IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monocular Videos via Explicit Ray Tracing

Shaofei Wang[1,2,3], Božidar Antić[2,3], Andreas Geiger[2,3], Siyu Tang[1]
[1]ETH Zürich [2]University of Tübingen [3]Tübingen AI Center

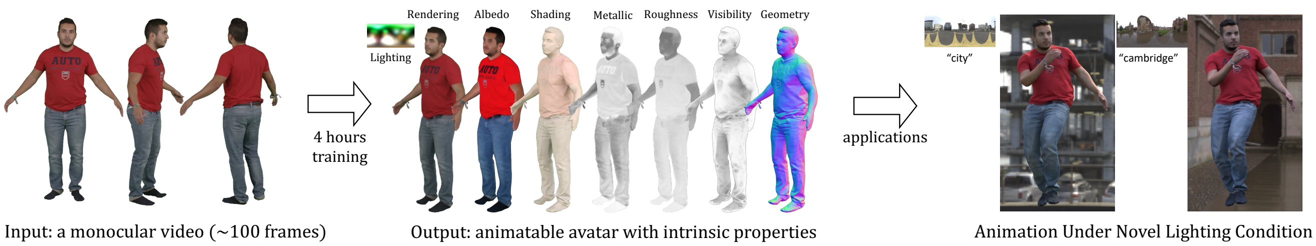

Figure 1. *IntrinsicAvatar* aims to achieve physically based inverse rendering of clothed humans from monocular videos. **Left:** Our model takes a monocular video as input and learns an avatar of the target person. **Middle:** We show decomposed properties of the learned avatar. Importantly, our model can produce such decomposition without any data-driven prior on geometry, albedo, or material. **Right:** With the learned avatar and intrinsic properties, we can animate and relight the avatar using arbitrary pose and arbitrary lighting condition.

Abstract

We present IntrinsicAvatar, a novel approach to recovering the intrinsic properties of clothed human avatars including geometry, albedo, material, and environment lighting from only monocular videos. Recent advancements in human-based neural rendering have enabled high-quality geometry and appearance reconstruction of clothed humans from just monocular videos. However, these methods bake intrinsic properties such as albedo, material, and environment lighting into a single entangled neural representation. On the other hand, only a handful of works tackle the problem of estimating geometry and disentangled appearance properties of clothed humans from monocular videos. They usually achieve limited quality and disentanglement due to approximations of secondary shading effects via learned MLPs. In this work, we propose to model secondary shading effects explicitly via Monte-Carlo ray tracing. We model the rendering process of clothed humans as a volumetric scattering process, and combine ray tracing with body articulation. Our approach can recover high-quality geometry, albedo, material, and lighting properties of clothed humans from a single monocular video, without requiring supervised pre-training using ground truth materials. Furthermore, since we explicitly model the volumetric scattering process and ray tracing, our model naturally generalizes to novel poses, enabling animation of the reconstructed avatar in novel lighting conditions.

1. Introduction

Photo-realistic reconstruction and animation of clothed human avatars is a long-standing problem in augmented reality, virtual reality, and computer vision. Existing solutions can achieve high-quality reconstruction for both geometry and appearance of clothed humans given dense multi-view cameras [24, 27, 61]. Recently, reconstruction of clothed humans from monocular videos has also been explored [23, 56, 73, 76]. While these approaches achieve satisfactory results, they model the appearance of clothed humans as a single neural representation. This makes it difficult to edit the physical properties of the reconstructed clothed human avatars, such as reflectance and material, or to relight the reconstructed clothed human avatars under novel lighting conditions. In this work, we aim to recover physically based intrinsic properties for clothed human avatars including geometry, albedo, material, and environment lighting from only monocular videos.

Physically based inverse rendering is a challenging problem in computer graphics and computer vision. Traditional approaches tackle this problem as a pure optimization problem with simplifying assumptions such as controlled, known illumination. On the other hand, recent advances in neural fields have enabled the high-quality reconstruction of geometry and surface normals from multi-view RGB images. Given this progress, physically based inverse rendering of static scenes under unknown natural illumination has been demonstrated [33, 88]. Most recently, various works have combined human body priors with the physically based inverse rendering pipeline to reconstruct clothed human avatars with disentangled geometry, albedo, material, and lighting from monocular videos [15, 28, 70]. However, these methods either ignore physical plausibility or model secondary shading effects via approximation, resulting in limited quality of reconstructed human avatars.

Two major challenges are present for physically based inverse rendering of clothed humans from monocular videos: *(1)* accurate geometry reconstruction, especially normal estimates are essential for high-quality inverse rendering. *(2)* Modeling secondary shading effects such as shadows and indirect illumination is expensive and requires a certain level of efficiency to query the underlying neural fields. Existing monocular geometry reconstruction methods of clothed humans all rely on large MLPs to achieve high-quality geometry reconstruction. However, using large MLPs negatively impacts the efficiency of secondary shading computation. Therefore, most existing methods are forced to rely on simple assumptions (no shadows, no indirect illumination) or approximations (pre-trained MLPs) to model secondary shading effects. More efficient neural field representations such as instant NGP (iNGP [48]) have proven to be effective for geometric reconstruction given multiple input views of a static scene [40, 62, 74], but it remains a challenge to extend such representation to dynamic humans under monocular setup.

In this paper, we employ iNGP with hashing-based volumetric representation and signed distance field (SDF) to achieve fast and high-quality reconstruction of clothed humans from monocular videos. The high-quality initial geometry estimation and efficiency of iNGP facilitate the modeling of inverse rendering via explicit Monte-Carlo ray tracing. Furthermore, traditional surface-based inverse rendering methods give ambiguous predictions at edges and boundaries. We propose to use volumetric scattering to model edges and boundaries in a more physically plausible way. Our experiments demonstrate that we can achieve high-quality reconstruction of clothed human avatars with disentangled geometry, albedo, material, and environment lighting from only monocular videos. In summary, we make the following contributions:

- We propose a model for fast, high-quality geometry reconstruction of clothed humans from monocular videos.
- We propose to combine volumetric scattering with the human body articulation for physically based inversed rendering of dynamic clothed humans. We use explicit Monte-Carlo ray tracing in canonical space to model the volumetric scattering process, enabling relighting for unseen poses.
- We demonstrate that our method can achieve high-quality reconstruction of clothed human avatars with disentangled geometry, albedo, material, and environment lighting from only monocular videos of clothed humans. We also show that our learned avatars can be rendered realistically under novel lighting conditions *and* novel poses.

We have made our code and models publicly available[1].

2. Related Work

Traditional Inverse Rendering: Traditional approaches to inverse rendering work on either single RGB images [4, 38, 39, 41, 64, 67, 75, 81] or multi-view, multi-modality inputs [21, 24, 35, 36, 50, 52, 59, 65, 83]. Recovering shape, reflectance, and illumination from a single RGB image is heavily underconstrained and often works poorly on real-world setups such as scene-level reconstruction and articulated object reconstruction. A more practical approach is to reconstruct shapes from multi-view RGB(D) images and make simplifying assumptions such as controlled lighting conditions [44, 50, 66]. This kind of approach often results in high-quality reconstruction of physical properties but lacks flexibility.

Physically Based Inverse Rendering with Neural Fields: Since the blossom of neural radiance fields (NeRF [47]), a variety of works have been proposed to tackle the inverse rendering problem using neural field representations. However, many works make use of simplifying assumptions such as known lighting conditions [68], ignoring shadowing effects [7, 8, 49, 82], or assuming constant material [82]. NeRFactor [84] was the first work that enabled full estimation of a scene's underlying physical properties (geometry, albedo, BRDF, and lighting) under a single unknown natural illumination while also taking shadowing effect into account. InvRender [85] builds upon the state-of-the-art shape and radiance field reconstruction methods [72, 79] and proposed to model indirect illumination by distilling a pre-trained NeRF into auxiliary MLPs. [45] learns a neural radiance transfer field to enable global illumination under novel lighting conditions, but relies on accurate geometry initialization and does not optimize it jointly with material and lighting. NVDiffRecMC [25] tackles the inverse rendering problem by exploring the combination of mesh-based Monte-Carlo ray tracing and off-the-shelf denoisers. However, the mesh-based representation of NVDiffRecMC gives less accurate reconstruction compared to [72, 79].

Most recently, TensoIR [33] takes advantage of fast radiance field data structures [10] and conducts explicit visi-

[1]https://neuralbodies.github.io/IntrinsicAvatar/

bility and indirect illumination estimation via ray marching. In comparison, we use an SDF representation and combine iso-surface search technique with volumetric scattering, resulting in better visibility modeling, especially for cloth wrinkles. Most importantly, we target dynamic, animatable clothed avatar reconstruction while TensoIR focuses on static scene reconstruction.

Microfacet fields [46] proposed to utilize volumetric scattering with a surface BRDF and ad-hoc sampling strategies. Concurrent to [46] , NeMF [86] also proposed to use volumetric scattering with microflake phase functions [26, 31] to replace surface-based BRDF for volume scattering, resulting in the ability to reconstruct thin structures and low-density volumes. Both methods focus on static scenes reconstruction and relighting while using density fields to represent the underlying geometry.

Neural Radiance Fields for Human Reconstruction: Neural radiance fields have been used for human reconstruction from monocular videos. Most works [19, 32, 37, 54, 55, 76] focus on appearance reconstruction while using density fields as a noisy geometry proxy. Some methods use SDFs to represent the geometry of humans and achieve impressive results in both geometry reconstruction and photorealistic rendering [23, 56, 73, 77]. However, these methods bake intrinsic properties such as albedo, material, and lighting all into the learned neural representations, preventing the application of these methods in relighting and material editing.

Physically Based Inverse Rendering of Humans: High-quality 3D relightable human assets can be obtained via a multi-view, multi-modality capture system with controlled lighting [6, 17, 24, 29, 63, 83] or by training regressors on high-quality digital 3D assets [3, 18, 87]. RANA [28] pre-trains a mesh representation on multiple subjects with ground truth 3D digital assets while using a simplified spherical harmonics lighting model, thus cannot handle secondary shading effects such as shadows and indirect illumination. [70] propose to model the secondary shading effects via spherical Gaussian approximations, which do not handle shadowing effects. Relighting4D [15] jointly estimates the shape, lighting, and the albedo of dynamic humans from monocular videos under unknown illumination by approximating visibility via learned MLPs. These learned MLPs are over-smoothed approximations to real visibility values, while also having the inherent problem of not being able to generalize to novel poses. In contrast, we employ fast, exact visibility querying via explicit ray tracing, and thus can generalize to any novel poses.

Concurrent Works: [5] and [34] respectively reconstruct relightable faces and hands from monocular videos. For full-body relightable avatars, [42] proposes to construct part-wise light visibility MLPs to achieve better novel pose generalization for relighting. However, it needs to train light visibility MLPs on additional unseen poses. In comparison, we use explicit ray tracing to compute secondary ray visibilities, which generalizes to novel poses without additional training. [78] designs a hierarchical distance query algorithm and extends DFSS [53] to deformable neural SDF, achieving efficient light visibility computation using sphere tracing. However, the use of sphere tracing and surface rendering results in visible artifacts around elbows and armpits, as sphere tracing does not guarantee convergence, especially when combined with human body articulation. In contrast, we use volumetric scattering to model the human body, which results in less visual artifacts.

3. Method

In this section, we first introduce basic concepts of neural radiance fields (NeRF [47]). Then we describe our framework of geometry reconstruction of clothed avatars from monocular videos. The clothed avatars are modeled as an articulated NeRF with SDF as its geometry representation. Next, we introduce the volumetric scattering process from computer graphics and draw a connection between it and NeRF. Finally, we describe our solution to secondary ray tracing of volumetric scattering, which combines the explicit ray-marching with iso-surface search and body articulation. The final outputs are intrinsic properties of clothed avatars including geometry, material, albedo, and lighting.

3.1. Background: Neural Radiance Fields

Given a ray $\mathbf{r} = (\mathbf{o}, \mathbf{d})$ defined by its camera center $\mathbf{o}$ and viewing direction $\mathbf{d}$, NeRF computes the output radiance (i.e. pixel color) of the ray via:

$$\begin{aligned} C_{rf}(\mathbf{r}) &= \int_{t_n}^{t_f} T(t_n, t)\sigma_t(\mathbf{r}(t))L(\mathbf{r}(t), -\mathbf{d})dt \\ \text{s.t} \quad & \mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \\ & T(t_n, t) = \exp\left(-\int_{t_n}^{t} \sigma_t(\mathbf{r}(s))ds\right) \end{aligned} \tag{1}$$

where (t_n, t_f) defines the near/far point for the ray integral. In practice, NeRF uses a ray marching algorithm to approximate the exact value of the integral:

$$\begin{aligned} C_{rf}(\mathbf{r}) &\approx \sum_{i=1}^{N} w^{(i)} L(\mathbf{r}(t^{(i)}), -\mathbf{d}) \\ \text{s.t} \quad & \mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \\ & w^{(i)} = T^{(i)} \left(1 - \exp(-\sigma_t(\mathbf{r}(t^{(i)}))\delta^{i}\right) \\ & T^{(i)} = \exp\left(-\sum_{j<i} \sigma_t(\mathbf{r}(t^{(j)}))\delta^{(j)}\right) \\ & \delta^{(i)} = t^{(i+1)} - t^{(i)} \end{aligned} \tag{2}$$

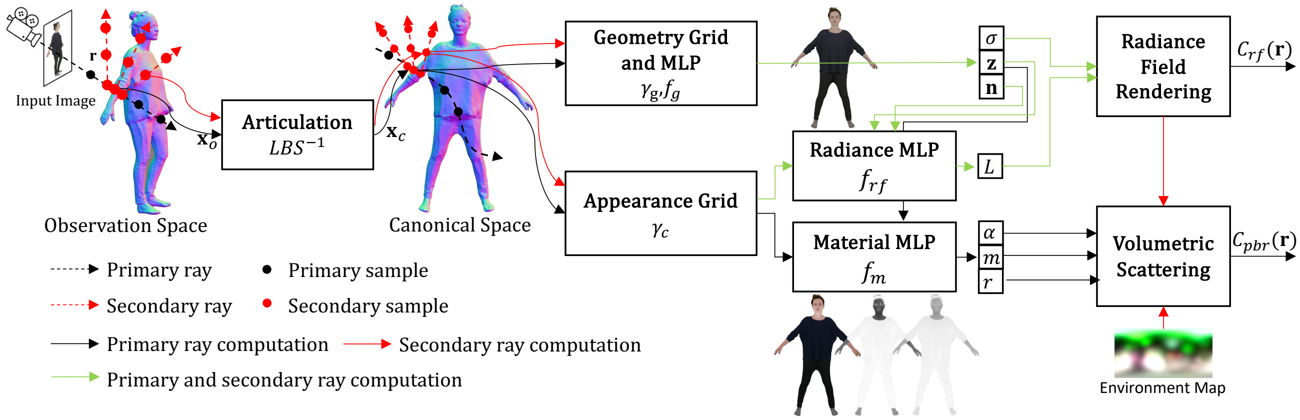

Figure 2. **Inverse Rendering of Clothed Avatars with Volumetric Scattering.** Given an input image and associated camera rays, we warp the rays to the canonical space and do both primary and secondary ray marching/tracing in canonical space. We model geometry with a geometry hash grid γ_g and MLP f_g, while also modeling volumetric radiance and material with an appearance grid γ_c and two additional MLPs f_{rf}, f_m. We supervise both C_{rf} and C_{pbr} using a L1 loss wrt. the input image.

where $\{t^{(1)}, \cdots, t^{(N)}\}$ are a set of sampled offsets on the ray. $\sigma_t(\cdot)$ and $L(\cdot, \cdot)$ are represented as either neural networks [47, 51, 72, 80], explicit grid data [1, 69], or a hybrid of both [10–12, 16, 20, 48, 60].

3.2. Clothed Humans Avatars as Articulated Neural Radiance Fields

We follow the recent approaches of modeling humans as articulated NeRF [32, 37, 73, 76]. We assume body articulations are based on the SMPL model [43]. Following previous works, we define the observation space as a space where the human is observed, and the canonical space as a space where the human is in a canonical pose. We apply inverse linear blend skinning (LBS) to transform 3D points in the observation space $\mathbf{x}_o = \mathbf{r}(t)$ to the point in canonical space $\mathbf{x}_c$. We model the radiance field, materials, and albedo all in the canonical space.

Articulation via Inverse LBS: In the SMPL model, the linear blender skinning (LBS) function is defined as:

$$\mathbf{x}_o = \mathrm{LBS}(\mathbf{x}_c, \{\mathbf{B}_b\}_{b=1}^{B}, w(\mathbf{x}_c)) = \left(\sum_{b=1}^{B} w(\mathbf{x}_c)_b \mathbf{B}_b\right) \mathbf{x}_c \tag{3}$$

where $\{\mathbf{B}_b\}_{b=1}^{B}$ are the rigid bone-transformations defined by estimated SMPL parameters. $w(\mathbf{x}_c)$ are skinning weights of point $\mathbf{x}_c$. We use Fast-SNARF [14] to model the canonical skinning weight function $w(\cdot)$ and the inverse skinning function:

$$\mathbf{x}_c = \mathrm{LBS}^{-1}(\mathbf{x}_o, \{\mathbf{B}_b\}_{b=1}^{B}, w(\mathbf{x}_c)) \tag{4}$$

For simplicity, we drop the dependency on $\{\mathbf{B}_b\}_{b=1}^{B}$ and $w(\mathbf{x}_c)$ for the remainder of the paper.

Geometry: We use iNGP [48] with SDF to represent the underlying canonical shape of clothed humans. Specifically, given a query point $\mathbf{x}_c$ in canonical space, we predict the SDF value of the point and a latent feature $\mathbf{z}$:

$$(\mathrm{SDF}(\mathbf{x}_c), \mathbf{z}) = f_g(\gamma_g(\mathbf{x}_c)) \tag{5}$$

where $\gamma_g(\cdot)$ is the iNGP hash grid feature of the input point, and f_g is a small MLP with a width of 64 and one hidden layer. We use VolSDF [80] to convert from SDF to density σ_t.

Radiance and Material: Radiance and materials are predicted as follows:

$$L(\mathbf{x}_c, \mathbf{d}) = f_{rf}(\gamma_c(\mathbf{x}_c), \mathbf{z}, \mathrm{ref}(\mathbf{d}, \mathbf{n}), \mathbf{n}) \tag{6}$$

$$\alpha(\mathbf{x}_c), r(\mathbf{x}_c), m(\mathbf{x}_c) = f_m(\gamma_c(\mathbf{x}_c), \mathbf{z}) \tag{7}$$

where $\gamma_c(\cdot)$ is the feature from another iNGP hash grid designed specifically for radiance and material prediction. The same strategy was also employed in [62] for learning better geometric details. f_{rf} and f_m are both MLPs with a width of 64 and two hidden layers. $\mathbf{n}$ is the analytical normal obtained from SDF fields. $\mathrm{ref}(\mathbf{d}, \mathbf{n})$ reflects the viewing direction $\mathbf{d}$ around the normal $\mathbf{n}$, similar to [71]. $L(\cdot, \cdot)$ will be used for Eq. (2) whereas α, r, and m are spatially varying *albedo*, *roughness*, and *metallic* parameters that will be used for physically based rendering.

For ray marching, we use 128 uniform samples and do two rounds of importance sampling, each time with 16 samples, to obtain a final set of 160 samples per ray.

With the aforementioned model, we can quickly reconstruct the detailed geometry of clothed human avatars from a single monocular video in less than 30 minutes.

3.3. Physically Based Inverse Rendering via Volumetric Scattering

With initial geometry and radiance estimation from previous sections, we now account intrinsic properties of clothed human avatars, i.e. material, albedo, and lighting conditions for the rendering process.

With the standard equation of transfer of participating media in computer graphics [30, 57], we reach the NeRF formula Eq. (1) by assuming all the radiance that reaches the camera is modeled by neural networks. On the other hand, if we think all the radiance that reaches the camera is *scattered* from some light sources (e.g. environment maps) by a volume of media, while the media itself does not emit any radiance, then we are tackling the volume scattering problem.

Formally, we have the following integral to compute the radiance scattered by the volume representing the human body along a certain camera ray $(\mathbf{o}, \mathbf{d})$:

$$\begin{aligned} C_{pbr}(\mathbf{r}) = \int_{t_n}^{t_f} & T(t_n, t)\sigma_s(\mathbf{r}(t))L_s(\mathbf{r}(t), -\mathbf{d})dt \\ \text{s.t} \quad & \mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \\ & T(t_n, t) = \exp\left(-\int_{t_n}^{t}\sigma_t(\mathbf{r}(s))ds\right) \\ & L_s(\mathbf{x}, -\mathbf{d}) = \int_{S^2} f_p(\mathbf{x}, -\mathbf{d}, \bar{\mathbf{d}})L_i(\mathbf{x}, -\bar{\mathbf{d}})d\bar{\mathbf{d}} \\ & \sigma_t(\mathbf{r}(t)) = \sigma_a(\mathbf{r}(t)) + \sigma_s(\mathbf{r}(t)) \end{aligned} \tag{8}$$

S^2 is the domain of a unit sphere. σ_s and σ_a are the *scattering* coefficient and the *absorption* coefficient, respectively. Their sum is the *attentuation* coefficient, which is also known as the *density* in NeRF literature. $f_p(\mathbf{x}, -\mathbf{d}, \bar{\mathbf{d}})$ is the *phase function* that describes the probability of light scattering from direction $\bar{\mathbf{d}}$ to $-\mathbf{d}$ at point $\mathbf{x}$. $L_i(\mathbf{x}, -\bar{\mathbf{d}})$ is the incoming radiance towards point $\mathbf{x}$ along the direction $-\bar{\mathbf{d}}$, it can be computed as a weighted sum of $C_{rf}(\mathbf{x}, \bar{\mathbf{d}})$ (Eq. (1)) and radiance from an environment map $\text{Env}(\bar{\mathbf{d}})$:

$$\begin{aligned} L_i(\mathbf{x}, -\bar{\mathbf{d}}) =& C_{rf}(\mathbf{x}, \bar{\mathbf{d}}) \\ &+ \exp\left(-\int_{t_{n'}}^{t_{f'}}\sigma_t(\mathbf{x} + s\bar{\mathbf{d}})ds\right)\text{Env}(\bar{\mathbf{d}}) \end{aligned} \tag{9}$$

where $t_{n'}$ and $t_{f'}$ are the near and far points of secondary rays. In traditional physically based rendering, the first term represents indirect illumination while the second term represents direct illumination. Instead of modeling indirect illumination with path tracing, we use the trained radiance field to approximate it. This is also done in various recent works [33, 85] for modeling static scenes from multi-view input images. For Monte-Carlo estimation of $C_{pbr}(\mathbf{r})$, we will have to sample the two integrals $\int_{t_n}^{t_f}$ and $\int_{S^2}$ separately.

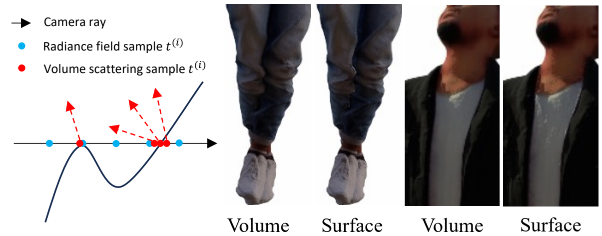

Figure 3. **Illustration of Volumetric Scattering.** Volumetric scattering can blend between multiple surfaces when a ray crosses edges (left). This results in smooth transitions of appearance at boundaries, avoiding noisy shadow (middle) and lighting (right) at these locations.

The first integral is estimated via quadrature as was done in standard NeRF rendering. We next describe how to sample the second integrals.

For approximating Eq. (8), we importance sample offsets $\{\bar{t}^{(1)}, \cdots, \bar{t}^{(M)}\}$ from the PDF estimated by radiance field samples that have been used to estimate Eq. (2) . The approximated Eq. (8) becomes:

$$\begin{aligned} C_{pbr}(\mathbf{r}) \approx \sum_{i=1}^{M} & w^{(i)}\frac{\sigma_s(\mathbf{r}(\bar{t}^{(i)}))}{\sigma_t(\mathbf{r}(\bar{t}^{(i)}))}L_s(\mathbf{r}(\bar{t}^{(i)}), -\mathbf{d}) \\ \text{s.t} \quad & \mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \\ & w^{(i)} = T^{(i)}\left(1 - \exp(-\sigma_t(\mathbf{r}(\bar{t}^{(i)})\delta^{(i)})\right) \\ & T^{(i)} = \exp\left(-\sum_{j<i}\sigma_t(\mathbf{r}(\bar{t}^{(j)}))\delta^{(j)}\right) \\ & L_s(\mathbf{r}(\bar{t}^{(i)}), -\mathbf{d}) = \frac{f_p(\mathbf{r}(\bar{t}^{(i)}), -\mathbf{d}, \bar{\mathbf{d}}^{(i)})}{\text{pdf}(\bar{\mathbf{d}}^{(i)})} \\ & \qquad \cdot L_i(\mathbf{r}(\bar{t}^{(i)}), -\bar{\mathbf{d}}^{(i)}) \\ & \sigma_t(\mathbf{r}(t)) = \sigma_a(\mathbf{r}(t)) + \sigma_s(\mathbf{r}(t)) \end{aligned} \tag{10}$$

in which $\frac{\sigma_s(\mathbf{r}(\bar{t}^{(i)}))}{\sigma_t(\mathbf{r}(\bar{t}^{(i)}))}$ corresponds to the spatially varying *albedo* and is analogous to that in surface-based rendering. $\text{pdf}(\bar{\mathbf{d}}^{(i)})$ is the PDF from which $\bar{\mathbf{d}}^{(i)}$ is sampled.

Essentially, we use quadrature to estimate the first integral $\int_{t_n}^{t_f}$, and Monte-Carlo sampling to estimate the second integral $\int_{S^2}$, all together with M samples. We refer readers to the Supp. Mat. for a detailed derivation of Eq. (10). During training $\bar{\mathbf{d}}^{(i)}$ is uniformly sampled from the unit sphere with $M = 512$ and stratified jittering [58]. For relighting, we use light importance sampling with $M = 1024$ to sample from a known environment map.

We note that when using an SDF-density representation, most of the samples $\bar{t}^{(i)}$ are concentrated around the surface of the human body. This makes the volumetric scattering process similar to a surface-based rendering process when there is a clear intersection between the ray and the surface.

On the other hand, rays at edges and boundaries may not have a well-defined surface as the corresponding pixels may cover multiple surfaces. For these rays, it would be difficult to employ surface-based rendering while volume scattering suits naturally for this case (Fig. 3).

We use a simplified version of Disney BRDF [9] to model the combined effect of the volumetric albedo $\frac{\sigma_s(\mathbf{r}(\bar{t}^{(i)}))}{\sigma_t(\mathbf{r}(\bar{t}^{(i)}))}$ and the phase function f_p. It takes predicted albedo α, roughness r and metallic m as inputs:

$$\frac{\sigma_s}{\sigma_t} f_p(\omega_o, \omega_i) = \mathrm{BRDF}(\omega_o, \omega_i, \alpha, r, m, \mathbf{n}) \max(\mathbf{n} \cdot \omega_i, 0)$$

We drop dependency on spatial locations for brevity. An extended implementation detail of the above BRDF can be found in the Supp. Mat. More physically accurate phase functions for rendering surface-like volumes, such as SGGX [26] can also be plugged into our formulation, but we empirically do not find them providing any advantage for our application.

3.4. Articulated Secondary Ray Tracing

Given the M samples $\{\bar{t}^{(i)}\}_{i=1}^{M}$ on a primary ray, we trace one secondary ray for each of the samples, and compute opacity (or visibility in a surface rendering setup) and radiance for each secondary ray. Formally, we trace a secondary ray $\bar{\mathbf{r}}^{(i)}$ from the corresponding sample, where $\bar{\mathbf{r}}^{(i)} = (\bar{\mathbf{o}}^{(i)}, \bar{\mathbf{d}}^{(i)})$ with $\bar{\mathbf{o}}^{(i)} = \mathbf{r}(\bar{t}^{(i)})$.

Secondary Ray Tracing: We note that traditional sphere tracing could lead to non-convergence rays when the SDF is not smooth. This is exacerbated when the SDF is approximated by neural networks and combined with body articulation. Furthermore, the sequential evaluation of SDF values on a ray is not amenable to parallelization, especially when a large number of secondary rays need to be evaluated and each evaluation involves neural networks.

Given the underlying NeRF representation, precise surface location is often not required to compute radiance, while the opacity is binary most of the time due to the SDF-density representation. This motivates us to use ray marching to compute secondary shading effects. However, we observe that the Laplace density function of [80] tends to assign non-negligible density values to small positive SDF values. This will cause the secondary ray marching to give non-zero weights to points that are very close to the surface, i.e. starting points $\bar{\mathbf{o}}$'s of secondary rays. While NeuS [72] is more well-behaved as it only assigns high weights for SDF zero-crossing intervals, estimating weights of ray segments requires estimation of analytical surface normals, which usually doubles the computation cost of ray marching.

Motivated by these facts, we propose a hybrid approach to secondary ray marching by searching for the first SDF zero-crossing point of a set of uniform samples on the secondary ray and only start accumulating importance weights from that point. Given the weights of uniform samples, we sample 4 additional samples on the secondary ray and compute the transmittance and radiance from these 4 samples. The computed transmittance and radiance are inputs to incoming radiance evaluation Eq. (9).

Algorithm 1 Zero-Crossing Search and Importance Weight Accumulation

Require: $\{\mathrm{SDF}(\bar{\mathbf{r}}(t'^{(i)}))\}_{i=1}^{64}$, $\bar{\mathbf{r}} = (\bar{\mathbf{o}}, \bar{\mathbf{d}})$
Ensure: Importance weights $\{w^{(i)}\}_{i=1}^{63}$

$s \leftarrow 1$
$\{w^{(i)}\}_{i=1}^{63} \leftarrow \mathbf{0}$
while $s < 63$ **do**
  if $\mathrm{SDF}(\bar{\mathbf{r}}(t'^{(s)})) \cdot \mathrm{SDF}(\bar{\mathbf{r}}(t'^{(s+1)})) < 0$ **then**
    break
  end if
  $s \leftarrow s + 1$
end while
$T(\bar{\mathbf{r}}) \leftarrow 1$
for $i = s$ **to** 63 **do**
  $\delta^{(i)} \leftarrow t'^{(i+1)} - t'^{(i)}$
  $w^{(i)} \leftarrow \left(1 - \exp(-\sigma_t(\bar{\mathbf{r}}(t'^{(i)}))\delta^{(i)})\right) T(\bar{\mathbf{r}})$
  $T(\bar{\mathbf{r}}) \leftarrow T(\bar{\mathbf{r}}) \exp\left(-\sigma_t(\bar{\mathbf{r}}(t'^{(i)}))\delta^{(i)}\right)$
end for
return $\{w^{(i)}\}_{i=1}^{63}$

Formally, given a secondary ray $\bar{\mathbf{r}}$, we first uniformly sample 64 offsets $\{t'^{(1)}, \cdots, t'^{(64)}\}$ on the ray between the near and far points, $t_{n'} = 0$ $t_{f'} = 1.5$. Each of the sampled offsets is transformed to the canonical space to query its SDF value:

$$\mathrm{SDF}(\bar{\mathbf{r}}(t')) = f_g(\gamma_g(\mathrm{LBS}^{-1}(\bar{\mathbf{r}}(t')))) \quad (11)$$

Alg. 1 describes the procedure of searching for the first zero-crossing point and accumulating weights for each of the points. This is similar to the traditional sphere tracing algorithm, with the difference that SDF values are evaluated uniformly in parallel instead of sequentially. We parallelize Alg. 1 together with importance sampling over rays with custom CUDA implementation.

3.5. Training Details

We use standard L1 loss wrt. input images on radiance predicted by both radiance field (RF loss) and volumetric scattering (PBR loss). We apply eikonal loss [22] (throughout training) and curvature loss [62] (only the first half of the training) to regularize the SDF field. We also apply Lipschitz regularization [62] and standard smoothness regularization [33, 84] to the material predictions. Details on losses and hyperparameters can be found in the Supp. Mat.

We train a total of 25k iterations with a learning rate of 0.001 decayed by a factor of 0.3 at 12.5k, 18.75k, $22.5k$, and 23.75k iterations, respectively. The first 10k iterations are trained with the RF loss only, while the rest of the iterations are trained on both the RF loss and the PBR loss. We use a batch size of 4096 rays. Training is done on a single NVIDIA RTX 3090 GPU in 4 hours.

4. Experimental Evaluation

4.1. Datasets

We utilize 3 different datasets to conduct our experiments

- **RANA [28]** To quantitatively evaluate our estimation of the physical properties of the reconstructed avatar, we use 8 subjects from the RANA dataset. The dataset is rendered using a standard path tracing algorithm, with ground truth albedo, normal, and relighted images available for evaluation. We follow protocol A in which the training set resembles a person holding an A-pose rotating in front of the camera under unknown illumination. The test set consists of images of the same subject in random poses under novel illumination conditions.
- **PeopleSnapshot [2]** In PeopleSnapshot, subjects always hold a simple A-pose and rotate in front of the camera under natural illumination. We use 6 subjects from the dataset with refined pose estimation from [13, 32].
- **SyntheticHuman-Relit** To additionally evaluate relighting on more complex training poses of continuous videos, we create a synthetic dataset by rendering two subjects from the SyntheticHuman dataset [56] with Blender under different illumination conditions. Due to space limits, we refer readers to the Supp. Mat. for details and results on this dataset.

4.2. Baselines

To our knowledge, Relighting 4D (R4D [15]) is the only baseline with publicly available code for the physically based inversed rendering of clothed human avatars under unknown illumination, without pretraining on any ground truth geometry/albedo/materials. RANA [28] only provides public access to their data at the time of our submission. Furthermore, RANA pretrains on ground truth albedo, which is not available in our setting.

We note that the original R4D implementation does not employ any mask loss. We therefore also report a variant of R4D (denoted as R4D*) that employs a mask loss. R4D* achieves overall better performance than R4D (Tab. 1) and thus we primarily compared our method to this improved version of R4D.

4.3. Evaluation Metrics

On synthetic datasets, we evaluate the following metrics:

- **Albedo PNSR/SSIM/LPIPS** we evaluate the standard image quality metrics on albedos rendered under training views. Due to ambiguity in estimating albedo and light intensity, we follow the practice of [84] to align the predicted albedo with the ground truth albedo. More details can be found in the Supp. Mat.
- **Normal Error** this metric evaluates normal estimation error (in degrees) between predicted normal images and the ground-truth normal images.
- **Relighting PSNR/SSIM/LPIPS** we also evaluate image quality metrics on images synthesized on novel poses with novel illumination. Relighting evaluation on training poses (i.e. SyntheticHuman-Relit dataset) is reported in the Supp. Mat.

On real-world datasets, i.e. PeopleSnapshot, we primarily present qualitative results including novel view/pose synthesis under novel illuminations.

4.4. Comparison to Baselines

We present the average metrics on the RANA dataset in Tab. 1. Our method significantly outperforms R4D and R4D* on all metrics, achieving 77% and 64% reduction in the normal estimation error, respectively. This combined with our explicit ray tracing technique also gives us a significant improvement in albedo-related metrics on training poses.

For relighting novel poses, we note that the SMPL model is not perfectly aligned with images in the RANA dataset, which could make the PSNR metric less meaningful. Thus we argue that SSIM and LPIPS can better reflect the quality of the relighting results. Nevertheless, R4D* fails to produce reasonable results due to its inability to generalize to novel poses. On the other hand, our method can produce high-quality re-posing and relighting results (Fig. 4).

4.5. Ablation Study

In this section, we ablate several of our design choices. We use subject 01 from the RANA dataset for this ablation study. We visualize average visibility (AV) maps which best reflect the quality of the reconstruction geometry and secondary ray tracing. The AV value of a primary ray $\mathbf{r}$ is defined as:

$$\mathrm{AV}(\mathbf{r}) = 2 * \frac{1}{M} \sum_{i=1}^{M} V(\bar{\mathbf{r}}^{(i)}) \tag{12}$$

where $V(\bar{\mathbf{r}}_i)$ is the visibility of the i-th secondary ray (1 for not occluded, 0 for occluded), and M is the number of secondary rays sampled for each primary ray. We multiply visibility by 2 as we sample secondary rays on a unit sphere instead of a hemisphere. The results are summarized in Fig. 5. We describe different variants in the following:

- **Ours:** Our full method with all the components described in Section 3.
- **Rendered Depth with Surface Scattering:** This variant corresponds to [33] that uses rendered depth and surface scattering.

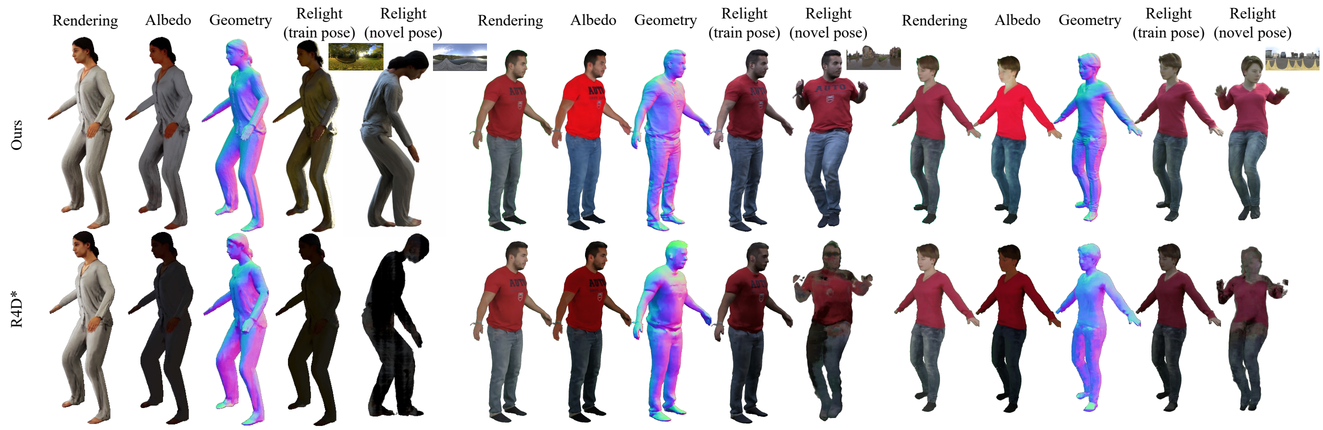

Figure 4. **Qualitative comparison to the baseline.** We show the results of our method and R4D* on both synthetic (left) and real (middle, right) datasets. As indicated, R4D* struggles to recover intrinsic properties of avatars and do not produce realistic relighting results. Furthermore, it fails to generalize to novel poses. Our method produces high-quality results on both synthetic and real datasets, while generalizing well to novel poses and illuminations. More qualitative results can be found in the Supp. Mat.

Method	Albedo			Normal	Relighting (Novel Pose)		
	PSNR ↑	SSIM ↑	LPIPS ↓	Error ↓	PSNR ↑	SSIM ↑	LPIPS ↓
R4D	18.24	0.7780	0.2414	42.69 °	14.37	0.8133	0.2017
R4D*	18.23	0.8254	0.2043	27.38 °	16.62	0.8370	0.1726
Ours	**22.83**	**0.8816**	**0.1617**	**9.96** °	**18.18**	**0.8722**	**0.1279**

Table 1. **Quantitative comparison to the baseline on the RANA dataset.**

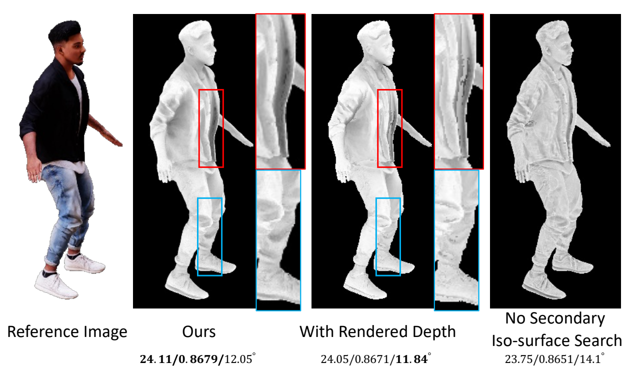

Figure 5. **Ablation study.** We visualize average visibility (AV) maps of each variant and report albedo PSNR (↑)/albedo SSIM (↑)/Normal Error (↓). Surface scattering with rendered depth results in discontinuities at boundaries and edges. Without our proposed iso-surface search for secondary ray tracing, the visibility map is much darker and does not reflect true visibility. We also refer readers to Fig. 3 for qualitative relighting results

- **No Iso-surface Search for Secondary Ray Tracing:** In this variant we do not perform the iso-surface search for secondary ray tracing (Sec. 3.4) and start accumulating weights from the first sample of the 64 samples on the secondary ray.

5. Conclusion

We have presented a novel approach to the inverse rendering of dynamic humans from only monocular videos. Our method can achieve high-quality reconstruction of clothed human avatars with disentangled geometry, albedo, material, and environment lighting from only monocular videos. We have also shown that our learned avatars can be rendered realistically under novel lighting conditions *and* novel poses. Experiment results show that our method significantly outperforms the state-of-the-art method both qualitatively and quantitatively. We discuss limitations and future work in the Supp. Mat.

6. Acknowledgment

We thank Umar Iqbal and Zhen Xu for the helpful discussion on setting up synthetic datasets. We thank Zijian Dong and Naama Pearl for constructive feedback on our early draft. SW, BA and AG were supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645. SW and ST acknowledge the SNSF grant 200021 204840.

References

[1] Alex Yu and Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[2] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[4] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 37(8):1670–1687, 2015. 2

[5] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. Flare: Fast learning of animatable and relightable mesh avatars. *ACM Trans. on Graphics*, 42(6):1–15, 2023. 3

[6] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason M. Saragih. Deep relightable appearance models for animatable faces. *ACM Trans. on Graphics*, 40 (4):89:1–89:15, 2021. 3

[7] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[8] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *Advances in Neural Information Processing Systems (NIPS)*, 2021. 2

[9] Brent Burley. Physically-based shading at disney. In *Proc. of SIGGRAPH*, 2012. 6

[10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2, 4

[11] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Factor fields: A unified framework for neural fields and beyond. *arXiv.org*, 2302.01226, 2023.

[12] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Dictionary fields: Learning a neural basis decomposition. *ACM Trans. on Graphics*, 42(4):1–12, 2023. 4

[13] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv.org*, 2106.13629, 2021. 7

[14] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 45(10): 11796–11809, 2023. 4

[15] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2, 3, 7

[16] Zhang Chen, Zhong Li, Liangchen Song, Lele Chen, Jingyi Yu, Junsong Yuan, and Yi Xu. Neurbf: A neural fields representation with adaptive radial basis functions. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 4

[17] Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas Simon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, He Wen, Lucas Evans, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, Shoou-I Yu, Javier Romero, Michael Zollhöfer, Yaser Sheikh, Ziwei Liu, and Shunsuke Saito. URhand: Universal relightable hands. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[18] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[19] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *Proc. of SIGGRAPH*, 2022. 3

[20] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 4

[21] Purvi Goel, Loudon Cohen, James Guesman, Vikas Thamizharasan, James Tompkin, and Daniel Ritchie. Shape from tracing: Towards reconstructing 3d object geometry and SVBRDF material from images via differentiable path tracing. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2020. 2

[22] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proc. of the International Conf. on Machine learning (ICML)*, 2020. 6

[23] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[24] Kaiwen Guo, Peter Lincoln, Philip L. Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Trans. on Graphics*, 38(6): 217:1–217:19, 2019. 1, 2, 3

[25] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using monte carlo rendering and denoising. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[26] Eric Heitz, Jonathan Dupuy, Cyril Crassin, and Carsten Dachsbacher. The SGGX microflake distribution. *ACM Trans. on Graphics*, 34(4):48:1–48:11, 2015. 3, 6

[27] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Trans. on Graphics*, 42(4):1–12, 2023. 1

[28] Umar Iqbal, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, and Jan Kautz. Rana: Relightable articulated neural avatars. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2, 3, 7

[29] Shun Iwase, Saito Saito, Tomas Simon, Stephen Lombardi, Bagautdinov Timur, Rohan Joshi, Fabian Prada, Takaaki Shiratori, Yaser Sheikh, and Jason Saragih. Relightablehands: Efficient neural relighting of articulated hand models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[30] Wenzel Jakob. *Light Transport On Path-Space Manifolds*. PhD thesis, Cornell University, 2013. 5

[31] Wenzel Jakob, Adam Arbree, Jonathan T. Moon, Kavita Bala, and Steve Marschner. A radiative transfer framework for rendering materials with anisotropic structure. *ACM Trans. on Graphics*, 29(4):53:1–53:13, 2010. 3

[32] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4, 7

[33] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6, 7

[34] Pratik Kalshetti and Parag Chaudhuri. Intrinsic hand avatar: Illumination-aware hand appearance and shape reconstruction from monocular rgb video. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[35] Pierre-Yves Laffont, Adrien Bousseau, and George Drettakis. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE Transactions on Visualization and Computer Graphic (TVCG)*, 19(2):210–224, 2013. 2

[36] Hendrik P. A. Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Trans. on Graphics*, 22(2):234–257, 2003. 2

[37] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3, 4

[38] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans. on Graphics*, 37(6):269:1–269:11, 2018. 2

[39] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[40] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[41] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W. Jacobs. Shape and material capture at home. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[42] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Relightable and animatable neural avatars from videos. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2024. 3

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. on Graphics*, 2015. 4

[44] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, pages 101–113. Wiley Online Library, 2021. 2

[45] Linjie Lyu, Ayush Tewari, Thomas Leimkuehler, Marc Habermann, and Christian Theobalt. Neural radiance transfer fields for relightable novel-view synthesis with global illumination. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[46] Alexander Mai, Dor Verbin, Falko Kuester, and Sara Fridovich-Keil. Neural microfacet fields for inverse rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 3

[47] 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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2, 3, 4

[48] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. on Graphics*, 2022. 2, 4

[49] Jacob Munkberg, Wenzheng Chen, Jon Hasselgren, Alex Evans, Tianchang Shen, Thomas Müller, Jun Gao, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[50] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. Practical SVBRDF acquisition of 3d objects with unstructured flash photography. *ACM Trans. on Graphics*, 37(6): 267:1–267:12, 2018. 2

[51] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 4

[52] Jeong Joon Park, Aleksander Holynski, and Steven M. Seitz. Seeing the world in a bag of chips. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[53] Steven Parker, Peter Shirley, and Brian Smitsa. Single sample soft shadow. Technical Report UUCS-98-019, Computer Science Department, University of Utah, 2018. 3

[54] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3

[55] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[56] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2024. 1, 3, 7

[57] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*, chapter 14. The MIT Press, 4th edition, 2023. 5

[58] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*, chapter 8. The MIT Press, 4th edition, 2023. 5

[59] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM Trans. on Graphics*, 38(4): 78:1–78:14, 2019. 2

[60] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 4

[61] Edoardo Remelli, Timur M. Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason M. Saragih, and Yaser Sheikh. Drivable volumetric avatars using texel-aligned features. In *Proc. of SIGGRAPH*, 2022. 1

[62] Radu Alexandru Rosu and Sven Behnke. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 6

[63] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[64] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and SVBRDF estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[65] Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, and Andreas Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[66] Carolin Schmitt, Božidar Antić, Andrei Neculai, Joo Ho Lee, and Andreas Geiger. Towards scalable multi-view reconstruction of geometry and materials. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 45:15850–15869, 2023. 2

[67] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[68] 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. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[69] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[70] Wenzhang Sun, Yunlong Che, Han Huang, and Yandong Guo. Neural reconstruction of relightable human model from monocular video. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2, 3

[71] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[72] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4, 6

[73] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 1, 3, 4

[74] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2

[75] Xin Wei, Guojun Chen, Yue Dong, Stephen Lin, and Xin Tong. Object-based illumination estimation with rendering-aware neural networks. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[76] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 4

[77] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[78] 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. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

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

[80] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4, 6

[81] Ye Yu and William A. P. Smith. Inverserendernet: Learning single image inverse rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[82] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[83] Xiuming Zhang, Sean Ryan Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip L. Davidson, Christoph Rhemann, Paul E. Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. Neural light transport for relighting and view synthesis. *ACM Trans. on Graphics*, 40(1):1–17, 2021. 2, 3

[84] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul E. Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor: neural factorization of shape and reflectance under an unknown illumination. *ACM Trans. on Graphics*, 40(6): 237:1–237:18, 2021. 2, 6, 7

[85] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[86] Youjia Zhang, Teng Xu, Junqing Yu, Yuteng Ye, Junle Wang, Yanqing Jing, Jingyi Yu, and Wei Yang. Nemf: Inverse volume rendering with neural microflake field. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 3

[87] Ruichen Zheng, Peng Li, Haoqian Wang, and Tao Yu. Learning visibility field for detailed 3d human reconstruction and relighting. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[88] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, et al. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12489–12498, 2023. 2