one training scene in People-Snapshot dataset, discovering that progressive training helps model reconstruct the given scene faster and better.

Table 2. Ablation Studies on the BlenderHuman Dataset. We take the average metrics on all 16 scenes. The top three techniques for each metric are highlighted in red, orange, and yellow respectively.

Method	Relighting			Normal Map	Diffuse Map		
	PSNR ↑	SSIM ↑	LPIPS ↓	Degree° ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Full model	26.1475	0.9118	0.1639	32.1803	28.9517	0.9279	0.1502
w/o $\psi_t(\boldsymbol{x})$	21.1163	0.8407	0.2372	36.5699	25.6806	0.9008	0.1896
w/o τ_V, τ_n	25.3504	0.8800	0.2061	32.9243	28.3329	0.9224	0.1660
w/o $\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{n}}$	22.4221	0.8559	0.2285	57.0452	27.6652	0.9165	0.1425
w/o H_A	27.7545	0.9042	0.1717	30.6685	24.0195	0.8950	0.1767
w/o progressive	25.5562	0.9031	0.1742	30.1662	24.2455	0.8958	0.1760

5 Discussion and Conclusion

Limitations. We have demonstrated the capability of *Relighting4D* on relighting dynamic humans with free viewpoints. Nevertheless, there are a few limitations. First, for tractable training and rendering, we consider only the one-bounce direct environment light, thus our method cannot relight furry appearances. Second, as we leverage a fully physically based renderer, the rendering quality is tied with the accuracy of geometry. Dense scenes with multiple people, which may negatively impact the estimation of geometry, will lead to poor performance. Finally, if the texture patterns are complicated or the lighting is harsh during training, the decomposition of reflectance and geometry is hard to solve due to the ambiguity of color scale, causing poor relighting quality. It can be alleviated by incorporating more information other than self-supervision from videos into the network (e.g., other supervision signals or data-driven priors).

In this paper, we present a principled rendering scheme called *Relighting4D*, a method that enables relighting with free viewpoints from only posed human videos under unknown illuminations. Our method exploits the physically based rendering pipeline and decomposes the appearance of humans into geometry and reflectance. All components are parameterized by MLPs based on the neural field conditioned on the deformable human model. Extensive experiments on synthetic and real datasets demonstrate that *Relighting4D* is capable of high-quality relighting of dynamic human performers with free viewpoints.

Acknowledgements. This work is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-019), NTU NAP, MOE AcRF Tier 2 (T2EP20221-0033), and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8387–8397. IEEE (2018). https://doi.org/10.1109/CVPR.2018.00875, https://ieeexplore.ieee.org/document/8578973/
2. Alldrin, N.G., Mallick, S.P., Kriegman, D.J.: Resolving the generalized bas-relief ambiguity by entropy minimization. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition.,pp. 1–7 (2007). https://doi.org/10.1109/CVPR.2007.383208
3. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 334–341. IEEE (2012). https://doi.org/10.1109/CVPR.2012.6247693, https://ieeexplore.ieee.org/document/6247693/
4. Bi, S., et al.: Deep relightable appearance models for animatable faces. ACM Trans. Graph. **40**(4), 1–15 (2021). https://doi.org/10.1145/3476576.3476647, https://dl.acm.org/doi/10.1145/3476576.3476647
5. Bi, S., et al.: Neural reflectance fields for appearance acquisition. arXiv:2008.03824 [cs] (2020)
6. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.P.A.: NeRD: neural reflectance decomposition from image collections. arXiv:2012.03918 [cs] (2021)
7. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation (2018). http://www.blender.org
8. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W.: Acquiring the Reflectance Field of a Human Face. In: SIGGRAPH (2000)
9. Finlayson, G.D., Drew, M.S., Lu, C.: Entropy minimization for shadow removal. Int. J. Comput. Vis. **85**(1), 35–57 (2009). https://doi.org/10.1007/s11263-009-0243-z
10. Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3D Shape. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4856–4865. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00491, https://ieeexplore.ieee.org/document/9157823/
11. Guo, K., et al.: The relightables: volumetric performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6), 1–19 (2019) https://doi.org/10.1145/3355089.3356571, https://dl.acm.org/doi/10.1145/3355089.3356571
12. Joo, H., Simon, T., Sheikh, Y.: Total capture: a 3D deformation model for tracking faces, hands, and bodies. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8320–8329 (2018). https://doi.org/10.1109/CVPR.2018.00868
13. Kanamori, Y., Endo, Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. ACM Trans. Graph. **37**(6), 1–11 (2019). https://doi.org/10.1145/3272127.3275104, https://arxiv.org/abs/1908.02714
14. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering. arXiv:2109.07448 [cs] (2021)
15. Lagunas, M., et al.: Single-image Full-body human relighting. arXiv:2107.07259 [cs] (2021). https://doi.org/10.2312/sr.20211300

16. LeGendre, C., et al.: DeepLight: learning illumination for unconstrained mobile mixed reality. CoRR abs/1904.01175 (2019), http://arxiv.org/abs/1904.01175
17. LeGendre, C., et al.: Learning illumination from diverse portraits. In: SIGGRAPH Asia 2020 Technical Communications. SA 2020, Association for Computing Machinery (2020). https://doi.org/10.1145/3410700.3425432
18. Li, J., Feng, Z., She, Q., Ding, H., Wang, C., Lee, G.H.: Mine: towards continuous depth MPI with nerf for novel view synthesis. In: ICCV (2021)
19. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. arXiv:2011.13084 [cs] (2021)
20. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: neural free-view synthesis of human actors with pose control. arXiv:2106.02019 [cs] (2021)
21. Liu, Y., Neophytou, A., Sengupta, S., Sommerlade, E.: Relighting images in the wild with a self-supervised siamese auto-encoder. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WaACV), pp. 32–40. IEEE (2021). https://doi.org/10.1109/WACV48630.2021.00008, https://ieeexplore.ieee.org/document/9423347/
22. Liu, Y., Li, Y., You, S., Lu, F.: Unsupervised learning for intrinsic image decomposition from a single image. arXiv:1911.09930 [cs] (2020)
23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)
24. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the Wild: Neural radiance fields for unconstrained photo collections. In: CVPR (2021)
25. Meka, A., et al.: Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. ACM Trans. Graph. **38**(4), 1–12 (2019). https://doi.org/10.1145/3306346.3323027, https://dl.acm.org/doi/10.1145/3306346.3323027
26. Meka, A., et al.: Deep relightable textures: volumetric performance capture with neural rendering. ACM Trans. Graph. **39**(6), 1–21 (2020)
27. Meka, A., Shafiei, M., Zollhoefer, M., Richardt, C., Theobalt, C.: Real-time global illumination decomposition of videos. ACM Trans. Graph. **40**(3), 1–16 (2021). https://doi.org/10.1145/3374753, http://arxiv.org/abs/1908.01961
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing Scenes as neural radiance fields for view synthesis. arXiv:2003.08934 [cs] (2020)
29. Oechsle, M., Peng, S., Geiger, A.: UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction. arXiv:2104.10078 [cs] (2021)
30. Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F.: Neural Scene Graphs for Dynamic Scenes. arXiv:2011.10379 [cs] (2021)
31. Pandey, R., et al.: Total relighting: learning to relight portraits for background replacement. ACM Trans. Graph. **40**(4), 1–21 (2021). https://doi.org/10.1145/3476576.3476588, https://dl.acm.org/doi/10.1145/3476576.3476588
32. Park, K., et al.: Nerfies: deformable neural radiance fields. arXiv:2011.12948 [cs] (2021)
33. Pavlakos, G., et al.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10975–10985 (2019)
34. Peng, S., et al.: Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. arXiv:2012.15838 [cs] (2021)

35. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: neural radiance fields for dynamic scenes. arXiv:2011.13961 [cs] (2020)
36. Raj, A., Tanke, J., Hays, J., Vo, M., Stoll, C., Lassner, C.: ANR: articulated neural rendering for virtual avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3722–3731 (2021)
37. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D., Kautz, J.: Neural Inverse rendering of an indoor scene from a single image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8597–8606. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00869, https://ieeexplore.ieee.org/document/9008823/
38. Shen, L., Yeo, C.: Intrinsic images decomposition using a local and global sparse representation of reflectance. In: CVPR 2011, pp. 697–704 (2011). https://doi.org/10.1109/CVPR.2011.5995738
39. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5444–5453. IEEE (2017). https://doi.org/10.1109/CVPR.2017.578, https://ieeexplore.ieee.org/document/8100061/
40. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Advances in Neural Information Processing Systems, vol. 33, pp. 7462–7473. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html
41. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: continuous 3D-structure-aware neural scene representations. arXiv:1906.01618 [cs] (2020)
42. Srinivasan, P.P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., Barron, J.T.: NeRV: neural reflectance and visibility fields for relighting and view synthesis. arXiv:2012.03927 [cs] (2020)
43. Sun, G., et al.: Neural free-viewpoint performance rendering under complex human-object interactions. arXiv:2108.00362 [cs] (2021)
44. Sun, T., et al.: Single Image Portrait Relighting. ACM Trans. Graph. **38**(4), 1–12 (2019). —DOIurl10.1145/3306346.3323008, https://arxiv.org/abs/1905.00824
45. Sun, T., Lin, K., Bi, S., Xu, Z., Ramamoorthi, R.: Nelf: Neural light-transport field for portrait view synthesis and relighting. CoRR abs/2107.12351 (2021). https://arxiv.org/abs/2107.12351
46. Suo, X., et al.: NeuralHumanFVV: real-time neural volumetric human performance rendering using RGB cameras. arXiv:2103.07700 [cs] (2021)
47. Tajima, D., Kanamori, Y., Endo, Y.: Relighting humans in the wild: Monocular full-body human relighting with domain adaptation (2021)
48. Tancik, M., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. arXiv:2006.10739 [cs] (2020)
49. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 548–557. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00063, https://ieeexplore.ieee.org/document/9156372/
50. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. In: Proceedings of the 18th Eurographics Conference on Rendering Techniques, pp. 195–206. EGSR 2007, Eurographics Association (2007). https://doi.org/10.2312/EGWR/EGSR07/195-206
51. Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Trans. Graph.

39(6), 1–13 (2020). https://doi.org/10.1145/3414685.3417824, https://dl.acm.org/doi/10.1145/3414685.3417824

52. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
53. Wang, Z., Philion, J., Fidler, S., Kautz, J.: Learning indoor inverse rendering with 3D spatially-varying lighting. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
54. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. arXiv:2011.12950 [cs] (2021)
55. Xu, D., Principe, J.: Learning from examples with quadratic mutual information. In: Neural Networks for Signal Processing VIII. Proceedings of the 1998 IEEE Signal Processing Society Workshop (Cat. No.98TH8378), pp. 155–164 (1998). https://doi.org/10.1109/NNSP.1998.710645
56. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: neural radiance fields from one or few images (2020)
57. Yu, Y., Smith, W.A.P.: InverseRenderNet: Learning single image inverse rendering. arXiv:1811.12328 [cs] (2018)
58. Zhang, J., et al.: Editable free-viewpoint video using a layered neural representation. ACM Trans. Graph. **40**(4), 1–18 (2021). https://doi.org/10.1145/3450626.3459756, https://arxiv.org/abs/2104.14786
59. Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: PhySG: inverse rendering with spherical gaussians for physics-based material editing and relighting. arXiv:2104.00674 [cs] (2021)
60. Zhang, L., Zhang, Q., Wu, M., Yu, J., Xu, L.: Neural video portrait relighting in real-time via consistency modeling. arXiv:2104.00484 [cs] (2021)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
62. Zhang, X., et al.: Neural light transport for relighting and view synthesis. ACM Trans. Graph. **40**(1), 1–17 (2021). https://dl.acm.org/doi/10.1145/3446328
63. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor. ACM Trans. Graph. **40**(6), 1–18 (2021). https://doi.org/10.1145/3478513.3480496, https://dx.doi.org/10.1145/3478513.3480496
64. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.: Deep single-image portrait relighting. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7193–7201. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00729, https://ieeexplore.ieee.org/document/9010718/