Neural Reconstruction of Relightable Human Model from Monocular Video

Wenzhang Sun
Beijing Institute of Technology
Beijing, China
swzlx1996@163.com

Yunlong Che
AI^2 Robotics
Beijing, China
yunlongche@126.com

Han Huang
AI^2 Robotics
Beijing, China
huang.h92@outlook.com

Yandong Guo
AI^2 Robotics
Beijing, China
yandong.guo@live.com

Abstract

Creating relightable and animatable human characters from monocular video at a low cost is a critical task for digital human modeling and virtual reality applications. This task is complex due to intricate articulation motion, a wide range of ambient lighting conditions, and pose-dependent clothing deformations. In this paper, we introduce a novel self-supervised framework that takes a monocular video of a moving human as input and generates a 3D neural representation capable of being rendered with novel poses under arbitrary lighting conditions. Our framework decomposes dynamic humans under varying illumination into neural fields in canonical space, taking into account geometry and spatially varying BRDF material properties. Additionally, we introduce pose-driven deformation fields, enabling bidirectional mapping between canonical space and observation. Leveraging the proposed appearance decomposition and deformation fields, our framework learns in a self-supervised manner. Ultimately, based on pose-driven deformation, recovered appearance, and physically-based rendering, the reconstructed human figure becomes relightable and can be explicitly driven by novel poses. We demonstrate significant performance improvements over previous works and provide compelling examples of relighting from monocular videos of moving humans in challenging, uncontrolled capture scenarios.

1. Introduction

Capturing the human appearance under varying poses, viewpoints, and environmental lighting is essential. This capability enables a range of applications from digital 3D human creation to immersive experiences for X-R experiences. Traditional pipelines, utilizing specialized equipment for multi-view human scanning [12, 10, 16], rely on expensive hardware and are not feasible in uncontrolled environments, making them unsuitable for individual users. In contrast, recent neural rendering methods such as NeRF [27] and its variants [51, 49, 14, 48, 28, 24, 24] have achieved significant progress in generating realistic human rendering effects. These methods are simple yet effective, offering a promising alternative to traditional pipelines. It has been demonstrated that the human body can be represented as neural fields, enabling control and relighting, which expands the potential for creating more flexible and adaptive virtual human models.

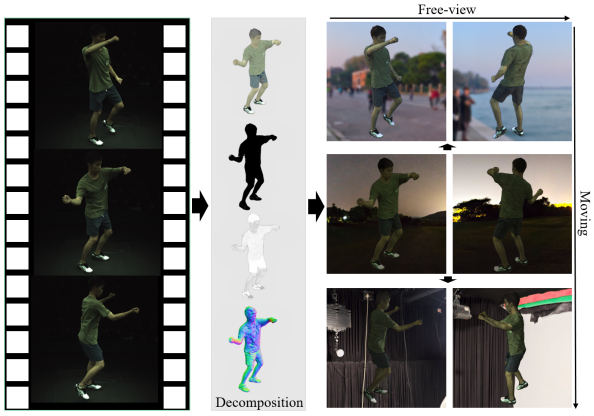

Figure 1. Given a monocular human motion video as input, our framework enables human relighting under novel illuminations and poses.

Different from existing methods that rely on multi-view static scans [21, 53, 8], we focus on the problem of relighting dynamic humans using only a single monocular video. As illustrated in Fig. 1, we model the dynamic human body by employing a neural human appearance field and a pose-

driven deformation field. The former encodes the dynamic human, accounting for varying illumination, into a canonical volume, while the latter allows explicit control of the canonical model using a condition code (such as SMPL or SMPL-X [23, 32]). Following the paradigm of recent work [5], we utilize a set of multi-layer perceptron (MLP) networks as an implicit representation to store the geometry and spatially varying BRDF (Bidirectional Reflectance Distribution Function) of the human body within the canonical volume. Geometry represents the human shape, encompassing characteristics like density, color, and normal. The spatially varying BRDF is broken down into three components: base color, roughness, and metalness. In summary, our work has the following contributions:

We present a principled framework, which is the first to build a relightable and animatable human in complex motion from a single video. we introduce a dense bidirectional mapping to modify rotation-related and rotation-unrelated parameters based on the deformation process.

We propose a progressive training strategy that enables learning BRDFs, surface normal, and ambient lighting in a self-supervised way under complex motions. Experiments show that our framework outperforms the state-of-the-art method in the human relighting task.

2. Related Work

Human Reconstruction. Collet et al. [10] recorded human performances using a dense array of RGB and IR video cameras, generating dynamic textured surfaces, and compressing them into a streamable 3D video format. Guo et al. [16] introduced a volumetric capture system designed for photorealistic and high-quality relightable full-body performance capture. However, these works demand complex multi-view configurations and costly hardware. Recent research [2, 54] has aimed at reconstructing detailed geometry and textures from color images through the use of parametric mesh fitting. For example, Alldieck et al. [2] fitted the SMPL model to all frames and optimized per-vertex offsets using dynamic human silhouettes. A core contribution of this paper lies in the transformation of a dynamic body into a canonical frame of reference, a method frequently employed to handle dynamic human bodies. Nevertheless, the main limitation of these techniques is their heavy reliance on a fixed base topology, resulting in poor generalization for loose clothing. Pifu [36], employing pixel-aligned implicit functions for human reconstruction, can predict high-resolution 3D shapes of individuals, including complex hairstyles and various clothing, even in largely unseen regions, from single or multi-view RGB input. Despite its capabilities, the reconstructed surfaces in Pifu are not rigid and thus cannot be relighted.

Human Neural Representation. The Neural Radiance Field (NeRF) is a cutting-edge method for 3D implicit representation that utilizes Multilayer Perceptron (MLP) to model a scene's geometry and view-dependent appearance. NeRF can be optimized from calibrated RGB images through differentiable volumetric rendering techniques [25]. Over the past two years, it has demonstrated exceptional performance across a variety of 3D applications, such as scene/object reconstruction [47, 41], relighting [5, 4, 38, 46], and generation [15, 19, 42]. Some researchers have extended NeRF to dynamic scenes by introducing a neural deformation field, allowing for the handling of deformations or pose synthesis through latent space interpolation [30, 40, 31, 35]. However, these works tend to struggle with rapid human motion and fail to generate images with specific pose inputs. To build a human model from motion sequences, several studies [45, 33, 22, 29] have represented a dynamic human using a neural field coupled with a pose-driven deformation field. This approach captures a human in a canonical volume through an implicit neural function that takes a position x as input and returns the corresponding human appearance value (such as geometry or color). A pose-driven deformation field depicts the deformation between the canonical space and the observed view, often decomposed into skeleton-driven deformation (for coarse body movements) and non-rigid deformation (for local deformities between the SMPL model and clothed human) [45, 33]. However, these methods, by incorporating all color information within the implicit functions, prove unsuitable for human relighting.

Human Relighting. Previous research [26, 18] approaches this task by framing it as an inverse rendering problem, with the objective of jointly recovering human geometry, reflectance, and illumination from images. Building on convolutional neural networks and extensive datasets that include labeled geometry and materials, LeGendre *et al.* [20] are able to predict high dynamic range, omnidirectional illumination from a single low dynamic range image. However, similar methods often heavily depend on the one-light-at-a-time image capture technique. This approach is not only challenging to execute but also limits the ability to relight subjects with novel poses, as it lacks appropriate 3D representations for such manipulations.

Recent research has explored the benefits of implicit NeRF representations, with significant advancements in various applications [5, 4, 38]. For example, Zhang et al. [50] utilized mixtures of spherical Gaussians to represent specular BRDFs and environmental illumination, while parameterizing geometry as a signed distance function. Boss *et al.* [5] introduced a neural reflectance decomposition framework, employing physically-based rendering to separate static scenes into shape and spatially varying BRDF material properties. This method, however, is limited to static scenarios and necessitates accurate extrinsic camera calibration. Relighting4D [9] takes a different approach,

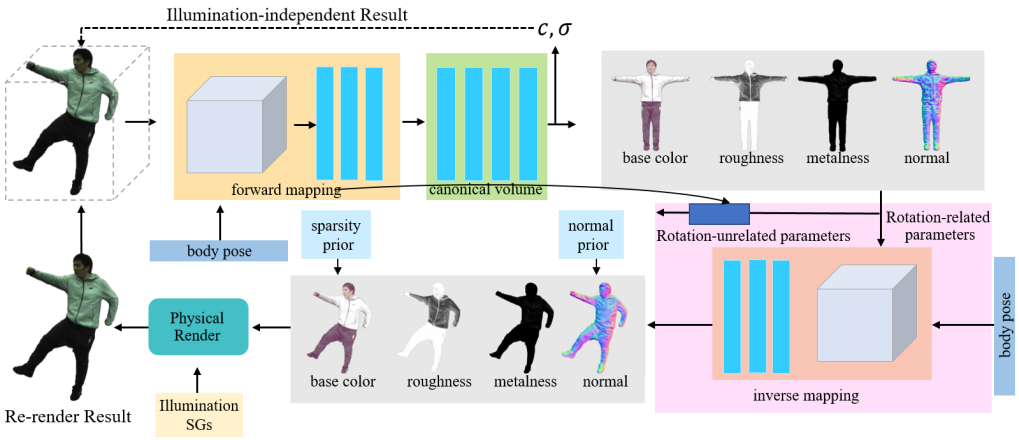

Figure 2. Overview of our proposed framework. Given an image with human poses, we sample points along the camera ray in frame view and transform these points to canonical volume with forward mapping. The canonical volume output the corresponding geometry and BRDFs. The photometric loss is imposed between the classical volume rendering results and input. Then, we warp back geometry and BRDFs from canonical volume to frame view with inverse mapping. We employ physically based rendering with these warped parameters to generate the re-render color. The re-render loss is constructed by distance between the re-render color and the input color. Our framework is optimized by minimizing photometric loss the re-render loss. Other priors are also employed during training..

disassembling the human model into geometry and 3D reflectance fields, and adopting the method from [29] to manage human motion. A limitation in this work is its reliance on articulated latent codes rather than explicit deformation, leading to difficulties for Relighting4D in producing accurate render results for complex motions. Additionally, the dependence on time steps further complicates the rendering process, causing the method to struggle when faced with poses that were not included in the training data.

Our reconstruction framework capitalizes on the strengths of multiple representations, harmoniously combining them to achieve specific goals. The implicit neural human appearance field facilitates realistic relighting effects across various illumination conditions, while the pose-driven deformation field offers precise articulation-based control. By incorporating the parameters of the SMPL model and illumination code, our framework has the capability to generate high-resolution images and finely-tailored 3D surface geometry through the utilization of physically-based rendering.

3. Method

Fig. 2 provides an overview of our framework. This section is organized as follows: We begin with a brief review of the basic NeRF formulation in Sec. 3.1. Next, we outline the human appearance field, denoted by $\mathcal{F}$, in Sec. 3.2, followed by an explanation of the pose-driven deformation field, represented by $\mathcal{W}$, in Sec. 3.3. We then introduce the specific priors employed during training in Sec. 3.4, and conclude this section with an overview of our training strategy in Sec. 3.5.

3.1. NeRF Overview

NeRF represents the target scene using a parameterized MLP, which takes the position x and views direction d as input, and outputs the density σ and radiance color c emitted by particles at that location along the given viewing direction. The density σ controls the accumulation of radiance by a ray passing through the point x. Differentiable volume rendering [17] is utilized to render the color C of a ray $r(t) = o + td$ within the range $[t_n, t_f]$, where o is the camera position, t represents the sampled step of the ray, and $[t_n, t_f]$ defines the bounding box of the scene. The accumulated transmittance along the ray is denoted by $T(t)$, and it can be expressed as follows:

$$T(t) = exp(-\int_{t-1}^{t} \sigma(t)dt) \tag{1}$$

Hence, the color C can be represent as:

$$\mathbf{C} = \int_{t_n}^{t_f} T(t) \cdot \sigma(t) \cdot c(t)dt \tag{2}$$

In practice, to more accurately represent high-frequency details, both the 3D position x and the viewing direction d are mapped into a higher-dimensional space using positional encoding [39].

3.2. Human Appearance Field

Following recent implicit representations of scene [5], we encode the geometry and BRDFs of the human in the canonical volume to a set of MLPs, which include color $c \in \mathbb{R}^3$, density $\sigma \in \mathbb{R}$, normal $n \in \mathbb{R}^3$ and BRDF $b \in \mathbb{R}^5$.

Instead of independently predicting diffuse and specular reflection colors, we employ the paradigm of Disney BRDF Base color Metallic parameterization [7]. The above canonical volumes are encoded as continuous filed $\mathcal{F}$:

$$\mathcal{F} : (x_c) \rightarrow [c, \sigma, n, b] \tag{3}$$

Where x_c is the point located in canonical space. Following Boss *et al.* [6], the $[\sigma, n, b]$ and illumination τ can be served into physically based rendering for relighting. The rendering equation is:

$$\mathbf{C} \approx \sum_{m=1}^{24} \rho_d(\omega_o, \tau_m, n, b) + \rho_s(\omega_o, \tau_m, n, b) \tag{4}$$

Where τ_m is the spherical Gaussian illumination map. ρ_d and ρ_s are the functions evaluated diffuse and specular lobes[43], which related with σ, we refer readers for more details in [6].

Illumination The environment map τ is represented by spherical Gaussian mixtures with parameters $\Gamma \in \mathbb{R}^{24\times7}$ (24 lobes). Each lobe is composed of axis $\in \mathbb{R}^3$, sharpness$\in \mathbb{R}^1$, and amplitude $\in \mathbb{R}^3$ which is initialized in a uniform distribution, the environment map τ as a parameter is also optimized. Consider a scene with fixed illumination; the incoming light from a certain viewpoint is related to the global rotation of the human. We use the global rotation at each frame to calculate illumination under different viewpoints during training. The axes of τ are corrected after each iteration to remain orthogonal. Moreover, rendering can produce an extensive value range depending on the incident light and the object's specularity, So sRGB curve is applied for this scenario.

3.3. Pose-driven Deformation Field

The pose-driven deformation field bridge the canonical volume and live frame view. It consists of two part: forward mapping $\overrightarrow{\mathcal{W}}$ and inverse mapping $\overleftarrow{\mathcal{W}}$. $\overrightarrow{\mathcal{W}}$ deflects the observation in frame view to canonical volume, and $\overleftarrow{\mathcal{W}}$ map BRDFs and normal back to frame view.

3.3.1 Forward Mapping

Following [45], we decouple the deformation $\vec{\mathcal{W}}$ which represents from frame view to canonical volume into two stages: skeleton motion $\vec{\mathcal{W}}_s(x, \theta)$ and non-rigid motion $\vec{\mathcal{W}}_{nr}(x, \theta)$. The warping process of point x_o can divided into two steps:

$$\vec{\mathcal{W}} : (x_o, \theta) \rightarrow \vec{\mathcal{W}}_s \rightarrow (x_s, \theta) \rightarrow \vec{\mathcal{W}}_{nr} \rightarrow (x_c) \tag{5}$$

where x_c is a 3D position in canonical volume, θ is the current human pose in SMPL format. The skeleton-driven deformation $\vec{\mathcal{W}}_s$, which represents the coarse deformation produced by joint rotation. It wraps point x_o to x_s(in canonical space). $\vec{\mathcal{W}}_{nr}$ starts from x_s and produces an offset $\triangle x$ to it, $\vec{\mathcal{W}}_{nr}$ provides the non-rigid effects caused by clothing. The $\vec{\mathcal{W}}_{nr}$ is considered as a offset $\triangle x$ to the skeleton-driven result x_s. To be specific, point x_o is warped by $\vec{\mathcal{W}}_s$ to the skeleton-driven position x_s. Then, the non-rigid motion MLP estimates the offset to the x_s and gets the final position $x_c = x_s + \triangle x$ in canonical space: $\vec{\mathcal{W}}_{nr} : (x_s, \theta) \rightarrow \triangle x$. Details are given in supplementary material.

3.3.2 Inverse Mapping

Based on the forward mapping $\vec{\mathcal{W}}$ and volume rendering equation Eq. (2), the estimated density σ and color c will generate illumination-independent color $\overrightarrow{C}_r$. During this process, only density and color back the propagation gradient. To optimize other decomposed parameters like BRDFs and normal, we introduce the inverse mapping, which warps point x_o in canonical space back to x_c in frame view. The warped BRDF b, density σ, normal n, and illumination τ are fed to a physically based rendering to generate the re-rendered color value $\overleftarrow{C}_r$.

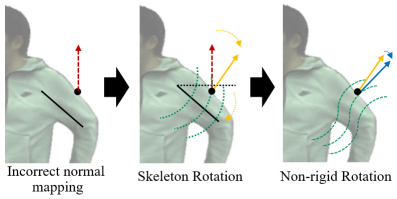

Figure 3. Visualization of the inverse mapping process of normal.

The inverse mapping $\overleftarrow{\mathcal{W}}$ warp the $[n, \sigma, b]$ back to the observation view. These decomposed parameters are divided into two categories: **rotation-unrelated** and **rotation-related**. The first category includes BRDFs b, density σ, and color c. The BRDFs, density, and color of the human body are consistent in observation and canonical spaces. We represent them with the queried result in the canonical volume. In the training iteration, we obtained the point-to-point mapping from frame view to canonical volume via forward mapping; in the inverse mapping process, the dense associated pairs can directly map the rotation-unrelated parameters back to observation. Hence, the $\overleftarrow{\mathcal{W}}$ for $[\sigma, b]$ are computed by inverse the $\overrightarrow{\mathcal{W}}$. The second category is the rotation-related parameter: normal. It cannot simply represent warping by adding translation offset(we show an example in Fig. 3). Therefore, we need to compute the normal of the human body in frame view. To distinguish, we use $\mathcal{R}$ to represent the $\overleftarrow{\mathcal{W}}$ for n. The inverse mapping of normal is similar to the forward mapping process, but the final output is the $\triangle\alpha$ which rotates the normal in canonical space into observation. We break it into two steps: inverse skeleton

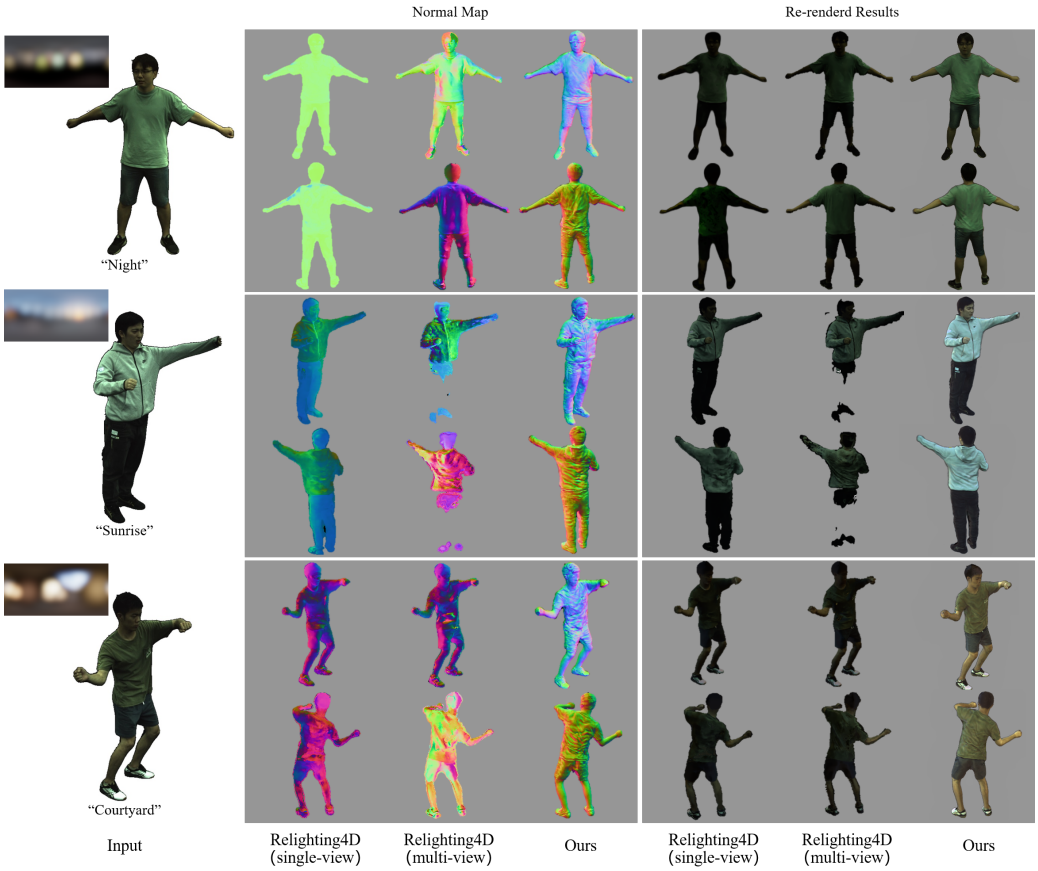

Figure 4. Novel view comparison with Relighting4D and a variant of Relighting4D (single view). We show the novel view of normal and re-rendered results on ZJU-Mocap dataset. Relighting4D and the variant method fail to give an accurate normal and physically incorporate the lighting.

rotation $\mathcal{R}_s$ and non-rigid rotation $\mathcal{R}_{nr}$. The final normal n_o in observation space can be compute as :

$$n_o = n_c \cdot \mathcal{R}_s\left(x_o, \theta\right) \cdot \mathcal{R}_{nr}\left(\mathcal{R}_s\left(x_o, \theta\right), \theta\right) \tag{6}$$

Where $n_c \in \mathbb{R}^3$ is the queried normal in canonical space, $R_s\left(x_o, \theta\right) \in \mathbb{R}^3$ is skeleton-driven rotation, $R_{nr}\left(\mathcal{W}_s\left(x_o, \theta\right), \theta\right) \in \mathbb{R}^3$ is the rotation caused by non-rigid effect. As shown in Fig. 3, we visualize how the incorrect normal in the canonical space is corrected by skeleton and non-rigid rotations.

Inverse Skeleton Rotation. The inverse skeletal rotation warps the direction of a given point in the canonical space to the frame view. The process can be denoted:

$$\mathcal{R}_s(x_o, \theta) = \frac{\sum_{i=1}^{24} \overleftarrow{w}_i(R_i x_o)}{\sum_{i=1}^{24} \overleftarrow{w}_i(x_o)} \tag{7}$$

Similar to the skeleton motion in forward mapping, the $\overleftarrow{w}_i(x_o)$ is the i-th blend weight of x_o corresponding to i-th bone which is stored in a weight volume $\overleftarrow{V}$. The current weight for a given point x_o in the observation space is calculated as $\overleftarrow{w}_i = \overleftarrow{V}(R_i x_o + t_i)$, $\hat{R}_i$ is the corresponding rotation matrix which is computed from pose θ.

Inverse Non-rigid Rotation. The inverse non-rigid motion output the offset rotation $\triangle\alpha$ to $\mathcal{R}_s(x_o, \theta)$. $\triangle\alpha$ is estimated using an MLP in the form of Euler angles. $\mathcal{R}_s(x_o, \theta)$ is encoded with sinusoidal positional encoding, it can be denoted as:

$$\mathcal{R}_{nr} : (x_s, \theta) \rightarrow \triangle\alpha \tag{8}$$

3.4. Priors of Decomposed Parameters

Normal Prior. Previous work [5] define normal as the normalized inverse gradient of the local density field:

$$\boldsymbol{n} = -\frac{\nabla_x \sigma}{\|\nabla_{\boldsymbol{x}} \sigma\|} \tag{9}$$

However, such expression leads to uneven surfaces in dynamic scenes. We employ n to initialize the network by Eq. (10), and adopt a MLP network $\mathcal{F}_{n_c}$ to predict the normal $\boldsymbol{n_c}$ in canonical space and using n as weak supervision:

Figure 5. Decomposition and relighting results of dynamic human. Our method produces plausible BRDFs and normal of a dynamic human. When relighting with unseen illuminations, the appearance is well reproduced.

$$\boldsymbol{n_c} = \begin{cases} -\frac{\nabla_x \sigma}{\|\nabla_{\boldsymbol{x}} \sigma\|} & \text{step} <= \text{t}; \\ \mathcal{F}_{n_c}(x, \theta) & \text{step} > \text{t}; \end{cases} \tag{10}$$

Local Smooth Prior. We measure the local smoothness of metalness by adding 3D perturbation ε which is sampled from a Gaussian distribution with zero mean and standard deviation 0.01. We regularize metalness m in BRDF by L1 penalty: $L_m = |m(x+\varepsilon) - m(x)|_1$.

Sparsity Prior. Previous research [13, 3, 37] has proved that global minimum-entropy sparsity prior can remove shadow effectively. Base color should be sparse enough. Given n sampled rays, the PDF of the base color $B(x)$ can be estimated using a Gaussian KDE (Kernel Density Estimator). The entropy of $B(x)$ is computed as an expectation:

$$L_s = \mathbb{E}[-log(\frac{1}{n}\sum_{i=1}^{n} K_G(B(x) - B_i(x)))] \tag{11}$$

Where K_G is the standard normal density function.

3.5. Progressive Training

The proposed framework concurrently optimizes the human's normal, shape, BRDFs, illumination, and deformation field. Achieving this is particularly challenging when only video input is provided, and the illumination is unknown. To address this, we utilize a progressive end-to-end training strategy. Specifically, we minimize the difference between the illumination-independent color and the ground truth color, a metric referred to as the photometric loss. In addition to the photometric loss, we leverage the re-rendered color for further supervision. Weak supervision is also provided through the use of a sparsity prior and predicted normal, which align with the inverse gradient of the density field. The total loss is defined as follows:

$$\begin{aligned} \mathcal{L} = &\overrightarrow{\lambda} \sum_{r \in R} ||\overrightarrow{C(r)} - C(r)|| + \overleftarrow{\lambda} \sum_{r \in R} ||\overleftarrow{C(r)} - C(r)|| \\ &+ \overleftarrow{\lambda}(\lambda_n L_n + \lambda_s L_s + \lambda_m L_m) \end{aligned} \tag{12}$$

where $\lambda_n = \sum ||\boldsymbol{n_c} - \boldsymbol{n}||$, R is the set of rays in each batch, and $C(r)$ is the ground-truth color. $\overrightarrow{C(r)}$ is illumination-independent colors which generate by human appearance field $\mathcal{F}$, $\overleftarrow{C(r)}$ is re-render colors by physically render. As shown in Eq. (12), we assign a weight $\overrightarrow{\lambda} = 0.9^{\frac{iter}{5000}}$ to the illumination-independent loss at the start of training, and this weight will gradually fade out throughout the training process. Details are given in supplementary material.

4. Experiment

To validate the efficiency of our approach in handling dynamic humans across a broad spectrum of motions, we conducted a series of experiments. These experiments focused on parameter decomposition, relighting, free-view rendering, and animation to thoroughly assess the capabilities of our method.

Comparison methods. We compare our method with the state-of-the-art human-relighting method **Relighting4D** [9], and dynamic human modeling method **HumanNeRF** [45]. **Relighting4D** predicts diffuse and specular characteristics to perform relighting but requires a pre-trained model as a geometry proxy and relies on multi-view videos from the ZJU-Mocap dataset. This method struggles with representing dynamic humans over a large range of motions and cannot ensure physical accuracy. Furthermore, we crafted a single-view version of Relighting4D, named **Relighting4DS**, which is trained on monocular videos to

provide a fair comparison with our method. On the other hand, **HumanNeRF** is adept at modeling dynamic humans exhibiting large motions but falls short in relighting them under varying lighting conditions (visual results are available in the supplementary material).

Datasets: We validate our method on the ZJU-Mocap dataset [34] qualitatively. This dataset captures dynamic humans engaged in complex motions using a multi-camera system; however, for our training data, we exclusively utilize images captured by camera 1. To further illustrate the efficacy of our approach, we generate a challenging synthetic dataset using the Blender engine [11] for both qualitative and quantitative assessments. This dataset is crafted around a 3D human character animated with intricate dance movements. We render the character under eight different lighting conditions, selecting one video for training and reserving the others for evaluation. Additional details are provided in the supplementary material.

Evaluation metrics. For quantitative analysis, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [44], and Learned Perceptual Image Patch Similarity (LPIPS) [52] as metrics. We use the mean square error (MSE) as an evaluation metric for the normal map.

Results are rendered in the ambient lighting and the OLAT setting. We use publicly available HDRi maps [1] for ambient lighting. We generate one-hot light probes given the incoming light directions as the HDRi maps.

4.1. Result on Real Dataset

Relighting with novel views. Fig. 4 presents the qualitative results on the ZJU-Mocap dataset. Each method trains a distinct model for each subject and re-renders the human in accordance with the input illuminations. Specifically, Relighting4D is trained using videos captured from four different viewpoints, and in the original paper, the input frame size is set to 60. A variant, Relighting4DS, is trained using a monocular video taken by camera 1, employing the entire video frames as input (1000 frames for Subject 313, 510 frames for Subject 387, and 554 frames for Subject 392). Our method and HumanNeRF adhere to the same training settings as those used for Relighting4DS.

Fig. 4 reveals that both Relighting4D and Relighting4DS struggle to accurately estimate the normal of the human in scenarios with extensive movement, resulting in re-rendering results that appear unreasonable. Despite these challenges, our method continues to produce satisfactory rendering results. A specific issue with Relighting4D occurs when the dynamic color range is narrow and the motion extensive (as with subject 387); the method fails to correctly distinguish the foreground from the background. This leads to the disappearance of the person's legs in the image. In contrast, our method consistently provides well-re-rendered images and accurately captures the normal of the dynamic human subject (a visual comparison with HumanNeRF is provided in the supplementary material).

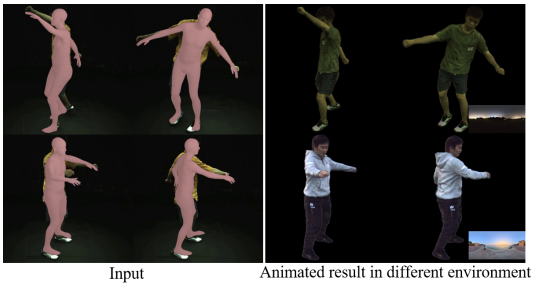

Figure 6. Animated results under different lighting conditions. Our method can be driven by pose data and re-rendered under unseen illuminations.

Decomposition and relighting of moving human. Fig. 5 presents the decomposition and relighting results for the dynamic human subjects within the ZJU-Mocap dataset. Our method successfully models dynamic humans with extensive movement and renders images under various lighting conditions. Although the estimated BRDF parameters may exhibit imperfections in certain areas, our method still manages to reproduce the images, even in a setup with purely passive unknown illumination. These deviations can be attributed to the inherent ambiguity of the dynamic decomposition problem and to differences in shading based on spherical Gaussians (SGs).

Animated results. Since our entire network is driven by human pose, it enables the use of pose-driven data for animation. We tested this capability on the ZJU-Mocap dataset by evaluating motions that were not included in the training set. Our approach produces convincing animated results, and the model can be re-rendered under diverse lighting conditions, as illustrated in Fig. 6. Additional results are provided in the supplementary material.

4.2. Results on Synthetic Dataset

Figure 7. OLAT results on the synthetic dataset. The shadow cast by limbs and clothes proves that our rendering results are physically correct.

We conducted a quantitative comparison with Relighting4D on the simulation dataset, as shown in Tab. 1. Due to slight differences in the normal coordinates, we focus

solely on comparing the effects of the re-rendered results. Our method is better able to capture the dynamic deformation process and surpasses Relighting4D across all evaluation metrics.

Method	Relighting		
	PSNR↑	SSIM↑	LPIPS*↓
Relighting4DS	26.58	0.9742	34.03
Relighting4D	27.57	0.9543	23.46
Ours	29.87	0.9889	22.65

Table 1. Comparison with Relighting4D on the synthetic dataset. Our method outperform Relighing4D in all evaluation metrics

We also present our qualitative results in the challenging OLAT setting, as illustrated in Fig. 7. The one-hot OLAT HDRi maps are represented using a set of SGs, and our method delivers plausible results, with convincing highlights on the limbs and clothes. Additional results are provided in the supplementary material.

4.3. Ablation Experiment

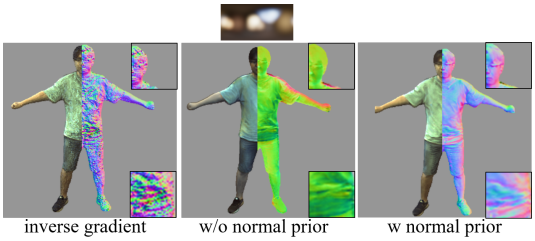

Figure 8. The normal obtained using the inverse gradient is uneven due to large human motions. Accurate normal cannot be obtained by MLP prediction because of lacking constraints. Using inverse gradient as weak supervision leads to better results.

Normal prior. We evaluated different methods for estimating the normal, including: 1. Directly utilizing the inverse gradient of the density field; 2. Predicting the normal using an MLP without a normal prior; 3. Estimating the normal with an MLP alongside a normal prior. We present the normals and re-rendered results for both the ZJU-Mocap and synthetic datasets in Fig. 8 and Tab. 2. Our strategy, employing the gradient as weak supervision, yields more precise results. Using the inverse gradient as the normal leads to a slight improvement in normal estimation (a 7% reduction in error) compared to direct estimation using the MLP. However, it also results in a worse re-rendered outcome (a 3% drop in PSNR performance). In comparison to the inverse gradient, our method produces more accurate normals (a 37% reduction in MSE error) and improved re-rendered results (a 11% enhancement in PSNR performance).

Sparsity prior. Utilizing the sparsity prior helps to eliminate most of the shadows in the base color, as shown in Fig. 9. While this approach may slightly reduce the accuracy of the normal estimation (resulting in a 3% increase in error), it enhances the overall re-rendering results (yielding a 2% improvement in PSNR performance).

Method	Relighting			Normal map
	PSNR↑	SSIM↑	LPIPS*↓	MSE↓
inverse gradient	26.76	0.9751	33.97	0.0979
w/o normal prior	27.63	0.9832	30.02	0.1051
w/o sparsity prior	29.41	0.9877	22.88	0.0598
full mode	29.87	0.9889	22.65	0.0617

Table 2. Ablation studies of normal and sparsity prior on the synthetic dataset. Estimating normal using MLP with normal prior achieves the best overall performance across all metrics. (best: red; second: yellow; LPIPS* = LPIPS $\times 10^3$.)

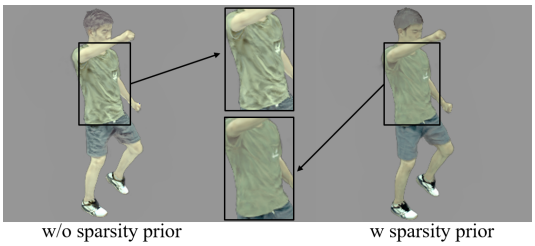

Figure 9. Visualization result of the base color. Using sparsity prior can effectively remove the shadow residual in the base color.

Inverse mapping network. Without inverse mapping, the results are constrained by the representation of point positions in canonical space. This limitation prevents accurate outcomes for a large range of motions and non-rigid body deformation. On the other hand, the use of inverse mapping enhances the normal's accuracy, as illustrated in Fig. 10. Moreover, we conducted a quantitative comparison on the synthetic dataset, detailed in Tab. 3. Fine-tuning with inverse mapping further elevates the performance, leading to more convincing re-rendered results (a 3% improvement in PSNR) and a more precise normal (6% MSE error drop).

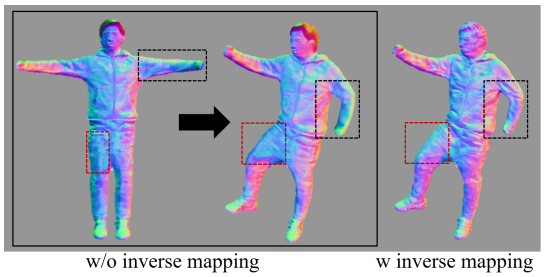

Figure 10. The inverse mapping method can modify the queried results in canonical space. The correction is made for the normal rotation caused by skeleton-driven motion and non-rigid deformation. This results in a more accurate normal in observation space.

Progressive training. Progressive training is instrumental within our framework, leading to refined estimation outcomes. As delineated in Tab. 3, the application of progressive training engenders an enhancement in re-rendered images (evidenced by a 6% increment in PSNR) and an

Method	Relighting			Normal map
	PSNR↑	SSIM↑	LPIPS*↓	MSE↓
w/o inverse mapping	28.40	0.9821	30.34	0.0648
w/o progressive	27.61	0.9731	35.01	0.0756
full mode	29.87	0.9889	22.65	0.0617

Table 3. Ablation studies of inverse mapping on the synthetic dataset. Using inverse mapping leads to better render results and more accurate normal.(best: red; second: yellow; LPIPS* = LPIPS $\times 10^3$.)

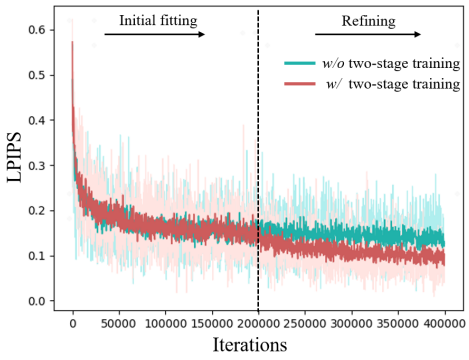

Figure 11. Visualization of the LPIPS during training on ZJU-Mocap dataset ($t = 200000$). Progressive training makes more plausible decomposition initialization, which leads to more precise results.

augmentation in the precision of normals (manifested by a 20% reduction in MSE error). Moreover, an illustration of the correlation between LPIPS and iterations during the training process on the ZJU-Mocap dataset is presented in Fig. 11. The initialization facilitated by progressive training steers the process towards more accurate results.

5. Discussion

Limitations: Our framework necessitates a relatively precise initial prediction of SMPL parameters. Any inaccuracies in these parameters may result in an imprecise reconstruction, rendering the decomposed parameters unrealistic, particularly in scenes characterized by continuous illumination changes and the absence of necessary constraints. Therefore, potential avenues for future research include the incorporation of a human pose estimation module or the modeling of more nuanced appearance aspects, such as shadows or high-frequency lighting effects.

Conclusion: In this work, We introduce a novel framework that facilitates the recovery of a relightable and animatable human model from a monocular video. This framework dissects the human appearance into geometry and reflectance, both represented as neural fields. Furthermore, it incorporates pose-driven deformation fields, which facilitate bidirectional mapping of human appearance between canonical space and observation. This deformation field aligns the surface normal with the human body during optimization, enabling our framework to realistically simulate relighting effects under arbitrary, unseen illuminations. Extensive experiments conducted on both synthetic and real datasets confirm that our approach is adept at high-quality relighting of dynamic human subjects, even when assuming novel poses.

References

[1] Poly haven. In *https://www.polyhaven.com*, 2020. 7

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[3] Neil G Alldrin, Satya P Mallick, and David J Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–7. IEEE, 2007. 6

[4] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sulkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *https://arxiv.org/abs/2008.03824*, 2020. 2

[5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural reflectance decomposition from image collections. *https://arxiv.org/abs/2012.03918*, 2020. 2, 3, 5

[6] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2020. 4

[7] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *ACM SIGGRAPH*, volume 2012, pages 1–7. vol. 2012, 2012. 4

[8] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos, 2021. 1

[9] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 606–623, Cham, 2022. Springer Nature Switzerland. 2, 6

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34:1 – 13, 2015. 1, 2

[11] Community. B.o.: Blender - a 3d modelling and rendering package. In *Blender Fundation*, 2018. 7

[12] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of ACM SIGGRAPH*, volume 2000, pages 145–156, 07 2000. 1

[13] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 6

[14] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903*, 2020. 1

[15] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis, 2021. 2

[16] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1, 2

[17] James T. Kajiya and Brian P Von Herzen. Ray tracing volume densities. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '84, page 165–174, New York, NY, USA, 1984. Association for Computing Machinery. 3

[18] Yoshihiro Kanamori and Yuki Endo. Relighting humans: Occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.*, 37(6), dec 2018. 2

[19] Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J. Rezende. Nerf-vae: A geometry aware 3d scene generative model. *https://arxiv.org/abs/2104.00587*, 2021. 2

[20] Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, S. Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul E. Debevec. Learning illumination from diverse portraits. *SIGGRAPH Asia 2020 Technical Communications*, 2020. 2

[21] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 1

[22] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM SIGGRAPH Asia*, 2021. 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[24] Ricardo Martin-Brualla, Noha Radwan, Mehdi Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. *https://arxiv.org/abs/2008.02268*, 2020. 1

[25] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 2

[26] Abhimitra Meka, Mohammad Shafiei, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. Real-time global illumination decomposition of videos. *ACM Trans. Graph.*, 40(3), aug 2021. 2

[27] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *The European Conference on Computer Vision (ECCV)*, 2020. 1

[28] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. *https://arxiv.org/abs/2011.12100*, 2020. 1

[29] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. *arXiv preprint arXiv:2104.03110*, 2021. 2, 3

[30] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. *https://arxiv.org/abs/2011.10379*, 2020. 2

[31] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan Goldman, Steven Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. *https://arxiv.org/abs/2011.12948*, 2020. 2

[32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[33] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. *arXiv preprint arXiv:2105.02872*, 2021. 2

[34] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 7

[35] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *https://arxiv.org/abs/2011.13961*, 2020. 2

[36] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 2

[37] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR 2011*, pages 697–704. IEEE, 2011. 6

[38] Pratul Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. *https://arxiv.org/abs/2012.03927*, 2020. 2

[39] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *CoRR*, abs/2006.10739, 2020. 3

[40] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a deforming scene from monocular video. *https://arxiv.org/abs/2012.12247*, 2020. 2

[41] Alex Trevithick and Bo Yang. GRF: Learning a general radiance field for 3D scene representation and rendering. *https://arxiv.org/abs/2010.04595*, 2020. 2

[42] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. *arXiv preprint arXiv:2112.05139*, 2021. 2

[43] Jiaping Wang, Peiran Ren, Minmin Gong, John Snyder, and Baining Guo. All-frequency rendering of dynamic, spatially-varying reflectance. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10. Association for Computing Machinery, 2009. 4

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[45] Chung–Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16189–16199, 2022. 2, 4, 6

[46] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[47] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronan Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[48] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 1

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. *https://arxiv.org/abs/2012.02190*, 2020. 1

[50] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5449–5458, 2021. 2

[51] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NERF++: Analyzing and improving neural radiance fields. *https://arxiv.org/abs/2010.07492*, 2020. 1

[52] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. 7

[53] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, June 2022. 1

[54] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G. Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video, 2020. 2