HR Human: Modeling Human Avatars with Triangular Mesh and High-Resolution Textures from Videos

Qifeng Chen*
Zhejiang University
caf7419@gmail.com

Kai Huang*
Institute of Computing Technology, CAS
ky.huang.null@gmail.com

Yuchi Huo
Zhejiang University
huo.yuchi.sc@gmail.com

Qi Wang
Zhejiang University
wanina1995@gmail.com

Wenting Zheng
Zhejiang University
wtzheng@cad.zju.edu.cn

Rong Li
Zhejiang University
skillzero.lee@gmail.com

Rengan Xie
Zhejiang University
rgxie@zju.edu.cn

Abstract

Recently, implicit neural representation has been widely used to generate animatable human avatars. However, the materials and geometry of those representations are coupled in the neural network and hard to edit, which hinders their application in traditional graphics engines. We present a framework for acquiring human avatars that are attached with high-resolution physically-based material textures and triangular mesh from monocular video. Our method introduces a novel information fusion strategy to combine the information from the monocular video and synthesize virtual multi-view images to tackle the sparsity of the input view. We reconstruct humans as deformable neural implicit surfaces and extract triangle mesh in a well-behaved pose as the initial mesh of the next stage. In addition, we introduce an approach to correct the bias for the boundary and size of the coarse mesh extracted. Finally, we adapt prior knowledge of the latent diffusion model at super-resolution in multi-view to distill the decomposed texture. Experiments show that our approach outperforms previous representations in terms of high fidelity, and this explicit result supports deployment on common renderers.

Keywords: Human modeling, Rendering, Texture super-resolution

1. Introduction

Digital avatars have been widely used across various applications, such as in the metaverse and film production. However, producing a high-fidelity digital avatar equipped with complex attributes, including geometry, texture parameters, and material baking, requires complex pipelines and expensive equipment [44, 55, 54, 13, 46], which limits the use of ordinary creators. Recently, research on neural implicit representation [32, 35, 50, 60, 33] has shown impressive results in multi-view reconstruction. Advances in neural volume rendering have soon fueled various exciting works on recovering digital avatars. For the implicit animatable human reconstruction, recent works [36, 37, 27, 53, 10] have solved the challenging task of multi-view reconstruction without the supervision of 3D information and present the inherent challenges of rendering non-rigid bodies and skins under dynamic motion. At the same time, inspired by neural reflectance decomposition[66, 7], Relighting4D [11] and Relightavatar [59] have attempted to recover human avatars with decoupled geometry and materials with implicit representation. However, the implicit geometry and texture are hard to edit, and the texture produced by those methods suffers from low clarity. In addition, the digital avatars represented implicitly cannot be applied in traditional graphics engines, which hinders their application in various fields.

Obviously, explicit representations appeal to us. Nvdiffrec [34] is dedicated to reconstructing general static objects in explicit representation that can be deployed in traditional graphics engines with triangle meshes and corresponding spatially-varying materials properties. The physical differentiable rasterization renderer has natural advantages for learning surface texture and fast rendering. However, Nvdiffrec struggles to reconstruct geometry and texture from sparse views. Furthermore, it fails to reconstruct a human in motion from monocular video. And recently 3D Gaussian(point-based rendering)[23] has shown great potential in dynamic human[28]. Although a unity compatibility plugin has been released in the community[5], the obvious

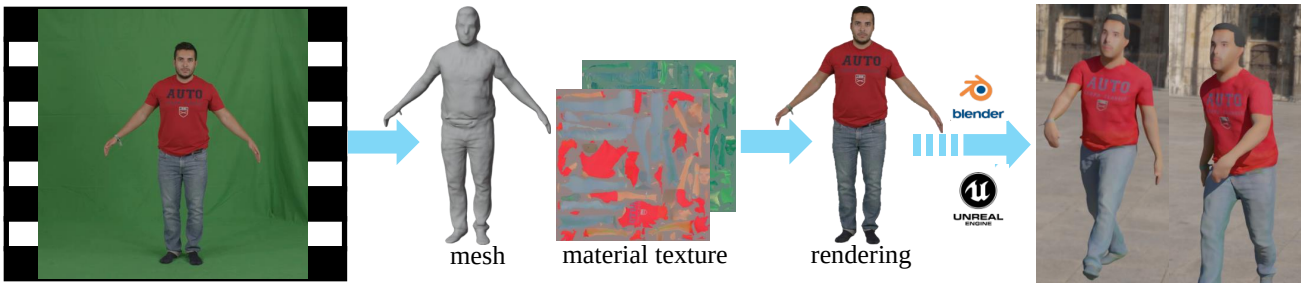

Figure 1: Given a monocular video of a performer, our method reconstructs a digital avatar equipped with high-quality triangular mesh and high-resolution corresponding PBR material textures. The result is compatible with standard graphics engines and can be edited.

software limitations and lack of editing are truly concerning. In contrast, our goal is to **reconstruct mesh digital avatars, taking as input the monocular video that records a human in motion, which is more compatible with traditional graphics engines and supports relighting and editing**.

In this work, we propose a novel framework for acquiring human avatars attached with high-resolution physically-based (PBR) material textures and triangular mesh from monocular video. We first reconstruct a human as a deformable neural implicit surface using volume rendering. In this process, the human information of frames in time sequence is integrated into the deformable neural implicit surface. From this, we extract the coarse triangle mesh and synthesize images of the human from dense virtual cameras. The synthesized images, alongside the input view, serve as supervisory data for subsequent training. We refer to this process as an information fusion strategy that combines the information from sequential video frames to compensate for the lack of spatial multi-view information. Then, we optimize the geometry mesh, material decomposition, and lighting using a differentiable rasterization renderer in the supervision of the multi-view images.In addition, We correct the bias for the boundary and size of the coarse mesh extracted from the implicit field. Finally, we introduce to adapt prior knowledge of the pretrained latent diffusion model [39] at super-resolution texture to distill the high-resolution decomposed texture. The LDM model has demonstrated superior performance and generalization compared to CNN pretrained network [9] in various vision tasks [39]. It fully fits our pipeline and enrichs texture space for details representation.

In summary, our main contributions are:

- We propose a novel framework that enables reconstructing a digital avatar equipped with triangular mesh and corresponding PBR material texture from monocular video. The digital avatars produced by our method are compatible with standard graphics engines.

- We propose an information fusion strategy to tackle the issue of lacking multi-view supervision in reconstructing explicit geometry and texture, which integrates information from all frames in the temporal sequence of video, transforming it into spatial supervision.

- We propose an approach to correct the bias for the triangular mesh and introduce the latent diffusion model to conduct distillation on super-resolution PBR texture. We result in a high-resolution texture and mesh with greater clarity.

2. Related Work

2.1. Scene Reconstruction

Recently, neural implicit representations have achieved impressive results in 3D reconstruction [32, 31, 58, 61, 15, 38]. These approaches represent a scene as a field of radiance and opacity, enabling the synthesis of photo-realistic novel viewpoints. However, directly using density-based methods for representation leads to numerous geometric artifacts. VolSDF [60] and Neus [50] proposed training the Signed Distance Function (SDF) field using volume rendering, which facilitates easy access to geometric surface normals. To further decouple the material properties, NeRD [7] is capable of learning geometry and spatially-varying Bidirectional Reflectance Distribution Functions parameters of objects from unconstrained environmental illumination. TensorIR [14] utilizes the low-rank tensors to simultaneously estimate the geometry and material of the scene. In addition, some methods integrate deep learning with texture-based techniques to model scenes [47, 43, 56]. However, the neural implicit representation cannot be applied in traditional graphics engines, which hinders their application in various fields. The recent trend is point-based rendering of 3DGS [23], and along with it comes a series of improvement work [28, 12]. However, point-based representation present challenges to editing and compatibility in complex graphic production. Therefore, mesh-based rendering still has its own unique advantages. DMTet [45] introduces a deformable tetrahedral

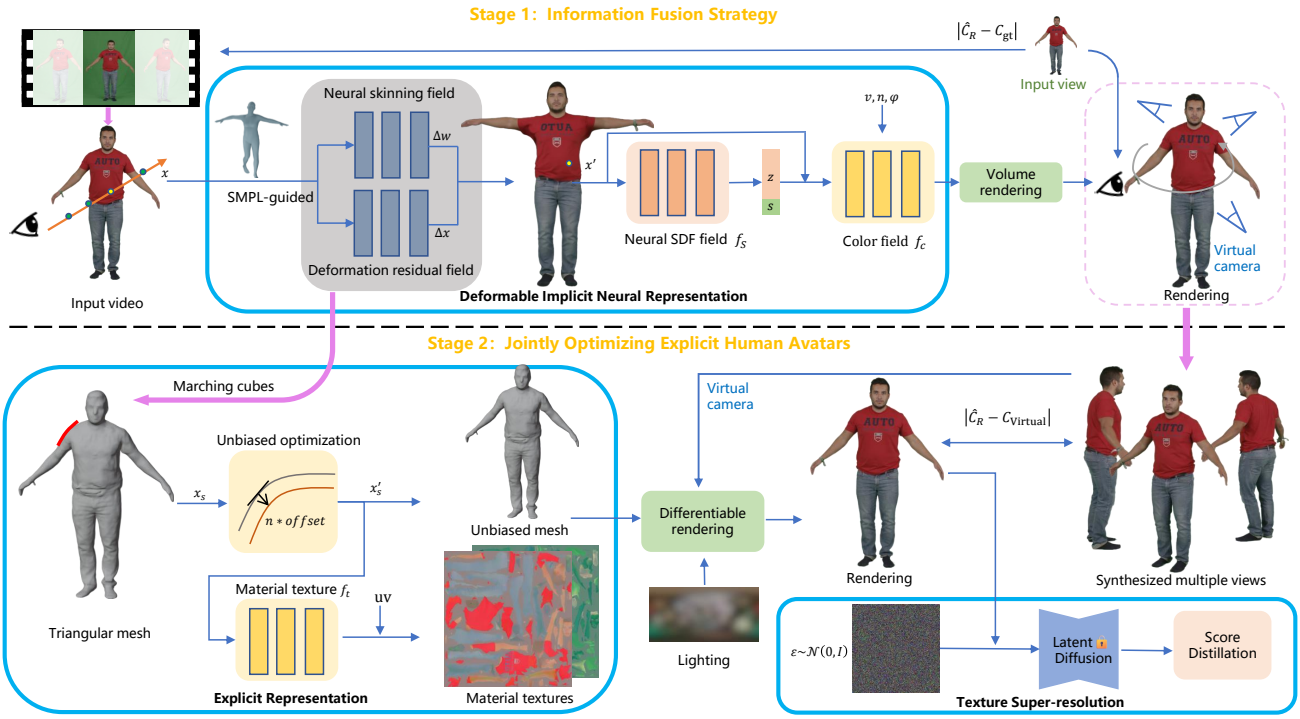

Figure 2: The overview of HR Human pipeline, which takes a video frame as input to reconstruct explicit avatars with triangular mesh and PBR texture. The pipeline includes deformable neural representations (used to extract volume surfaces and enhance spare input view), explicit representations (texture and geometry are jointly optimized), and super-resolution texture modules (introduced to generate high-resolution textures).

grid with learned mesh topology and vertex positions and utilizes coordinate-based networks to represent volumetric texturing. And Nvdiffrec [34] extends DMTet to 2D multi-view supervision in static scenes, jointly optimizing mesh and corresponding PBR Texture. Decomposing geometry and appearance from images contributes significantly to the progression of downstream tasks.

2.2. Image and Texture SR

Super-resolution is a commonly used approach to enhance detail expression in images or textures before being applied to downstream tasks of computer vision. The landscape of image super-resolution research encompasses a variety of influential works. The early super-resolution neural networks evolved from convolutional neural networks [6] to GAN [63], and later Transformer-based super-resolution [26] has achieved astonishing results. Super-resolution models are present in many application scenarios, such as mobisr [25], TexSR [9], NeRFSR [49]. Recently, diffusion models have also shown talent in the domain of super-resolution [39, 41], diffusion models achieve highly competitive performance on various super-resolution tasks, and exploring their practicality in graphics seems valuable. In this work, we introduce to adapt prior knowledge of the pretrained latent diffusion model [39] at super-resolution texture to distill the high-resolution decomposed texture.

2.3. Human Reconstruction

On the one hand, the single-image-based human body reconstruction method with implicit functions, such as PIFu [42] and ECON [57], have demonstrated promising outcomes. However, these approaches are not fully adaptable to video inputs nor capable of generating assets that include physically-based rendering (PBR) textures. On the other hand, advances in neural volume rendering have fueled various exciting works on recovering digital avatars. There have been numerous efforts to reconstruct humans in motion using multi-view videos [37, 27, 46, 16, 51, 10, 53]. Further, StylePeople [17] and NeuTex [56] introduce neural texture to restore complex texture features. And Relighting4D and Relightavatar [59, 11] have attempted to recover human avatars with decoupled geometry and materials implicit representation. Ani-GS[28] demonstrates the potential of 3DGS in dynamic human reconstruction. However, the methods mentioned above still have a long way to go before being directly edited and illuminated on traditional graphics engines.

3. Method

Figure 2 shows the overview of our method. The framework takes as input a **monocular video** of a human to reconstruct triangular mesh and corresponding high-resolution physically-based texture, which is compatible with traditional graphics engines and supports fast **editing and relighting**. To achieve this goal, we propose an information fusion strategy to combine human information from sequential video frames, resulting in coarse geometry mesh and multi-view synthesized images. Specifically, we first reconstruct humans as deformable neural implicit volume surfaces with the supervision of monocular video and extract corresponding high-quality triangle mesh. Then, we refine the human mesh and optimize the decomposition of corresponding materials texture using a differentiable rasterization renderer in the supervision of the dense and cross-view images that are synthesized by the first stage. Finally, to acquire texture with high fidelity, we adapt prior knowledge of the latent diffusion model at super-resolution in multi-view to distill the decomposed texture. Next, we will provide a more detailed introduction to the various parts of the method.

3.1. Information Fusion Strategy

For monocular video, there is only a single view in each frame, and the human exhibits different poses across different video frames. Existing methods [34, 45, 64], which designed for reconstructing static object from mult-view supervison, are unable to reconstruct high-quality trangular mesh and physical material texture of human body from a monocular video. Therefore, we propose an information fusion strategy to extract multi-view supervision from monocular video to augment the parse input view. Information fusion strategy is composed of two primary components: firstly, optimizing a deformable neural surface of human body from video which is aimed at fuse the sequential frame information of the human body in motion; and secondly, generating pseudo multi-view images from virtual viewpoints, which act as supervison for subsequent stages.

Reconstruct Deformable Neural Surface. Inspired by volsdf [60], we use a SDF-based neural network f_s to represent the human model in the canonical space:

$$f_s : (x') \rightarrow (s(x'), z(x')), \tag{1}$$

where $s(x')$ denotes the value of signed distance field (SDF) in canonical space, and $z(x') \in \mathbb{R}^{256}$ is a feature vector that represents implicit geometric information for further learning of appearance fields. The deformation from point x in pose space to point x' in canonical space can be divided into the sum of rigid deformation and non-rigid deformation:

$$x' = \hat{x} + D_i(\hat{x}, p(i)), \hat{x} = T_i(x) \tag{2}$$

where $D_i(\hat{x}, p(i))$ denotes non-rigid deformation field that is usually limited to finetune within a small range. And $\hat{x}$ is a preliminary result of $T_i(x)$, i.e., we add an offset to the result of rigid deformation in canonical space. $p(i)$ is the SMPL [29] pose parameter of the i-th frame. Specifically, $D_i(\hat{x}, p(i))$ is also implemented using a MLP $f_{Tn} : (\hat{x}, p(i)) \rightarrow \Delta x$. In addition, $T_i(x)$ denotes a motion field that maps point x in the pose space to canonical space. And points x in pose space can be projected to canonical space based on skinning weights:

$$T_i(x) = \sum_{j}^{K} w_i^j (R_i^j x + t_i^j), \tag{3}$$

where R_i^j and t_i^j denote the rotation and translation at each joint j, and w_i^j is the blend weight for the j-th joint. Following peng *et al*. [36], we use the initial blend weight from SMPL to guide the rigid deformation. So we calculate the w_i^j using the sum of the blend weight of SMPL and neural blend weight as:

$$w_i^j(x) = \Delta w + \hat{w}_i^j(x), \tag{4}$$

where $\hat{w}_i^j$ is the coarse blend weight, calculated based on the nearest points on the surface of SMPL. Δw is the deviation used to finetune blend weights, predicted by a MLP $f_{\Delta w}$: $(x, \psi_i) \rightarrow \Delta w$, and ψ_i is a latentcode of frame.

As a result, the underlying surface of the human in pose space or canonical space can be easily defined as a zero-level set of f_s:

$$S = \{x : f(x) = 0\} \tag{5}$$

Following volsdf [60], we optimize the implicit SDF field and color field of the human in canonical space using volume rendering end to end. We define the density σ and color c as:

$$\sigma(x') = \begin{cases} \alpha\Big(1 - \frac{1}{2} exp\Big(\frac{s(x')}{\beta}\Big)\Big) & \text{if } s(x') < 0, \\ \frac{1}{2}\alpha exp\Big(-\frac{s(x')}{\beta}\Big)\Big) & \text{if } s(x') \geq 0, \end{cases} \tag{6}$$

$$c_i(x) = f_c(x', n(x'), z(x'), v(x'), \psi_i), \tag{7}$$

where $\alpha, \beta > 0$ are learnable parameters and $s(x')$ is the signed distance value of point x'. The color field f_c takes as input the sample points x', normal $n(x')$, view direction $v(x')$, geometry feature code $z(x')$, latent code ψ_i and predicts the color of each point in canonical space. Then the expected color $C(r)$ of a pixel along ray r can be calculated using:

$$C(r) = \sum_{n=1}^{N} (\prod_{m=1}^{n-1} (1 - \alpha_m)\alpha_n c_n), \alpha_n = 1 - exp(-\sigma_n \delta_n), \tag{8}$$

where $\delta_n = t_{n+1} - t_n$ is the interval between sample n and $n + 1$, c_n is the color of sampled point along the ray.

Generating Pseudo Multi-view Images After successfully constructing the deformable neural surface, the information from sequential video frames is effectively integrated into the neural surface, which captures global body information and can be used to synthesize images of the human from any viewpoint under multiple poses. To train an overall texture of the explicit mesh of a human in the subsequent stage, we uniformly sample 50 viewpoints around the person, aiming to cover all observable surfaces of the human body as much as possible. On the other hand, to ensure that as many areas of the human body surface are observed as possible, we select training poses that are as stretched out as possible. Even though the viewpoints are set in a single well-behaved pose, after training to acquire the explicit mesh in the subsequent stage, this human body can be animated into any pose. Then, we utilize volumetric rendering to generate images from these viewpoints. The texture information from the video is encoded into those images and optimized into the overall texture of explicit mesh in the subsequent stage. Therefore, the information fusion strategy enables recovering explicit mesh and corresponding texture from monocular video.

3.2. Optimizing Explicit Human Avatars

As mentioned in Section 3.1, we can extract a mesh from the deformable implicit surface using the marching cubes algorithm [30]. However, the mesh obtained through marching cubes tends to be coarse due to the inherent bias of the signed distance field (SDF). To address this, we propose an unbiased optimization method to refine the mesh. We then jointly optimize the physically-based material texture and the mesh using inverse rendering.

Unbiased Optimization for Mesh. To create a triangle mesh from the neural SDF field, we first create a 256^3 resolution 3D grid with the same size as the pose bounding box. We only select points within a certain range of space around the bone joints to query the SDF value in the MLP network, which accelerates the time required for the Marching Cubes [30], to extract mesh in the 3D grid. We also use maximum pooling to discard small floating objects that may exist near the real human body surface. However, We observe that the neural implicit surfaces may converge in a biased range. Specifically, the Signed Distance Function (SDF) value of a well-defined surface often deviates from 0, such as ranging between 0.001 and 0.003, when the Marching Cubes algorithm is applied. This results in the extracted mesh not matching the human shape and being fatter than the real human body, which hinders texture optimization. We introduce a stable and easily trainable offset that works directly on the extracted explicit mesh. We further observe that the majority of the SDF bias consistently aligns with the normal of the human surface, as shown in Figure 4. Consequently, we constrain the offset along the normal direction of the vertex.

$$x_s' = x_s - f_0 \cdot n_s \tag{9}$$

where x_s is a set of biased vertices, f_0 are learnable parameters, n_s are normal vector direction of vertices. x_s' is the result vertices applied trainable offset. Specifically, we jointly optimize the bias in the early epochs of the second stage and remove it later.

Material Model. Inspired by Nvdiffrec [34], we represent the material properties of the human surface as a physically-based material model from Disney [8] and directly optimize it using differentiable rendering [24]. We use MLP to parameterize the decomposed material properties including a diffuse term k_d and an isotropic and specular GGX lobe [48]:

$$f_t : (x_s') \rightarrow (k_d, r, m), \tag{10}$$

where k_d denotes albedo color, r is roughness value, m is metalness factor. And specular color can be defined as:

$$k_s = (1 - m) \cdot 0.04 + m \cdot k_d, \tag{11}$$

Specifically, we parameterize the uv texture mapping for surface mesh using Xatlas [62] and sample the material model on the surface to create learnable 2D textures. This is beneficial for us to continue optimizing details with high-resolution textures and make edits to the texture (Section 3.3).

Physically-based Rendering. In our implementation, we follow the general rendering equation [22]:

$$L_o(x_s, \omega_o) = \int_\Omega f_r(x_s, \omega_i, \omega_o) L_i(x_s, \omega_i)(\omega_i \cdot n(x_s)) d\omega_i, \tag{12}$$

where ω_i is the incident direction, ω_o is the ougoing direction, x_s is a surface point of humans, $f_r(x_s, \omega_i, \omega_o)$ is the BRDF term, $L_i(x_s, \omega_i)$ is the incident radiance from direction ω_i, and the integration domain is the hemisphere Ω around the surface normal $n(x_s)$ of the intersection point.

In order to fast the performance of differentiable rendering, we use split sum approximation [19] for lighting representation. And the Equation (12) is approximated as:

$$\begin{aligned} L_o(x_s, \omega_o) = &\int_\Omega f_r(x_s, \omega_i, \omega_o)(\omega_i \cdot n(x_s)) d\omega_i \\ &* \int_\Omega L_i(x_s, \omega_i) D(\omega_i, \omega_o)(\omega_i \cdot n(x_s)) d\omega_i \end{aligned} \tag{13}$$

where D is a function representing the GGX [48] normal distribution (NDF), the first term represents the integral of the specular BSDF with a solid white environment light, and the second term represents the integral of the incoming radiance with the specular NDF. Both of them can be pre-integrated and represented by a filtered cubemap following Karis [1]. Further, we employ a differentiable version of the split sum shading model to optimize the lighting represented in a learnable trainable cubemap and the material properties.

3.3. Super-Resolution Texture.

Inspired by the impressive performance of the Latent diffusion model(LDM) [39] in the distillation task, we introduce it to help produce super-resolution texture with more detail. We first interpolate a coarse high-resolution 2D texture mapping (2048^2) from the low-resolution texture mapping(512^2) learned from RGB render loss. The coarse high-resolution 2D texture mapping and explicit mesh are utilized to render images R in each view using differentiable rendering. Then, the images R are fed into the LDM model as low-resolution input, and the pretrained super-resolution LDM [39] is used as a teacher model. Following the score distillation, images R is noised to a randomly drawn time step t,

$$R_t = \sqrt{\bar{\alpha}_t}R + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{14}$$

where $\epsilon \in N(0, I)$, and $\bar{\alpha}_t$ is a time-dependent constant specified by diffusion model. The score distillation loss will be calculated, and gradients will be propagated from the rendering pixel to learnable 2d texture.

$$\nabla_x L_{SDS} = w(t)(\epsilon_\phi(R_t, t, R) - \epsilon) \tag{15}$$

where ϵ_ϕ is the denoising U-Net of the diffusion model, $w(t)$ is a constant multiplier that depends on $\bar{\alpha}_t$.

In this manner, the prior knowledge learned from extensive common datasets within the LDM model is gradually distilled into human textures, resulting in a super-resolution texture.

4. Training

In the first stage, to optimize the information fusion strategy without 3D supervision, we march the ray from the camera at each frame and minimize the difference between the rendered color and the ground truth color. The loss function $\mathcal{L}_1$ is defined as:

$$\mathcal{L}_1 = \mathcal{L}_{color} + \mathcal{L}_{eik} + \mathcal{L}_{curv} + \mathcal{L}_{offset} + \mathcal{L}_w \tag{16}$$

where $\mathcal{L}_{color}$ is a L_1 loss between images. The $\mathcal{L}_{eik}$ and $\mathcal{L}_{curv}$ are the eikonal loss and curve loss applied to smooth the geometry. In addition, $\mathcal{L}_{offset}$ is a regularization term, which constrains non rigid deformation within a small range. $\mathcal{L}_w$ is a consistency regularization term to consistent the neural blend field.

In the second stage, we aim to refine the mesh extracted from the first stage and produce a high-fidelity decomposed PBR texture of humans.The loss function $\mathcal{L}_2$ consists of the following parts:

$$\mathcal{L}_2 = \mathcal{L}_{render} + \mathcal{L}_{bias} + \mathcal{L}_{SDS} + \mathcal{L}_{smooth} + \mathcal{L}_{light}, \tag{17}$$

where the $\mathcal{L}_{render}$ is the color loss of images. $\mathcal{L}_{bias}$ is applied to optimize the residual of biased surface. $\mathcal{L}_{light}$ is regularization term [34] designed to penalizes color shifts. $\mathcal{L}_{smooth}$ is a smooth term that smooths the texture in human surface points. Specially, $\mathcal{L}_{SDS}$ is the score distillation loss defined as Equation (15). Please refer to Section A and Section B in the appendix for more details.

5. Experiment

5.1. Datasets

We evaluate our method on multiple datasets, including real-world data and synthesized data.

Real-World Datasets. We validate our method on two real-world datasets, including ZJU-MoCap [37] and People-Snapshot [4]. ZJU-MoCap contains multiple dynamic human videos captured by a multi-camera system. People-Snapshot contains monocular videos recording humans in rotation. In addition, the approach [21] is applied to obtain the SMPL parameters within poses. We chose the most commonly used subjects to train the experiment model, including "ZJU313" and "ZJU377" from ZJU-CoMap, "M2C" and "M3C" from People-Snapshot with a monocular camera.

Synthesized Datasets. In order to more accurately evaluate and ablate the proposed method, we follow Renderpeople[3] to capture videos under a virtual monocular camera in the blender[2], including "Megan", "Josh", "Brain" and "Manuel". Each human in the synthesized dataset is equipped with the reference geometry mesh and material textures. Meanwhile, synthesized data are allowed to generate videos with more complex actions as input.

5.2. Baseline Comparisons

Comparison Methods We compare our method with other SOTA methods that focus on reconstructing the relighting human body from videos without 3D supervision, including Relighting4D [11], PhySG [64], SDF-PDF [36]. Relighting4d aims to decompose the surface material and geometry, as well as the environment lighting, from the videos. PhySG focuses on recovering the geometry and material properties of static objects from dense input views. Thus, we feed 120 multiview images sampled from the specific video frame as input. Even if SDF-PDF does not involve material decomposition, we also compared our method with it, which demonstrated good performance in reconstructing the dynamical human surface. Specifically, the works mentioned above reconstruct objects or humans in the neural implicit representation. Additionally, we compare our method with other popular approaches, such as ECON [57], InstantAvatar [20], and Ani-3DGS [28]. Further details can be found in the appendix.

Metrics Our main evaluation metrics for images include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM)[52], Learned Perceptual Image Patch Similarity (LPIPS)[65]. In addition, we follow [42] to

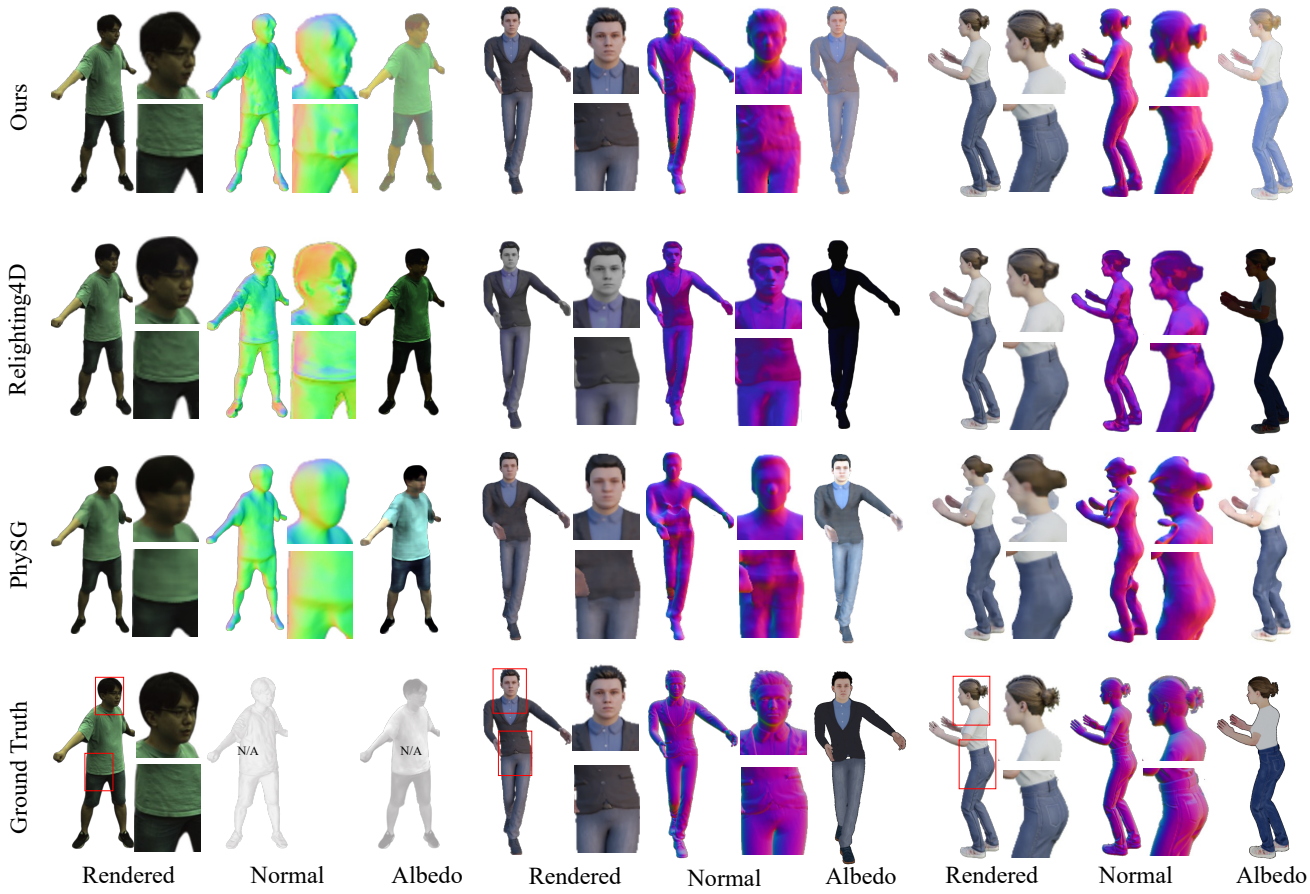

Figure 3: Qualitative comparison results of comparison methods, including albedo, geometric normal, and rendered image. Specifically, the real-world data from ZJU-MoCap only have the ground truth of the rendering result.

use 3D metrics, including Chamfer Distance(CD), Point-to-Surface Distance(P2S) and the angle degree difference between reference normals and predicted normals, which are applied to evaluate the quality of reconstructed geometry.

Comparison Results As shown in Figure 3, our method outperforms the state-of-the-art works both on geometry and color appearance. Our method recovers geometry that is smoother with accurate geometric details. In addition, our method recovers convincing and clear texture details, such as the human face and the accessories of clothes, which benefit from the correct geometry and the distillate knowledge from the pretrained LDM model. Table 1 reflects the stronger capability of our method in reconstructing human body geometry and texture materials in quantity. We show more results about materials and geometry Figure 6 and Figure 7. Meanwhile, to demonstrate the compatibility of our results in standard graphics engines, Figure 8 shows the results of **relighting, texture editing, novel poses synthesis**.

Method	PSNR↑	SSIM↑	LPIPS↓	Normal Degree°↓	CD (cm)↓	P2S (cm)↓
PhySG	18.67	0.722	0.284	49.308	-	-
SDF-PDF	24.99	0.919	0.096	32.712	0.76	0.70
Re4D	25.09	0.920	0.121	38.508	1.38	1.42
Ours	**27.08**	**0.941**	**0.027**	**24.401**	**0.70**	**0.55**

Table 1: Quantitative comparison results of various methods. The abbreviation "Re4D" means "Relighting4D". The result metrics are the average of all comparison results.

5.3. Ablation Study

We conduct ablation experiments from three aspects, including the effectiveness of the information fusion strategy, optimization of geometric bias, and super-resolution texture distillate. Below, we provide detailed quantitative and qualitative results.

Information Fusion Strategy. As previously mentioned, without an information fusion strategy, it would be impossible to directly train explicit mesh and PBR textures from monocular videos. Therefore, we take a step back to as-

sess the impact of a coarse mesh on the outcomes. Figure 4 shows that the performance of the differentiable renderer significantly deteriorates when the coarse mesh, which is extracted from neural implicit surfaces, is removed. There are a significant amount of self-intersecting triangles in the reconstructed mesh. In contrast, our method obtained a smooth and high-quality mesh and rendering. As shown in Table 2, our method performs better in both synthesized and real-world data. Furthermore, we conducted ablation experiments to evaluate the impact of the number of synthesized virtual viewpoints on the reconstruction results. For more details, please refer to the appendix.

DATA	METHOD	PSNR ↑	SSIM ↑	LPIPS ↓	Normal Degree°↓	CD (cm)↓	P2S (cm)↓
ZJU313	w/o Fusion	24.26	0.87	0.141	-	-	-
	w/ Fusion	30.68	0.96	0.028	-	-	-
Josh	w/o Fusion	24.80	0.91	0.081	100.25	0.78	0.58
	w/ Fusion	30.14	0.96	0.026	24.086	0.75	0.58

Table 2: Quantitative comparison for the effectiveness of information fusion strategy on real-world data "ZJU313" and synthesized data "Josh."

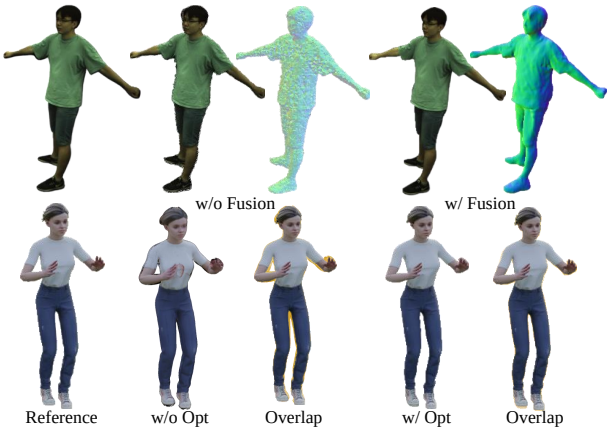

Figure 4: Qualitative comparison of the effectiveness of information fusion strategy and unbiased optimization. The highlight generated after mesh overlapping represents geometric bias.

Unbiased Optimization. The neural SDF method has always exited bias when representing zero-level surfaces. Figure 4 shows that the mesh extracted directly from the implicit SDF field does not match with the true contour, which will affect the learning of appearance, especially at the boundary. After unbiased optimization, the mesh aligns well with the real shape, yielding better color prediction results. Table 3 further reflects the improvement of rendering results by unbiased optimization.

Super-Resolution Texture. We design three ablation experiments to investigate the performance improvement of introducing texture super-resolution in the explicit 2D texture mapping space. Figure 5 shows introducing super-resolution textures reduces noise in the images and restores more details. In addition, we try to optimize neural implicit representation directly under the supervision of super-resolution images. However, Figure 5 shows that it will blur local details (such as the buttons). This is because the corresponding projection within the deformable field from pose space to canonical space is not stable, which results in the multiple points with different colors in pose space projected to a single point in canonical space and further results in the blurring of textures. We perform super-resolution in the explicit texture space, which ensures a stable correspondence between geometry and color and produces more clearer rendering result as shown in Table 3.

	PSNR↑	SSIM↑	LPIPS↓
w/o Unbiased Opt	23.71	0.891	0.100
w/ Unbiased Opt	24.60	0.910	0.043
NeRF SR	31.01	0.955	0.070
w/o SR	30.15	0.910	0.125
w/ SR	30.81	0.951	0.071

Table 3: Quantitative comparison for the effectiveness of unbiased optimization on synthesized data "Megan". And Quantitative comparison for the effectiveness of the super-resolution texture on synthesized data "Josh".

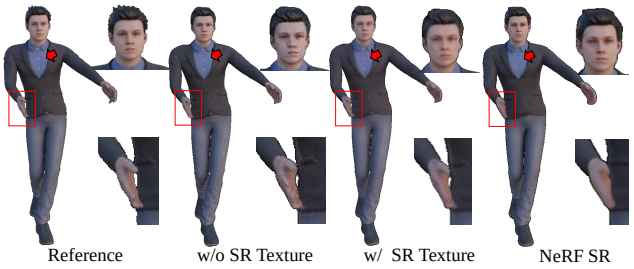

Figure 5: Qualitative comparison for the effectiveness of super-resolution. The GT and the rendering results for optimized textures at 512^2 resolution, optimized textures at 2048^2 resolution and implicit neural field are shown from left to right separately.

6. Conclusion

This paper proposes HR human, a novel framework that enables reconstructing a digital avatar equipped with triangular mesh and corresponding PBR material texture from a monocular camera. We introduce a novel information fusion strategy to combine the information from the monocular video and synthesize virtual multi-view images to compensate for the missing spatial view information. In addition, we correct the bias for the boundary and size of the mesh

extracted from the implicit field. Finally, we introduce a pre-trained latent diffusion model to distill the super-resolution texture when jointly optimizing the mesh and texture. The high-quality mesh and high-resolution texture produced by our method are compatible with common modern engines and 3D tools, which simplify the modeling process of digital avatars in various downstream applications and can be directly edited and reilluminated.

Acknowledgement

This work is supported by National Key R&D Program of China(No. 2024YDLN0011), NSFC(No. 62441205), and NSFC(No. U22B2034).

References

[1] Real shading in unreal engine 4. SIGGRAPH Course: Physically Based Shading in Theory and Practice, 2013. 5

[2] Blender. Blender, 2021. 6

[3] renderpeople. renderpeople, 2022. 6

[4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 6

[5] aras p. Gaussian splatting playground in unity, 2023. 1, 14

[6] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 3

[7] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 1, 2

[8] B. Burley and W. D. A. Studios. Physically-based shading at disney. In *Acm Siggraph*, volume 2012, pages 1–7. vol. 2012, 2012. 5

[9] B. Chaudhuri, N. Sarafianos, L. Shapiro, and T. Tung. Semi-supervised synthesis of high-resolution editable textures for 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7991–8000, 2021. 2, 3

[10] X. Chen, Y. Zheng, M. J. Black, O. Hilliges, and A. Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 1, 3

[11] Z. Chen and Z. Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. 1, 3, 6

[12] K. Cheng, X. Long, K. Yang, Y. Yao, W. Yin, Y. Ma, W. Wang, and X. Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. *arXiv preprint arXiv:2402.14650*, 2024. 2

[13] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 1

[14] S. Feng, B. Hou, H. Jin, W. Lin, J. Shao, R. Lai, Z. Ye, L. Zheng, C. H. Yu, Y. Yu, et al. Tensorir: An abstraction for automatic tensorized program optimization. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 804–817, 2023. 2

[15] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 2

[16] C. Geng, S. Peng, Z. Xu, H. Bao, and X. Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 3

[17] A. Grigorev, K. Iskakov, A. Ianina, R. Bashirov, I. Zakharkin, A. Vakhitov, and V. Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 3

[18] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 13

[19] S. Hill, S. McAuley, L. Belcour, W. Earl, N. Harrysson, S. Hillaire, N. Hoffman, L. Kerley, J. Patry, R. Pieké, et al. Physically based shading in theory and practice. In *ACM SIGGRAPH 2020 Courses*, pages 1–12. 2020. 5

[20] T. Jiang, X. Chen, J. Song, and O. 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. 6, 14, 15

[21] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 6

[22] J. T. Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 5

[23] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1, 2

[24] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 5

[25] R. Lee, S. I. Venieris, L. Dudziak, S. Bhattacharya, and N. D. Lane. Mobisr: Efficient on-device super-resolution through heterogeneous mobile processors. In *The 25th annual international conference on mobile computing and networking*, pages 1–16, 2019. 3

[26] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

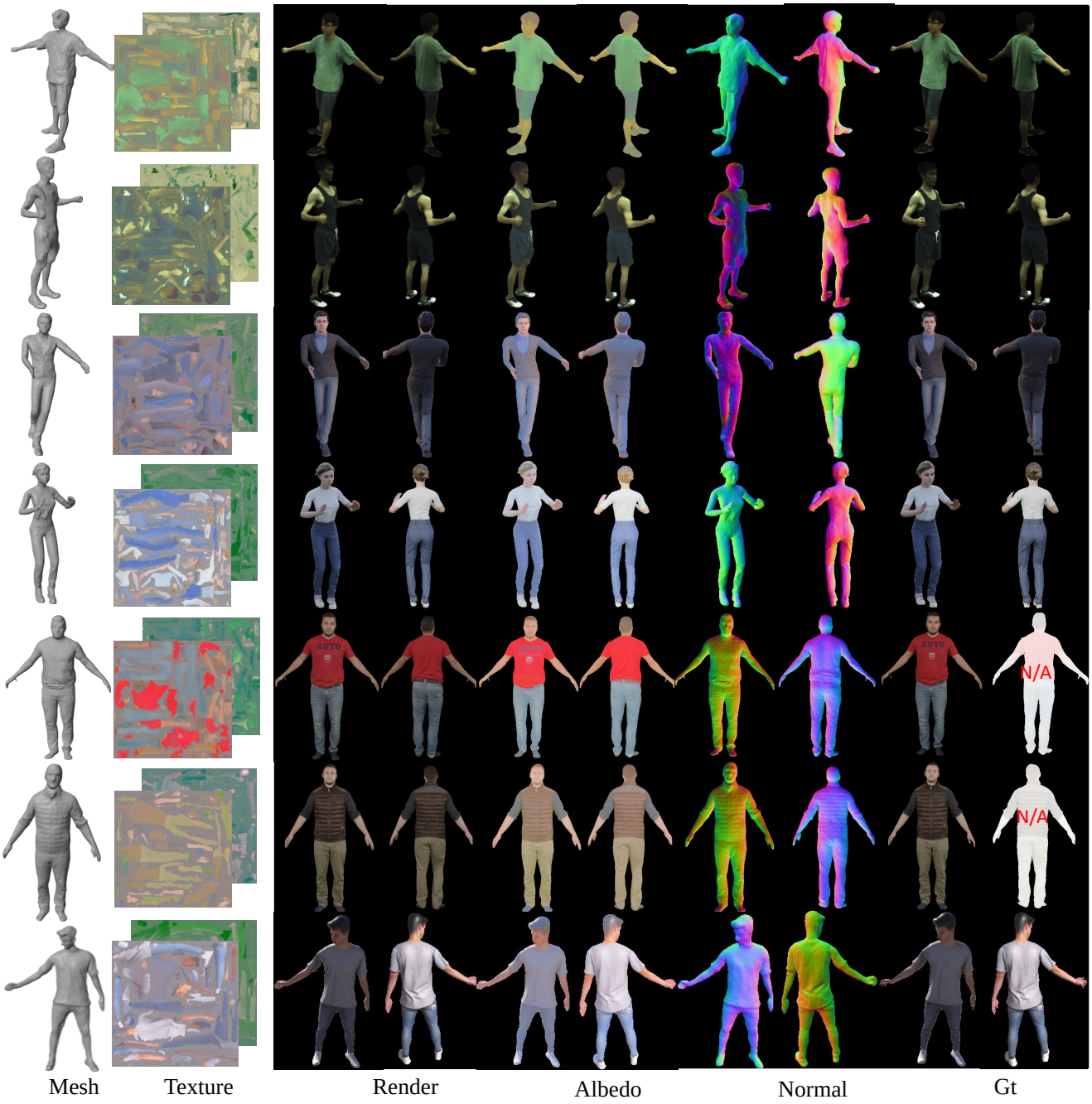

Figure 6: More high-quality reconstruction results of human body geometry and texture materials, including real-world dataset ZJU-MoCap and PeopleSnapshot, synthesized dataset Renderpeople. From top to bottom, they are "ZJU313", "ZJU377", "Josh", "Megan", "M3C", "M2C", "Manuel".

[27] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 1, 3

[28] Y. Liu, X. Huang, M. Qin, Q. Lin, and H. Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. *arXiv preprint arXiv:2311.16482*, 2023. 1, 2, 3, 6, 14, 15

[29] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

[30] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–

Figure 7: Reconstruction results of different frames in a dance video of “Brain”. From left to right, they are the input videos, the mesh corresponding to the pose, and the four surrounding rendering corresponding to the pose.

Figure 8: The rendering result of texture edit, relighting and novel poses synthesis, that work on graphics engines.

353. 1998. 5

[31] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[32] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[33] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 14

[34] J. Munkberg, J. Hasselgren, T. Shen, J. Gao, W. Chen, A. Evans, T. Müller, and S. Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, pages 8280–8290, 2022. 1, 3, 4, 5, 6, 13

[35] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[36] S. Peng, Z. Xu, J. Dong, Q. Wang, S. Zhang, Q. Shuai, H. Bao, and X. Zhou. Animatable implicit neural representations for creating realistic avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022. 1, 4, 6

[37] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 1, 3, 6

[38] A. Pumarola, E. Corona, G. Pons-Moll, and F. 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. 2

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 6

[40] R. A. Rosu and S. Behnke. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2023. 13

[41] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 3

[42] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 3, 6

[43] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5144–5153, 2017. 2

[44] S. Saito, J. Yang, Q. Ma, and M. J. Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 1

[45] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 2, 4

[46] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 1, 3

[47] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[48] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007. 5

[49] C. Wang, X. Wu, Y.-C. Guo, S.-H. Zhang, Y.-W. Tai, and S.-M. Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022. 3

[50] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2

[51] S. Wang, K. Schwarz, A. Geiger, and S. Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European conference on computer vision*, pages 1–19. Springer, 2022. 3

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[53] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. 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. 1, 3

[54] D. Xiang, T. Bagautdinov, T. Stuyck, F. Prada, J. Romero, W. Xu, S. Saito, J. Guo, B. Smith, T. Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 1

[55] D. Xiang, F. Prada, T. Bagautdinov, W. Xu, Y. Dong, H. Wen, J. Hodgins, and C. Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 1

[56] F. Xiang, Z. Xu, M. Hasan, Y. Hold-Geoffroy, K. Sunkavalli, and H. Su. Neutex: Neural texture mapping for volumetric neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7128, 2021. 2, 3

[57] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 3, 6, 14

[58] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2

[59] Z. Xu, S. Peng, C. Geng, L. Mou, Z. Yan, J. Sun, H. Bao, and X. Zhou. Relightable and animatable neural avatar from sparse-view video. *arXiv preprint arXiv:2308.07903*, 2023. 1, 3

[60] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2, 4

[61] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman. Multiview neural surface reconstruction

by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2

[62] J. Young. Xatlas:mesh parameterization / uv unwrapping library, 2022. 5

[63] K. Zhang, J. Liang, L. Van Gool, and R. Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 3

[64] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. 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. 4, 6

[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[66] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 1

Appendices

A. Training Strategy

We divide the complete training into two stages. In the first stage, we trained the deformable implicit neural representation based on volume rendering. We typically took 100k iterations for the first stage. Then, we use marching cubes to extract a mesh in a well-behaved pose from deformable implicit neural fields. Further, we take 15k iterations for the second stage of training. In the second stage, we aim to optimize the PBR material textures, lighting, and triangular mesh using a differentiable PBR-based render layer. Specifically, we first apply an unbiased optimization to adjust the coarse mesh extracted from the first stage($\sim$ 1k iters), resulting in a finetuned mesh aligned with the real human. Equipped with the finetuned mesh, we optimized the corresponding PBR texture under the supervision of sparse real camera views and dense synthesis views. After the coarse texture had converged($\sim$ 10k iters), we adapted prior knowledge of the latent diffusion model at super-resolution in multi-view rendering to distill the texture. In practice, the Adam optimizer was employed to optimize all networks and parameters. We set the learning rate to 5×10^{-4} with an exponential falloff during the optimization. The entire experiment was trained on an NVIDIA A100 GPU.

B. Loss Function

The definition of the loss functions mentioned in the main paper for the training of the first stage includes, $\mathcal{L}_{color}$ follows the $L1$ loss:

$$\mathcal{L}_{color} = \sum_{r \in R} \left\| \hat{C}_i(r) - C_i(r) \right\|_1 . \tag{18}$$

$\mathcal{L}_{eik}$ is the Eikonal term [18] encouraging f_s to approximate a signed distance function, and we set λ_1 as 0.1:

$$\mathcal{L}_{eik} = \lambda_1 \sum_{x'} (\|\nabla f_s(x')\|_2 - 1)^2, \tag{19}$$

$\mathcal{L}_{curv}$ is the curvature term [40] encouraging to recover smoother surfaces in reflective or untextured areas:

$$\mathcal{L}_{curv} = \lambda_2 \sum_{x'} (n \cdot n_\epsilon - 1)^2, \tag{20}$$

where $n = \nabla f_s(x')$ are the normal at the points x', $n_\epsilon = \nabla f_s(x'_\epsilon)$ are the normal at perturbed points x'_ϵ.The perturbed points x'_ϵ are sampled randomly in the tangent plane, $x'_\epsilon = x' + \epsilon(n \times \tau)$. τ is a random unit vector. And we set λ_2 as 0.65.

$\mathcal{L}_{offset}$ is the regularization term, which constrains the non-rigid deformation within a small range. We set λ_3 as 0.02, and the loss is defined as:

$$\mathcal{L}_{offset} = \lambda_3 \overline{\|\Delta x\|}_2 \tag{21}$$

In addition, we use a consistency regularization term $\mathcal{L}_w$ to minimize the difference between blend weights of the canonical and observation spaces, which are supposed to be the same. The loss is defined as:

$$\mathcal{L}_w = \sum_x \left\| w_i(x) - w_i^{can}(x') \right\|_1 , \tag{22}$$

The definition of the loss functions mentioned in the main paper for the training of the second stage includes, $\mathcal{L}_{render}$ also follows the $L1$ loss:

$$\mathcal{L}_{render} = \sum_{r \in R} \left\| \hat{C}_i(x_s) - C_i(r) \right\|_1 . \tag{23}$$

$\mathcal{L}_{mask}$ is applied in early epochs(such as 10), to estimate the residual of biased surface. It is defined as:

$$\mathcal{L}_{mask} = \sum_{r \in R} \left\| \hat{M}_i(x_s) - M_i(r) \right\|_2 . \tag{24}$$

where $\hat{M}_i(x_s)$ is the mask after rasterization.

$\mathcal{L}_{light}$ is regularization term [34] designed to penalizes color shifts. λ_4 is set as 0.005. Given the per-channel average intensities $\hat{c}_i$, we define it as:

$$L_{light} = \lambda_4 \frac{1}{3} \sum_{i=0}^{3} \left| \hat{c}_i - \frac{1}{3} \sum_{i=0}^{3} \hat{c}_i \right| \tag{25}$$

$\mathcal{L}_{smooth}$ is a smooth term that calculates texture differences between surface points x_s and its random displacement $x_s + \epsilon$. λ_5 is set as 0.002. And we define it as:

$$\mathcal{L}_{smooth} = \lambda_5 \sum_{x_s} |k_d(x_s) - k_d(x_s + \epsilon)| \tag{26}$$

$\mathcal{L}_{SDS}$ is defined as eq.15 of the main paper and is activated after the $\mathcal{L}_{render}$ converges.

C. More Results and Application

We present more results in the video on a standard graphics engine, including texture editing, relighting, novel pose synthesis, and novel view synthesis.

D. Additional Experiments

The effectiveness of the number of synthetic views used in information fusion strategy. We designed four controlled to find the most suitable number of synthetic views for the fusion training. Figure 9 shows the comparison of training results from different numbers of synthetic views in novel view. Table 4 shows that increasing the number of synthesized views is beneficial for learning more detailed textures because unknown surfaces are reduced. We usually choose 50 views as the training baseline to balance training efficiency and effectiveness.

VIEWS	PSNR	SSIM	LPIPS
2	19.71	0.794	0.136
10	22.54	0.853	0.110
50	23.13	0.910	0.057
100	23.27	0.921	0.045

Table 4: Quantitative comparison for the effectiveness of the number of synthesized views used in fusion strategy on "Megan".

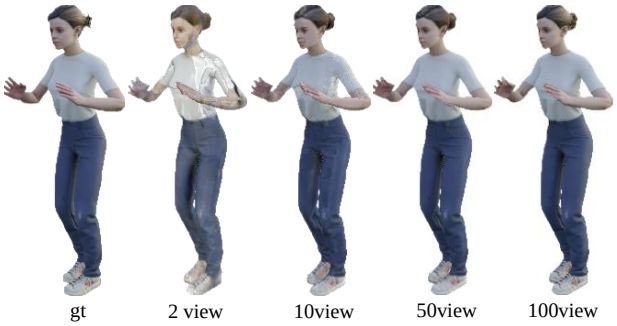

Figure 9: Qualitative comparison of the effectiveness of the number of synthesized views used in fusion strategy. From left to right, the number of training views is increasing.

Comparison with image-based explicit human reconstruction. We designed a comparative experiment with the SOTA method of the single-image-based model, ECON [57]. ECON is an image-based explicit human reconstruction method that combines implicit representation and explicit body regularization. Unlike ECON, which only reconstructs geometry, our approach further delivers triangular meshes and PBR textures, both of which are highly valued as 3D assets in the industry. As shown in Figure 10. Our method offers accurate full-body geometry, including details of the face, back, and legs.

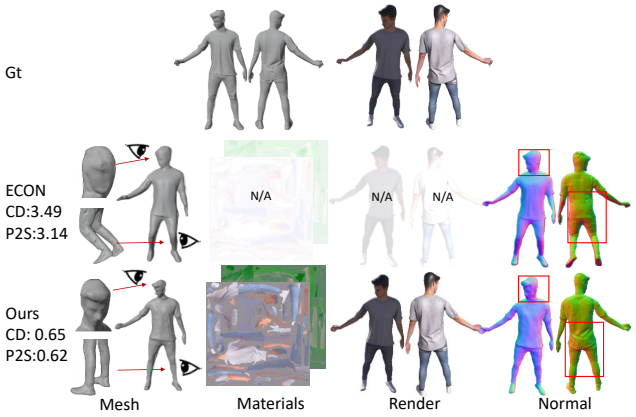

Figure 10: Comparison result with ECON [57] on RenderPeople dataset.

Comparison with fast human reconstruction methods. We compare our method with the state-of-the-art (SOTA) method that focuses on reconstruction and rendering acceleration. InstantAvatar [20] and Ani-3DGS [28]. InstantAvatar represents the human body as an Instant-NGP [33], and Ani-3DGS represents the human body as 3D Gaussian points. Both of these are based on surface point priors and achieve real-time performance. However, they all ignore the decoupling of PBR materials and high-precision geometry. Because of the deformation residual field and fine-tuning of mesh, our approach does not rely on surface initialization and pose. Thus, as reflected in the metrics, our approach has a more accurate surface and multi-view consistent textures in real-world data. Meanwhile, our approach further delivers triangular meshes and PBR textures, which support direct editing and relighting in a common graphics engine. Although a unityGS[5] compatibility plugin has been released in the community, the obvious software limitations and lack of editing and relighting are truly concerning. As shown in Figure 11 and Table 5, we have to admit that we have more train cost, but getting clearer, editable, and relightable textures and geometry are worth it.

	PSNR	SSIM	LPIPS	Train Cost	Output
InstantAvatar	26.61	0.930	0.121	$<$ 5min	implicit
Animatable GS	29.81	0.974	0.023	$<$ 30min	point+gaussian
Ours	32.40	0.971	0.017	$>$ 1h	mesh+texture

Table 5: Quantitative comparison of different rendering methods on the PeopleSnapshot dataset.

Figure 11: Comparison result with InstantAvatar [20] and Ani-3DGS [28] on PeopleSnapshot dataset.

E. Limitation

Our method still has the following limitations. Firstly, we model the human without distinguishing clothes and the human body. Thus, our method does not apply to humans wearing complex or loose clothing. In addition, our method still lacks competitiveness in terms of training costs, so we will consider introducing CUDA acceleration or another strategy. Finally, we choose to accelerate optimization without considering global lighting. Therefore, there is still room for improvement in the decoupling of our materials.