# Illumination-Consistent Human-Scene Reconstruction from Monocular Video

Rongbin Zheng[1] Wensheng Li[1] Lingzhe Zeng[1] Dong Wang[2] Chengying Gao[1*]
[1]Sun Yat-Sen University [2]South China Agricultural University

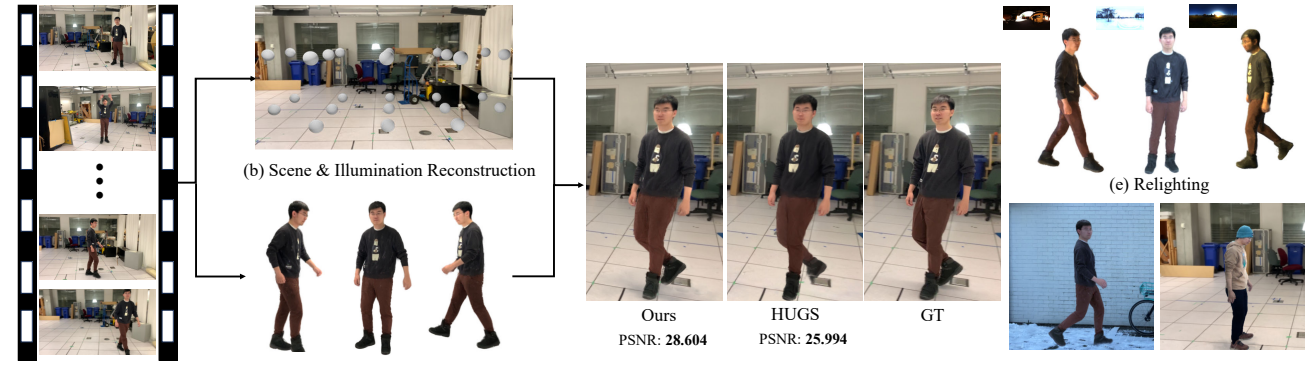


Figure 1. Given a monocular video (a), our method can reconstruct the static scene (b) and the animatable human (c). With our proposed scene lighting representation, we achieve novel view synthesis with novel human poses (d), facilitate human relighting (e), and enable rendering the human into the novel scene with corresponding lighting conditions (f).

## Abstract

*Reconstructing 3D humans and scenes from monocular videos is a challenging task, particularly due to human motion, varying illumination, and dynamic scene shadows. While recent works have explored scene disentanglement by jointly modeling humans and their surrounding scenes, they often overlook illumination and shadow effects—resulting in inconsistent human appearance and degraded scene realism. To address this gap, we propose a photometrically consistent integration of human and scene reconstruction based on 3D Gaussian Splatting, with a key focus on modeling spatially-varying illumination and shadows. Central to our method is a learnable light volume that provides localized lighting cues to human Gaussians, enabling more realistic and consistent appearance synthesis. To further ensure accurate human geometry and alignment, we adopt a two-stage reconstruction strategy: we first optimize a human mesh and then anchor Gaussians to the refined surface. In addition, we introduce an implicit shadow estimation module that disentangles cast shadows from the scene, thus supporting plausible human shadow synthesis. Our framework also facilitates human relighting and compositing into novel scenes with contextually appropriate lighting. Quantitative and qualitative results demonstrate that our method achieves state-of-the-art performance, producing consistent appearances, realistic illumination, and enhanced overall scene realism.*

## 1. Introduction

Reconstructing 3D humans from monocular videos is important for applications in film production, gaming, and VR/AR. Recent methods [6, 18, 20, 29, 31, 34, 41, 42, 47, 56, 68, 76] leverage NeRF [40] or 3D Gaussian Splatting (3DGS) [24] to reconstruct dynamic humans in well-controlled environments. However, these methods are often limited in in-the-wild settings due to uncontrolled lighting and scene complexity. Other approaches that jointly model full scenes [4, 9, 10, 30, 43, 48, 51, 72, 74] support human-scene reconstruction but struggle to animate humans, as they lack explicit human modeling. Alternatively, some methods [16, 21, 25] reconstruct humans and scenes independently and then merge them in a straightforward manner. These approaches fail to capture the intricate relationships between humans and their environments, particularly the interplay of illumination and shadows, which is crucial for realistic human-scene reconstruction.

In real-world environments, dynamic human motion,

*Corresponding author.

camera movement, occlusions, and varying light sources introduce complex illumination changes that affect both human appearance and scene shadows. To handle these effects, disentangling human material properties and scene lighting is necessary. However, recent relightable human avatar methods [5, 8, 33, 67, 75, 81] typically rely on environment maps that assume infinitely distant lighting, making them unsuitable for modeling spatially variant illumination. Moreover, human movement naturally produces dynamic shadows. Although existing inverse rendering techniques [14, 22, 54] employ ray tracing to model shadows, applying them to millions of Gaussians in large-scale scenes is computationally prohibitive. Furthermore, since these methods are designed for static scenes, they struggle to handle the complex shadow effects caused by dynamic humans.

In this work, we take a step forward by addressing photometrically consistent human–scene reconstruction, where geometry, material, and spatially-varying illumination are jointly inferred within a unified framework. Unlike prior works that focus on one aspect, such as human reconstruction or relighting, we specifically model the interplay of lighting and shadows between dynamic humans and their environments, a crucial yet underexplored task for realistic reconstruction. To our knowledge, this is the first exploratory study toward illumination-consistent human–scene reconstruction from in-the-wild videos.

To this end, we introduce a light-aware reconstruction framework comprising three components: a light volume, a two-stage human reconstruction pipeline, and an implicit shadow estimation module. The light volume defines a spatially-distributed set of probes, each storing spherical harmonic coefficients for radiance and a learnable latent feature for shadow reasoning. This compact representation captures spatially-varying illumination, enabling coherent and photorealistic rendering. Based on this lighting representation, we reconstruct relightable humans in two stages: first, by optimizing the SMPL [38] mesh to capture geometry, and second, by estimating material properties through PBR-based optimization on human Gaussians. Finally, the implicit shadow module leverages learned lighting features to efficiently model human-cast shadows on nearby scene regions, avoiding costly full-scene ray tracing. Together, these components allow our framework to achieve coherent lighting, realistic shadows, and photometrically consistent human–scene integration under challenging in-the-wild conditions. In summary, the main contributions are:

- We introduce a framework for illumination-consistent human–scene reconstruction, jointly modeling geometry, appearance, and spatially-varying illumination.
- We present the light volume, a new lighting representation that captures spatially-varying illumination and enables the renderings of fine-grained lighting details.
- Based on the light volume, we introduce an implicit scene shadow estimation module that effectively and efficiently disentangles shadow effects from scene and enables shadow-consistent rendering.
- Our framework supports a wide range of downstream applications, including relightable human avatars and illumination-consistent human–scene transfer.

## 2. Related Work

### 2.1. Human reconstruction

Traditional methods require specialized capture systems such as depth cameras [17, 57] or dense multi-camera rigs [65, 66], which limits their practicality in daily use. Previous works utilized parametric models [1, 38] to represent humans, enabling reconstruction across various poses and shapes, but they struggle to capture fine details such as clothing textures and hair. In recent years, numerous methods [16, 20, 21, 26, 31, 32, 37, 46, 47, 58, 63, 83] have used NeRF [40] for implicit human representation, allowing human reconstruction from multi-view inputs and producing high-quality renderings. However, NeRF-based methods often suffer from low processing speed and substantial memory consumption. With the advancement of 3DGS technology [24], recent works [6, 18, 25, 29, 34, 41, 42, 45, 49, 56, 62, 68, 76, 82] employ explicit point clouds to represent dynamic humans, achieving both high rendering quality and efficiency. However, most of these methods rely on accurate background pre-segmentation and overlook the influence of the surrounding scene.

### 2.2. Human-Scene Reconstruction

Current 4D reconstruction approaches can produce high-quality results of dynamic human and static background. Some works [43, 44, 48, 71] reconstruct dynamic scenes by decoupling the scenes into a canonical space and a temporal deformation field. Another line of work exploits 4D grid-based representations [4, 11–13] to reconstruct the scene. These methods cannot animate humans as they lack explicit human modeling.

Performing separate reconstructions of the human and the scene can effectively handle both camera and human motion. Neuman [21] renders remarkable results in reconstructing humans and scenes from monocular video input, while Vid2Avatar [16] introduces SDF to reconstruct finer human geometric details. HUGS [25], based on 3DGS, achieves higher quality and higher speed dealing with the time-consuming problem of NeRF-based methods. HSR[70] and ODHSR[80] mainly focus on addressing occlusion handling or camera localization issues within a unified framework. However, these works emphasize the details of human reconstruction while neglecting the interactions between humans and the scene, including lighting conditions and shadows.

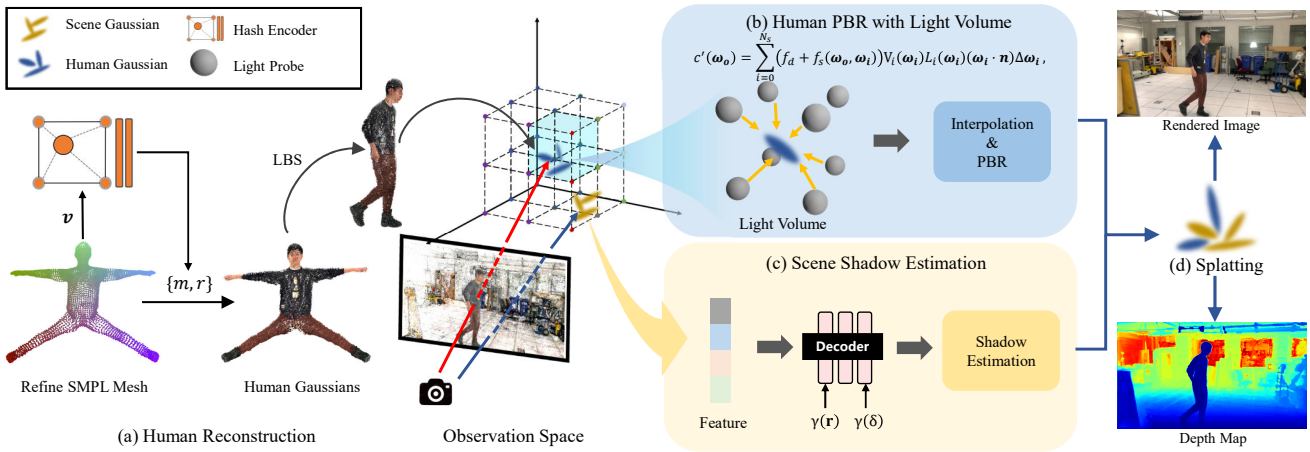


Figure 2. **Overview of our method.** a) First, we transform the human Gaussians from canonical space to observation space. b) Next, before splatting, we query the incoming light radiance from the k-nearest probes in light volume. Then we interpolate the light radiance and perform physically-based rendering on human Gaussians. c) We decode the lighting features extracted from the light volume to estimate human shadows in the scene. d) Finally, the human and the scene Gaussians are splatting together to render the image and the depth map.

### 2.3. Material and Lighting Estimation

Disentangling material and lighting from images remains challenging. Traditional approaches [2, 3, 15, 27, 28, 39, 50, 61] typically assume controlled illumination or known geometry. NeRF-based methods [22, 54, 64, 77–79] relax these assumptions by jointly estimating geometry and material properties under arbitrary lighting. Recent 3DGS-based approaches [14, 35, 52, 55, 84] estimate material properties directly on Gaussian primitives for high-quality reconstruction, but they are unsuitable for dynamic humans.

To generate relightable humans, some methods [8, 19, 59, 67, 69] leverage NeRF to estimate material properties and produce corresponding material maps through volume rendering. Some of them [67, 69] further utilize SDF to reconstruct. Nevertheless, these implicit-field-based works are challenging to apply practically due to slow rendering and training speeds. Another work [7] reconstructs triangular human avatars based on mesh, which is compatible with traditional graphics pipeline. Recent works [5, 33, 34, 36, 75] have introduced 3DGS and 2DGS [81] into the relighting task. Most of them focus on decoupling material properties for relighting, but using simplified environment maps limits their ability to capture scene illumination accurately.

## 3. Method

Our approach aims to reconstruct relightable humans and static scenes, while jointly learning scene illumination and shadows. As shown in Fig. 2, we employ 3DGS to reconstruct human and scene separately. The human Gaussians are defined in canonical space based on refined SMPL mesh and warped to posed space using LBS (Sec. 3.1). We perform PBR on human Gaussions with light volume, which achieves spatially-variant scene illumination reconstruction (Sec. 3.2). To achieve scene shadow estimation, for the scene Gaussians around human, we query features from the light volume, and leverage a decoder to estimate the occlusion (Sec. 3.3). Several additional constraints are applied during training (Sec. 3.4).

### 3.1. Human Reconstruction

To reconstruct more detailed human geometry, we adopt a two-stage training strategy as shown in Fig. 3.

**Stage 1: Geometry and Color Initialization.** In the first stage, we reconstruct the human directly without involving physically-based rendering. Following [45], we up-sample the SMPL mesh surface and distribute human Gaussians over the mesh faces. Then, we employ two hash encoders to learn vertex-wise offsets $\Delta_v$ and color attributes. Specifically, the refined vertices $\boldsymbol{v'}$ and colors $\boldsymbol{c}$ are obtained as:

$$\boldsymbol{v'} = \boldsymbol{v} + \mathcal{F}_{\Delta}(\boldsymbol{v}), \quad \boldsymbol{c} = \mathcal{F}_c(\boldsymbol{v}). \tag{1}$$

In this stage, we extract Gaussians from mesh surface through barycentric interpolation and indirectly optimize the mesh by refining these Gaussians.

**Stage 2: Physically-Based Appearance Modeling.** In the second stage, we incorporate the PBR pipeline to enable relightable human appearance. Each human Gaussian, defined in the canonical space, carries basic attributes including rotation $\boldsymbol{q} \in \mathbb{R}^4$, scale $\boldsymbol{s} \in \mathbb{R}^3$, position $\boldsymbol{x} \in \mathbb{R}^3$, albedo $\boldsymbol{b} \in \mathbb{R}^3$, and opacity $\alpha \in \mathbb{R}$. These parameters are directly optimized and further adjusted by densification and pruning based on KL divergence [18]. To enable realistic shading effects, we extract normals $\boldsymbol{n} \in \mathbb{R}^3$ from the refined human mesh and assign the roughness $r \in \mathbb{R}$ and metallic $m \in \mathbb{R}$ attributes with a hash encoder:

$$\{m, r\} = \mathcal{F}_m(\boldsymbol{v}). \tag{2}$$

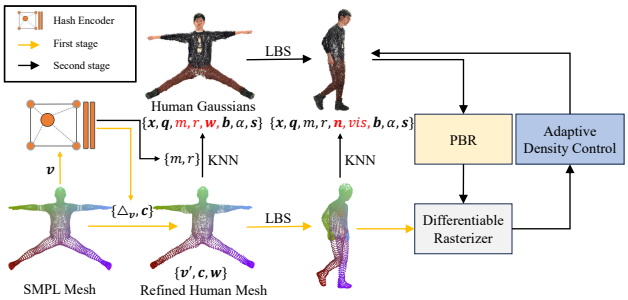


Figure 3. **Pipeline of human reconstruction**. We use a two-stage strategy to model the human, with the yellow and black arrows representing the first and second stages, respectively.

Self-occlusion introduces ambiguity in disentangling material and lighting effects. To address this, we introduce a pose-aware visibility estimator $\mathcal{F}_{vis}$ [36], which takes the mesh vertex $\boldsymbol{v}$, pose vector $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and view direction $\boldsymbol{\phi} \in \mathbb{R}^3$ as input to predict per-vertex visibility $vis$.

$$vis = \mathcal{F}_{vis}(\boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{\phi}). \tag{3}$$

The final material attributes $(m, r)$, normal attribute $\boldsymbol{n}$, visibility $vis$ and LBS weights $\boldsymbol{w} \in \mathbb{R}^{24}$ for each human Gaussian are obtained from their corresponding mesh vertices via k-nearest neighbors (KNN).

### 3.2. Illumination Reconstruction

To enable the decoupling of scene illumination, we apply physically-based rendering to each human Gaussian to simulate lighting effects and achieve more realistic rendering. The physically-based rendering equation [23] we employ can be represented as follows:

$$L_o(\boldsymbol{\omega_o}, \boldsymbol{x}) = \int_{\Omega} f(\boldsymbol{\omega_o}, \boldsymbol{\omega_i}, \boldsymbol{x}) V(\boldsymbol{\omega_i}, \boldsymbol{x}) L_i(\boldsymbol{\omega_i}, \boldsymbol{x}) (\boldsymbol{\omega_i} \cdot \boldsymbol{n}) d\boldsymbol{\omega_i}, \tag{4}$$

where $L_i$ and $L_o$ represent the incoming and outgoing light radiance in directions $\boldsymbol{\omega_i}$ and $\boldsymbol{\omega_o}$. $V$ is the visibility term modeled in Sec. 3.1. $\Omega$ denotes the hemispherical domain around the normal $\boldsymbol{n}$. And $f$ indicates the simplified Disney BRDF model in [14], divided into diffuse reflection term $f_d = \frac{\boldsymbol{b}}{\pi}$, and specular reflection term $f_s$:

$$f_s(\boldsymbol{\omega_o}, \boldsymbol{\omega_i}) = \frac{D(\boldsymbol{h}; r) \cdot F(\boldsymbol{\omega_o}, \boldsymbol{h}) \cdot G(\boldsymbol{\omega_i}, \boldsymbol{\omega_o}, \boldsymbol{h}; r)}{(\boldsymbol{n} \cdot \boldsymbol{\omega_i}) \cdot (\boldsymbol{n} \cdot \boldsymbol{\omega_o})}, \tag{5}$$

where $\boldsymbol{h}$ represents the half vector. $D$, $F$, $G$ denote the microfacet distribution function, Fresnel reflection and geometric shadowing factor, respectively.

Existing human relighting methods typically assume that the light is emitted from infinity and use a single environment map, which cannot handle spatially-varying or occluded illumination in complex scenes. Unlike most works which set the lighting conditions as an environment map, we introduce light volume to model the finer level illumination. As shown in Fig. 2(b), the light volume is designed as a grid where each vertex is represented as a light probe. We define the light probes as spherical harmonics instead of environment maps. Since we reconstruct the scene from a single view, certain incoming light remains unobserved, making it challenging for the environment map to capture information from unseen directions. However, low-degree harmonics can smooth the light radiance across all directions, which is beneficial for novel-view rendering.

Given a human Gaussian in observation space, we locate the $n$-nearest light probes. For each light probe $p_k$, we perform the same operation: Fibonacci sampling [73] of the incoming light $\boldsymbol{\omega_i}$ based on the normal $\boldsymbol{n}$ of the Gaussians, then derive the incoming radiance $L_k$ using spherical harmonics. The final radiance $L_i$ for each incoming direction is obtained via interpolation:

$$L_i(\boldsymbol{x}, \boldsymbol{\omega_i}) \approx \frac{\sum_k^n w_k(\boldsymbol{x}) L_k(p_k, \boldsymbol{\omega_i})}{\sum_k^n w_k(\boldsymbol{x})}, \tag{6}$$

where $n$ denotes the number of probes, $w_k(\cdot)$ denotes the weight of the probe $p_k$.

Finally, the PBR color of human Gaussians can be given by Monte Carlo integration:

$$c'(\boldsymbol{\omega_o}) = \sum_{i=0}^{N_l} (f_d + f_s(\boldsymbol{\omega_o}, \boldsymbol{\omega_i})) V(\boldsymbol{\omega_i}) L_i(\boldsymbol{\omega_i}) (\boldsymbol{\omega_i} \cdot \boldsymbol{n}) \Delta\boldsymbol{\omega_i}, \tag{7}$$

where $N_l$ represents the number of sampled light rays.

### 3.3. Scene Shadow Estimation

To model the shadows cast by dynamic humans onto the static scene, we estimate local occlusion effects in an implicit manner, as shown in Fig. 2(c).

Each light probe $p_i$ in the light volume is assigned a latent lighting feature $\boldsymbol{z_i}$, which captures localized illumination context. To efficiently estimate human-induced shadows, we first compute an axis-aligned bounding box around the human in observation space and identify scene Gaussians located within this region before splatting.

For each selected scene Gaussian, we retrieve the relevant lighting feature $\boldsymbol{z}$ from the surrounding light probes via interpolation. We additionally consider two spatial descriptors to model the influence of lighting on shadow formation: the distance $\delta$ and orientation $\mathbf{r}$ of the scene Gaussian relative to the center of the human bounding-box. These descriptors are concatenated with the lighting feature $\boldsymbol{z}$ and passed into a shadow-weight decoder network $\mathcal{F}_{ao}$ to predict an occlusion factor $ao$. The resulting scalar is then used to modulate the spherical harmonics of each scene Gaussian, simulating the soft shadow effect introduced by the nearby human. The decoder can be formally expressed as:

Table 1. **Quantitative comparison on the NeuMan dataset [21].** Our method achieves state-of-the-art performance in all sequences.

| | Jogging | | | Bike | | | Lab | | | Parkinglot | | | Citron | | | Seattle | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| 4DGS | 19.277 | 0.554 | 0.483 | 20.457 | 0.726 | 0.318 | 18.841 | 0.727 | 0.364 | 21.317 | 0.699 | 0.392 | 16.971 | 0.677 | 0.432 | 18.532 | 0.577 | 0.329 |
| Vid2Avatar | 15.043 | 0.414 | 0.609 | 14.862 | 0.514 | 0.582 | 13.962 | 0.601 | 0.566 | 21.561 | 0.688 | 0.344 | 14.324 | 0.622 | 0.503 | 17.406 | 0.558 | 0.480 |
| NeuMan | 22.697 | 0.681 | 0.273 | 25.551 | 0.830 | 0.166 | 24.960 | 0.862 | 0.149 | 25.434 | 0.800 | 0.201 | 24.757 | 0.812 | 0.178 | 23.987 | 0.782 | 0.194 |
| HUGS | 23.746 | 0.778 | 0.177 | 25.454 | 0.844 | 0.097 | 25.994 | 0.915 | 0.070 | 26.859 | 0.849 | 0.135 | 25.539 | 0.859 | 0.095 | 25.934 | 0.852 | 0.093 |
| Ours | 26.125 | 0.862 | 0.120 | 29.02 | 0.922 | 0.046 | 28.604 | 0.934 | 0.055 | 29.859 | 0.900 | 0.095 | 27.2070 | 0.880 | 0.068 | 29.56 | 0.931 | 0.051 |

Table 2. **Quantitative comparison of ours method with baseline mehods on the NeuMan dataset [21] over the human regions.** Our method significantly outperforms NeRF-based and Gaussian-based baselines on all metrics.

| | Jogging | | | Bike | | | Lab | | | Parkinglot | | | Citron | | | Seattle | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Vid2Avatar | 16.360 | 0.465 | 0.278 | 12.444 | 0.392 | 0.482 | 15.987 | 0.616 | 0.290 | 18.513 | 0.652 | 0.214 | 15.964 | 0.594 | 0.242 | 16.902 | 0.512 | 0.250 |
| NeuMan | 17.572 | 0.538 | 0.274 | 19.049 | 0.660 | 0.174 | 18.756 | 0.726 | 0.193 | 17.663 | 0.660 | 0.206 | 18.589 | 0.638 | 0.161 | 18.417 | 0.578 | 0.186 |
| HUGS | 17.452 | 0.588 | 0.256 | 19.476 | 0.673 | 0.158 | 18.789 | 0.761 | 0.152 | 19.437 | 0.729 | 0.151 | 19.157 | 0.706 | 0.134 | 19.060 | 0.668 | 0.142 |
| Ours | 20.475 | 0.684 | 0.152 | 22.586 | 0.77 | 0.084 | 22.106 | 0.808 | 0.108 | 22.375 | 0.788 | 0.102 | 20.955 | 0.726 | 0.096 | 22.902 | 0.771 | 0.092 |

$$ao = \mathcal{F}_{ao}(\gamma(\mathbf{r}), \gamma(\delta), \boldsymbol{z}), \tag{8}$$

where $\gamma(\cdot)$ denotes the position encoding. This implicit formulation enables efficient approximation of dynamic shadowing without full-scene ray tracing. The updated spherical harmonics for the scene Gaussians are then defined as:

$$SH' = ao \cdot SH. \tag{9}$$

Finally, the scene Gaussians are combined with human Gaussians and sent to the Gaussian splatting rasterizer.

### 3.4. Training Objectives

**Image Loss.** Similar to previous work, we supervise our rendered image of both the human and the scene by directly applying L1 loss, SSIM loss [60], and perceptual loss with VGG [53]. The image loss is defined as:

$$\mathcal{L}_{image} = \lambda_h \mathcal{L}_{human} + \lambda_s \mathcal{L}_{scene}. \tag{10}$$

**Regularization of Depth.** Accurate scene geometry is essential for stable human movement and reducing depth ambiguity. We render depth maps via splatting and supervise them using an L1 loss, represented as $\mathcal{L}_{depth}$.

**Mesh loss.** To smooth the human mesh, we apply a Laplacian regularizer $\mathcal{L}_{mesh}$ in the first stage.

**Smooth Loss.** To estimate materials correctly, we apply a smooth constraint [14] on the materials as:

$$\mathcal{L}_{materials} = \|\nabla R\| \exp(-\|\nabla C_{gt}\|). \tag{11}$$

where $R$ denotes the rendered materials map.

To smooth the scene shadow, we apply $\mathcal{L}_{shadows}$ as:

$$\mathcal{L}_{shadows} = \sum_{i=1}^{N_s} \sum_{k \in \mathcal{P}_s^i} \parallel ao^i - ao^k \parallel_1, \tag{12}$$

where $N_s$ is the number of scene Gaussians, $\mathcal{P}_s^i$ is the k-nearest domain of scene Gaussians $x_s^i$.

We expect that the lighting of probe neighbors will not change drastically, then we introduce a probes loss to smooth the local light conditions:

$$\mathcal{L}_{probe} = \sum_{i=1}^{N_p} \parallel L_i - \frac{1}{K} \sum_{j=1}^{K} L_j \parallel_1, \tag{13}$$

where $N_p$ is the probes in light volume, $K$ is the k-nearest probes count, $L_i$ is the radiance of the probes $p_i$. Then the smooth loss involves $\mathcal{L}_{materials}$, $\mathcal{L}_{probe}$ and $\mathcal{L}_{shadow}$:

$$\mathcal{L}_{smooth} = \lambda_m \mathcal{L}_{materials} + \lambda_p \mathcal{L}_{probe} + \lambda_s \mathcal{L}_{shadow}. \tag{14}$$

**Scale Loss.** We find that large Gaussians might lead to artifacts in novel poses synthesizes, then we apply scale loss $\mathcal{L}_{scale}$ [75] on the size property of human Gaussians.

The overall training objective function is formulated as :

$$\mathcal{L} = \lambda_1 \mathcal{L}_{image} + \lambda_2 \mathcal{L}_{depth} + \lambda_3 \mathcal{L}_{smooth} + \lambda_4 \mathcal{L}_{scale}. \tag{15}$$

## 4. Experiments

### 4.1. Dataset and Metrics

**NeuMan [21].** A collection of 6 videos captured by mobile phone, in which a person walks in different scenes, including the indoor scene *Lab* and outdoor scenes *Citron*, *Seattle*, *Parkinglot*, *Bike*, and *Jogging*.

**ZJU-MoCap [47].** An indoor dataset without scene background. We select 6 human subjects (*377, 386, 387, 392, 393, 394*) to conduct experiments.

**Metrics.** We follow previous works to evaluate our method using three standard metrics, *i.e.*, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

### 4.2. Comparison

#### 4.2.1. Quantitative Results

We follow HUGS [25] and validate our method on two aspects: entire scene and human reconstruction.

(a) Vid2Avatar (b) NeuMan (c) HUGS (d) Ours (e) Ground Truth

Figure 4. **Qualitative results comparing our method with baseline methods.** Our work shows better reconstruction quality both human and scene. Furthermore, our approach captures more human details (green boxes), lighting (blue boxes) and shadow effects (red boxes).

Table 3. **Quantitative comparison of our method with SOTA methods on the ZJU-MoCap dataset [47].**

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| NB | 29.31 | 0.9628 | 0.05208 |
| HumanNeRF | 30.53 | 0.9695 | 0.03131 |
| Intrinsic Avatar | 28.43 | 0.9582 | 0.04894 |
| HUGS | 30.56 | 0.9703 | 0.03089 |
| Ours | 30.73 | 0.9705 | 0.02840 |

Tab. 1 and Tab. 2 compare our method with 4DGS [72], Vid2Avatar [16], NeuMan [21], and HUGS [25] on the NeuMan dataset. Tab. 1 reports scene-level reconstruction quality, and Tab. 2 reports human-region quality. Our method outperforms all baseline methods in both evaluations.

In Tab. 3, we also compare our method with four baselines on novel view synthesis on the ZJU-MoCap dataset. We train all methods using a single view as input, and the results show that our method also achieves state-of-the-art performance.

#### 4.2.2. Qualitative Results

**Human-Scene Reconstruction.** In Fig. 4, our method shows better reconstruction quality than the baselines in the background regions of the scene. Vid2Avatar models the scene in a human-centered coordinate system, resulting in the lack of global 3D consistency across frames. Therefore, it fails to render a high-quality background. Thanks to the advancements of 3DGS, both our method and HUGS are able to reconstruct more detailed background information. In contrast, our reconstructed scene exhibits enhanced clarity and finer detail, as evidenced by the clearly rendered

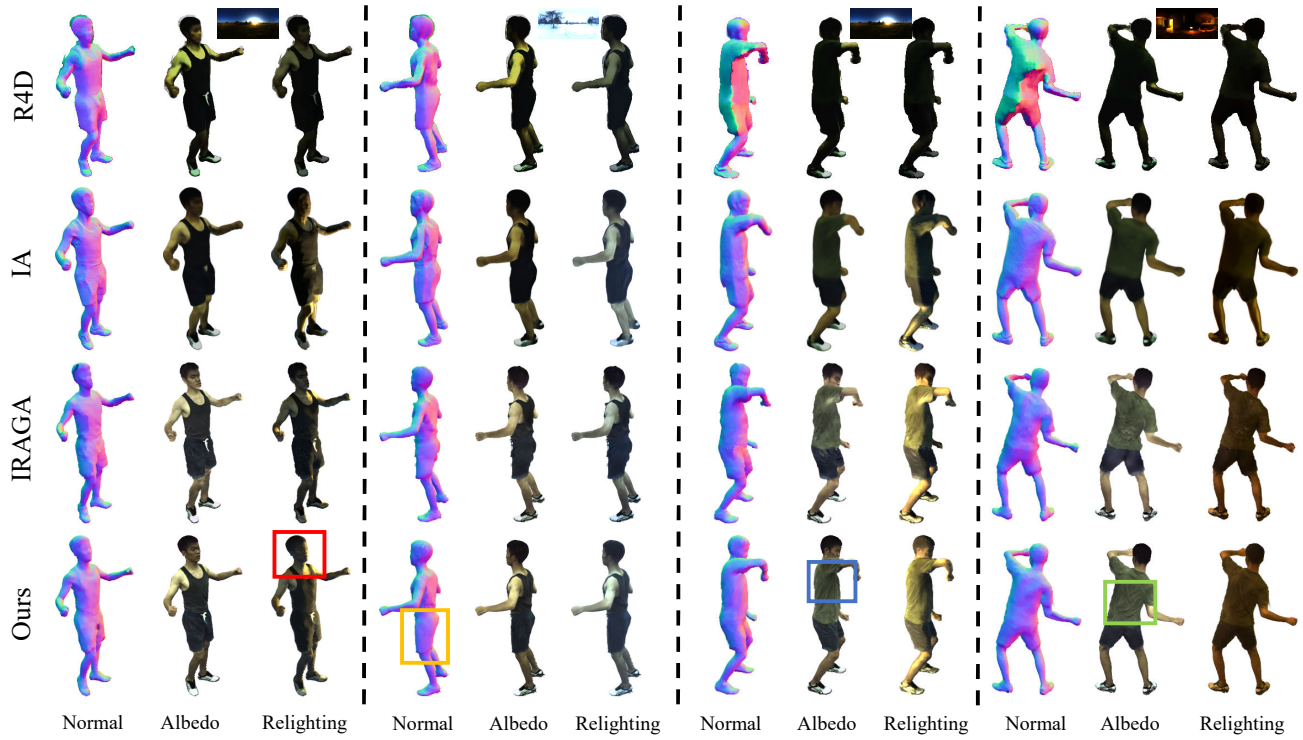


Figure 5. **Qualitative results comparing our method with human relighting baseline methods.**


Figure 6. **Qualitative results of ablation study on the Lab sequence of NeuMan dataset.** "w/o light volume" means that we don't perform PBR on human Gaussians.

floor tiles (row 1, 2). Additionally, owing to our scene shadow estimation module, our method can predict the human shadow within the scene (red boxes in row 2, 3).

In human regions, our method shows more details in geometry and lighting. Compared with Vid2Avatar and Neuman, we can preserve more geometry details on human heads, hands, and shoes. Due to the lack of lighting and material decoupling, NeuMan and HUGS display incorrect human appearance (blue box in row 1). Although Vid2Avatar also shows the correct appearance, it relies on human pose parameters instead of lighting. Thanks to light volume, we can reconstruct appearances with lighting features such as the highlights on the hair and localized lighting effects on the clothing (blue boxes in row 1, 3).

**Human Relighting.** As shown in Fig. 5, we present qualitative comparisons of human relighting on the ZJU-MoCap dataset, compared with the baselines R4D [8], IA [59], and IRAGA [75]. Due to limited material disentanglement, R4D fails to relight effectively. IA over-smooths the geometry, causing noticeable loss of appearance details, especially in facial regions (red box). IRAGA relies on a 3DGS reconstruction with a pre-extracted mesh, but often introduces geometric artifacts, while our two-stage reconstruction produces smoother and more stable geometry (orange box). In addition, IRAGA's albedo shows regional inconsistencies, leading to unrealistic relighting (blue box), whereas our disentangled albedo preserves richer fine-grained details, such as clothing wrinkles (green box).

Table 4. **Ablation study on several designs.** All methods are trained and rendered on *Lab* sequence of NeuMan dataset.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| w/o light volume | 21.97 | 0.8055 | 0.1061 |
| w/o shadow | 21.59 | 0.8100 | 0.1068 |
| w/o $\mathcal{L}_{probe}$ | 21.87 | 0.8035 | 0.1121 |
| w/o two-stage | 22.08 | 0.8054 | 0.1130 |
| Ours | 22.18 | 0.8104 | 0.1049 |

Reference full model w/o shadow estimation

Figure 7. **Ablation study on scene shadow estimation.** Results show that our method can effectively disentangle shadows from the scene, thereby improving the quality of scene reconstruction.

### 4.3. Ablation study

We conduct ablation studies on the NeuMan dataset to evaluate the contribution of key components in our framework.

**Ablation studies on Light Volume.** As shown in Tab. 4, light volume and probe loss significantly improve all metrics. In Fig. 6, without light volume, the human appearance lacks significant lighting detail, such as highlights on the hair and diffuse reflections on the clothing, which suggests that the light volume is effective for local lighting reconstruction in the scene. We also exhibit the effect of removing the probe loss from the light volume, which results in unrealistic lighting in certain areas (face, shoes) in Fig. 6. This occurs because some light probes within the light volume are insufficiently trained.

**Ablation studies on Scene Shadow Estimation.** Tab. 4 shows that scene shadow estimation can produce better results. In Fig. 6, the integration of shadow estimation module enables the synthesis of plausible shadows beneath the human, which in turn enhances the perceptual realism of the reconstructed human-scene. Furthermore, the results in Fig. 7 present that the model without shadow estimation might learn dynamic human shadows as a part of the scene, decreasing the scene reconstruction quality.

### 4.4. Applications

Our framework supports a wide range of downstream applications. As shown in Fig. 8, we place humans into new scenes and apply the corresponding illumination, resulting in consistent appearance changes across scenes, facilitating applications in scene replacement. Fig. 9 shows that our method can perform human relighting from monocular in-the-wild video.


Figure 8. **Qualitative results of human scene transfer.** We transfer humans from the *Lab* and *Bike* sequences to different scenes with the correspondent lighting condition.

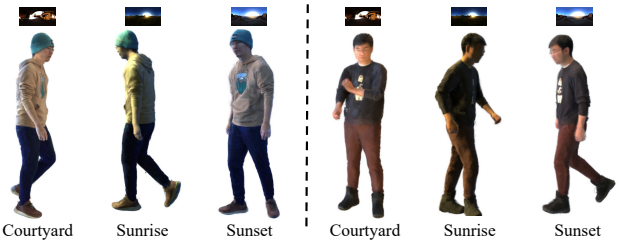


Figure 9. **Qualitative results of relighting.** We present renderings under different environment lighting conditions.

## 5. Discussion

**Conclusion.** We propose a novel framework for 3D human-scene reconstruction from monocular videos, with a specific focus on modeling spatially-varying illumination and dynamic shadows. To achieve this, we adopt a two-stage reconstruction strategy to build a relightable human with refined geometry. We further introduce a light volume representation that encodes localized lighting information across the scene. Leveraging these lighting features, we develop an implicit shadow estimation module that predicts occlusion factors to model cast shadows. Extensive experiments demonstrate that our approach achieves state-of-the-art performance in novel-view and novel-pose synthesis, and enables applications such as scene and illumination transfer.

**Limitation.** Our method achieves human-scene modeling through an optimization-based approach, indicating that we cannot perform novel human or scene prediction in a feed-forward manner. Moreover, our method assumes a relatively static scene environment, which limits its ability to model mutual interactions between multiple dynamic humans and complex scene dynamics.

**Acknowledgement.** This work was supported by the Guangxi Science and Technology Major Program (GuikeAA23073007).

## References


[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM siggraph 2005 papers*, pages 408–416. Association for Computing Machinery, 2005. 2

[2] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 294–311. Springer, 2020. 3

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

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 1, 2

[5] Diogo Carbonera Luvizon, Vladislav Golyanik, Adam Kortylewski, Marc Habermann, and Christian Theobalt. Relightable neural actor with intrinsic decomposition and pose control. In *European Conference on Computer Vision*, pages 465–483. Springer, 2024. 2, 3

[6] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10723–10734, 2025. 1, 2

[7] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *European Conference on Computer Vision*, pages 250–269. Springer, 2024. 3

[8] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. 2, 3, 7

[9] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[10] Runze Fan, Jian Wu, Xuehuai Shi, Lizhi Zhao, Qixiang Ma, and Lili Wang. Fov-gs: Foveated 3d gaussian splatting for dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1

[11] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2

[12] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[13] Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4D: voxel for 4d novel view synthesis. *IEEE Trans. Vis. Comput. Graph.*, 30(2):1579–1591, 2024. 2

[14] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. 2, 3, 4, 5

[15] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1060–1071, 2009. 3

[16] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6

[17] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 38(2):14:1–14:17, 2019. 2

[18] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20418–20431, 2024. 1, 2, 3

[19] Umar Iqbal, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, and Jan Kautz. Rana: Relightable articulated neural avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23142–23153, 2023. 3

[20] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 1, 2

[21] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 1, 2, 5, 6

[22] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 2, 3

[23] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 4

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[25] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 5, 6

[26] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 2

[27] Jason Lawrence, Szymon Rusinkiewicz, and Ravi Ramamoorthi. Efficient brdf importance sampling using a factored representation. *ACM Transactions on Graphics (ToG)*, 23(3):496–505, 2004. 3

[28] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5918–5928, 2019. 3

[29] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19876–19887, 2024. 1, 2

[30] Deqi Li, Shi-Sheng Huang, and Hua Huang. Mpgs: Multi-plane gaussian splatting for compact scenes rendering. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1

[31] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, 2022. 1, 2

[32] Wensheng Li, Lingzhe Zeng, Chengying Gao, and Ning Liu. Efficient integration of neural representations for dynamic humans. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[33] Zhe Li, Yipengjing Sun, Zerong Zheng, Lizhen Wang, Shengping Zhang, and Yebin Liu. Animatable and relightable gaussians for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096*, 2023. 2, 3

[34] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19711–19722, 2024. 1, 2, 3

[35] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 3

[36] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Relightable and animatable neural avatars from videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3486–3494, 2024. 3, 4

[37] 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 transactions on graphics (TOG)*, 40(6):1–16, 2021. 2

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

[39] Abhimitra Meka, Mohammad Shafiei, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. Real-time global illumination decomposition of videos. *ACM Transactions on Graphics (ToG)*, 40(3):1–16, 2021. 3

[40] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2

[41] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 1, 2

[42] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 788–798, 2024. 1, 2

[43] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 1, 2

[44] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6):238:1–238:12, 2021. 2

[45] Pramish Paudel, Anubhav Khanal, Danda Pani Paudel, Jyoti Tandukar, and Ajad Chhatkuli. ihuman: Instant animatable digital humans from monocular videos. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024. 2, 3

[46] Bo Peng, Jun Hu, Jingtao Zhou, Xuan Gao, and Juyong Zhang. Intrinsicngp: Intrinsic coordinate based hash encoding for human nerf. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):5679–5692, 2023. 2

[47] 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, 2, 5, 6

[48] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1, 2

[49] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 2

[50] Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, and Andreas Geiger. On joint estimation of pose,

geometry and svbrdf from a handheld scanner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3493–3503, 2020. 3

[51] Mingwen Shao, Yuanjian Qiao, Kai Zhang, and Lingzhuang Meng. Frequency-aware uncertainty gaussian splatting for dynamic scene reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1

[52] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jian Zhang, Bin Zhou, Errui Ding, and Jingdong Wang. Gir: 3d gaussian inverse rendering for relightable scene factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[54] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2, 3

[55] Hanxiao Sun, Yupeng Gao, Jin Xie, Jian Yang, and Beibei Wang. Svg-ir: Spatially-varying gaussian splatting for inverse rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16143–16152, 2025. 3

[56] Haocheng Tang, Ruoke Yan, Xinhui Yin, Qi Zhang, Xinfeng Zhang, Siwei Ma, Wen Gao, and Chuanmin Jia. Hgc-avatar: Hierarchical gaussian compression for streamable dynamic 3d avatars. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8125–8134, 2025. 1, 2

[57] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. Scanning 3d full human bodies using kinects. *IEEE transactions on visualization and computer graphics*, 18(4):643–650, 2012. 2

[58] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 2

[59] Shaofei Wang, Bozidar Antic, Andreas Geiger, and Siyu Tang. Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1877–1888, 2024. 3, 7

[60] 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. 5

[61] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. in 2021 ieee. In *CVF International Conference on Computer Vision, ICCV*, pages 10–17, 2021. 3

[62] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 2

[63] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 2

[64] Tong Wu, Jia-Mu Sun, Yu-Kun Lai, and Lin Gao. De-nerf: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 3

[65] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6): 1–15, 2021. 2

[66] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41 (6):1–15, 2022. 2

[67] Junjin Xiao, Qing Zhang, Zhan Xu, and Wei-Shi Zheng. Neca: Neural customizable human avatar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20091–20101, 2024. 2, 3

[68] Wangze Xu, Yifan Zhan, Zhihang Zhong, and Xiao Sun. Sequential gaussian avatars with hierarchical motion context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13592–13603, 2025. 1, 2

[69] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 990–1000, 2024. 3

[70] Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinghua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Otmar Hilliges. Hsr: holistic 3d human-scene reconstruction from monocular videos. In *European Conference on Computer Vision*, pages 429–448. Springer, 2024. 2

[71] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 2

[72] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 6

[73] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based material estimation. In *European Conference on Computer Vision*, pages 700–716. Springer, 2022. 4

[74] Meng You and Junhui Hou. Decoupling dynamic monocular

videos for dynamic view synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[75] Youyi Zhan, Tianjia Shao, He Wang, Yin Yang, and Kun Zhou. Interactive rendering of relightable and animatable gaussian avatars. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2, 3, 5, 7

[76] Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26297–26307, 2025. 1, 2

[77] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 3

[78] 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 Transactions on Graphics (ToG)*, 40(6):1–18, 2021.

[79] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2022. 3

[80] Zetong Zhang, Manuel Kaufmann, Lixin Xue, Jie Song, and Martin R Oswald. Odhsr: Online dense 3d reconstruction of humans and scenes from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21824–21835, 2025. 2

[81] Yiqun Zhao, Chenming Wu, Binbin Huang, Yihao Zhi, Chen Zhao, Jingdong Wang, and Shenghua Gao. Surfel-based gaussian inverse rendering for fast and relightable dynamic human reconstruction from monocular videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 3

[82] Yihao Zhi, Wanhu Sun, Jiahao Chang, Chongjie Ye, Wensen Feng, and Xiaoguang Han. Strugauavatar: Learning structured 3d gaussians for animatable avatars from monocular videos. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[83] Tiansong Zhou, Jing Huang, Tao Yu, Ruizhi Shao, and Kun Li. Hdhuman: High-quality human novel-view rendering from sparse views. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):5328–5338, 2023. 2

[84] Zuo-Liang Zhu, Jian Yang, and Beibei Wang. Gaussian splatting with discretized sdf for relightable assets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25155–25164, 2025. 3