Gaussian Shadow Casting for Neural Characters

Luis Bolanos[1] Shih-Yang Su[1] Helge Rhodin[1,2]
[1]The University of British Columbia [2]Bielefeld University

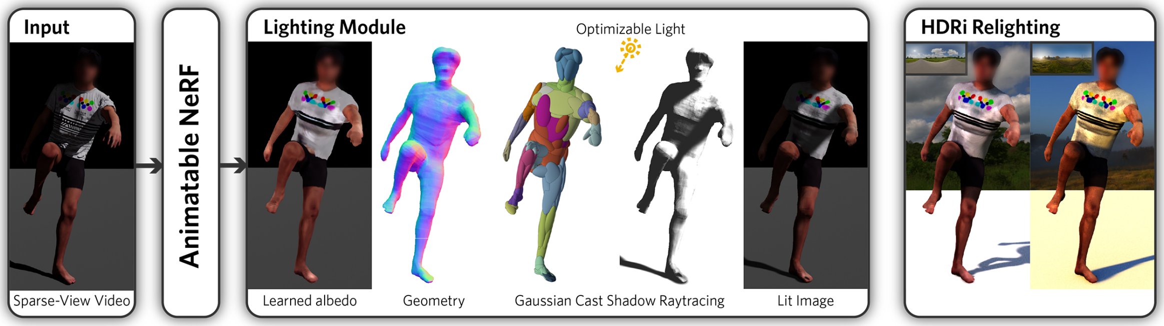

Figure 1. **Gaussian Shadow Casting (GSC):** Our method is able to reconstruct 3D neural characters from a sparse set of videos in settings with strong directional illumination. GSC uses a sum of Gaussians density model to cast secondary shadow rays efficiently with an analytic formula. Our method learns to remove shadows from the neural color field, allowing us to relight in novel illuminations. **All faces are blurred for anonymity.**

Abstract

Neural character models can now reconstruct detailed geometry and texture from video, but they lack explicit shadows and shading, leading to artifacts when generating novel views and poses or during relighting. It is particularly difficult to include shadows as they are a global effect and the required casting of secondary rays is costly. We propose a new shadow model using a Gaussian density proxy that replaces sampling with a simple analytic formula. It supports dynamic motion and is tailored for shadow computation, thereby avoiding the affine projection approximation and sorting required by the closely related Gaussian splatting. Combined with a deferred neural rendering model, our Gaussian shadows enable Lambertian shading and shadow casting with minimal overhead. We demonstrate improved reconstructions, with better separation of albedo, shading, and shadows in challenging outdoor scenes with direct sun light and hard shadows. Our method is able to optimize the light direction without any input from the user. As a result, novel poses have fewer shadow artifacts, and relighting in novel scenes is more realistic compared to the state-of-the-art methods, providing new ways to pose neural characters in novel environments, increasing their applicability. Code available at: `https://github.com/LuisBolanos17/GaussianShadowCasting`

1. Introduction

It is now possible to reconstruct animatable 3D neural avatars from video but methods do not account for accurate lighting and shadows. They have to rely on recordings that have soft uniform lighting, which precludes recording outdoors in direct sun light and on film sets with spotlights, and most are unable to relight characters in novel environments, limiting their applicability in content creation.

The most recent body models [17,20,28,36,37,40] which are based on neural radiance fields (NeRFs) [26], approximate the light transport by casting primary rays between the camera and the scene, sampling the underlying neural network dozens of times along each ray to obtain the density and color. As they do not include an illumination model, the color that the NeRF learns includes lighting, shadow, and view-dependent effects. Learning a body model in a challenging scene with a strong directional light source, such as the sun, leads to the neural field overfitting to the observed shadows. It does not generalize to novel poses, as the cast shadows are global effects where movement of a joint could affect the appearance of other distant areas of the body. Figure 1 shows such setting. This is in contrast to local shading effects such as wrinkles in clothing which current body models can successfully reconstruct.

To cast shadows within NeRF, secondary ray tracing from the reconstructed body model to the light source is an option. Although the predominant NeRF formulation enables casting secondary rays without change, it comes with a massive computational cost. For each sample along the primary ray, an equal number of secondary rays would have to be computed, each with multiple samples, leading to a quadratic, instead of linear, complexity in the number of samples per pixel. As a result, current re-lighting models only support diffuse reflection [16], hard shadows that do not generalize to novel poses [7], and soft dynamic shadow maps by approximate sphere tracing [43].

Our core contribution is introducing an additional volumetric density field that is approximate but significantly speeds up dynamic shadow casting while still maintaining differentiability and smoothness for gradient-based optimization. We introduce an anisotropic Gaussian density model and associated rendering functions that approximate the fine-grained density of the NeRF. The Gaussians have the beneficial property that we can integrate their density along a ray in closed form, thereby avoiding any sampling steps. Our derivation and implementation differs significantly from existing work using Gaussians for rendering. Compared to Gaussian Splatting [18, 31, 33], we neither require an affine approximation nor back-to-front ordering. Compared to Gaussian density models we alleviate their sampling [31] with an analytic integration and extend the existing analytic integration [30] to apply to anisotropic Gaussians. Notably, the Gaussian density is optimized alongside the NeRF without requiring a reference mesh such as SMPL [25]; it is template-free.

To further reduce runtime, we use a deferred shading approach [7] in which the first rendering pass computes the albedo, depth, and normal for each pixel. The second pass casts only one secondary ray per pixel from the estimated surface point to the light source. This makes shadow computations independent of the number of samples in the NeRF, avoiding the mentioned quadratic complexity.

Our experiments with strong directional light and cast shadows demonstrate that our explicit lighting reduces the occurrence of artifacts in novel-view and novel-pose synthesis tasks. Figure 1 shows how our method is able to disentangle lighting and shadows from the avatar's albedo given sparse-view data from only a single illumination. We take advantage of the dynamic aspect of the data where we can observe the same body part in multiple illuminations as the subject moves. We further demonstrate the ability to optimize the unknown light directions without any user input or careful initialization. Moreover, relighting of the neural character enables us to composite recorded motions into novel scenes realistically, making them directly applicable in computer graphics and entertainment industries, as demonstrated by the HDRi re-lighting in Figure 1-right.

2. Related Work

We build on neural body models using NeRF [26], which we introduce briefly. The subsequent discussion focuses on relighting methods for 3D scenes and body models as well as how Gaussians are used in rendering and reconstruction.

Neural avatars model dynamic performances by conditioning the neural rendering model on a template mesh driven by skeleton motion [4, 16, 20, 23, 24, 40, 49, 50] or template-free by linking neural fields directly to a skeleton [22,27,36–38]. Our implementation uses the more flexible template-free approach but it is general enough to extend to any NeRF-based model.

Static NeRF scene relighting approaches can be categorized by either implicit [6, 9, 32, 34, 48] or explicit [13, 42] implementations. In implicit methods, the NeRF's MLP is extended to further output illumination data such as shadow, direct and indirect illumination or occlusion maps [9, 32, 34], or decompose the scene into material properties such as metallicity and roughness which can be used in a Bidirectional Reflectance Distribution Function (BRDF) lighting model [6, 48]. These extended MLPs are conditioned at training and test time on lighting information such as spherical harmonics coefficients [32], or light direction [9]. Implicit methods require large amounts of data in both multi-view and multiple illuminations with lighting information known or estimated [9, 32]. Explicit methods simulate how real light interacts with the environment which improves the generalizability to novel illuminations but are difficult to extend to dynamic scenes or objects. These methods either utilize a secondary data structure such as proxy geometries where lighting computations can be done using established methods [42], or attempt to cast the necessary secondary rays within the neural field's volume which comes with a significant computational burden [13].

Dynamic neural character relighting has been built on top of volume rendering methods [5, 7, 21, 29, 43, 44, 51] as well as 2D CNN based models [16]. Implicit methods again require large amounts of data which can only be captured using light stages with known illumination [5, 21, 44, 51] or, across multiple subjects for faces that are self-similar, each captured in a different setting with in-the-wild illumination [29]. Our model provides dynamic and explicit shading and is most closely related to the following three methods.

RANA [16] uses SMPL+D [3] to estimate the coarse geometry of a person and extract an albedo texture map using TextureNet [15]. Given a target pose, they render person-specific neural features alongside coarse albedo and normals from the SMPL-D model. These are passed through two CNNs to refine the albedo and normal maps. Finally, they generate a light map using spherical harmonics and the normal map which is multiplied by the albedo map to obtain the final lit image. While spherical harmonics allow a wide array of lighting conditions to be simulated, cast shadows are not present, e.g., an arm casting a shadow on the body. Our work implements a Gaussian density model [30, 31] to facilitate fast and efficient *secondary* ray tracing to compute these cast shadows.

Likewise, Relighting4D [7] uses SMPL [25] to condition a 4D neural field of latent features which are trilinearly interpolated based on the nearby vertices to the query location. The latent features are passed through an MLP to obtain geometry, occlusion, and reflectance properties which are fed through a BRDF to get the final lit image. It is able to estimate the light probe, and at inference time, be able to switch the light probe to a new illumination. However, Relighting4D was not designed to work with hard shadows in novel poses, which is the focus point of our work.

Finally, Xu et al. [43] utilize a signed distance field (SDF) based approach to learn a neural human avatar which utilizes SMPL-based inverse Linear Blender Skinning (LBS) and a displacement field to obtain canonical features. They utilize Hierarchical Distance Queries (HDQ) to compute minimum distances from world space to surface locations and perform sphere tracing to obtain material and surface properties for each camera ray. They further take advantage of HDQ through the SDF by computing soft visibility maps towards a learned light probe. While HDQ allows for fast occlusion checks, their solution focuses on soft approximate shadows whereas our work enables hard shadow casting.

Gaussians have been used in rendering applications as differentiable methods for computing visibility and occlusions [30, 31, 35], as components of environment maps [46], or as a means to improve rendering efficiency for neural scenes [18]. Most methods are limited to spherical Gaussians [30, 31], while Gaussian Splatting uses an affine approximation that is only accurate when many small Gaussians are used [18], and Sridhar et al. use an approximation by perspective projection of ellipsoids [33]. Our work extends Rhodin et al. [30, 31] to use anisotropic Gaussians, without introducing any approximation, and tailors the analytic formulas and implementation towards shadow casting.

3. Method

Our method reconstructs a neural character from a set of N images of width W and height H, $\{\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{N}$, and corresponding character poses $\theta_t \in \mathbb{R}^{J\times 4\times 4}$. The pose is represented as a skeleton with one 4×4 local-to-world transformation matrix for each of the J joints. Figure 2 gives an overview of our method. A key element of our design is a deferred illumination model [39] that separates the rendering into computing albedo, $\mathbf{a} \in \mathbb{R}^3$, surface normal, $\hat{\mathbf{n}} \in \mathbb{R}^3$, and depth, $\mathbf{d} \in \mathbb{R}$, in a first pass and subsequently adding shading and shadow, $s \in [0,1]$, in a second pass. Our key contribution is the closed form formula for the shadow s.

3.1. Deferred Neural Illumination

Our volumetric body model is optimized on a reconstruction objective, $\mathcal{L}_{\text{RGB}}$ that minimizes the squared difference between the input images $\mathbf{I}_t$ and the rendering of the model. We test our method using DANBO [36]. It outputs a color and density for samples $\mathbf{x}$ along the primary view rays. These are subsequently integrated to compute a color, which we interpret as the albedo, $\mathbf{a}$. The illuminated color for a given pixel of the reconstructed image, $\hat{\mathbf{c}}$, is computed by a Lambertian reflectance model,

$$\hat{\mathbf{c}} = \mathbf{a}(\theta_t)\left(\hat{\mathbf{L}}_{\text{amb}} + s(\theta_t)\mathbf{L}_{\text{col}}(\hat{\mathbf{L}}_{\text{dir}} \cdot \hat{\mathbf{n}}(\theta_t))\right). \qquad (1)$$

This diffuse shading model illuminates the entire body with an ambient light $\hat{\mathbf{L}}_{\text{amb}}$ and a directional light with color $\mathbf{L}_{\text{col}}$. The directional light intensity is attenuated by the cast shadows s and the cosine angle between the surface normal $\hat{\mathbf{n}}$ and light direction $\hat{\mathbf{L}}_{\text{dir}}$.

Shading extensions. The benefit of the deferred rendering approach is that it lets us compute lighting information only once for each pixel, as opposed to at every sample location of the volumetric ray tracing leading to significantly faster computation. To be applicable, we extend DANBO to yield surface normals $\hat{\mathbf{n}}$ and depth $\mathbf{d}$ for a given view ray. The former we attain by switching the density formulation to a signed distance function with an Eikonal loss. The normal is then readily estimated by differentiating the distance with respect to the original query location $\mathbf{x}$ as in [45]. We compute $\mathbf{d}$ likewise to albedo $\mathbf{a}$ by integrating the sample's $\mathbf{x}$ positions along the ray, weighted by their density and transmittance. Furthermore, we fix the intensity of the directional light to white with a magnitude of 1.5. Without fixing the directional light intensity, the equation would be over parametrized and lead to ambiguities. Our model is invariant to light intensities between 1.0 and 2.0, as shown by Table 1.

Table 1. **Light intensity modulation:** Our method is invariant to moderate light intensity values.

Novel View & Pose Light Intensity:	1.0	1.25	1.50	2.0	5.0
PSNR ↑	24.46	**24.52**	24.37	24.00	18.98
LPIPS ↓	0.165	0.164	0.165	**0.163**	0.185

3.2. Gaussian Shadow Casting

For modeling shadows more efficiently, we represent the body shape with a set of Gaussians rigidly attached to the skeleton model. The relative positions, orientation, and size

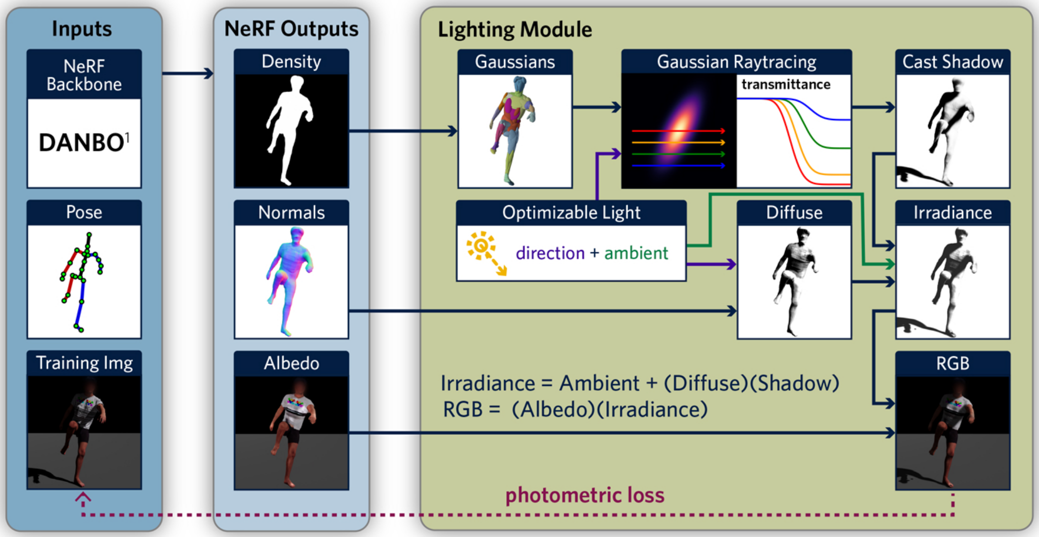

Figure 2. **Method Overview.** Our method takes as input images and poses of a person. Using a neural radiance field as a backbone [36][1], density, normals, and albedo values are volumetrically reconstructed and rendered. We fit a sum of 3D anisotropic Gaussian density model to approximate the neural density field and compute shadow maps using our novel anisotropic Gaussian ray occlusion equations. The shadow map is combined with a diffuse shading pass to produce the lit image. The whole model is optimized with a photometric loss against the training images. Our method is able to optimize the light direction and ambient intensity without any initialization. It also separates albedo from shading and shadow, allowing us to relight the model.

of the Gaussians are optimized to approximate the density of the neural field and to allow for a fast, efficient, and closed-form solution for integration along a ray (occlusion checking). Our model extends previous work [30,31] by using anisotropic Gaussians (variable scale and rotation along each axis) and avoids the need for sampling during integration as in NeRF.

Anisotropic Gaussian body model. We define the anisotropic Gaussian density model as the matrix $\mathbf{G} \in \mathbb{R}^{J \times K \times 13}$, with K being the number of Gaussians per joint, typically $\sim$ 8, and the columns representing the 3D mean (μ_x, μ_y, μ_z), the axis aligned standard deviations $(\sigma_x, \sigma_y, \sigma_z)$, the rotation defined using the 6 DOF representation $(R_{0,0}, R_{0,1}, R_{0,2}, R_{1,0}, R_{1,1}, R_{1,2})$ [52], and density $(\mathcal{C})$. Figure 3 gives examples with varying numbers of Gaussians.

The 3D density function, $\mathbf{G}(\mathbf{x})$, defines the density of the Gaussian model at the query location $\mathbf{x}$ in world space. We define the density function of a single 3D anisotropic Gaussian as

$$\mathbf{G}_i(\mathbf{x}) = \mathcal{C} \exp\left[-0.5\left((\mu - \mathbf{x})^T)\Sigma^{-1}(\mu - \mathbf{x})\right)\right], \quad (2)$$

where the precision matrix $\Sigma^{-1} = \mathbf{R}^T\mathbf{D}\mathbf{R}$ and $\mathbf{R}$ is the rotation matrix computed from the 6 DOF representation and $\mathbf{D} = \text{diagonal}(1/\sigma_x^2, 1/\sigma_y^2, 1/\sigma_z^2)$.

The density of the entire Gaussian model is the sum of the density of each. The query location is transformed to the local space of the given Gaussian's joint j at time-step t using the world-to-local transformation matrix $\theta_{t,j}^{-1}$, rigidly attaching the Gaussians to the underlying skeleton and facilitating animation,

$$\mathbf{G}(\mathbf{x}) = \sum_{i=0}^{J \times K} \mathbf{G}_i(\theta_{t,j}^{-1}\mathbf{x}). \quad (3)$$

We jointly fit the parameters of the Gaussian density model to the neural field by minimizing $\mathcal{L}_{\text{gDensity}}$, the L2 error between the density function $\mathbf{G}(\mathbf{x})$ and the target neural density field at query location $\mathbf{x}$. We detach the gradients of the neural density field to optimize the Gaussians, fitting the

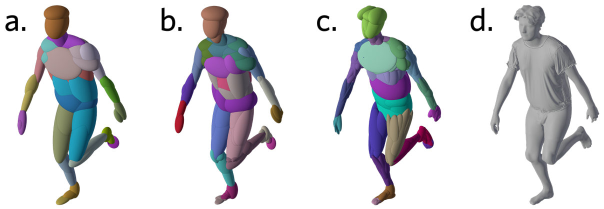

Figure 3. **Gaussian Density Model.** The approximation to the NeRF's density field using a sum of 3D anisotropic Gaussians using: a) 2 Gaussians per bone, b) 4 Gaussian per bone, and c) 8 Gaussian per bone; d) is the groundtruth mesh. *Note: ellipses are scaled to 2.5 STD of the Gaussians (99th percentile)*

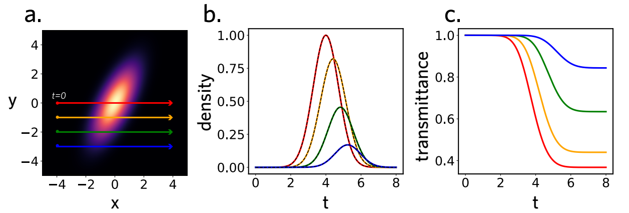

Figure 4. **3D Anisotropic Gaussian Raytracing.** a) A cross-section of a 3D anisotropic Gaussian with rays passing through the Gaussian. b) The computed 1D Gaussians resulting from our derivation in Section 3.2 (colored solid), compared to sampling the 3D Gaussian directly (dashed), with their exact match validating the correctness. c) The transmittance along each ray which is used as the shadow map value.

Gaussians to the neural field and not the other way around.

Gaussian Ray Tracing. Figure 4 shows how casting a ray, r, with ray origin $\mathbf{r}_o \in \mathbb{R}^{3\times 1}$ and direction $\mathbf{r}_d \in \mathbb{R}^{3\times 1}$, through a 3D anisotropic Gaussian results in a 1D Gaussian density along the ray. Through the Gaussian body model, this equates to a sum of 1D Gaussians for which analytic integrals can be computed. The amount of occlusion these rays experience is equal to the sum of the integrals of each of the 1D Gaussians across the rays. The transmittance value, $\mathcal{T}$, used as the shadow map value, s, is the exponential of the negative integral from the start of the ray, $t = 0$, to the length of the ray, $t = l$,

$$s = \mathcal{T}_r = \exp\left[-\sum_{i=0}^{J\times K}\int_0^l \mathbf{G}_i^r\right]. \tag{4}$$

G_i^r is the 1D Gaussian created by the ray, r, going through the 3D anisotropic Gaussian, G_i with mean $\mu \in \mathbb{R}^{3\times 1}$ and precision matrix $\Sigma^{-1} \in \mathbb{R}^{3\times 3}$. We derive in the supplemental how the 1D Gaussian's density function takes the form

$$G_i^{r^s} = \bar{\mathcal{C}} \cdot \exp\left(-\frac{(\bar{\mu} - x)^2}{2\bar{\sigma}^2}\right), \tag{5}$$

where

$$\begin{aligned}
\bar{\mathcal{C}} &= \mathcal{C}\exp\left(-0.5\left((\mu - \mathbf{r}_o)^T\Sigma^{-1}(\mu - \mathbf{r}_o) - \frac{\bar{\mu}^2}{\bar{\sigma}^2}\right)\right),\\
\bar{\mu} &= \frac{\mathbf{r}_d^T\Sigma^{-1}(\mu - \mathbf{r}_o)}{\mathbf{r}_d^T\Sigma^{-1}\mathbf{r}_d}, \text{ and}\\
\bar{\sigma} &= \sqrt{\frac{1}{\mathbf{r}_d^T\Sigma^{-1}\mathbf{r}_d}}.
\end{aligned} \tag{6}$$

This formula is more complex than in [30], as it now accounts for anisotropic Gaussians with an arbitrary covariance instead of isotropic Gaussians. The comparison to sampling the 3D Gaussian in Figure 4 validates their correctness. It also lets us compute the cumulative density function (CDF) analytically, thereby avoiding the sampling in classical NeRFs,

$$\int_0^x \mathbf{G}_i^{r^s} = \bar{\mathcal{C}} \cdot 0.5 \cdot \left(1 + \operatorname{erf}\left(\frac{x - \bar{\mu}}{\bar{\sigma}\sqrt{2}}\right)\right). \tag{7}$$

Together with Equation 4, this integral computes the shadow s when applied to the secondary ray with origin r_o^s, as the point on the subject's surface computed from the depth map $\mathbf{d}$, and direction r_d^s towards the light.

3.3. Optimization

In addition to the introduced reconstruction loss $\mathcal{L}_{\text{RGB}}$, $\mathcal{L}_{\text{Eikonal}}$ for SDF regularization as in [12], and Gaussian fitting $\mathcal{L}_{\text{gDensity}}$, we regularize the training with i) a $\mathcal{L}_{\text{mask}} = |\hat{\rho} - \rho|$ that regularizes density by minimizing the difference between integrated accumulation, $\hat{\rho}$, and the foreground mask, ρ, ii) $\mathcal{L}_{\text{amb}} = ||\hat{\mathbf{L}}_{\text{amb}} - 0.1||^2$ preferring small ambient light values, iii) $\mathcal{L}_{\text{gSigma}}$ that prevent too large or small Gaussians, and iv) $\mathcal{L}_{\text{gMean}}$ that pulls Gaussians closer to the center of the bones.

Training proceeds in three stages. In stage I, the reconstruction loss is replaced with one that encourages predicting gray inside the silhouette, to learn a rough body shape without illumination effects. In stage II, $\mathcal{L}_{\text{gDensity}}$ and its regularizers are introduced, allowing the Gaussian density model to fit. Finally, in stage III, the $\mathcal{L}_{\text{RGB}}$ takes over to optimize the light direction and learn the albedo. Additional training details are provided in the supplemental.

4. Results

We evaluate our method on synthetic sequences, as done in prior work [16]. However, this does not test performance in real world conditions. Hence, we captured a new dataset in direct sunlight and compare to the most closely related baselines, showing significantly improved relightable body models. The supplemental video and supplemental document provide additional qualitative comparisons, including relighting with HDRi environment maps.

Synthetic datasets. We test our model on the RANA dataset [16], and further create our own synthetic sequence by obtaining a textured mesh of a subject with a 3D full body scanner (VITUS 3D Body Scanner). A Blender [8] cloth simulation was applied to a shirt over the scan and the character was automatically rigged and animated using Mixamo [2]. We use the 'swing-dance' animation as the driving motion as it contains a variety of poses from all body angles. Four (3 train, 1 test) cameras are placed around the subject at 90 degrees from each other. A directional light source illuminates the scene with a slight ambient contribution such that the shadowed areas were not fully black. The animations are rendered using the Cycles render engine. In addition, ground-truth pose and segmentation masks are exported. We split the dataset into 600/114 images for the train/test sets, with the test set including 57 novel view images and 57 novel poses, out of which 15 have a strong hard cast shadows that we test separately.

Outdoor sunlight dataset. We recorded two sequences of real human motion in an outdoor scene during a sunny day. This setting has largely been unexplored in neural body models and few publicly available datasets are available. We capture the data using 3 cameras (Canon EOS R8, Canon EOS 70D, iPhone12) and obtain SMPL estimates using EasyMocap [1,10,11]. Segmentation masks are obtained using the Segment Anything Model (SAM) [19]. We divide the frames into 600/200 images for the train/test splits, using all three cameras for training.

Baselines We evaluate our method using the hard illumination dataset against Relighting4D [7]. Due to code being unavailable, we were not able to compare against RANA [16] and Xu et al. [43] using our datasets. We instead quantitatively compare albedo estimates between our model and RANA on their dataset as these results were kindly made available. We further compare our work with other template-less neural body models [36, 37], highlighting the drawback when not explicitly modeling lighting. DANBO [36] is our neural field backbone. NPC [37] forms the current state-of-the-art template-less neural character model.

4.1. Albedo Estimates (training poses)

We ran the official implementation of Relighting4D (R4D) [7] on our outdoor dataset, providing the same segmentation masks and SMPL body model as to our method (our method only uses the skeleton, not the surface). As the first stage of R4D is NeuralBody [28] which does not take shadowing into account, it produces dark floaters in the space to approximate the hard shadow, hindering their subsequent relighting module from estimating shadow and shading correctly as seen in Figure 5a.

To compare against RANA [16], we run our method on subject 1 of their synthetic dataset. Even with the dataset being monocular (light comes from the same direction relative to the camera), our method was able to accurately estimate the light direction (error of 9.9 degrees) and obtain albedo estimates with fewer lighting artifacts compared to RANA, see the back of the left leg in Figure 5b.

4.2. Novel-Pose Rendering with Shadows

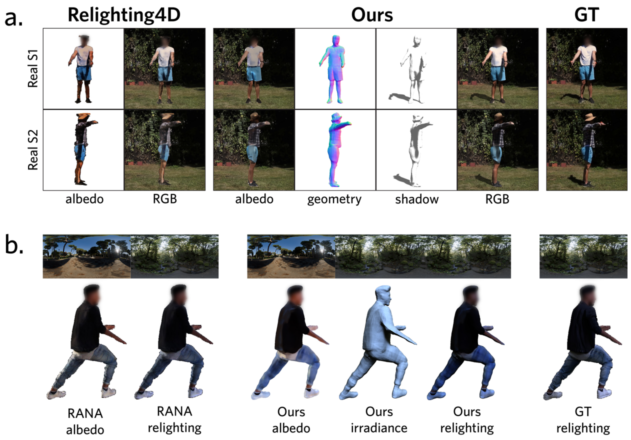

Figure 5. **Albedo Estimation.** Our method can better separate shadows and lighting from training images to obtain better albedo estimates without lighting artifacts compared to a) Relighting4D [7] and b) RANA [16].

In this setting, the camera and illumination are unchanged and only novel-poses are tested. These poses created new shadow casts that resulted in large appearance changes distant to the changed body part. As expected, existing methods (NPC, DANBO) overfit the training poses and the shadows created in the novel poses are highly inaccurate. Figure 6 shows how for frames that had body parts casting shadows on other regions, our method produced more accurate shadows. Table 2 quantifies the gains across novel poses and Table 3 across the subset of the novel poses that has a shadow cast across the body.

Table 2. **Novel-pose synthesis (all test frames).** Our Gaussian Shadow Casting model achieves consistently better PSNR scores for novel pose renderings as it properly models the hard shadows cast by the limbs in novel positions. Existing methods only shine in perceptual metrics (SSIM and LPIPS) as these normalize contrast and hence lessen the impact of proper shadows and shading.

	Synthetic ($N = 57$)			Real S1 ($N = 200$)			Real S2 ($N = 200$)			Average		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DANBO [36]	17.52	0.756	0.195	16.57	**0.599**	**0.328**	17.69	**0.588**	**0.325**	17.26	0.648	**0.283**
NPC [37]	17.57	0.758	0.188	16.33	0.590	0.334	17.47	0.575	0.328	17.12	0.641	**0.283**
Ours	**22.04**	**0.829**	**0.166**	**17.57**	0.592	0.356	**18.29**	0.577	0.351	**19.30**	**0.666**	0.291

Table 3. **Novel-pose synthesis (subset of test set with observed self-casting shadows).** Our Gaussian Shadow Casting renders novel poses with strong hard shadows well. Our scores drop marginally on these hard frames compared to all frames in Tab. 2, while the baselines drop significantly.

	Synthetic ($N = 15$)			Real S1 ($N = 41$)			Real S2 ($N = 36$)			Average		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DANBO [36]	17.78	0.740	0.209	15.11	0.559	**0.354**	16.55	**0.547**	**0.353**	16.48	0.615	0.305
NPC [37]	17.81	0.741	0.201	14.88	0.553	0.357	16.51	0.538	0.355	16.40	0.611	0.304
Ours	**22.13**	**0.821**	**0.175**	**16.88**	**0.572**	0.365	**17.40**	0.544	0.371	**18.81**	**0.646**	**0.303**

For the synthetic sequence, improvements were consistent across all three metrics. In the real outdoor sequence, all methods attain a lower quality as the cameras are spaced further apart and the segmentation masks and 3D input

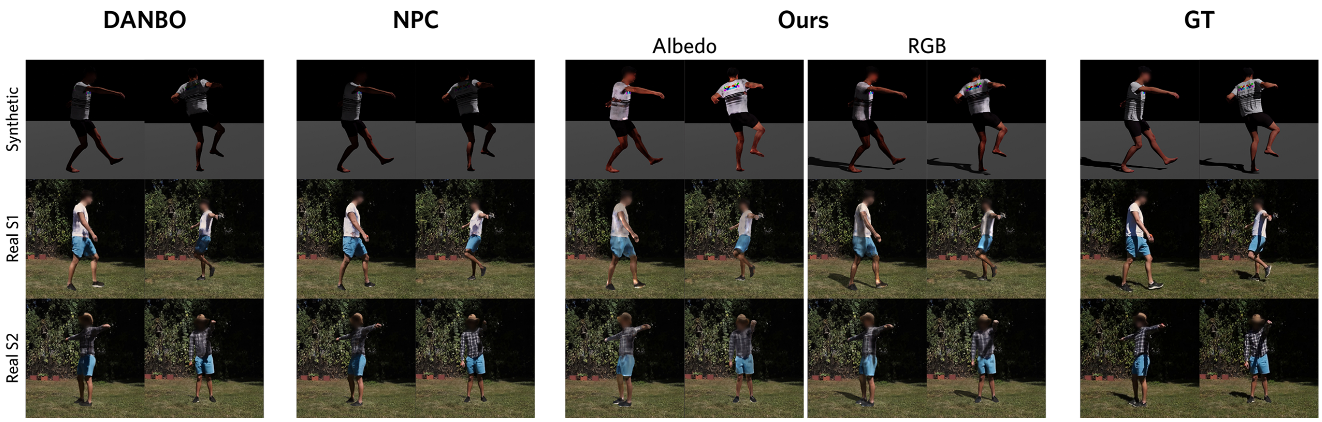

Figure 6. **Novel Pose Rendering.** Our method can more accurately reproduce the shadow in novel poses compared to the baselines.

pose, estimated with off-the-shelf 2D pose detection and lifting methods, are less reliable. Nevertheless, Figure 6 shows how our model can accurately optimize the light direction and predict realistic shadows, including on the ground. To map shadows to the ground, we estimate the ground plane from the reconstructed foot positions and cast the Gaussian shadow on it by modulating the static background with the ground shadow map.

Table 2 and Table 3 show that our method consistently improves the PSNR while the perceptual metrics SSIM [41] improves only in one and the baselines perform better for LPIPS [47]. This lower performance in perceptual metrics is expected because these metrics normalize for brightness and contrast differences, thereby lessening the importance of producing proper shading and shadowing. In addition, the texture and geometry detail of our method is slightly lower, which we attribute to the separation into shading and albedo imposing additional constrains, thereby leading to slightly less detailed reconstructions.

4.3. Render Time Comparison

Table 4 lists the render time of our baseline compared to our full model. Casting shadows with GSC has minimal overhead (0.3s for one ray, only 2% of the entire render time), enabling efficient training alongside NeRF optimization. Casting a shadow with the NeRF baseline requires processing twice the number of samples by the NeRF. The deferred shading model creates one occlusion ray and each of these secondary rays requires a similar number of samples as for the primary ray. Already with a single light source, this increases runtime by 25%, a ten-fold difference to GSC.

Method	render time [s]
DANBO + DS	17.13
DANBO + DS + GSC	17.47
DANBO + DS + NeRFSC	21.4
DANBO + DS + GSC-HDRi-8	20.70
DANBO + DS + GSC-HDRi-16	23.57
DANBO + DS + GSC-HDRi-32	29.49
DANBO + DS + GSC-HDRi-64	41.22

Table 4. **Render time.** The overhead of Gaussian Shadow Casting (GSC) is minimal on DANBO with diffuse shading (DANBO + DS) and enables casting many rays (64 for GSC-HDRi-64). By contrast, NeRF shadow casting (NeRFSC) doubles the runtime with every light source, making training prohibitively slow and HDRi relighting impractical.

4.4. Relighting with Environment Maps

The shadow computation not only benefits training time but also enables computing irradiance maps for environment maps. Figure 1 shows relighting with two different HDRi maps (obtained from Poly Haven [14]) by casting 64 secondary light rays towards the environment map for each pixel through importance sampling. In both cases, the bright sun casts a strong shadow while the colored light from the environment leads to natural shading that matches the character with the environment. This enables placing the reconstructed characters into new environments and giving them a natural and consistent look with respect to the rest of the scene while still containing cast shadows.

Our method is plug-and-play with other neural body models due to the deferred rendering approach. It can be used with higher quality volumetric neural models without degraded quality when training on uniformly lit data and using GSC for relighting as shown in Figure 7.

4.5. Ablation Study

We test a variety of implementation details in our model, including using only diffuse shading on top of DANBO (DANBO + Diff. Shading), only using the Gaussians to cast shadows (Ours w/o Diff. Shading), providing ground truth lighting (Ours-GT Light), and detaching the normals prior to the diffuse shading (Ours-Detached Normals). The

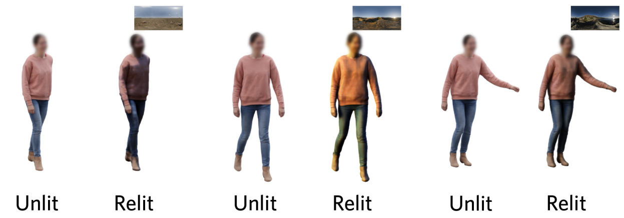

Figure 7. **HDRi Relighting on MonoPerfCap using NPC [37] as the backbone.** Our Gaussian relighting method can be utilized at inference with higher quality models on data that is uniformly lit.

results of which can be seen in Table 5, which shows that each of our contributions improves reconstruction quality at test time.

Table 5. **Ablation on Synthetic Sequence.**

	Training			Novel Pose			Novel View		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DANBO	**27.66**	**0.913**	**0.133**	17.52	0.756	0.195	18.85	0.773	0.179
DANBO + Diff. Shading	24.82	0.860	0.187	18.60	0.785	0.208	24.40	0.875	0.167
Ours w/o Diff. Shading	27.17	0.879	0.161	21.22	0.802	0.178	25.08	0.867	0.163
Ours-Detached Normals	25.90	0.863	0.182	20.43	0.814	0.192	26.18	0.893	0.158
Ours-GT Light	25.22	0.861	0.180	21.23	**0.830**	0.184	26.68	**0.895**	0.155
Ours	26.60	0.876	0.165	**22.30**	0.827	**0.176**	**27.32**	0.882	**0.154**

Diffuse Shading. Shadowing alone does not account for accurate shading based on how incident the light hits the surface. Moreover, the Gaussians cast long-range shadows, but their smooth and low-resolution approximation to the NeRF's density prevents them from representing finer details such as small extremities (nose, fingers). As a result, finer shading details are missed as seen in Figure 8c. On the other hand, using only shading (DANBO + Diff. Shading) already reduces texture detail quality as seen by the training scores, but achieves improved performance in novel pose and novel view conditions. However, the missing cast shadows play the most significant role in improving test scores.

Detached Normals. We compare results between a model where network gradients could, Figure 8d, and could not, Figure 8b, backpropagate through the surface normals used in the diffuse shading to see whether or not artifacts in the shading would smoothen out the geometry. We find that the surface is indeed affected by the gradients backpropagating through the diffuse computation and observe a smoother geometry reconstruction.

GT Light. Our method is able to fit the direction of the light source and the ambient intensity with little user input. We observe accurate light recovery when the light is initialized randomly, e.g. when coming from the back the angle error is only 1.36 degrees on our synthetic sequence. We found providing the ground truth light direction did not improve results and may have hindered the model due to the added constraints, Figure 8a.

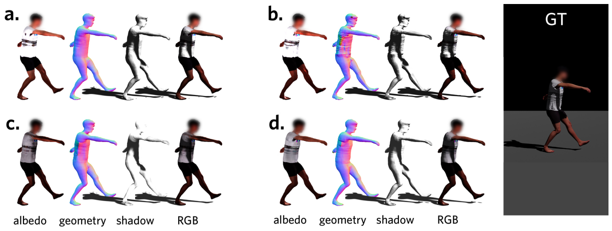

Figure 8. **Ablation Comparisons (novel-pose).** a) The model trained with the groundtruth light direction. b) The model trained while detaching the gradients from the surface normals during diffuse shading. c) The model trained without diffuse shading. d) Our full model.

4.6. Limitations

The Gaussian cast shadows model long-range effects, such as the arm casting a shadow on the leg but the smooth Gaussians lack high frequency details. This is a minor drawback since the diffuse shading already faithfully reproduces the light intensity fall-off as the light direction becomes more incident with the surface and therefore shades the back side of small extremities (i.e. nose and fingers) well. A future extension could be to integrate mid-scale effects with screen-space ambient occlusion and shading.

Moreover, we noticed that disentangling color into shading and albedo, compared to the original DANBO backbone, leads to slightly lower image reconstruction metrics when shading effects are minimal. We attribute this to the additional constraints that are imposed on the model. However, the overall performance in novel light conditions is still improved significantly by our model.

5. Conclusion

We enabled the 3D reconstruction of human motions in uncontrolled environments with a Gaussian-based shadow model that applies to dynamic scenes and is differentiable for iterative refinement. The reconstructed characters support reposing and relighting in novel environments. They are equipped with global shadow computation, diffuse shading, geometric reconstruction, and a consistent albedo, much like hand-crafted computer graphics models would provide. The deferred lighting approach allows our method to be combined with other neural body models with efficient shadow computations.

6. Acknowledgements

This work was supported in part by an NSERC Discovery Grant and the computational resources and services provided by Advanced Research Computing at The University of British Columbia.

References

[1] Easymocap - make human motion capture easier. https://github.com/zju3dv/EasyMocap, 2021. 6

[2] Adobe. Mixamo. www.mixamo.com, 2023. 5

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, 2019. 2

[4] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM TOG (Proc. SIGGRAPH)*, 2021. 2

[5] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM TOG (Proc. SIGGRAPH)*, 2021. 2

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 2

[7] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *ECCV*, 2022. 2, 3, 6

[8] Blender Online Community. Blender - a 3d modelling and rendering package. http://www.blender.org, 2023. 5

[9] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. In *CVPR*, 2021. 2

[10] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. In *TPAMI*, 2021. 6

[11] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *ECCV*, 2020. 6

[12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020. 5

[13] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering, 2020. 2

[14] Poly Haven. Poly haven. www.polyhaven.com/hdris, 2023. 7

[15] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *CVPR*, 2019. 2

[16] Umar Iqbal, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, and Jan Kautz. Rana: Relightable articulated neural avatars. *arXiv preprint arXiv:2212.03237*, 2022. 2, 5, 6

[17] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 1

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG (Proc. SIGGRAPH)*, 2023. 2, 3

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 6

[20] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 2021. 1, 2

[21] Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. Megane: Morphable eyeglass and avatar network. In *CVPR*, 2023. 2

[22] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *ECCV*, 2022. 2

[23] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM TOG (Proc. SIGGRAPH)*, 2023. 2

[24] 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 TOG (Proc. SIGGRAPH Asia)*, 2021. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG (Proc. SIGGRAPH Asia)*, 2015. 2, 3

[26] 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 *ECCV*, 2020. 1, 2

[27] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 2

[28] 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 *CVPR*, 2021. 1, 6

[29] Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, and Oncel Tuzel. Facelit: Neural 3d relightable faces. In *CVPR*, 2023. 2

[30] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016. 2, 3, 4, 5

[31] H. Rhodin, N. Robertini, C. Richardt, H. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*, 2015. 2, 3, 4

[32] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *ECCV*, 2022. 2

[33] Srinath Sridhar, Helge Rhodin, Hans-Peter Seidel, Antti Oulasvirta, and Christian Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *3DV*, 2014. 2, 3

[34] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021. 2

[35] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. *ICCV*, 2011. 3

[36] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *ECCV*, 2022. 1, 2, 3, 4, 6

[37] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. In *ICCV*, 2023. 1, 2, 6, 8

[38] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 2

[39] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG (Proc. SIGGRAPH)*, 2019. 3

[40] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 1, 2

[41] 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

[42] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *CVPR*, 2023. 2

[43] Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Relightable and animatable neural avatar from sparse-view video, 2023. 2, 3, 6

[44] Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. Towards practical capture of high-fidelity relightable avatars. In *ACM TOG (Proc. SIGGRAPH Asia)*, 2023. 2

[45] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 3

[46] 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 *CVPR*, 2021. 3

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[48] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM TOG (Proc. SIGGRAPH)*, 2021. 2

[49] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, 2022. 2

[50] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM TOG (Proc. SIGGRAPH)*, 2023. 2

[51] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations. In *CVPR*, 2023. 2

[52] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4