

# Relightable and Animatable Neural Avatars from Videos

Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, Feng Xu

School of Software and BNRist, Tsinghua University

lwb20@mails.tsinghua.edu.cn, zhengcw18@gmail.com, yongjh@tsinghua.edu.cn, xufeng2003@gmail.com

## Abstract

Lightweight creation of 3D digital avatars is a highly desirable but challenging task. With only sparse videos of a person under unknown illumination, we propose a method to create relightable and animatable neural avatars, which can be used to synthesize photorealistic images of humans under novel viewpoints, body poses, and lighting. The key challenge here is to disentangle the geometry, material of the clothed body, and lighting, which becomes more difficult due to the complex geometry and shadow changes caused by body motions. To solve this ill-posed problem, we propose novel techniques to better model the geometry and shadow changes. For geometry change modeling, we propose an invertible deformation field, which helps to solve the inverse skinning problem and leads to better geometry quality. To model the spatial and temporal varying shading cues, we propose a pose-aware part-wise light visibility network to estimate light occlusion. Extensive experiments on synthetic and real datasets show that our approach reconstructs high-quality geometry and generates realistic shadows under different body poses. Code and data are available at <https://wenbin-lin.github.io/RelightableAvatar-page/>.

## 1 Introduction

Human digitizing has been rapidly developed in recent years, in which the reconstruction and animation of 3D clothed human avatars have many applications in telepresence, AR/VR, and virtual try-on. One important goal here is to render the human avatar in desired lighting environment with desired poses. Therefore, the human avatars need to be both relightable and animatable and achieve photorealistic rendering quality. Usually, the generation of these high-quality human avatars relies on high-quality data like the ones recorded by Light Stages (Debevec et al. 2000) which are complicated and expensive.

Recently, the emergence of Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) opens a new window to generate animatable and relightable 3D human avatars just from the daily recorded videos. NeRF-based methods have achieved remarkable success in 3D object representation and photorealistic rendering of both static and dynamic objects including human bodies (Peng et al. 2021b,a; Xu,

Alldieck, and Sminchisescu 2021; Weng et al. 2022; Jiang et al. 2022a,b; Wang et al. 2022; Peng et al. 2022; Yu et al. 2023; Su, Bagautdinov, and Rhodin 2023). Also, NeRF can be used for intrinsic decomposition to achieve impressive relighting results for static objects (Zhang et al. 2021; Yao et al. 2022; Boss et al. 2021a; Srinivasan et al. 2021; Boss et al. 2021b; Zhang et al. 2022; Jin et al. 2023). However, NeRF-based dynamic object relighting is rarely studied. One key challenge is that the dynamics cause dramatic changes in object shading, which is hard to model with the current NeRF techniques.

In this work, we propose to reconstruct both relightable and animatable 3D human avatars from sparse videos recorded under uncalibrated illuminations. To achieve this goal, we need to reconstruct the body geometry, material, and environmental light. The dynamic body geometry is modeled by a static geometry in a canonical space and the motion to deform it to the shape in the observation space of each frame. We propose an invertible neural deformation field that builds a bidirectional mapping between points of the canonical space and all observation spaces. With this bidirectional mapping, we can easily leverage the body mesh extracted in the canonical pose to better solve the inverse linear blend skinning problem, thus achieving high-quality geometry reconstruction. After the geometry reconstruction of all frames, we propose a light visibility estimation module to better model the dynamic self-occlusion effect for material and light reconstruction. We transfer the global pose-related visibility estimation task into multiple, part-wise, local ones, which dramatically simplifies the complexity of light visibility estimation. This model has good generalization capability with limited training data benefiting from the part-wise architecture, and thus successfully estimates the light visibility under various body poses and lighting conditions. Finally, we optimize the body material and lighting parameters, and then our method can render photorealistic images under any desired body pose, lighting, and viewpoint. In summary, the contributions include:

- the first method that is able to reconstruct both relightable and animatable human avatars with plausible shadow effects from sparse multiview videos,
- an invertible deformation field that better solves the inverse skinning problem, leading to accurate dense correspondence between different body poses,

- part-wise light visibility networks that better estimate pose and light-related shading cues with high generalization capability.

## 2 Related Work

### 2.1 Neural Human Avatars

In recent years, neural radiance fields (NeRF) (Mildenhall et al. 2020) have shown great abilities in photorealistic rendering. And many methods have successfully combined NeRF with human parametric models for human body reconstruction (Peng et al. 2021b; Weng et al. 2022) and animatable human body modeling (Wang et al. 2022; Jiang et al. 2022b; Zheng et al. 2022; Chen et al. 2021; Peng et al. 2021a, 2022; Jiang et al. 2022a; Yu et al. 2023) with sparse videos. For dynamic body motion modeling, people usually leverage linear blend skinning (LBS) (Lewis, Corder, and Fong 2000) to drive the body to different poses and use neural displacement fields to model the non-rigid deformations. Among these works, the deformation fields only model single-direction displacement, either forward deformation (canonical to observation) (Wang et al. 2022; Li et al. 2022) or backward deformation (observation to canonical) (Peng et al. 2021a; Chen et al. 2021; Peng et al. 2022). Different from them, our method proposes an invertible deformation field to solve the correspondence between canonical and observation space bidirectionally, which helps to better solve the inverse skinning problem, and leads to better geometry reconstruction. Recent work MonoHuman (Yu et al. 2023) also models bidirectional deformations, but unlike the compact single invertible network in our approach, they use two non-invertible neural networks to model the deformations separately. Additionally, these methods model body appearance using view-dependent color without decomposing it into lighting and reflectance. In contrast, our method enables relighting by reconstructing the environment lighting and the surface material.

### 2.2 Human Relighting

Some methods have been proposed to enable relighting of human images (Sun et al. 2019; Wang et al. 2020; Zhou et al. 2019; Kanamori and Endo 2018; Pandey et al. 2021; Ji et al. 2022). However, these image-based methods do not support changing the viewpoints and human poses. To further enable novel view relighting, 3D reconstruction techniques have been leveraged to model the human geometry (Guo et al. 2019). For video-based human relighting, Relighting4D (Chen and Liu 2022) enables free-viewpoint relighting from only human videos under unknown illuminations by using a set of neural fields of normal, occlusion, diffuse, and specular maps. But it is hard to relight the human with novel poses as it involves per-frame latent features which are not generalizable for novel poses. RANA (Iqbal et al. 2023) proposed a generalizable relightable articulated neural avatars creation method based on SMPL+D (Alldieck et al. 2018) model with albedo, normal map refinement techniques. But their method did not model specular reflection and cast shadows. In this paper, we present the first method that can reconstruct relightable and animatable

human avatars from videos under unknown illuminations, while providing physically correct shadows.

### 2.3 Invertible Neural Network

Invertible Neural Networks (INNs) (Dinh, Krueger, and Bengio 2015; Dinh, Sohl-Dickstein, and Bengio 2017; Behrmann et al. 2019; Chen et al. 2018; Kingma and Dhariwal 2018) are capable of performing invertible transformations between the input and output space. They are widely used in generative models like Normalizing Flows (Kobyzev, Prince, and Brubaker 2020) for density estimation. Moreover, the ability of INNs to maintain cycle consistency between two spaces makes them suitable for modeling the deformation field of 3D objects. As a result, INNs have been used for 3D shape completion (Niemeyer et al. 2019; Jiang et al. 2020; Paschalidou et al. 2021; Lei and Danilidis 2022), geometry processing (Yang et al. 2021), dynamic scenes reconstruction (Cai et al. 2022), and building animatable avatars with 3D scans (Kant et al. 2023). However, for video-based dynamic body deformation modeling, existing works only use non-invertible single-directional deformation. In this work, we leverage the invertibility of the INNs to model the dynamic body motions and reconstruct high-quality dynamic body geometry.

## 3 Method

Given multi-view videos of a user with some arbitrary motions, our goal is to reconstruct a relightable and animatable avatar of the user. The key challenge of this task is disentangling the geometry, material of the clothed body, and lighting, which is a highly ill-posed problem. To tackle this problem, we first reconstruct the body geometry from the input videos using the neural rendering techniques, where the geometry is modeled by a neural signed distance function (SDF) field and the dynamics of the human body are modeled with a rigid bone transformation of SMPL (Loper et al. 2015) model plusing an invertible neural deformation field (Sec.3.1, top left of Fig.1). Then, with the reconstructed geometry, we train a pose-aware part-wise light visibility estimation network, which is able to predict the light visibility of any query point under any light direction and body pose (Sec.3.2, bottom left of Fig.1). Finally, with the visibility information, we achieve the disentangling of the material of the human body and the illumination parameters (Sec.3.3, top right of Fig.1). Therefore, we can render a free-viewpoint video of the human with any target pose and illumination.

### 3.1 Geometry and Motion Reconstruction

Dynamic body deformation consists of articulated rigid motion and neural non-rigid deformation. Correspondingly, we propose a mesh-based inverse skinning method and an invertible neural deformation field to map points between the canonical and the observation space bidirectionally.

**Mesh-based Inverse Skinning.** The rigid motion is computed using linear blend skinning (LBS) algorithm (Lewis, Corder, and Fong 2000). For point  $\mathbf{x}_c$  in the canonical space, we use the bone transformation matrices  $\{\mathbf{B}_b\}_{b=1}^{24}$  of

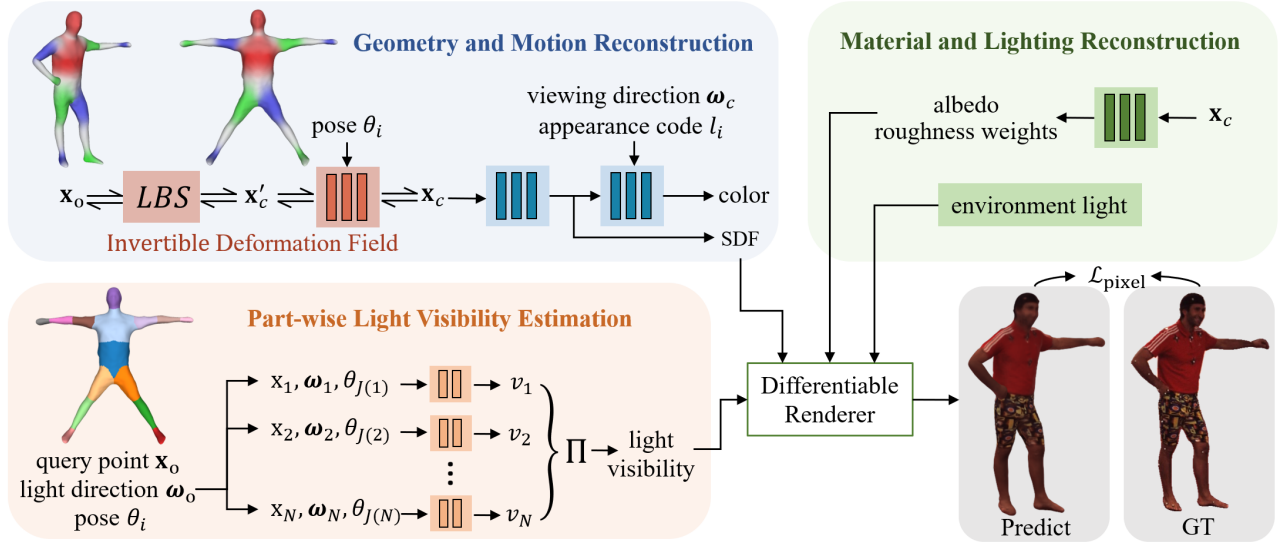


Figure 1: The pipeline of our method. The invertible deformation field in *Geometry and Motion Reconstruction* contributes to reconstruct more accurate dynamic body geometry (Sec.3.1). Then the networks in *Part-wise Light Visibility Estimation* are trained to estimate pose-aware light visibility in an effective manner (Sec.3.2). With these two parts fixed, the networks and lighting coefficients in *Material and Light Estimation* are trained and optimized by the photometric losses (Sec.3.3).

the SMPL (Loper et al. 2015) model to transform  $\mathbf{x}_c$  to  $\mathbf{x}_o$  in the observation space (we omit the transformations of homogeneous coordinates for simplicity of notation):

$$\mathbf{x}_o = \sum_{b=1}^{24} (w_b(\mathbf{x}_c) \mathbf{B}_b) \mathbf{x}_c \quad (1)$$

where  $w_b(\mathbf{x}_c)$  is the skinning weights of  $\mathbf{x}_c$ , and  $\sum_{b=1}^{24} w_b(\mathbf{x}_c) = 1$ . Similarly, for  $\mathbf{x}_o$  in the observation space, we can transform it back to the canonical space by:

$$\mathbf{x}_c = \sum_{b=1}^{24} (w_b(\mathbf{x}_o) \mathbf{B}_b)^{-1} \mathbf{x}_o \quad (2)$$

Similarly, for a query view direction  $\omega_o$  in the observation space, we can apply the same backward transformation to get the view direction  $\omega_c$  in the canonical space.

In volume rendering, we need to transform sampled ray points in the observation space to the canonical space (i.e. solve the inverse skinning problem) to query their SDF and color values. However, determining the skinning weights of points in the observation space is non-trivial as  $w_b$  is calculated in the canonical space rather than the observation space. Many existing works, such as (Peng et al. 2021a, 2022; Weng et al. 2022), rely on the posed SMPL mesh and use the skinning weights of neighboring SMPL mesh points to compute the inverse skinning weights of the ray points. However, the naked SMPL mesh differs from the body surface, resulting in inaccurate weights. Differently, we leverage the extracted explicit body mesh to compute the inverse skinning weights. We first extract an explicit mesh of the body in the canonical space and compute the skinning weights of mesh vertices. Then, we use the LBS algorithm to deform the mesh to the observation space. For any

points in the observation space, we compute their skinning weights by the skinning weights of the nearest neighbor on the deformed body mesh. As the deformed body mesh fits the actual body surface better than the naked SMPL mesh, our method does not suffer from the inaccuracies of skinning weights.

**Invertible Deformation Field.** Since only rigid bone transformation is not enough for modeling the body motion, we use an invertible neural displacement field to model the non-rigid motions. As shown in Fig.1, on the one hand, we apply non-rigid motion to the explicit mesh in canonical space. On the other hand, for sampled ray points in observation space, we need to map them back to the canonical space. Therefore, the neural displacement field should be able to transform points bidirectionally and ensure the cycle consistency of the transformation. So, we involve an invertible neural network to represent the non-rigid motion. For a point  $\mathbf{x} = [u, v, w]$  in the canonical space, we use the invertible network  $D$  to apply displacement to it:

$$\mathbf{x}' = D(\mathbf{x}) = [u', v', w'] \quad (3)$$

Besides, the invertible network  $D$  can also transform  $\mathbf{x}'$  back while keep the cycle consistency:

$$\mathbf{x} = D^{-1}(\mathbf{x}') = D^{-1}(D(\mathbf{x})) \quad (4)$$

To keep the cycle consistency, we design a network similar to Real-NVP (Dinh, Sohl-Dickstein, and Bengio 2017). Specifically, we split the coordinates  $[u, v, w]$  into two parts, for example,  $[u, v]$  and  $[w]$ . During forward deformation, we assume the displacement of  $[w]$  is decided by the value of  $[u, v]$ :

$$[w'] = [w] + f([u, v]) \quad (5)$$

and then the displacement of  $[u, v]$  is decided by  $[w']$ :

$$[u', v'] = [u, v] + g([w']) \quad (6)$$

With this two-step forward deformation  $D$ , we can directly get an invertible backward deformation  $D^{-1}$  which deforms point  $[u', v', w']$  to  $[u, v, w]$  as follows:

$$\begin{aligned} [u, v] &= [u', v'] - g([w']) \\ [w] &= [w'] - f([u, v]) \end{aligned} \quad (7)$$

The functions  $f(\cdot), g(\cdot)$  are implemented as MLPs, they form a transformation block of the invertible network  $D$ . As the aforementioned  $f(\cdot)$  makes the deformation decided by  $[u, v]$  only, we stack more transformation blocks and change the split of  $[u, v, w]$  in these blocks. We assume that the non-rigid deformations are pose-dependent, for the  $i$ th frame, we use the body pose  $\theta_i$  as the condition of the network  $D$ . Besides, we found that it is hard to learn the deformation using only the pose and coordinates as conditions. Thus, we use the skinning weights of the query points  $\mathbf{W}(\mathbf{x}) \in \mathbb{R}^{24}$  as an additional condition, which leads to better results. The displacement field  $D$  can be formulated as  $\mathbf{x}' = D(\mathbf{x}, \theta_i, \mathbf{W}(\mathbf{x}))$ . The use of skinning weights will slightly affect the cycle consistency of the deformation network, as  $\mathbf{W}(\mathbf{x})$  is not strictly equal to  $\mathbf{W}(\mathbf{x}')$ . But we found the skinning weights field in the canonical pose is smooth, and the deformations are relatively small, so they are almost the same, the sacrifice on cycle consistency is negligible.

**Network Training.** To supervise these neural fields with videos, we use the technique proposed in VolSDF (Yariv et al. 2021) to convert the SDF values to density and conduct volume rendering. For the color field, we introduce learnable per-frame appearance latent codes  $\{l_i\}_{i=1}^N$  to model the dynamic appearance, where  $N$  is the number of frames. Besides, we optimize the pose vectors  $\{\theta_i\}_{i=1}^N$ , as the initial poses may not be accurate. In sum, the training parameters contain the SDF network, the color network, the deformation network  $D$ , the appearance latent codes  $\{l_i\}_{i=1}^N$  and the pose vectors  $\{\theta_i\}_{i=1}^N$ . The training loss consists of rendering photometric loss and multiple regularizers:

$$\mathcal{L} = \lambda_{\text{pixel}} \mathcal{L}_{\text{pixel}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}} + \lambda_{\text{disp}} \mathcal{L}_{\text{disp}} \quad (8)$$

where  $\mathcal{L}_{\text{pixel}}$  is an L2 pixel loss for predicted color,  $\mathcal{L}_{\text{mask}}$  is a binary cross-entropy loss for the rendering object mask and input mask,  $\mathcal{L}_{\text{eik}}$  is the Eikonal regularization term (Gropp et al. 2020),  $\mathcal{L}_{\text{disp}}$  is an L2 regularizer for the output displacements. For more details about network architecture and training, please refer to the supplemental document.

### 3.2 Part-wise Light Visibility Estimation

With the reconstructed geometry, we then conduct pose-aware light visibility estimation. Modeling the visibility allows for the extraction or generation of shadows on images, which helps to better disentangle material and lighting from input images as well as produce physically plausible shadow effects in rendered images. Given a query point  $\mathbf{x}$  and a query light direction  $\omega$ , our goal is to train a network to predict whether the query point will be lighted or occluded by the body in a certain a pose and light direction.

Traditionally, estimating light visibility is solved by performing ray tracing. However, for implicit neural network-based methods, tracing a path of light requires numerous queries, as we need to trace all possible lighting directions for one 3D point, which is very time-consuming. Thus, existing methods (Zhang et al. 2021, 2022; Chen and Liu 2022) use MLPs to re-parameterize and speed up this process as  $V(\mathbf{x}, \omega) \mapsto v$ , where  $v = 1$  indicates the point is visible to the light from  $\omega$  direction. However, with the motion of the human body, light visibility changes dramatically. Relighting4D (Chen and Liu 2022) leverages temporally-varying latent codes to model these changes, but it is limited to seen poses as there is no latent code for unseen motions. To solve this problem, we need to make it pose-aware for light visibility estimation. A naive approach is to use the pose vectors as the condition of the visibility network, but we found this approach does not work well as the relationship among pose, lighting, and shadow is too complex to be modeled.

Our observation is that how light rays are blocked is determined by the object geometry, even though the human body as a whole can be in different complex shapes caused by pose changes, for a single body part, its geometry changes are relatively small among different poses. So, we divide the human body into  $N (= 15)$  parts as shown in the orange rectangle in Fig.1, where different colors denote different body parts. Then for each body part, we train a neural network respectively to predict how the body part blocks the lights. Finally, we combine the light visibility of all body parts by multiplying all the predicted visibility. Thus, our method achieves light visibility prediction of any query points, light directions, and body poses.

To be specific, given the query point  $\mathbf{x}_o$  and light direction  $\omega_o$  in observation space, we first transform them to the local coordinate of each body part:

$$\mathbf{x}_i = \mathbf{B}_i^{-1} \mathbf{x}_o, \quad \omega_i = \mathbf{B}_i^{-1} \omega_o \quad (9)$$

where  $\mathbf{B}_i$  is the bone transformation of the  $i$ th body part. Besides, although the geometry changes of body parts are relatively small, there are still some pose-dependent deformations. And the geometry of a body part is majorly affected by the poses of its neighboring joints, so we use them as the networks' condition. We denote the neighboring joints of body part  $i$  as  $J(i)$ . So, the visibility network of a body part  $i$  can be formulated as:

$$V_i(\mathbf{x}_i, \omega_i, \theta_{J(i)}) \mapsto v_i \quad (10)$$

For network training, we sample different query points, light directions, and body poses, then we perform ray tracing to compute the ground truth light visibility of each body part. We impose binary cross-entropy loss to train the networks for visibility estimation.

### 3.3 Material and Light Estimation

At this stage, we fix the geometry and light visibility estimation modules and optimize the material network and light parameters as shown in the green rectangle in Fig.1. Here, we parameterize the material using the Disney BRDF (Burley and Studios 2012) model and use albedo and roughness to represent the material. However, we found that directly

optimizing the roughness is difficult. Similar to (Hui and Sankaranarayanan 2017; Li and Li 2022; Yang et al. 2022), we use a weighted combination of specular bases. Each basis is defined by a different roughness value. For a query point in the canonical space, we use an implicit neural network  $M$  to predict its albedo value and roughness weights. For environment light, we parameterize it using  $L = 128$  spherical Gaussians (SGs) (Xu et al. 2013):

$$E(\omega_i) = \sum_{j=1}^L G(\omega_i; \xi_j, \lambda_j, \mu_j) \quad (11)$$

where  $\omega_i \in \mathbb{S}^2$  is the query lighting direction,  $\xi_j \in \mathbb{S}^2$  is the lobe axis,  $\lambda_j \in \mathbb{R}_+$  is the lobe sharpness,  $\mu_j \in \mathbb{R}^3$  is the lobe amplitude. To compute the visibility of an SG light, we sample 4 directions around the lobe axis based on the distribution defined by  $\lambda_j$ . Then we predict their visibilities by the light visibility estimation network and use the weighted sum of these samples as the visibility of the SG light.

With geometry, material, environment light, and light visibility, we can render images of the human body using a differentiable renderer. The rendering equation computes the outgoing radiance  $L_o$  at point  $\mathbf{x}$  viewed from  $\omega_o$ :

$$L_o(\mathbf{x}, \omega_o) = \int_{\Omega} L_i(\mathbf{x}, \omega_i) R(\mathbf{x}, \omega_i, \omega_o, \mathbf{n}) (\omega_i \cdot \mathbf{n}) d\omega_i \quad (12)$$

where  $L_i(\mathbf{x}, \omega_i)$  is the incident radiance of point  $\mathbf{x}$  from direction  $\omega_i$ , which is determined by the environment light  $E$  and masked by the light visibility.  $R(\mathbf{x}, \omega_i, \omega_o, \mathbf{n})$  is the Bidirectional Reflectance Distribution Function (BRDF) which is determined by the albedo values and roughness weights predicted by  $M$ .

To train the material network  $M$  and the light parameters of  $E$ , we use L1 pixel loss between the rendered images and the recorded images. However, there are strong ambiguities in solving material and lighting, so we apply some regularization strategies. First, the material network is designed as an encoder-decoder architecture following (Zhang et al. 2022), so that we can impose constraints on the latent space to ensure the sparsity of albedo and roughness weights. We denote the encoder and decoder of  $M$  as  $M_E$  and  $M_D$ , for a query point  $\mathbf{x}$  in the canonical space, its latent vector is  $\mathbf{z} = M_E(\mathbf{x}) \in \mathbb{R}^N$ . For  $K$  latent codes in a batch  $\{\mathbf{z}_i\}_{i=0}^K$ , we impose Kullback-Leibler divergence loss to encourage the sparsity of the latent space:

$$\mathcal{L}_{kl} = \sum_{j=1}^N \text{KL}(\rho || \hat{\rho}_j) \quad (13)$$

where  $\hat{\rho}_j$  is the average of the  $j$ th channel of  $\{\mathbf{z}_i\}_{i=0}^K$ .  $\rho$  is set to 0.05. Moreover, we apply smooth loss to both the latent vectors and the output albedo and roughness weights by adding perturbations:

$$\mathcal{L}_{\text{smooth}} = \lambda_z \|M_D(\mathbf{z}) - M_D(\mathbf{z} + \xi_z)\|_1 + \lambda_x \|M(\mathbf{x}) - M(\mathbf{x} + \xi_x)\|_1 \quad (14)$$

where  $\xi_z$  and  $\xi_x$  are the perturbations of the latent code  $\mathbf{z}$  and the query point  $\mathbf{x}$ , which is sampled from a Gaussian distribution with zero mean and 0.01 variance.

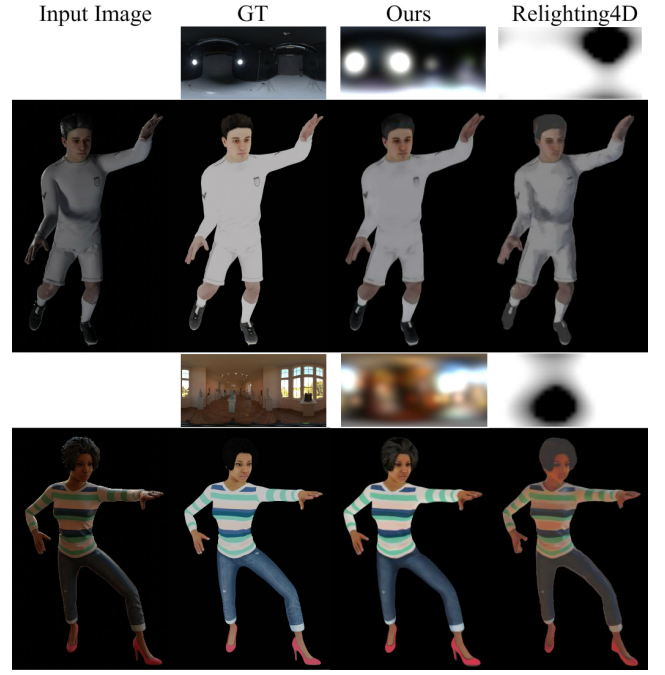


Figure 2: Qualitative comparison of the reconstructed albedo and lighting on synthetic data. Environment lighting is shown on top of the albedo in each result.

In sum, the full training loss for this stage is:

$$\mathcal{L} = \lambda_{\text{pixel}} \mathcal{L}_{\text{pixel}} + \lambda_{kl} \mathcal{L}_{kl} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} \quad (15)$$

With the trained geometry field, the deformation field, the light visibility estimation networks  $V$ , and the material network  $M$ , we can render the avatar in novel poses, lightings, and viewpoints. Thus we achieve a relightable and animatable neural avatar.

## 4 Experiments

In this section, we evaluate the performance of our method qualitatively and quantitatively. First, we introduce the used datasets. Then we compare our method with the state-of-the-art human relighting method Relighting4D (Chen and Liu 2022). Since our geometry reconstruction is improved by the proposed invertible deformation field, we also compare our method with the state-of-the-art video-based human geometry reconstruction methods ARAH (Wang et al. 2022) and Peng et al. (2022). Next, we perform ablation studies to validate our key design choices. Finally, we show the synthesized results on various characters with various body motions under various lightings, and video results can be seen in the supplemental video.

### 4.1 Datasets

We use both real and synthetic datasets for comparisons and evaluations. For the real dataset, we use multi-view dynamic human datasets including the ZJU-MoCap (Peng et al. 2021b), Human3.6M (Ionescu et al. 2014), DeepCap



| Method                           | Albedo Map      |                 |                    | Relighting (Training poses) |                 |                    | Relighting (Novel poses) |                 |                    |
|----------------------------------|-----------------|-----------------|--------------------|-----------------------------|-----------------|--------------------|--------------------------|-----------------|--------------------|
|                                  | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$             | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$          | SSIM $\uparrow$ | LPIPS $\downarrow$ |
| Relighting4D (Chen and Liu 2022) | 21.5103         | 0.8320          | 0.2299             | 19.7323                     | 0.7568          | 0.2721             | 16.7475                  | 0.6729          | 0.3330             |
| Ours w/o visibility              | 24.7611         | 0.8918          | 0.1655             | 23.7758                     | 0.8376          | 0.2223             | 18.9768                  | 0.7333          | 0.2638             |
| Ours w/o part-wise visibility    | <b>25.2150</b>  | <b>0.8921</b>   | 0.1652             | 24.7064                     | 0.8462          | 0.2173             | 19.7119                  | 0.7452          | 0.2580             |
| Ours                             | 25.1666         | 0.8919          | <b>0.1645</b>      | <b>25.3477</b>              | <b>0.8546</b>   | <b>0.2124</b>      | <b>19.8622</b>           | <b>0.7518</b>   | <b>0.2533</b>      |

Table 1: Quantitative comparison of the reconstructed albedo and the relighting results on synthetic data.

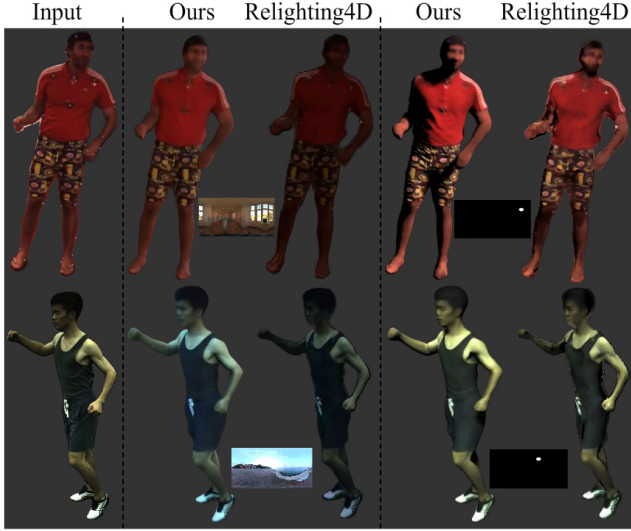


Figure 3: Qualitative comparison of relighting results on real data. The environment lighting of the rendered results is shown at the bottom.

(Habermann et al. 2020) and PeopleSnapshot (Alldieck et al. 2018) dataset. To perform a quantitative evaluation, we create a new synthetic dataset. We leverage 4 rigged characters from Mixamo<sup>1</sup> and transfer the body motion from the ZJU-MoCap dataset to generate motion sequences. Each sequence contains 100 frames. Then, we use Blender<sup>2</sup> to render multi-view videos under different illuminations with HDRI environment maps from PolyHaven<sup>3</sup>. Besides, we use 4 OLAT light sources for relighting evaluations.

## 4.2 Comparisons

Since Relighting4D (Chen and Liu 2022) is the state-of-the-art for video-based human motion relighting, we compare our full method with it on albedo estimation, lighting reconstruction, and relighting under training/novel poses. Body geometry is an intermediate result of our method, we also compare it with the state-of-the-art video-based human geometry reconstruction methods Peng et al. (2022) and ARAH (Wang et al. 2022).

**Material Estimation and Relighting.** The comparison results with Relighting4D (Chen and Liu 2022) are shown

in Fig.2. Relighting4D cannot disentangle the lighting and appearance very well, as we can see noticeable errors on both the estimated environment map and the reconstructed albedo. For example, there are shadows wrongly baked into albedo in the result on the top right side. Besides, some lighting information is backed into albedo in the result on the bottom right side. For numerical comparisons, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Measure (SSIM) (Wang et al. 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) as metrics. In Tab.1, we show the numerical albedo estimation result on synthetic data, which also indicates our improvement in albedo estimation. Tab.1 also shows the final relighting results for both training poses and novel poses, both show that we achieve noticeably better results than Relighting4D. Note that for novel poses, there are misalignments in the geometry between the animated geometry and the ground truth geometry, which leads to noticeable performance drops for novel poses. Besides, qualitative results for relighting on real datasets are shown in Fig.3 (please zoom in for better comparison). The overall lighting effect is better rendered by our method as shown on the left, and the spatial variant effect caused by point light is also correctly generated by our method as shown on the right due to the success of visibility modeling. More video results can be seen in our supplementary video. Notice that as Relighting4D relies on per-frame latent codes to model the dynamics, it does not support novel poses synthesis by design. So, when performing relighting for a novel pose, we find the closest pose in its training poses and use its latent code for inference.

**Geometry Reconstruction.** We evaluate different methods on our synthetic dataset with ground truth geometry and use point-to-surface distance (P2S) and Chamfer Distance (CD) as metrics. The results as shown in Tab.2, our method outperforms the compared two methods on all test sequences. We also show qualitative comparisons of rendering images in the real dataset in Fig.4. We can find that there are obvious artifacts in the results of Peng et al. (2022) and ARAH in the elbow and hand regions. While the result of our method does not suffer from the artifacts, as our mesh-based inverse skinning helps to find accurate correspondences between the observation space and the canonical space. In contrast, Peng et al. (2022) use posed SMPL models to compute the backward skinning weights which leads to worse correspondences, especially for regions with body contacts. ARAH involves iterative root-finding to compute the correspondences, but the optimization sometimes fails to converge, thus also leading to artifacts.

<sup>1</sup><https://www.mixamo.com/>

<sup>2</sup><https://www.blender.org/>

<sup>3</sup><https://polyhaven.com/>

|      | Method            | S1           | S2           | S3           | S4           | Avg          |
|------|-------------------|--------------|--------------|--------------|--------------|--------------|
| P2S↓ | Peng et al.       | 0.387        | 0.359        | 0.339        | 0.339        | 0.356        |
|      | ARAH              | 0.317        | 0.340        | 0.325        | 0.280        | 0.316        |
|      | Ours w/o MIS      | 0.241        | 0.230        | 0.234        | 0.241        | 0.237        |
|      | Ours w/o <b>W</b> | 0.246        | 0.247        | 0.243        | 0.256        | 0.248        |
|      | Ours              | <b>0.185</b> | <b>0.179</b> | <b>0.182</b> | <b>0.184</b> | <b>0.182</b> |
| CD↓  | Peng et al.       | 0.656        | 0.864        | 0.521        | 0.528        | 0.642        |
|      | ARAH              | 0.531        | 0.714        | 0.477        | 0.441        | 0.541        |
|      | Ours w/o MIS      | 0.462        | 0.666        | 0.423        | 0.391        | 0.485        |
|      | Ours w/o <b>W</b> | 0.479        | 0.700        | 0.427        | 0.424        | 0.507        |
|      | Ours              | <b>0.395</b> | <b>0.609</b> | <b>0.358</b> | <b>0.363</b> | <b>0.431</b> |

Table 2: Quantitative comparison of the reconstructed geometry on synthetic data.

### 4.3 Ablation Study

Here, we evaluate our two key components: mesh-based inverse skinning (MIS) and part-wise visibility estimation. The MIS based on the invertible deformation makes it possible to deform the more accurate mesh in the canonical space to the observation space to calculate skinning weights. Otherwise, the naked SMPL mesh with large geometry errors has to be used. So, we compare our method with using the SMPL mesh in weights calculation. Besides, using the part-wise design achieves accurate light visibility estimation, which is crucial to generate self-occlusion effects on bodies. To evaluate it, we compare it with two alternatives, removing the light visibility module and using only one neural network to predict the light visibility.

**Mesh-based Inverse Skinning.** As shown in Tab.2, the reconstruction errors without the mesh-based inverse skinning are consistently larger. We also show the qualitative result in Fig.4, using SMPL mesh to compute the skinning weights leads to artifacts in the contact regions.

Besides, we evaluate the effect of the condition of skinning weights **W** (in Sec. 3.1) in the invertible deformation network. As shown in Tab.2, we can also find that removing the condition of **W** also leads to worse results.

**Part-wise Visibility Estimation.** We show quantitative comparisons in Tab.1, the results show that although the albedo map reconstruction qualities are similar, our method with part-wise visibility estimation achieves the best results on relighting. Furthermore, We show qualitative results in Fig.5. We can see that without light visibility modeling (results in the fourth column), the self-occlusion effect cannot be generated at all. With the baseline light visibility modeling, self-occlusion can be partly generated for some poses (results of the third column). For our final solution, the received lighting for different body regions on the novel poses are well modeled and thus the relighting results are consistent with the ground truth rendering.

## 5 Limitations

The network training takes about 2.5 days in total on a single RTX 3090 GPU, and it takes about 40 seconds to render an image with a resolution of  $512 \times 512$  during inference (more details in the supplemental document). Integrating instant training techniques like Instant-NGP (Müller et al. 2022) may improve the efficiency of our technique. It is still hard

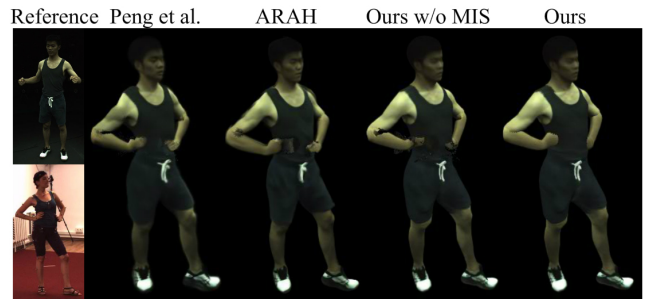


Figure 4: Qualitative results of novel poses synthesis on real data. This novel pose results reflect the accuracy of the reconstructed geometry to a certain extent.

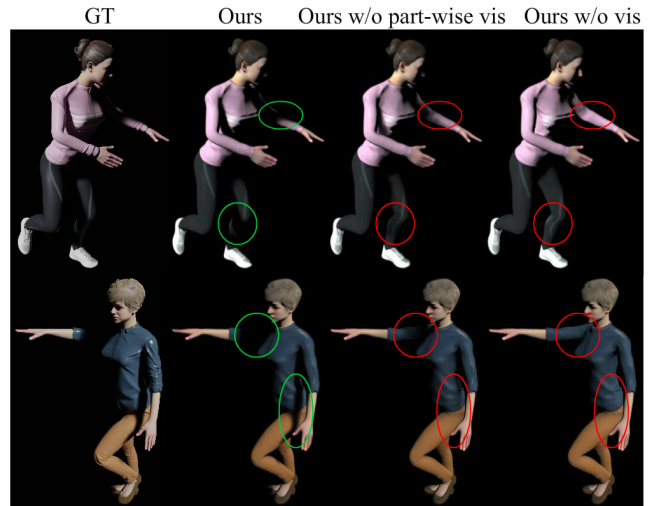


Figure 5: Ablation study on part-wise light visibility. See our method synthesizing plausible self-occlusions.

for our method to animate pose-dependent wrinkle deformations (especially for loose clothing) or generate global illumination effects, which are also open problems in this topic. Our method only considers the body motion rather than the face and hands, while recent works (Zheng et al. 2023; Shen et al. 2023) provide possibilities to handle them.

## 6 Conclusion

This is the first work that reconstructs relightable and animatable neural avatars with plausible shadow effects from sparse human videos. For dynamic body geometry modeling, the proposed invertible deformation field provides a novel and effective way to solve the inverse skinning problem. Besides, the part-wise light visibility modeling solves the problem of dynamic object relighting based on neural fields. Benefiting from the two techniques, our method succeeds in disentangling the geometry, material of the clothed body, and lighting, thus building a relightable and animatable neural avatar in a lightweight setting.

## Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFA0704000), the NSFC (No.62021002), and the Key Research and Development Project of Tibet Autonomous Region (XZ202101ZY0019G). This work was also supported by THUICS, Tsinghua University, and BLBCI, Beijing Municipal Education Commission. Jun-Hai Yong and Feng Xu are the corresponding authors.

## References

- Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018. Video based reconstruction of 3d people models. In *CVPR*, 8387–8397.
- Behrmann, J.; Grathwohl, W.; Chen, R. T.; Duvenaud, D.; and Jacobsen, J.-H. 2019. Invertible residual networks. In *International Conference on Machine Learning*, 573–582. PMLR.
- Boss, M.; Braun, R.; Jampani, V.; Barron, J. T.; Liu, C.; and Lensch, H. P. A. 2021a. NeRD: Neural Reflectance Decomposition from Image Collections. In *ICCV*, 12664–12674.
- Boss, M.; Jampani, V.; Braun, R.; Liu, C.; Barron, J. T.; and Lensch, H. P. A. 2021b. Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition. In *NeurIPS*, 10691–10704.
- Burley, B.; and Studios, W. D. A. 2012. Physically-based shading at disney. In *ACM SIGGRAPH*, volume 2012, 1–7. vol. 2012.
- Cai, H.; Feng, W.; Feng, X.; Wang, Y.; and Zhang, J. 2022. Neural Surface Reconstruction of Dynamic Scenes with Monocular RGB-D Camera. In *NeurIPS*.
- Chen, J.; Zhang, Y.; Kang, D.; Zhe, X.; Bao, L.; Jia, X.; and Lu, H. 2021. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*.
- Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.
- Chen, Z.; and Liu, Z. 2022. Relighting4D: Neural Relightable Human from Videos. In *ECCV*, volume 13674, 606–623.
- Debevec, P. E.; Hawkins, T.; Tchou, C.; Duiker, H.; Sarokin, W.; and Sagar, M. 2000. Acquiring the reflectance field of a human face. In *SIGGRAPH 2000*, 145–156. ACM.
- Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Non-linear Independent Components Estimation. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings*.
- Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using Real NVP. In *5th International Conference on Learning Representations, ICLR 2017*.
- Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit Geometric Regularization for Learning Shapes. In *International Conference on Machine Learning*, 3789–3799. PMLR.
- Guo, K.; Lincoln, P.; Davidson, P.; Busch, J.; Yu, X.; Whalen, M.; Harvey, G.; Orts-Escolano, S.; Pandey, R.; Dourgarian, J.; et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6): 1–19.
- Habermann, M.; Xu, W.; Zollhöfer, M.; Pons-Moll, G.; and Theobalt, C. 2020. DeepCap: Monocular Human Performance Capture Using Weak Supervision. In *CVPR*, 5051–5062.
- Hui, Z.; and Sankaranarayanan, A. C. 2017. Shape and Spatially-Varying Reflectance Estimation from Virtual Exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(10): 2060–2073.
- Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7): 1325–1339.
- Iqbal, U.; Caliskan, A.; Nagano, K.; Khamis, S.; Molchanov, P.; and Kautz, J. 2023. RANA: Relightable Articulated Neural Avatars. In *ICCV*, 23142–23153.
- Ji, C.; Yu, T.; Guo, K.; Liu, J.; and Liu, Y. 2022. Geometry-aware single-image full-body human relighting. In *ECCV*, 388–405.
- Jiang, B.; Hong, Y.; Bao, H.; and Zhang, J. 2022a. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *CVPR*, 5605–5615.
- Jiang, C.; Huang, J.; Tagliasacchi, A.; and Guibas, L. J. 2020. Shapeflow: Learnable deformation flows among 3d shapes. *Advances in Neural Information Processing Systems*, 33: 9745–9757.
- Jiang, W.; Yi, K. M.; Samei, G.; Tuzel, O.; and Ranjan, A. 2022b. NeuMan: Neural Human Radiance Field from a Single Video. In *ECCV*.
- Jin, H.; Liu, I.; Xu, P.; Zhang, X.; Han, S.; Bi, S.; Zhou, X.; Xu, Z.; and Su, H. 2023. TensorIR: Tensorial Inverse Rendering. In *CVPR*, 165–174.
- Kanamori, Y.; and Endo, Y. 2018. Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.*, 37(6): 270.
- Kant, Y.; Siarohin, A.; Guler, R. A.; Chai, M.; Ren, J.; Tulyakov, S.; and Gilitshenski, I. 2023. Invertible Neural Skinning. In *CVPR*, 8715–8725.
- Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.
- Kobyzev, I.; Prince, S. J.; and Brubaker, M. A. 2020. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 3964–3979.
- Lei, J.; and Daniilidis, K. 2022. Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In *CVPR*, 6624–6634.
- Lewis, J. P.; Corder, M.; and Fong, N. 2000. Pose space deformation: a unified approach to shape interpolation and



- skeleton-driven deformation. In *SIGGRAPH 2000*, 165–172. ACM.
- Li, J.; and Li, H. 2022. Neural Reflectance for Shape Recovery with Shadow Handling. In *CVPR*, 16200–16209.
- Li, R.; Tanke, J.; Vo, M.; Zollhöfer, M.; Gall, J.; Kanazawa, A.; and Lassner, C. 2022. Tava: Template-free animatable volumetric actors. In *ECCV*, 419–436.
- Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.
- Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.
- Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2019. Occupancy flow: 4d reconstruction by learning particle dynamics. In *ICCV*, 5379–5389.
- Pandey, R.; Escolano, S. O.; Legendre, C.; Haene, C.; Bouaziz, S.; Rhemann, C.; Debevec, P.; and Fanello, S. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4): 1–21.
- Paschalidou, D.; Katharopoulos, A.; Geiger, A.; and Fidler, S. 2021. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *CVPR*, 3204–3215.
- Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; and Bao, H. 2021a. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *ICCV*, 14294–14303.
- Peng, S.; Zhang, S.; Xu, Z.; Geng, C.; Jiang, B.; Bao, H.; and Zhou, X. 2022. Animatable Neural Implicit Surfaces for Creating Avatars from Videos. *arXiv preprint arXiv:2203.08133*.
- Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021b. Neural Body: Implicit Neural Representations With Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*, 9054–9063.
- Shen, K.; Guo, C.; Kaufmann, M.; Zarate, J.; Valentin, J.; Song, J.; and Hilliges, O. 2023. X-Avatar: Expressive Human Avatars. In *CVPR*, 16911–16921.
- Srinivasan, P. P.; Deng, B.; Zhang, X.; Tancik, M.; Mildenhall, B.; and Barron, J. T. 2021. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *CVPR*, 7495–7504.
- Su, S.-Y.; Bagautdinov, T.; and Rhodin, H. 2023. NPC: Neural Point Characters from Video. In *ICCV*, 14795–14805.
- Sun, T.; Barron, J. T.; Tsai, Y.-T.; Xu, Z.; Yu, X.; Fyfe, G.; Rhemann, C.; Busch, J.; Debevec, P. E.; and Ramamoorthi, R. 2019. Single image portrait relighting. *ACM Trans. Graph.*, 38(4): 79–1.
- Wang, S.; Schwarz, K.; Geiger, A.; and Tang, S. 2022. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *ECCV*.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.
- Wang, Z.; Yu, X.; Lu, M.; Wang, Q.; Qian, C.; and Xu, F. 2020. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39(6): 1–13.
- Weng, C.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. In *CVPR*, 16189–16199.
- Xu, H.; Alldieck, T.; and Sminchisescu, C. 2021. H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion. In *NeurIPS*, 14955–14966.
- Xu, K.; Sun, W.-L.; Dong, Z.; Zhao, D.-Y.; Wu, R.-D.; and Hu, S.-M. 2013. Anisotropic Spherical Gaussians. *ACM Transactions on Graphics*, 32(6): 209:1–209:11.
- Yang, G.; Belongie, S.; Hariharan, B.; and Koltun, V. 2021. Geometry processing with neural fields. *Advances in Neural Information Processing Systems*, 34: 22483–22497.
- Yang, W.; Chen, G.; Chen, C.; Chen, Z.; and Wong, K.-Y. K. 2022. PS-NeRF: Neural Inverse Rendering for Multi-view Photometric Stereo. In *ECCV*.
- Yao, Y.; Zhang, J.; Liu, J.; Qu, Y.; Fang, T.; McKinnon, D.; Tsin, Y.; and Quan, L. 2022. NeILF: Neural Incident Light Field for Physically-based Material Estimation. In *ECCV*, volume 13691, 700–716.
- Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34: 4805–4815.
- Yu, Z.; Cheng, W.; Liu, X.; Wu, W.; and Lin, K.-Y. 2023. MonoHuman: Animatable Human Neural Field from Monocular Video. In *CVPR*, 16943–16953.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 586–595.
- Zhang, X.; Srinivasan, P. P.; Deng, B.; Debevec, P. E.; Freeman, W. T.; and Barron, J. T. 2021. NeRFactor: neural factorization of shape and reflectance under an unknown illumination. *ACM Trans. Graph.*, 40(6): 237:1–237:18.
- Zhang, Y.; Sun, J.; He, X.; Fu, H.; Jia, R.; and Zhou, X. 2022. Modeling Indirect Illumination for Inverse Rendering. In *CVPR*, 18622–18631.
- Zheng, Z.; Huang, H.; Yu, T.; Zhang, H.; Guo, Y.; and Liu, Y. 2022. Structured Local Radiance Fields for Human Avatar Modeling. In *CVPR*.
- Zheng, Z.; Zhao, X.; Zhang, H.; Liu, B.; and Liu, Y. 2023. AvatarRex: Real-time Expressive Full-body Avatars. *ACM Transactions on Graphics (TOG)*, 42(4).
- Zhou, H.; Hadap, S.; Sunkavalli, K.; and Jacobs, D. W. 2019. Deep single-image portrait relighting. In *ICCV*, 7194–7202.