iHuman: Instant Animatable Digital Humans From Monocular Videos

Pramish Paudel[1], Anubhav Khanal[1,2], Danda Pani Paudel[2,3,4], Jyoti Tandukar[1], and Ajad Chhatkuli[2,3,4]

[1] Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal
[2] Nepal Applied Mathematics and Informatics Institute for research (NAAMII)
[3] ETH Zürich, Switzerland
[4] INSAIT, Sofia University, Bulgaria

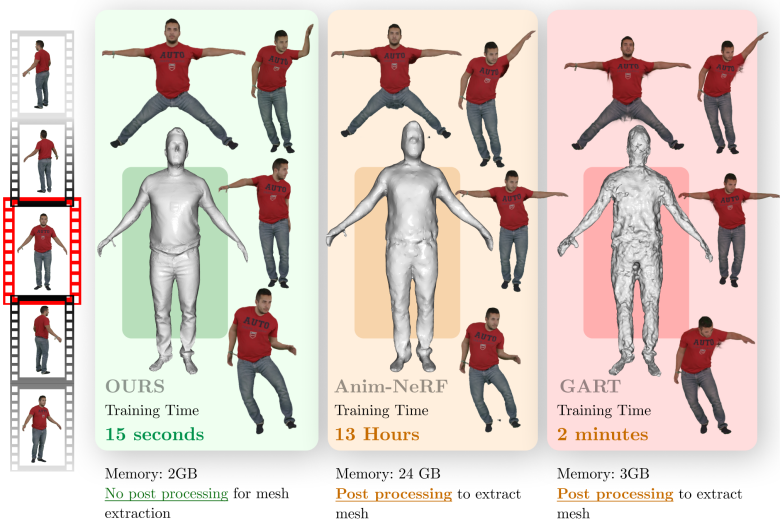

Fig. 1: Instant Digital Humans. Our method provides detailed and accurate 3D mesh and renderable Gaussian Splats instantly in 15 seconds of training time, from a monocular video. In contrast, the existing methods Anim-NeRF [58] and GART [33] provide lower quality mesh and rendered images, even after using more training time and compute. Input video (left) and rendered poses around the recovered meshes.

Abstract. Personalized 3D avatars require an animatable representation of digital humans. Doing so instantly from monocular videos offers scalability to broad class of users and wide-scale applications. In this paper, we present a fast, simple, yet effective method for creating animatable 3D digital humans from monocular videos. Our method utilizes the efficiency of Gaussian splatting to model both 3D geometry and appearance. However, we observed that naively optimizing Gaussian splats results in inaccurate geometry, thereby leading to poor animations.
This work achieves and illustrates the need of accurate 3D mesh-type modelling of the human body for animatable digitization through Gaussian splats. This is achieved by developing a novel pipeline that benefits

from three key aspects: (a) implicit modelling of surface's displacements and the color's spherical harmonics; (b) binding of 3D Gaussians to the respective triangular faces of the body template; (c) a novel technique to render normals followed by their auxiliary supervision. Our exhaustive experiments on three different benchmark datasets demonstrates the state-of-the-art results of our method, in limited time settings. In fact, our method is faster by an order of magnitude (in terms of training time) than its closest competitor. At the same time, we achieve superior rendering and 3D reconstruction performance under the change of poses. Our source code will be made publicly available.

Keywords: Digital Humans· Gaussian Splats · Surface reconstruction

1 Introduction

Instant and accurate creation of personalized 3D avatars is highly sought-after for digital human representation, to enable vast applications in virtual reality (VR), augmented reality (AR), gaming, and telepresence. A key component in this regard is the animatable representation [12, 19, 56]. On the other hand, reconstructing animatable digital humans instantly from monocular videos can immediately facilitate wide-scale applications serving a broad class users. Most of the existing monocular video based methods focus on either the real-time rendering solutions [1] (using long training/reconstruction time), or only mesh-level reconstruction [15] without the possibility of realistic re-rendering under the change in pose. These solutions eventually hinder the broad-scale applicability, which we aim to address in this paper by developing a novel method for instant and accurate modelling of animatable digital humans from monocular videos.

In recent year, building on the remarkable success of representing radiance fields implicitly (i.e. NeRF [45]), several methods [8–10,13,14,16,18,18, 21–23, 32, 32, 37, 40, 42, 48, 57, 58, 61, 64, 65, 69, 71, 73, 75, 76, 79, 82, 84, 87], have been developed to capture high-fidelity humans from multiple frames of videos. However, the high computational demand of the volume rendering in the NeRF-based methods [63, 77, 85] creates a major bottleneck for the aimed instant animatable digitization. Therefore, some very recent methods [33, 35, 36, 80] have been developed by leveraging the rendering efficiency of the Gaussian Splats [27]. However, these methods do not meet some or all of the required criteria in capturing (i) from monocular videos; (ii) in instant manner; (iii) animatable avatar; and (iv) high quality re-rendering under change in pose v) get mesh representation.

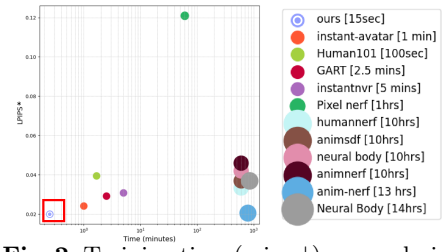

Fig. 2: Training time (mins ↓) vs. rendering (↓) comparison for different methods.

In this paper, we propose an efficient pipeline to convert a monocular video, with known pose, to animatable digital humans instantly (training time on par

with that of capture) in few seconds, using Gaussian splats based modelling. Doing so faithfully is particularly challenging primarily due to the difficulty of (i) inaccurate initialization of the Gaussian splats and (ii) ensuring the animatable nature of the output. Our proposed method addresses the challenges by introducing contributions on three aspects: (a) implicit modelling of surface's displacements and the color's spherical harmonics; (b) binding of 3D Gaussians to the respective triangular faces of the body template; (c) a novel technique to render normals followed by their auxiliary supervision. The proposed method is intuitive, simple, fast, yet effective. The qualitative and quantitative benefits of our method are highlighted in Figure 1 & 2, respectively.

The main contributions of this paper are:
– We propose a complete pipeline to obtain accurate 3D mesh bound Gaussian splats suitable for avatar animation, from monocular videos in instant manner.
– The proposed technical contributions involve; implicit representation, binding of gaussians to triangular faces, normal derivation for the auxiliary supervision.
– We conduct exhaustive experiments for comparisons, where our method achieves superior representation quality with an order of magnitude faster training speed.

2 Related Work

Mesh based reconstruction methods. Most methods that represent the human body as a mesh make use of SMPL [43] or other parametric body models [11,50, 55]. Methods in this category predict the parameters for the parametric body model either by regression [25,49,54] or by optimization [7]. Kolotouros et al. [31] and similar methods [39, 46] directly regress the 3D vertices. Although the output meshes here can be animated they do not contain the clothing details and personalized facial features. Methods which extend the parametric body model with a deformation layer [2–5] can model clothing as well but are unable to accurately model personalized geometric details.

Implicit functions based approaches. Implicit functions based reconstruction methods [44, 45, 52, 68, 78] use an MLP to learn an implicit function such as occupancy, signed distance fields or density fields to describe geometry. They can represent and render the geometric details of static scenes but suffer from high training time. Anim-NeRF [8] and other similar methods [9,10,16,18,19,21, 23,37,40,56,58,69,71,73] extend NERF to dynamic scenes by using SMPL [43] guided deformations between the observed space and a static canonical space allowing for explicit control. Instant Avatar [22] and similar approaches [24, 83] use [66] to speed up the training time but still have high memory requirements.

Gaussian Splat based approaches. Recently introduced 3D gaussian Splatting(3D-GS) [28] uses 3D Gaussians and its projections to represent a static scene. 3D-GS achieves significantly faster training and rendering time over NERF-based approaches. Recent works [20,30,33,36,38,47,51,60,86] extend 3D-GS to represent

dynamic scenes using SMPL guided deformations. They produce an animatable representation of the human body at faster speed compared to previous methods. However, the Gaussians obtained from the standard 3D-GS optimization process are unstructured and may not correspond to the surface, thus the Poission's reconstruction [26] does not result in accurate geometry. A 3D-GS extension, SuGaR [17] introduces an approach to align Gaussians to the surface geometry. Consequently, a mesh can be extracted using Poission's reconstruction [26]. Other methods [67], [59] use a mesh prior to initialize the 3D Gaussians to obtain well structured 3D Gaussians. Binding Gaussians to the mesh surface produces better geometry than approaches that only use 3D Gaussians but they still do not capture surface details. [61,74,75] use normal maps to capture high frequency details. However, [61, 74, 75] uses a single image to infer geometry, thereby resulting in relatively less accurate geometry. Our method uses Gaussians that are binded to the surface of a mesh as well as a novel normal guidance to produce animatable mesh of the human body while capturing body details.

3 Method

Provided a monocular video sequence with a dynamic human and the body poses, our goal is to generate a personalized colored mesh 3D model of a subject consisting of body shape, hair and clothing geometry, and underlying skeleton. Given an n frame video sequence $(I_t)_{t=1}^{n}$ of a single subject in front of a fixed camera (camera pose and intrinsics), along with the respective body poses $\{\theta_t\}$ we output a personalized animatable representation of the human subject. The keyword 'animatable' implies that we should be able to render the underlying representations in novel body poses $\{\theta_j\}$. Additionally, we want to complete the challenging training process in seconds, in favor of scalability.

We achieve our goal of obtaining an animatable 3D human using 3D Gaussian Splatting (3D-GS) [17, 27]. Below we explain the 3D-GS and its deformations as preliminaries in §3.1. We then introduce our iHuman representation and describe the details of our method. iHuman initializes 3D-GS in the canonical SMPL pose, see §3.2. We bind each 3D Gaussian to a triangle face as described in §3.3. We then proceed onto deforming the 3D-GS consistent to the posed space, corresponding to the real image in §3.4. Taking advantage of explicit 2D Gaussians embedded in 3D [17, 67], we encode normal for each Gaussian in §3.5.

3.1 Preliminaries

3D Gaussian Splatting. 3D Gaussian Splatting (3D-GS) has recently become the state-of-the-art tool for novel view synthesis. Important for our application, different from NERF, 3D-GS also uses explicit 3D representation using anisotropic 3D Gaussians.

A 3D Gaussian can be written in terms of its full 3D covariance matrix $\Sigma \in \mathbb{R}^{3\times3}, \Sigma \succeq 0$ and position in space $y \in \mathbb{R}^3$ along with center $\mathsf{x} \in \mathbb{R}^3$.

$$G(y) = \exp\left(-\frac{1}{2}(y-x)^\top \Sigma^{-1}(y-x)\right). \tag{1}$$

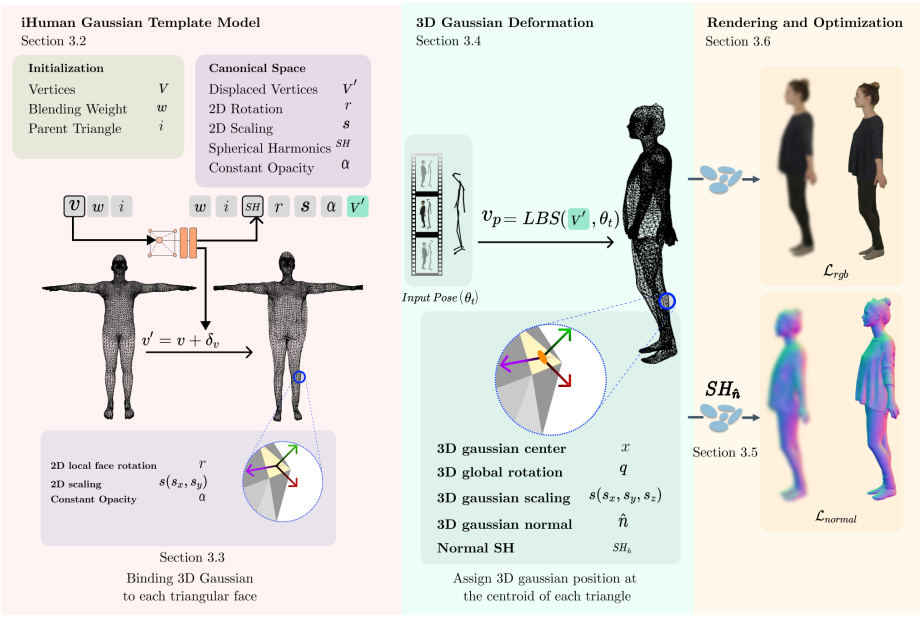

Fig. 3: Our method represents the human body in canonical space with gaussians parameterized by 3D gaussian centers x, rotations q, scales S, opacity α_o, colors SH, skinning weight w and its associated parent triangle i_x. It takes body pose θ_t of t^{th} frame as input and applies forward linear blend skinning to transform v' to posed space v_p. We compute gaussian center x from the posed space vertices v_p of i_x. The normal of parent triangle i_x is encoded to $SH_{\hat{n}}$ and rasterized to obtain the normal map $I_{\hat{n}}$. Then, we apply photometric loss and normal map loss to recover both geometry and color. The GT normal map $(\bar{I}_{\hat{n}})$ is obtained from monocular RGB image (I_n) using pix2pixHD [70] network.

Kerbl et al., [28] however represents each 3D Gaussian to be splatted by the Gaussian 3D center position $x \in \mathbb{R}^3$, color $c \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}$, its orientation parametrized by a 3D rotation written as a quaternion $q \in \mathbb{R}^4$, and anisotropic 3D scaling factor $s \in \mathbb{R}^3$ [28]. Instead of directly assigning a color value c, we use the Spherical Harmonics function, denoted as SH to model appearance on the projection ray originating from x. Thus the representation can be denoted as,

$$\mathcal{G} = \{x, q, S, \alpha_0, SH\}. \tag{2}$$

Note that Eq. (2) can be used with compact sets by approximating Gaussians with ellipsoids, thus simplifying both rendering and optimization [28]. Nonetheless, equivalence between the two exists where Σ can be written explicitly in terms of the scale s and orientation $R(q) \in SO_3$ as

$$\Sigma = RSS^\top R^\top. \tag{3}$$

In order to project the 3D Gaussians onto the camera for rendering, given a viewing transformation W, the covariance matrix Σ' in camera coordinates is

given as follows:

$$\Sigma' = \mathsf{J}W\Sigma W^\top \mathsf{J}^\top, \tag{4}$$

where, the Jacobian J is the affine approximation of the projective transformation.

Finally, the projected Gaussians are depth sorted and rasterized. During the process of rasterization, each Gaussian contributes to the pixel color through an opacity value α and coefficients $\mathbf{c}$ which encapsulate the color information. The resulting volumetric rendering equation that each Gaussian adds to the pixel is represented as follows:

$$C = \sum_{i \in N} c_i \alpha_i, \tag{5}$$

where, C is the final pixel color and N is the number of Gaussians (ellipsoids) projected on to the pixel.

3.2 Gaussian Human Template Model

Our iHuman approach uses a Gaussian template model on the canonical pose of the standard SMPL shape [43]. We denote this canonical mesh as $\mathcal{M}$ composed of vertices $V_{\mathcal{C}} = \{v_0, v_1, ..., v_m\}$ and triangles $F = \{i_x\}$, thus $\mathcal{M} = (V_{\mathcal{C}}, F)$. We then bind the Gaussians in 3D to the canonical mesh $\mathcal{M}$ as described in Eq. (2). This process is described in §3.3, where each Gaussian is tied to a specific face, *i.e.*, triangle i_x. We thus obtain the Gaussian Splat representation for the subject in the canonical SMPL pose by extending Eq. (2) as follows:

$$\mathcal{G}_{\text{skinned}} = \{x, q, S, \alpha, SH, w, \delta_x, i_x\}. \tag{6}$$

In contrast to Eq. (2), in Eq. (6), each Gaussian center x is in fact the centroid of a triangle i_x. Additionally, we introduce new parameters where w is the skinning weights obtained from the standard parametric body model [43]. Importantly, δ_v is the vertex displacement for the canonical shape vertex v to the clothed subject shape v':

$$v' = v + \delta_v. \tag{7}$$

The displacement vectors δ_v are obtained vertex-wise from a continuous Hash Encoder whose output is fed to a 3-layer MLP (multi-linear perceptron). We denote this as:

$$\delta_v = f_\delta\left(h(v)\right) \tag{8}$$

where f is a 3-layer MLP and $h(.)$ is the hash encoder similar to the instant NGP [66].

After the displacements, we bind the 3D Gaussians to the parent triangle i_x by centering it on the centroid of the face i_x at x. The rotation q and scale S are 2D rotation and scale as explained in 3.3. Each triangle center x for i_x is obtained as,

$$x = \frac{i_x[1] + i_x[2] + i_x[3]}{3}. \tag{9}$$

In order to obtain the new Gaussian representation of Eq. (6), we need to obtain the set of parameters $\{q, S, \alpha, SH, \delta_v\}$. Using Figure 3, we now explain how we obtain these quantities.

For each mesh face i_x, rotation q, scale values S and Gaussian opacity α are optimized as free variables. We further use optimization for the rotation q and SH coefficients for appearance rendering. Similarly, the SH coefficients are optimized as a function of $h(v)$.

3.3 Binding Gaussians To Mesh Surface

In the original works, the Gaussian splats are initialized on the point clouds output by a Structure-from-Motion (SfM) method such as COLMAP [62]. A recent work SuGaR [17] proposes using surface aligned Gaussians for the optimization, such that one of the axes of the Gaussian covariance Σ is aligned to the surface normal n_i, with the corresponding scale as 0. In iHuman, we have the advantage of mesh initialization through the SMPL canonical model. Therefore, we propose to align all Gaussians according to the following steps:

1. Compute the surface normal n_i for each face i_x.
2. For each Gaussian, assign n_i as one of the directions of its covariance matrix Σ, with the corresponding scale in S, *i.e.*, $S_3 = \epsilon$. In practice we keep $\epsilon = 1mm$, an extremely small value.
3. Assign the other two directions according to the major directions of the triangular face.

As a consequence of having a 2D Gaussian, we further reduce the learnable parameters required for obtaining the posed as well as canonical human mesh. For each Gaussian we can directly set $S_3 = \epsilon$ and the quaternion is reduced to a complex number of a single degree of freedom in order to keep the Gaussian aligned with the triangle.

3.4 3D Gaussian Deformation

Together with the Gaussian binding and the template model described in §3.2, §3.3, we are able to precisely represent a human surface in the canonical pose. In this section, we deform the Gaussian Splat model in order to represent any pose of a human subject. Given the input pose θ_t, we achieve the deformation using forward linear blend skinning [34].

Thus, we compute the transformation of each vertex v' in posed space with blend skinning $\omega(\theta_t)$. The transformation of each point v is calculated with blend skinning $\omega(\theta)$ and target bone transformation $B(\theta_t) = \{B_1(\theta_t), \ldots, B_{n_b}(\theta_t)\}$. The skinning weight field is defined as:

$$w(v_c) = \{w_1, \ldots, w_{n_b}\} \tag{10}$$

where v' is a point in canonical space and n_b is the number of bones.

Target bone transformations $B_t = \{B_t^1, ..., B_t^{n_b}\}$ in frame t can be calculated from the input poses and the corresponding skeleton as follows:

$$S_t, J, T_t \mapsto B_t, \tag{11}$$

where $S_t = \{\omega_t^1, ..., \omega_t^{n_b}\}$ refers to the rotation Euler Angle of each joint in frame t (world rotation for ω_t^1 and local rotation for the rest), T_t is the world translation in frame t, and $J = \{J_1, ..., J_{n_b}\}$ is the local position of each joint in canonical space. We transform a vertex in canonical space to posed space:

$$v_p = \sum_{i=1}^{n_b} w_i \cdot B_i^t \cdot v' \tag{12}$$

where v_p represents the direct mapping of v' in posed space. From v_p, we calculate the 3D Gaussian center x as given in Eq. (9) and 3D rotation q from the 2D rotation as explained in §3.3.

3.5 Normal Map from 3D Gaussian

Gaussian Splats are generally optimized using RGB photometric loss [27, 59]. However, we note that this approach results in a poor mesh, low on details. Our goal is to compute details of human surface, *e.g.*, facial attributes, wrinkles and hair [See Supp]. To that end, we take advantage of two crucial facts:

1. We have explicit representation of vertices(v_t) and faces (i_x) available for each Gaussian to obtain its normal without ambiguity from Eq. (13).
2. SOTA methods like ECON [61, 74, 75] rely on normal map prediction from RGB to produce SOTA results.

One can therefore use the depth gradient ∇Depth in order to compute the surface normals. However, such measurements tend to inherently noisy as it relies on the alpha blending of the gaussians which can introduce noise [see Supp]. On the other hand, the normal image $\bar{I}_{\hat{n}}$ should also equal to the aligned normals obtained from the posed vertices $\{v_p\}$. We first compute the mesh/Gaussian normals using Eq. (13).

$$\hat{n} = \frac{(v_p[i_x[1]] - v_p[i_x[0]]) \times (v_p[i_x[2]] - v_p[i_x[0]])}{\|(v_p[i_x[1]] - v_p[i_x[0]]) \times (v_p[i_x[2]] - v_p[i_x[0]])\|} \tag{13}$$

where $v_p[i_x[j]]$ refers to the j-th posed vertex of the triangle in the face i_x.

In order to obtain the normal map image from the estimated normals of Eq. (13), we again make use of the Gaussian splatting rasterizer. In order to preserve smoothness and accuracy, the Gaussian Splatting rasterizer already provides a highly efficient approach for normal map computation. For that purpose, we encode the normal $\hat{n}$ into a second spherical harmonics function $SH_{\hat{n}}$ of degree 0 by representing the components of the normal as: $\hat{n}_r$, $\hat{n}_g$ and $\hat{n}_b$ related

to the rgb values of the rasterizer. The Spherical harmonics again operates on the hash encoding of the vertices $\{v_p\}$ for efficiency and can be evaluated as:

$$SH_{\hat{n}} = \frac{1}{\sqrt{4\pi}} \begin{bmatrix} \hat{n}_r \\ \hat{n}_g \\ \hat{n}_b \end{bmatrix} \tag{14}$$

Thus we obtain the final normal prediction $I_{\hat{n}}$ using another pass of the Gaussian rasterizer with $SH_{\hat{n}}$. The final normal loss is therefore obtained as: $\mathcal{L}_{normal} = \bar{I}_{\hat{n}} - I_{\hat{n}}$. We obtain the ground truth normal map directly using a pre-trained pix2pixHD [70] network in order to obtain $(\bar{I}_{\hat{n}})$ for every frame t.

3.6 Training

Given a set of training images and input poses, we learn our Gaussian Human Template Model iHuman by optimizing the following objective function:

$$\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{normal} + \mathcal{L}_{reg} \tag{15}$$

where $\mathcal{L}_{rgb}$ is the photometric loss, $\mathcal{L}_{normal}$ is the normal map loss and $\mathcal{L}_{reg}$ is the 3D regularization term for normal consistency. We employ a combination of ℓ_1 and D-SSIM term Eq. (16) for both the $\mathcal{L}_{rgb}$ and the $\mathcal{L}_{normal}$, with the hyperparameter $\lambda = 0.2$:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D_{\text{SSIM}}}. \tag{16}$$

4 Experiments

4.1 Implementation Details

We use PyTorch [53] for the implementation and we choose Adam [29] as the optimizer. We conduct all experiments on a single NVIDIA RTX 4090. We use standard skinned human body template model, SMPL [43] as initial mesh template and also use its blend skinning weights. We upsample the mesh to obtain 165K faces in order to initialize our model. To obtain the ground truth normal maps for normal supervision, we use same pix2pixHD [70] network as used in PIFuHD [61]. Our method runs at 20 iterations per second (optimization on 20 images in 1 second for 1 epoch) during training with $>$ 100 fps during inference.

4.2 Datasets and Baselines

Datasets. We conduct experiments on 3 different datasets.

PeopleSnapshot [3]. It comprises of various monocular RGB videos of different subjects recorded in natural settings. In these videos, individuals assume an A-pose and rotate in place facing a stationary camera. We follow the same evaluation protocol as Instant Avatar [22] by training our model with the pose parameters optimized by Anim-NeRF [8]. We keep the poses frozen throughout training for a fair comparison.

UBC-Fashion [81]. PeopleSnapshot [3] contains tight clothings and all subjects assume an A pose. In order to evaluate our method on in-the-wild long clothing variations, we use videos from UBC-Fashion. As shown in Fig. 6, the subjects in UBC Fashion [81] turn around in-front of a stationary camera in loose clothing. We use the SOTA 3D human pose estimator ReFit [72] to obtain the SMPL poses. As the obtained poses can be misaligned, we enable pose optimization during training with our method.

Multi-Garment dataset [6]. Due to lack of high-quality geometry data of human body in general clothing, we synthesize several sequences from Multi-Garment Network (MGN) [6] dataset. The MGN dataset features 3D scanned models of the human body complete with textures, along with corresponding SMPL-D models that are registered for use in animation. For the creation of the videos, we chose 4 human body models of different body types and clothing variations. These models were animated based on motion sequences from the People-Snapshot dataset [3], where subjects rotate in an A-pose.

The synthetic data from the MGN dataset are mainly used to evaluate the quality of the 3D reconstructions. PeopleSnapshot is used for quantitative evaluation of novel view synthesis. Finally we use both PeopleSnapshot and UBC-Fashion for qualitative evaluation of novel view synthesis and 3D reconstructions.

Baselines. We use the recent works **GART** [33], **Anim-NeRF** [8] and **Instant-Avatar** [22] as our baselines. GART represents the human body in canonical pose represented by the 3D Gaussian parameters and is therefore relatively fast.**Anim-NeRF** [8] utilizes a multi-layer perceptron (MLP) based Neural Radiance Fields (NeRF) [45] to represent human features in a canonical domain and therefore naturally requires longer to optimize. Finally, **Instant-Avatar** [22] also employs NeRF [45] based method that improves the speed of Anim-NeRF. They achieve this speed up by the use of Instant-NGP [66] for radiance field representation and rendering, Fast SNARF [9] for articulation and by the use of occupancy grid for empty space skipping.

4.3 Evaluations

3D Mesh reconstruction. SOTA 3D human mesh reconstruction based methods such as vid2Avatar [16], selfRecon [21] requires more than 1 day of training for a single avatar reconstruction. In contrast, our iHuman is orders of times faster (15 seconds of training time) and uses lower memory. In our experimental setup, we limit the running time of all methods to **hard limit of 5 minutes**. So, we compare our method with current radiance-field based methods; GART, Anim-NeRF with relatively faster training time and from which meshes can be extracted. Specifically, we use marching cubes [41] for mesh extraction from Anim-NeRF and poisson reconstruction [26] on the point cloud obtained from GART to obtain the meshes.

Metrics. We use bi-directional vertex to vertex (v2v) distances (in mm) calculated by uniformly sampling the predicted and the ground truth mesh on

Fig. 4: Qualitative results: we obtain fully rigged colored mesh using iHuman since the reconstructed mesh share the same topology with SMPL body model. The obtained meshes are watertight and accurate.

the MGN dataset. Following standard evaluation protocols, our first step involves aligning the centers of the meshes and fixing their scales. Following this alignment process, vertex-to-vertex (v2v) computations are conducted while the meshes are in the canonical T-pose configuration. We also report normal consistency for surface reconstruction comparison.

Subject	**Ours**				**Anim-NeRF**				**GART**			
	v2v	**NC**	**PSNR**	**LPIPS**	**v2v**	**NC**	**PSNR**	**LPIPS**	**v2v**	**NC**	**PSNR**	**LPIPS**
Subject-1	11.15	0.0309	30.50	0.0172	71.22	0.3294	23.16	0.0745	147.81	0.0401	35.24	0.0140
Subject-2	12.81	0.0327	29.18	0.0267	84.62	0.3795	24.24	0.0744	138.37	0.0457	35.72	0.0160
Subject-3	11.78	0.0301	31.63	0.0159	70.41	0.3310	24.21	0.0691	134.43	0.0429	36.27	0.0146
Subject-4	13.12	0.0302	31.58	0.0178	68.52	0.3428	24.04	0.0667	152.99	0.0370	35.66	0.0162

Table 1: Numerical evaluation on MGN. We report v2v error (mm), mesh normal consistency (NC), PSNR and LPIPS by our method, Anim-NeRF [8] and GART [33].

Comparisons. The quantitative results are shown in Tab. 1. Our method achieves significantly better results on surface reconstruction compared to GART and Anim-NeRF demonstrating the superiority of our approach in producing accurate geometry.

In Fig. 5, we show side by side comparison of the ground truth subject and the predicted mesh of our method along with the v2v error heatmap. Our iHuman robustly handles mesh reconstruction for different clothing and body types for different subjects. In Fig. 4 we show some example reconstructions on PeopleSnapshot dataset and MGN dataset. We show more qualitative results along with comparison in Fig. 7. In only 15s, our method recovers high frequency details such as face structures while other methods struggle with coarse body geometry. For evaluation on more in-the-wild dataset, we show reconstruction on UBC-Fashion dataset in Fig. 6. We provide more evaluations in the supplementary.

Novel View Synthesis. We quantitatively evaluate our method on novel view synthesis and report PSNR, SSIM, and LPIPS metrics on the test frames of PeopleSnapshot. We note that all of our baselines require masked input image

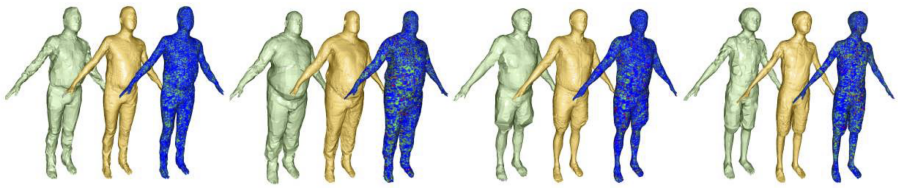

Fig. 5: Results on MGN dataset [6]: We compare the ground truth shapes (green) and prediction (yellow) along with corresponding error heatmaps with respect to ground truth shapes (blue represents errors ≤ 1 cm and red represents errors ≥ 3 cm).

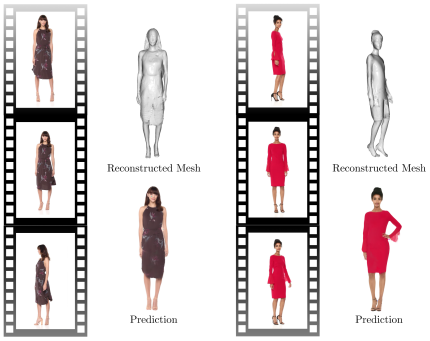

Fig. 6: View Synthesis and 3D Mesh Reconstruction on challenging UGB-Fashion dataset. Faithful reconstruction on UBC-Fashion shows robustness of our method on variety of clothing and poses.

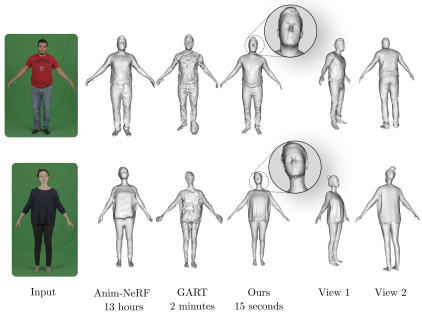

Fig. 7: Qualitative results of 3D Mesh reconstruction on PeopleSnapshot [3]. Our method produces high fidelity mesh even capturing subtle facial details like hair, ear in 15 seconds of computational budget.

sequences and the corresponding SMPL pose parameters, which are costly to obtain. Thus, iHuman benefits greatly by requiring less video frames in two ways: i) less amount of pre-processing ii) faster training. In this experiment, we limit the training time budget to maximum of 5 minutes for all the methods.

In Tab 2, we show novel view synthesis results for the proposed method and the baselines under different number of views. Our iHuman method achieves better LPIPS compared to all the baselines and report second best PSNR.

In Fig. 8, we show novel view and novel pose synthesis on PeopleSnapshot. The ability of our method to accurately model surface geometry helps in better novel poses without artifacts. For evaluation on more in-the-wild setting, we show results on UBC-Fashion in Fig. 6. As SMPL template based deformation model cannot account for loose clothing deformation, we can replace SMPL template with any flexible model that supports i) forward skinning and ii) has faces and vertices information like GART template model [33] to handle clothing deformations. This flexible architecture of our method allows for easy

Methods	male-3			male-4			female-3			female-4		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Anim-Nerf(6 views, 5 mins)	22.01	0.9211	0.0810	22.60	0.9285	0.0826	22.18	0.9370	0.0686	24.44	0.9372	0.0524
Instant Avatar(6 views, 5 mins)	15.21	0.8228	0.2653	16.61	0.8035	0.3010	16.30	0.8210	0.2693	21.88	0.9221	0.0880
GART(6 views, 100 secs)	27.28	0.9593	0.0393	25.84	0.9527	0.0543	23.30	0.9440	0.0548	27.18	0.9593	0.0339
Ours(6 views, 6 secs)	25.27	0.9483	0.0301	22.64	0.9283	0.0525	22.44	0.9368	0.0426	25.08	0.9471	0.0332
Anim-Nerf(12 views, 5 mins)	23.81	0.9847	0.0624	23.10	0.9333	0.0789	22.38	0.9384	0.0623	25.87	0.9466	0.0463
Instant Avatar(12 views, 5 mins)	22.24	0.9226	0.1017	17.31	0.8152	0.2774	21.75	0.9265	0.0783	19.19	0.8789	0.1598
GART(12 views, 105 secs)	29.58	0.9733	0.0315	25.79	0.9540	0.0572	25.15	0.9597	0.0429	28.89	0.9664	0.0327
Ours(12 views, 12 secs)	26.84	0.9586	0.0219	24.93	0.9455	0.0372	23.22	0.9444	0.0359	26.01	0.9549	0.0281
Anim-Nerf(20 views, 5 mins)	23.46	0.9288	0.0680	23.14	0.9340	0.0798	23.91	0.9491	0.0568	24.92	0.9408	0.0494
Instant Avatar(20 views, 5 mins)	26.68	0.9531	0.0333	24.14	0.9383	0.0568	22.52	0.9306	0.0784	26.25	0.9516	0.0238
GART(20 views, 110 secs)	29.99	0.9760	0.0327	27.07	0.9635	0.0537	25.60	0.9623	0.0427	28.78	0.9711	0.0321
Ours(20 views, 20 secs)	27.48	0.9616	0.0196	25.67	0.9506	0.0337	23.58	0.9478	0.0330	27.20	0.9631	0.0244

Table 2: Qualitative Comparison with SoTA on the PeopleSnapshot [3] dataset. We report PSNR, SSIM, and LPIPS between real images and the images generated by Anim-NeRF [8], InstantAvatar [22] and GART [33] with computational budget of maximum 5 minutes. We compare all three methods and ours at different number of inputs sequences (views). The best and second best methods on each metrics are marked on the table.

extension to benefit from better deformation handling models. We show example in supplementary material.

Ablation To study the effectiveness of our mesh binding strategy, hash based SH, displacement encoder and our novel normal map prediction on 3D reconstruction and novel view synthesis, we conduct the following ablations: i) removing the hash based SH Encoder ii) removing the hash based Displacement Encoder iii) removing the mesh binding of the Gaussians and iv) removing the normal map supervision.

Table 3: Ablation Study for Novel View Synthesis. We evaluate novel view synthesis by disabling key components. The results are averaged on 3 subjects of PeopleSnapshot [1].

Methods	LPIPS	PSNR
Full	0.0271	26.08
w/o SH Encoder	0.0341	24.55
w/o Displacement Encoder	0.0344	25.03
w/o Mesh Binding	0.0463	24.68
w/o Normal Loss	0.0523	24.64

Table 4: Ablation Study for 3D Reconstruction. We evaluate 3D reconstruction performance by disabling key components of our method. The results are averaged on 4 subjects of MGN [6].

Methods	v2v	NC
Full	12.21	0.0310
w/o SH Encoder	12.82	0.0311
w/o Displacement Encoder	19.75	0.5918
w/o Mesh Binding	27.51	0.7303
w/o Normal Loss	20.82	0.5725

We show 3D reconstruction results in Tab. 4. We observe that displacement encoder helps in better geometric modeling as shown by the normal consistency loss (NC). Without binding of Gaussians to the mesh, both the v2v error and normal consistency degrades. As shown in Fig. 9, only binding Gaussians to the mesh surface cannot produce accurate geometry of surface details with the photometric loss. This highlights the importance of our approach to encode and optimize normals in Gaussian Splatting [28]. In novel view synthesis, we achieve

the best PSNR and LPIPS with our full method as shown in Tab. 3. The Hash based SH Encoder improves convergence even for less number of views in Fig. 9.

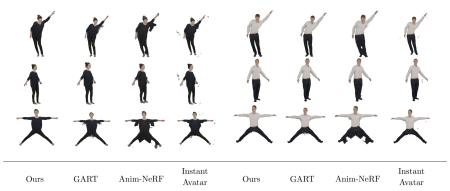

Fig. 8: Results on PeopleSnapshot [1] for 20 frames. Our method is robust to poses as it does not contain artifacts even in novel poses. Better viewed zoomed.

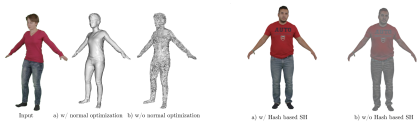

Fig. 9: Without normal supervision surface details are not captured (left). Without Hash based SH Encoder, rendered colors are inaccurate (right). Better viewed zoomed.

Number of Views The performance gain on 3D reconstruction saturates around 50 number of input sequences. For the case of novel view synthesis, 20 views are enough to achieve PSNR of above 25 and very low LPIPS than any other methods under same number of input views in Fig. 10.

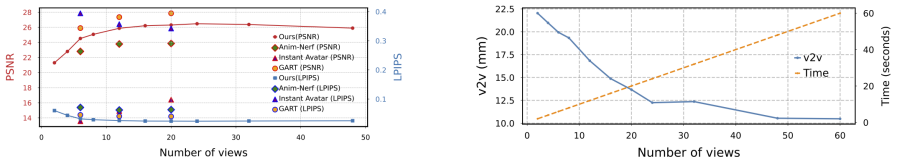

Fig. 10: Left: Performance on Novel View Synthesis of iHuman across different input sequences. Right: Performance on 3D Human Mesh Reconstruction of iHuman and training times for varying input views.

Limitations One major limitation of our method is ability to handle pose or view dependent dynamic appearances. Another limitation is the requirement of accurate 3D input poses for good quality reconstruction like other methods [33], [22], [8]. With SMPL [43] as template model, loose clothing can't be modeled accurately. So, better template model should be used.

5 Conclusion

In this work, we proposed a new method to obtain high fidelity animatable human model in record time. We obtain state-of-the-art performance in limited computational budget. To that end, we used mesh binded Gaussians, explicit normal rasterization and optimization through normal supervision providing fast and accurate results. Through experiments, we also illustrate the need of accurate surface representation, while using Gaussian splats, for faithful rendering under the change in pose. Extending our method to model the per-frame deformations for enabling fast monocular volumetric performance capture can be an interesting frontier to explore.

Acknowledgements

We would like to express our sincere gratitude to Mr. Kobid Upadhyay for his invaluable assistance in creating the figure and rendering the 3D models in blender. His contributions have significantly enhanced the clarity and quality of our visual presentations.

We would also like to thank Alternative Technology (https://alternative.com.np) for providing us with an RTX 4090 for experimentation, which greatly facilitated our research.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: 2018 International Conference on 3D Vision (3DV). pp. 98–109. IEEE (2018)
2. Alldieck, T., Magnor, M.A., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1175–1186. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00127, http://openaccess.thecvf.com/content_CVPR_2019/html/Alldieck_Learning_to_Reconstruct_People_in_Clothing_From_a_Single_RGB_CVPR_2019_paper.html
3. Alldieck, T., Magnor, M.A., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8387–8397. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00875, http://openaccess.thecvf.com/content_cvpr_2018/html/Alldieck_Video_Based_Reconstruction_CVPR_2018_paper.html
4. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12347, pp. 311–329. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_19, https://doi.org/10.1007/978-3-030-58536-5_19
5. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/970af30e481057c48f87e101b61e6994-Abstract.html
6. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5420–5430 (2019)
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image.

In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
8. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular rgb videos. arXiv preprint arXiv:2106.13629 (2021)
9. Chen, X., Jiang, T., Song, J., Rietmann, M., Geiger, A., Black, M.J., Hilliges, O.: Fast-snarf: A fast deformer for articulated neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
10. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11594–11604 (2021)
11. Dragomir, A., Praveen, S., Daphne, K., Sebastian, T., Jim, R., James, D.: Scape. ACM Transactions on Graphics (TOG) (2005). https://doi.org/10.1145/1073204.1073207
12. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)
13. Gafni, G., Thies, J., Zollhofer, M., Niessner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8649–8658 (June 2021), https://openaccess.thecvf.com/content/CVPR2021/html/Gafni_Dynamic_Neural_Radiance_Fields_for_Monocular_4D_Facial_Avatar_Reconstruction_CVPR_2021_paper.html
14. Geng, C., Peng, S., Xu, Z., Bao, H., Zhou, X.: Learning neural volumetric representations of dynamic humans in minutes. CVPR (2023)
15. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023)
16. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12858–12868 (2023)
17. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering (2023)
18. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11046–11056 (October 2021), https://openaccess.thecvf.com/content/ICCV2021/html/He_ARCH_Animation-Ready_Clothed_Human_Reconstruction_Revisited_ICCV_2021_paper.html
19. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2020)
20. Jena, R., Iyer, G.S., Choudhary, S., Smith, B., Chaudhari, P., Gee, J.: SplatArmor: Articulated Gaussian splatting for animatable humans from monocular RGB videos (Nov 2023), http://arxiv.org/abs/2311.10812, arXiv:2311.10812 [cs]
21. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Selfrecon: Self reconstruction your digital avatar from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5605–5615 (2022)
22. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. CVPR (2023)

23. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII. Lecture Notes in Computer Science, vol. 13692, pp. 402–418. Springer (2022). https://doi.org/10.1007/978-3-031-19824-3_24, https://doi.org/10.1007/978-3-031-19824-3_24
24. Jiang, Y., Yao, K., Su, Z., Shen, Z., Luo, H., Xu, L.: Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. CVPR (2023)
25. Kanazawa, A., Black, M.J., Jacobs, D., Malik, J.: End-to-end recovery of human shape and pose. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017). https://doi.org/10.1109/CVPR.2018.00744
26. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing. vol. 7, p. 0 (2006)
27. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139:1–139:14 (2023). https://doi.org/10.1145/3592433, https://doi.org/10.1145/3592433
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2015)
30. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: HUGS: Human Gaussian Splats (Nov 2023), http://arxiv.org/abs/2311.17910, arXiv:2311.17910 [cs]
31. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. CVPR (2019)
32. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021)
33. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: GART: Gaussian Articulated Template Models (Nov 2023), http://arxiv.org/abs/2311.16099, arXiv:2311.16099 [cs]
34. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 811–818 (2023)
35. Li, M., Yao, S., Xie, Z., Chen, K., Jiang, Y.G.: Gaussianbody: Clothed human reconstruction via 3d gaussian splatting. arXiv preprint arXiv:2401.09720 (2024)
36. Li, M., Tao, J., Yang, Z., Yang, Y.: Human101: Training 100+FPS Human Gaussians in 100s from 1 View (Dec 2023), http://arxiv.org/abs/2312.15258, arXiv:2312.15258 [cs]
37. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: TAVA: template-free animatable volumetric actors. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII. Lecture Notes in Computer Science, vol. 13692, pp. 419–436. Springer (2022). https://doi.org/10.1007/978-3-031-19824-3_25, https://doi.org/10.1007/978-3-031-19824-3_25
38. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling (Nov 2023), http://arxiv.org/abs/2311.16096, arXiv:2311.16096 [cs]

39. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. ICCV (2021)
40. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM transactions on graphics (TOG) **40**(6), 1–16 (2021)
41. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
42. Marc, H., Lingjie, L., Weipeng, X., Gerard, P.M., Michael, Z., Christian, T.: Hd humans. Proceedings of the ACM on Computer Graphics and Interactive Techniques (2023). https://doi.org/10.1145/3606927, https://dl.acm.org/doi/10.1145/3606927
43. Matthew, L., Naureen, M., Javier, R., Gerard, P.M., J., B.M.: Smpl. ACM Transactions on Graphics (TOG) (2015). https://doi.org/10.1145/2816795.2818013, https://dl.acm.org/doi/10.1145/2816795.2818013
44. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. CVPR (2019)
45. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. European Conference on Computer Vision (2020). https://doi.org/10.1007/978-3-030-58452-8_24
46. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. European Conference on Computer Vision (2020). https://doi.org/10.1007/978-3-030-58571-6_44
47. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human Gaussian Splatting: Real-time Rendering of Animatable Avatars (Nov 2023), http://arxiv.org/abs/2311.17113, arXiv:2311.17113 [cs]
48. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5762–5772 (October 2021), https://openaccess.thecvf.com/content/ICCV2021/html/Noguchi_Neural_Articulated_Radiance_Field_ICCV_2021_paper.html
49. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. International Conference on 3D Vision (2018). https://doi.org/10.1109/3DV.2018.00062
50. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: sparse trained articulated human body regressor. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 12351, pp. 598–613. Springer (2020). https://doi.org/10.1007/978-3-030-58539-6_36, https://doi.org/10.1007/978-3-030-58539-6_36
51. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering (Dec 2023), http://arxiv.org/abs/2312.05941, arXiv:2312.05941 [cs]
52. Park, J.J., Florence, P.R., Straub, J., Newcombe, R.A., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 165–174. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00025, http://openaccess.thecvf.com/content_CVPR_2019/html/Park_DeepSDF_Learning_Continuous_

Signed_Distance_Functions_for_Shape_Representation_CVPR_2019_paper.html
53. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
54. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/CVPR.2018.00055
55. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
56. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021)
57. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 14294–14303 (2021). https://doi.org/10.1109/ICCV48922.2021.01405
58. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 9054–9063. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00894, https://openaccess.thecvf.com/content/CVPR2021/html/Peng_Neural_Body_Implicit_Neural_Representations_With_Structured_Latent_Codes_for_CVPR_2021_paper.html
59. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. arXiv preprint arXiv: 2312.02069 (2023)
60. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting (Dec 2023), http://arxiv.org/abs/2312.09228, arXiv:2312.09228 [cs]
61. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. CVPR (2020)
62. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
63. Shahbazi, M., Ntavelis, E., Tonioni, A., Collins, E., Paudel, D.P., Danelljan, M., Van Gool, L.: Nerf-gan distillation for efficient 3d-aware generation with convolutions. arXiv preprint arXiv:2303.12865 (2023)
64. Su, S.Y., Bagautdinov, T.M., Rhodin, H.: Danbo: Disentangled articulated neural body representations via graph neural networks. European Conference on Computer Vision (2022). https://doi.org/10.48550/arXiv.2205.01666
65. Su, S.Y., Yu, F., Zollhoefer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. NEURIPS (2021)
66. Thomas, M., Alex, E., Christoph, S., Alexander, K.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (TOG) (2022). https://doi.org/10.1145/3528223.3530127, https://dl.acm.org/doi/10.1145/3528223.3530127

67. Waczyńska, J., Borycki, P., Tadeja, S., Tabor, J., Spurek, P.: Games: Mesh-based adapting and modification of gaussian splatting. arXiv preprint arXiv:2402.01459 (2024)
68. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. NEURIPS (2021)
69. Wang, S., Schwarz, K., Geiger, A., Tang, S.: ARAH: animatable volume rendering of articulated human sdfs. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII. Lecture Notes in Computer Science, vol. 13692, pp. 1–19. Springer (2022). https://doi.org/10.1007/978-3-031-19824-3_1, https://doi.org/10.1007/978-3-031-19824-3_1
70. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
71. Wang, Y., Gao, Q., Liu, L., Liu, L., Theobalt, C., Chen, B.: Neural novel actor: Learning a generalized animatable neural representation for human actors. IEEE Transactions on Visualization and Computer Graphics (2022). https://doi.org/10.48550/arXiv.2208.11905, https://arxiv.org/abs/2208.11905v2
72. Wang, Y., Daniilidis, K.: Refit: Recurrent fitting network for 3d human recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14644–14654 (2023)
73. Weng, C., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 16189–16199 (2022). https://doi.org/10.1109/CVPR52688.2022.01573
74. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. Computer Vision and Pattern Recognition (2022). https://doi.org/10.1109/CVPR52729.2023.00057
75. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. CVPR (2022)
76. Xu, H., Alldieck, T., Sminchisescu, C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. NEURIPS (2021)
77. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems **34**, 4805–4815 (2021)
78. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 4805–4815. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/25e2a30f44898b9f3e978b1786dcd85c-Paper.pdf
79. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: Monohuman: Animatable human neural field from monocular video. CVPR (2023)
80. Yuan, Y., Li, X., Huang, Y., De Mello, S., Nagano, K., Kautz, J., Iqbal, U.: Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. arXiv preprint arXiv:2312.11461 (2023)
81. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. arXiv preprint arXiv:1910.09139 (2019)
82. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics (TOG) **40**(4), 1–18 (2021)

83. Zhao, F., Jiang, Y., Yao, K., Zhang, J., Wang, L., Dai, H., Zhong, Y., Zhang, Y., Wu, M., Xu, L., Yu, J.: Human performance modeling and rendering via neural animated mesh. ACM Trans. Graph. **41**(6), 235:1–235:17 (2022). https://doi.org/10.1145/3550454.3555451, https://doi.org/10.1145/3550454.3555451
84. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15893–15903 (June 2022), https://openaccess.thecvf.com/content/CVPR2022/html/Zheng_Structured_Local_Radiance_Fields_for_Human_Avatar_Modeling_CVPR_2022_paper.html
85. Zhu, H., Zhan, F., Theobalt, C., Habermann, M.: Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. arXiv preprint arXiv:2312.05161 (2023)
86. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3D Gaussian Avatars (Nov 2023), http://arxiv.org/abs/2311.08581, arXiv:2311.08581 [cs]
87. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 4574–4584. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00444, https://doi.org/10.1109/CVPR52729.2023.00444