InstantGeoAvatar: Effective Geometry and Appearance Modeling of Animatable Avatars from Monocular Video

Alvaro Budria[1], Adrian Lopez-Rodriguez[2], Òscar Lorente*[3], and Francesc Moreno-Noguer*[4]

[1] Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
alvaro.francesc.budria@upc.edu
[2] Vody
[3] Floorfy
[4] Amazon

Abstract. We present InstantGeoAvatar, a method for efficient and effective learning from monocular video of detailed 3D geometry and appearance of animatable implicit human avatars. Our key observation is that the optimization of a hash grid encoding to represent a signed distance function (SDF) of the human subject is fraught with instabilities and bad local minima. We thus propose a principled geometry-aware SDF regularization scheme that seamlessly fits into the volume rendering pipeline and adds negligible computational overhead. Our regularization scheme significantly outperforms previous approaches for training SDFs on hash grids. We obtain competitive results in geometry reconstruction and novel view synthesis in as little as five minutes of training time, a significant reduction from the several hours required by previous work. InstantGeoAvatar represents a significant leap forward towards achieving interactive reconstruction of virtual avatars.

Keywords: 3D Computer Vision · Human Avatars · Neural Radiance Fields · Clothed Human Modeling

1 Introduction

Enabling the reconstruction and animation of 3D clothed avatars is a key step to unlock the potential of emerging technologies in fields such as augmented reality (AR), virtual reality (VR), 3D graphics and robotics. Interactivity and fast iteration over reconstructions and intermediate results can help speed up workflows for designers and practitioners. Different sensors are available for learning clothed avatars, including monocular RGB cameras, depth sensors and 4D scanners. RGB videos are the most widely available, yet provide the weakest supervisory signal, making this configuration elusive.

*Work done while at Institut de Robòtica i Informàtica Industrial (CSIC-UPC).
The project website is at github.com/alvaro-budria/InstantGeoAvatar.

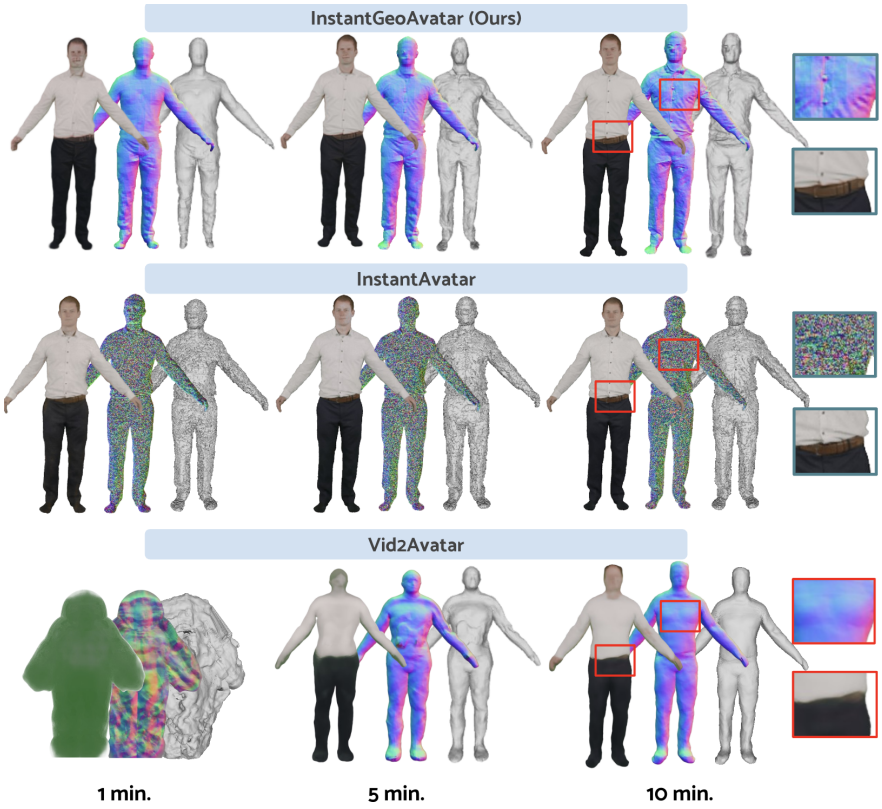


Fig. 1: InstantGeoAvatar. We introduce a system capable of reconstructing the geometry and appearance of animatable human avatars from monocular video in less than 10 minutes. In order to attain high quality geometry reconstructions, we propose a smoothing term directly on the learned signed distance field during optimization, requiring no extra computation or sampling and delivering noticeable qualitative improvements.

Traditional depth-based methods fusing depth measurements over time [8, 50, 73, 74] produce compelling reconstructions but need complex and expensive setups to avoid sensor noise. Dense multi-view capture systems [1, 9, 20, 30, 34, 41, 49, 64, 79, 91, 94, 96, 97, 105] offer detailed 3D reconstructions by leveraging multi-view images and cues like silhouette or stereo. However, they can only be applied in controlled camera studios. Moreover, neither depth-based nor dense multi-view approaches can effectively work from RGB inputs alone or produce satisfactory results within short fitting times.

Mesh-based approaches [2–4, 6, 7, 10, 12, 28, 44, 53, 71] struggle with garment deformations and are restricted to low-resolution topologies. Point cloud-based techniques [59–61, 66, 68, 80, 98, 116, 117, 120, 121] have shown promising outcomes, but are not yet capable of fast-training times with RGB supervision only.

The advent of neural radiance fields (NeRFs) [70] enabled techniques for novel view synthesis and animation of human avatars from RGB image supervision only [26, 38, 93, 109]. Volume rendering-based approaches typically learn a canonical representation that is deformed with linear blend skinning [65] and an additional non-rigid deformation [77, 104, 115]. Despite producing good renderings of human avatars, these techniques lack awareness of the underlying geometry. In parallel, some works have adopted a signed distance function (SDF) [111] as basic primitive for learning clothed human avatars from 3D scans [5, 85, 107]. To remove the need for 3D supervision, some works have embedded SDFs within a volumetric rendering pipeline [29, 77, 103]. Significant steps have been taken to speed up training of NeRF-based approaches [26, 38] by leveraging an efficient hash-grid spatial encoding [72]. Subsequent work has tried to improve training on such unstable and noisy hash grids [23, 56] with some success. Up to date, however, fast geometry learning in general and effective use of hash grid-based representations in particular for clothed human avatars with RGB supervision only remains elusive.

The challenges faced by NeRF- and SDF-based approaches trained with volume rendering can be succinctly reduced to effectively capturing realistic non-rigid deformations, dealing with noisy pose and camera estimates, slow training, and unstable training in the case of hash grid-based methods. In this paper, we specifically focus on the last two challenges, and aim at significantly advancing towards the realization of interactive use of human avatar modelling. We propose InstantGeoAvatar, a system capable of yielding good rendering and reconstruction quality in as little as 5 minutes of training, down from several hours as in prior work. Building on recent advances for fast training of NeRF based systems [38, 72] and efficient training of hash grid encodings [23, 56], we demonstrate that even in combination such prior improvements and techniques are insufficient for fast and effective learning of 3D clothed humans. Thus we propose a simple yet effective regularization scheme that imposes a local geometric consistency prior during optimization, effectively removing undesired artifacts and defects on the surface. The proposed approach, which effectively constrains surface curvature and torsion over continuous SDFs along ray directions, is easy to implement, fits neatly within the volume rendering pipeline, and delivers noticeable improvements over our base model without additional cost.

Our experiments demonstrate the effectiveness of the proposed method for effective and fast learning of animatable 3D human avatars from monocular video. At the short-training regime, InstantGeoAvatar yields superior geometry reconstruction and rendering quality compared to previous work in less than 10 minutes (see Fig. 1). While SoTA methods can yield more accurate reconstructions after several hours upon convergence, InstantGeoAvatar still shows comparable and even superior results with out of distribution (OOD) poses. In addition, the presented ablation demonstrates that previous work on improving training of hash grid-based representations is insufficient for obtaining satisfying geometry reconstructions, highlighting the suitability of our proposal.

2 Related Work

2.1 Reconstructing Humans with Multi-View and Depth

Traditional depth-based approaches for human shape reconstruction fuse depth measurements over time [73, 74], and subsequent works improve robustness by incorporating additional priors [8, 50, 112, 114], and by using custom-design sensors [22, 113]. Although effective, these methods require accurate depth information, which can only be obtained with laborious setups. Methods based on calibrated dense multi-view capture setups [1, 9, 20, 30, 34, 41, 49, 64, 79, 91, 94, 96, 97, 105] are capable of producing high-fidelity 3D reconstruction of human subjects by using multi-view images and other cues such as silhouette, stereo or shading. They are expensive and require skilled labor to operate, which hinders their applicability outside of camera studios. Both depth-based and dense multi-view cameras cannot work from monocular RGB inputs alone by design.

2.2 Reconstructing Humans from Monocular & Sparse Multi-views

Template and mesh-based approaches can yield reasonable results even in the low-data regime by leveraging a low-rank human shape prior. On the other hand, implicit representations are continuous by design and have been used to produce detailed reconstructions of a clothed human body. A specific subline of work aims at speeding up training of radiance field-based methods while maintaining good reconstruction quality. More recently, methods for human body modelling based on Gaussian Splats have shown compelling results at fast rendering times.
Explicit Representations. Mesh-based techniques typically represent cloth deformations as deformation offsets [2–4, 6, 7, 10, 12, 28, 44, 53, 71] added to a minimally clothed parametric human body model prior (e.g. SMPL [65]) that has been previously rigged to a particular subject. Despite being highly compatible with parametric human body models, these approaches struggle with garment deformations and are limited to a low resolution topology. Point cloud-based methods [59–61, 66, 68, 80, 98, 116, 117, 120, 121] have shown promising results by combining the advantages of a representation that is explicit and simultaneously allows to model varying topologies.
Neural Volumetric Rendering and Implicit Representations. Previous works have successfully applied neural radiance fields [70] and signed distance functions (SDFs) [75, 101, 111] as basic primitives for modelling human avatars. Image based methods trained on 3D scans [5,31,32,36,78,85,86,107,108,122] and methods based on RGB-D and pointcloud data [6,10,19,21,51,55,67,90,100,118,119] have difficulties with out-of-distribution poses, can fail to produce temporally consistent reconstructions, and suffer from incomplete observations in the case of RGB-D based methods. While some methods relying solely on 2D images attempt to train models that generalize to multiple subjects and can perform inference of novel views of a human directly [15, 18, 46], most such methods involve volumetric rendering and learn a per-subject representation in a canonical space which can then be deformed to a given pose [11, 13, 16, 17, 24–26, 29, 38,

42,54,57,62,69,77,81,88,88,89,92,93,102–104,106,109,110,115,123,124,127] or simply train a per-subject model without any canonical space [21,37,47,48,125].

Significant effort has been devoted to improving the performance of volume rendering-based methods on various fronts. [29,39,93,103,104] jointly optimize neural network and body pose parameters. [11,16,17,42,69,81,88,103,110] focus on the computation of correspondences between posed and canonical space. To improve the posing of the canonical representation, [16,17,24,25,54,54,77,81,115] learn or refine skinning weights during training. [24, 25, 29, 39, 54, 77, 103, 104, 109] add a non-rigid deformation module on top of the rigid deformation of a parametric body model. [25, 57, 69, 89, 92, 93, 115] leverage pose encodings that help improve the results of pose-conditioned components. [29] allows segmenting a human on a video without mask supervision. Despite their success, these works do not demonstrate high geometry reconstruction quality at fast speeds.

An additional line of work aims at speeding up the fitting of models to each scene. Some works [24,26,38,102,106,127] have employed a multiresolution hash grid spatial encoding [72]. Despite its effectiveness in accelerating training, it has been observed [23,33,56,63,126] that it lacks the implicit regularization towards lower frequencies that MLPs enjoy [83]. Even if the commonly used coordinate-based MLPs partially offset this bias [84] with a Fourier positional encoding, one can modulate the frequency bands of such encoding to still maintain a desirable level of smoothness, as done in common practice by works previously cited. Some works have attempted to offset the lack of regularization in hash grids by considering the behaviour of gradients [33, 56], implementing coarse-to-fine strategies [56, 63] and including a hybrid positional encoding to recover the regularizing effect [23, 126]. As shown in our experiments, these improvements are insufficient for obtaining quality reconstructions, hence the need for a more suitable training scheme, which we deliver in the form of an additional computational scheme in volume rendering.

Gaussian Splats [43] have emerged as a powerful alternative primitive for volumetric rendering. Recent work has shown impressive results at the task of novel view synthesis of human avatars [35, 40, 45, 58, 76, 82, 87, 99], even achieving remarkably short training times [52], but we are unaware of any works specifically focusing on modeling the geometry of human avatars from monocular video.

3 Method

InstantGeoAvatar learns a person-specific representation of geometry and appearance of a human subject from monocular video, given a set of input images with corresponding camera parameters, body masks and body poses. We learn a parameterization of the canonical 3D geometry and the texture of a clothed human as an implicit signed distance field (SDF) and a texture field. A specialized canonicalization module finds rigid correspondences between posed and canonical spaces, and a non-rigid deformation module learns non-rigid clothing deformation and pose-dependent effects. The SDF and texture modules are learned via differentiable volume rendering, which is accelerated with an empty

space skipping grid, as in previous work. We incorporate a surface regularization term within the volume rendering pipeline which not only ensures a nice surface smoothness and appearance, but also leads to watertight meshes, which other works struggle with.

3.1 Preliminaries

We next describe the fundamental building blocks of the proposed pipeline, including an accelerated signed distance and texture field to model shape and appearance in canonical space and a canonicalization module combined with a non-rigid deformation to map this canonical representation into deformed space.
Canonical Signed Distance Field. We model human geometry in a canonical space with a signed distance function $\boldsymbol{f}_{sdf}$ that assigns a distance value and a feature vector to each point in the 3D canonical space:

$$\mathbf{f}_{sdf} : \mathbb{R}^3 \rightarrow \mathbb{R}, \mathbb{R}^{16} \tag{1}$$

$$x_c \mapsto \mathbf{d}, \mathbf{v} \tag{2}$$

The body shape $\mathcal{S}$ in canonical space is then the zero-level set of $\boldsymbol{f}_{sdf}$:

$$\mathcal{S} = \{x_c \,|\, \boldsymbol{f}_{sdf}(x_c) = 0\} \tag{3}$$

Following [38] we use the multiresolution hash feature grid encoding from Instant-NGP [72] to parameterize $\boldsymbol{f}_{sdf}$.
Canonical Texture Field. We learn a texture field $\boldsymbol{f}_{rgb}$ in canonical space that models the subject's appearance, conditioned on the SDF's predicted features:

$$\mathbf{f}_{rgb} : \mathbb{R}^3, \mathbb{R}^{16} \rightarrow \mathbb{R}^3 \tag{4}$$

$$x_c, \mathbf{v} \mapsto \mathbf{c} \tag{5}$$

Articulating the Canonical Representation. We leverage the SMPL [65] parametric body model to map a canonical point x_c to a deformed point x_d via linear blend skinning (LBS), according to a set of bone transformations $\boldsymbol{B}_i$ which are derived from body pose θ:

$$x_d = LBS(x) = \sum_{i=1}^{n_b} w_i \boldsymbol{B}_i x_c \tag{6}$$

However, the canonical correspondence x_c^* of a deformed point x_d is defined by the inverse mapping of Eq. 6. Thus it is necessary to compute the one-to-many mapping from points in posed space x_d to their correspondences x_c^* in canonical space. Fast-SNARF [16] efficiently establishes such correspondences.

We add a pose-dependent offset to the rigid correspondence found by SNARF:

$$\boldsymbol{f}_{\Delta x} : \mathbb{R}^3, \mathbb{R}^{69} \rightarrow \mathbb{R}^3 \tag{7}$$

$$x_c, \theta \mapsto \Delta x, \tag{8}$$

3.2 Volume Rendering of SDF-based Radiance Fields

We learn the canonical representation end-to-end via differentiable volume rendering of our SDF function [111]. As in previous works [29, 103] we map the distance value to a density σ by squashing it with the Laplacian Cumulative Distribution Function:

$$\sigma(x_c) = \alpha\Big(\frac{1}{2} + \frac{1}{2} sgn(-\boldsymbol{f}_{sdf}(x_c))(1 - \exp -\frac{|\boldsymbol{f}_{sdf}|}{\beta})\Big), \tag{9}$$

where β is a learnable parameter and we set $\alpha = \frac{1}{\beta}$.

For a given pixel, we cast a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{l}$ from the optic center $\mathbf{o}$ with ray direction $\mathbf{l}$. We sample N_p points between the near and far bounds in occupied space taking into account an occupancy grid, as presented in [38]. At each point, we query color $\boldsymbol{c}_i$ and density σ_i from the canonical representation by warping ray points $\{x_d^i\}^{N_p}$ from deformed to canonical space.

We then integrate the queried radiance and color values along each ray to get the rendered pixel color $\hat{C}$ as

$$\hat{C} = \sum_{i=1}^{N_p} \alpha_i \prod_{i<j} (1 - \alpha_j)\boldsymbol{c}_i, \text{ with } \alpha_i = 1 - \exp(\sigma_i \delta_i), \tag{10}$$

where $\delta_i = ||x_c^{i+1} - x_c^i||$ is the distance between consecutive samples.

3.3 Training Objectives

We optimize our model against multiple weighted loss functions, including a smooth surface regularization term $\mathcal{L}_{smooth}$ that significantly boosts reconstruction quality:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_\alpha \mathcal{L}_\alpha + \lambda_{Eik}\mathcal{L}_{Eik} + \lambda_{smooth}\mathcal{L}_{smooth}. \tag{11}$$

$\mathcal{L}_{rgb}$ is the photometric loss for rendered pixel color, $\mathcal{L}_\alpha$ is an L1 loss between the ground-truth and rendered masks and helps guide the reconstruction, and $\mathcal{L}_{Eik}$ corresponds to the Eikonal loss term [27], which encourages the learned SDF to have a well-behaved wave-like transition between consecutive isosurface slices. The loss weightings are $(\lambda_{rgb},\ \lambda_\alpha,\ \lambda_{Eik},\ \lambda_{smooth}) = (10,\ 0.1,\ 0.1,\ 1.0)$. $\mathcal{L}_{smooth}$ can be flexibly set within the range [0.5, 1.5] with good smoothing and details balance.

Surface Regularization. Directly substituting the NeRF [70] rendering module with the VolSDF [111] rendering scheme explained in Sec. 3.2 produces noisy, stripped surfaces (Fig. 7a). The photometric loss only constrains the ray integral to render the desired color, but there are infinitely many shape variations that satisfy this condition. MLP-based architectures [29, 103] are naturally biased to low-energy solutions [83], but a hash-grid based representation lacks such implicit bias.

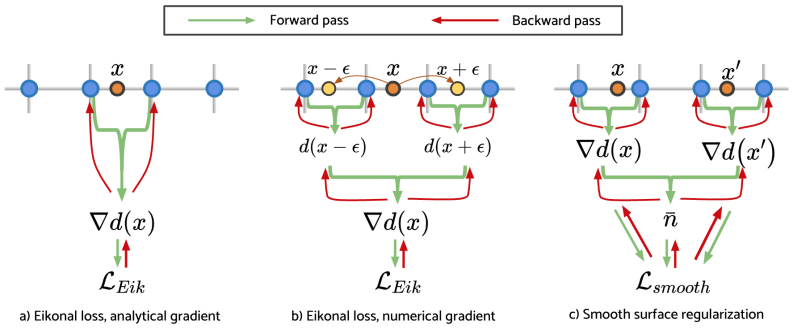


Fig. 2: Non-local updates of the hash grid features. We consider a 1D hash grid encoding segment to illustrate how the proposed regularization affects backpropagation updates. Vanilla Eikonal loss (a) performs backpropagation updates on a single local hash grid cell resulting in discontinuous and spatially disconnected updates. (b) [56] used numerical gradients to distribute backpropagation updates to other cells in the grid, resulting in more spatially coherent learned features. Our proposed smooth surface regularization (c) also distributes backpropagation updates.

We devise a regularization procedure that effectively imposes a local coherence prior to the surface, by constraining surface normals at consecutive points on a ray to have the same direction and magnitude, i.e. that the isosurface be flat along the ray direction at each interval. If we locally approximate an arc parameterization of the surface with a straight line along the ray direction $\mathbf{l}$ at ray depth t, this scheme amounts to taking the directional derivative $\frac{\mathrm{d}\mathbf{n}}{\mathrm{d}t}$ of the normal of the surface $\mathbf{n}$ along $\mathbf{l}$ at t. By the Frenet–Serret formulas [95] this term is related to the curvature κ and torsion τ as

$$\frac{\mathrm{d}\mathbf{n}}{\mathrm{d}t} = -\kappa\mathbf{T} + \tau\mathbf{B}, \tag{12}$$

with $\mathbf{T}$ the surface tangent vector and $\mathbf{B}$ the surface binormal vector. Thus, by minimizing this term, we are directly imposing a penalty on the amount of curvature (how much the surface deviates from a straight line) and torsion (how much the surface deviates from following a regular path) that the surface presents along this direction, effectively enforcing the surface to be well-behaved.

We incorporate this regularization strategy into volume rendering without additional sampling by computing finite differences on the estimated normals $\nabla_x d(x_c)$ at each sampled point x_c in canonical space, which we already obtained for the Eikonal loss term $\mathcal{L}_{Eik}$:

$$\mathcal{L}_{smooth} = \frac{1}{N_r}\sum_{i=1}^{N_r}\frac{1}{N_s^i}\sum_{j=1}^{N_s^i-1} ||\overline{\mathbf{n}}_i^j - \hat{\mathbf{n}}_i^j||_2 + ||\overline{\mathbf{n}}_i^j - \hat{\mathbf{n}}_i^{j+1}||_2 \tag{13}$$

where $\hat{\mathbf{n}}_i^j = \frac{\nabla_x d(x)}{||\nabla_x d(x)||_2}$ is the normalized estimated surface normal at point x_i^j, $\overline{\mathbf{n}}_i^j = \frac{1}{2}(\hat{\mathbf{n}}_i^j + \hat{\mathbf{n}}_i^{j+1})$ is the estimated normal at the midpoint of the interval

(x_i^j, x_i^{j+1}), N_s^i is the number of samples on the i-th ray and N_r is the number of rays.

Moreover, it has been noted [56] that surface normals $\nabla_x d(x)$ computed on hash grid-based representations only depend on the voxel features that immediately surround the point x. When learning with gradients from the Eikonal loss $\mathcal{L}_{Eik}$, this results in spatially discontinuous and incoherent updates that are limited to a single grid cell (Fig. 2a). Thus, by combining normals computed at different points with our approach (Fig. 2c), a greater number of adjacent and nearby grid cells are involved in the computations, and the resulting updates from the loss term are more stable spatially coherent.

Discussion. Both on a conceptual and computational level, our proposed term is significantly different from the Eikonal loss [27], the use of finite differences for surface normal computation [56], and the curvature loss [56] from previous works. The Eikonal loss $\mathcal{L}_{Eik}$ has a variational motivation and is derived as the solution of a wave propagation PDE. It acts as a global constraint on the SDF, ensuring it propagates through space as an onion, without silos or isolated components. Our constraint is geometric and differential as it works on local derivatives, and does not involve the magnitude of the gradient at a fixed point, but the differences of the normals along ray directions.

Our approach is similar to that of [56] in that we also propagate gradient updates to multiple cells for more stable training. But unlike them, we achieve this effect by considering the surface normals at consecutive ray points, and use more precise analytical gradients to compute those normals. We only take finite differences along ray directions to compute the derivative of the surface normal.

The curvature loss from [56] computes the Laplacian considering how much an SDF value at each sample point diverges from the SDF values of neighbors at a fixed distance, which in three-dimensional space is not explicitly related to the curvature and torsion of the surface. Moreover, computing it involves 6 additional samples and relies on a fixed scale on the finite differences which needs to be scheduled over training. On the other hand, our approach explicitly relates the directional derivative of the surface normals to geometric quantities of curvature and torsion. Since the rays' direction of incidence on the surface varies with pixel location and camera location, the regularization is effectively applied multi-scale over training (Fig. 4), without the need for any scale scheduling nor any additional samples.

4 Experiments

We evaluate the rendering and geometry reconstruction quality of our proposed method and compare it with other SoTA methods both on in-distribution and out-of-distribution poses. Additionally, we present an ablation that investigates the effect of different techniques proposed for stable training with hash-grid encodings and validate the effectiveness of our approach. Further results are presented in the supplementary video.

Table 1: Geometry Reconstruction. We report L2 Chamfer Distance (CD), Normal Consistency (NC) and Intersection over Union (IoU) on the X-Humans dataset [88].

Sequence	Metric	V2A	V2A^{10}	IA	Ours
25	CD ↓	**0.304**	0.602	0.901	0.579
	NC ↑	**0.919**	0.894	0.688	0.910
	IoU ↑	**0.977**	0.920	0.842	0.974
28	CD ↓	**0.393**	0.842	0.851	0.600
	NC ↑	**0.922**	0.871	0.727	0.914
	IoU ↑	**0.963**	0.902	0.755	0.947
35	CD ↓	**0.417**	0.641	1.073	0.557
	NC ↑	**0.912**	0.879	0.673	0.891
	IoU ↑	**0.944**	0.902	0.798	0.923
36	CD ↓	**0.572**	0.748	1.494	0.691
	NC ↑	**0.844**	0.817	0.576	0.838
	IoU ↑	**0.950**	0.915	0.771	0.936
58	CD ↓	**0.311**	0.512	1.693	0.597
	NC ↑	**0.930**	0.887	0.681	0.875
	IoU ↑	**0.973**	0.964	0.704	0.950

Table 2: Geometry Reconstruction. We report L2 Chamfer Distance (CD), Normal Consistency (NC) and Intersection over Union (IoU) on OOD poses of the X-Humans dataset [88].

Sequence	Metric	V2A	V2A^{10}	IA	Ours
25	CD ↓	**0.751**	0.892	1.72	0.881
	NC ↑	**0.889**	0.871	0.699	0.879
	IoU ↑	**0.931**	0.928	0.845	0.930
28	CD ↓	1.096	1.445	2.452	**1.089**
	NC ↑	0.860	0.842	0.720	**0.885**
	IoU ↑	**0.911**	0.901	0.761	0.909
35	CD ↓	**0.705**	0.804	1.213	0.800
	NC ↑	**0.883**	0.869	0.678	0.871
	IoU ↑	**0.945**	0.912	0.794	0.896
36	CD ↓	**0.760**	0.903	2.23	0.871
	NC ↑	0.842	0.830	0.579	**0.845**
	IoU ↑	0.910	0.899	0.768	**0.922**
58	CD ↓	**0.732**	0.785	1.892	0.842
	NC ↑	0.835	0.824	0.682	**0.838**
	IoU ↑	**0.898**	0.863	0.705	0.877

4.1 Datasets

PeopleSnapshot [4] contains videos of humans rotating in front of a camera, and is the main dataset employed by InstantAvatar [38]. For a fair comparison, we follow its same evaluation protocol, taking the pose parameters optimized by Anim-NeRF [14] and keeping them frozen throughout training.

X-Humans [88] provides several video sequences of humans in various poses and performing different actions. The dataset contains accurate ground-truth detailed clothed 3D meshes and SMPL fits for each subject at each frame. We select a subset of 5 sequences to perform our experiments. Each sequence contains several tracks or subsequences. For each sequence we pick a set of tracks amounting to between 150 to 300 consecutive frames for training and another set of tracks for evaluation, containing both in-distribution as well as out-of-distribution poses that significantly depart from those poses seen during training.

4.2 Baselines

InstantAvatar (IA) [38] is capable of modeling the appearance of a human subject in short training times. In our experiments, we use the publicly available code to train this baseline for 10 minutes. Despite its ability to produce high-quality renderings, IA lacks geometric awareness as it models radiance.

Vid2Avatar (V2A) [29] is able to produce high-quality geometry reconstructions without the need for ground-truth masks but requires several hours of training to reach convergence. We compare our method against Vid2Avatar after 24 hours (V2A) and 10 minutes (V2A^{10}) of training, to show that our method can achieve good rendering and reconstruction quality much faster.

Table 3: Rendering Quality. We report PSNR, SSIM and LPIPS on in-distribution poses of the X-Humans dataset [88].

Sequence	Metric	V2A	V2A^{10}	IA	Ours
25	PSNR ↑	29.82	25.12	**30.05**	29.87
	SSIM ↑	**0.991**	0.960	0.986	0.983
	LPIPS ↓	0.016	0.033	**0.012**	0.020
28	PSNR ↑	**28.20**	24.35	27.63	27.74
	SSIM ↑	0.980	0.952	**0.982**	0.976
	LPIPS ↓	**0.017**	0.025	0.019	0.032
35	PSNR ↑	29.01	25.49	**29.47**	28.68
	SSIM ↑	0.984	0.961	**0.988**	0.979
	LPIPS ↓	0.019	0.032	**0.011**	0.028
36	PSNR ↑	29.59	26.77	**31.74**	30.97
	SSIM ↑	0.973	0.957	**0.981**	0.974
	LPIPS ↓	0.015	0.030	**0.010**	0.021
58	PSNR ↑	30.32	26.93	**31.05**	30.54
	SSIM ↑	0.986	0.959	**0.989**	0.985
	LPIPS ↓	0.018	0.026	**0.009**	0.016

Table 4: Rendering Quality. We report PSNR, SSIM and LPIPS on out-of-distribution poses of the X-Humans dataset [88].

Sequence	Metric	V2A	V2A^{10}	IA	Ours
25	PSNR ↑	23.09	19.66	23.32	**23.56**
	SSIM ↑	**0.965**	0.937	0.957	0.964
	LPIPS ↓	**0.025**	0.045	0.032	0.027
28	PSNR ↑	24.50	20.93	24.56	**24.77**
	SSIM ↑	0.963	0.942	0.964	**0.966**
	LPIPS ↓	**0.028**	0.038	0.030	0.033
35	PSNR ↑	**25.68**	21.34	25.23	25.21
	SSIM ↑	**0.960**	0.952	0.955	0.958
	LPIPS ↓	0.031	0.048	**0.027**	0.039
36	PSNR ↑	**26.41**	22.05	25.07	26.33
	SSIM ↑	0.953	0.947	0.950	**0.955**
	LPIPS ↓	0.036	0.040	**0.028**	0.031
58	PSNR ↑	23.88	21.40	23.82	**24.05**
	SSIM ↑	**0.964**	0.951	0.957	0.960
	LPIPS ↓	**0.025**	0.040	0.033	0.029

Table 5: Rendering Quality. We report PSNR, SSIM and LPIPS on in-distribution poses of the PeopleSnapshot dataset [4]. Note that unlike in [38], we do not overfit the poses prior to testing.

Sequence	Metric	V2A	V2A^{10}	IA	Ours
f-3-c	PSNR ↑	**24.96**	22.91	24.24	24.89
	SSIM ↑	**0.957**	0.916	0.949	0.952
	LPIPS ↓	**0.030**	0.062	0.033	0.046
m-3-c	PSNR ↑	**28.37**	20.93	27.96	28.00
	SSIM ↑	**0.963**	0.942	**0.963**	0.961
	LPIPS ↓	**0.017**	0.038	0.021	0.035
f-4-c	PSNR ↑	**27.44**	21.34	26.89	27.16
	SSIM ↑	**0.968**	0.952	0.959	0.962
	LPIPS ↓	0.026	0.048	**0.023**	0.040
m-4-c	PSNR ↑	**26.53**	22.05	25.38	25.46
	SSIM ↑	**0.955**	0.947	0.953	0.951
	LPIPS ↓	**0.039**	0.040	0.042	0.059

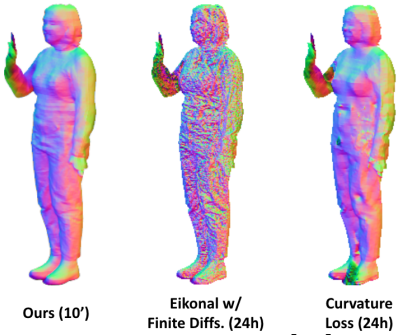


Fig. 3: Neuralangelo's [56] SDF training scheme at longer training regime. Our approach beats Neuralangelo's proposal even after 24 hours of training.

4.3 Geometry Reconstruction Quality

First, we compare our proposed human shape reconstruction approach to the baselines (Table 1). IA [38] fails to model meaningful surfaces. V2A can very accurately reconstruct human shapes, but with slow training times. V2A^{10} is initialized with a SMPL body shape, which gives it a significant boost in geometric accuracy at the beginning of training. However, as seen qualitatively in Figures 1 and 6 it fails to model clothing details. Our method strikes a good balance between reconstruction accuracy and speed, obtaining reasonably good results within 10 minutes of training.

Table 6: Quantitative comparison of SDF regularization schemes. We report PSNR, SSIM, Chamfer Distance (CD) and Normal Consistency (NC) on the X-Humans dataset [88].

Component	PSNR ↑	SSIM ↑	CD ↓	NC ↑
Base	28.90	0.961	0.827	0.827
Eikonal w/ Finite Diffs. [56]	28.60	0.960	2.45	0.657
Curvature Loss [56]	28.51	0.957	2.03	0.536
Eik. w/ F. Diffs. and Curv. Loss [56]	28.47	0.959	5.03	0.52
Hybrid Pos. Encoding [23]	26.48	0.954	**0.752**	0.803
$\mathcal{L}_{smooth}$ (ours)	**29.24**	**0.964**	0.772	**0.850**

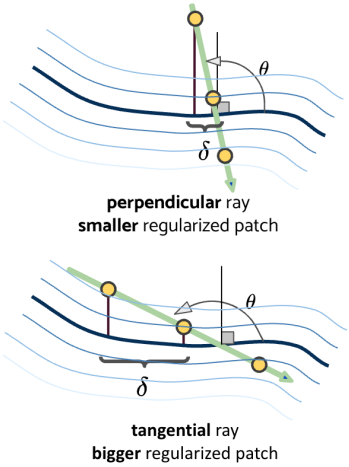


Fig. 4: Multiscale effect of the proposed loss term.

Table 7: Computational demand of SDF regularization schemes.

	Base	Ours	Hybrid Pos. Encoding [23]	Eikonal w/ Finite Diffs. [56]	Curvature Loss [56]
training speed	**8.0 it./s**	7.7 it./s	4.1 it./s	5.3 it./s	4.8 it./s
peak memory	∼ **15 GB**	∼ 15 GB	∼ 16 GB	∼ 20 GB	∼ 20 GB

4.4 View Synthesis Quality

Tables 3 and 5 show our method produces quality renderings on both datasets. Since Vid2Avatar is not designed for fast training, V2A^{10} fails to synthesize good renders, while both IA and InstantGeoAvatar achieve the same rendering quality as V2A in a much shorter time. As seen qualitatively in Figures 1 and 6, V2A^{10} yields flat-textured renders, whereas IA and InstantGeoAvatar can model texture patterns such as shirt wrinkles.

4.5 Synthesis & Reconstruction Quality on OOD Poses

As seen in Table 2, the geometry produced by V2A degrades significantly presumably because it conditions its SDF representation on pose parameters. Our method performs almost identically as V2A, and outperforms InstantAvatar and V2A^{10}. This can also be observed in Fig. 6 and in the supplementary video.

4.6 Ablation Study

We compare our approach against using numerical gradients and the Laplacian in the curvature loss [56], and against a hybrid positional encoding [23]. The

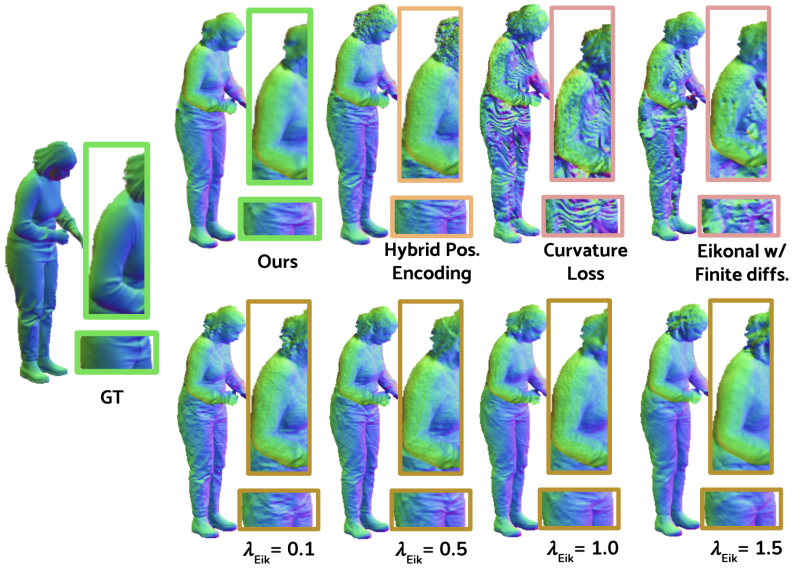


Fig. 5: Qualitative comparison of SDF regularization schemes. From left and right, top to bottom: ours, hybrid positional encoding [23], curvature loss and finite differences derivatives [56], and varying weights for Eikonal loss [27].

proposed SDF smoothing technique significantly boosts quality (see Table 6, and Figures 5 and 7). Even after training for several hours, the numerical gradients and the curvature loss fail to yield satisfactory results (Figure 3). Tweaking the weight of the Eikonal loss term is insufficient for good reconstruction quality (Figure 5). Table 7 shows that our approach keeps the same speed as the base model and incurs less memory overhead than other approaches.

5 Conclusions

We proposed InstantGeoAvatar, a method that can model geometry and appearance of animatable human avatars from monocular videos in less than 10 minutes. Building on previous work, we notice that hash grid-based representations lack the implicit regularization that MLP-based architectures enjoy, and introduce a surface regularization term in the optimization that effectively enhances the learned representation without additional computational cost.

Acknowledgements

This work has been supported by the project MOHUCO PID2020-120049RB-I00 funded by MCIU/AEI/10.13039/501100011033, and by the project GRAVATAR PID2023-151184OB-I00 funded by MCIU/AEI/10.13039/501100011033 and by ERDF, UE.

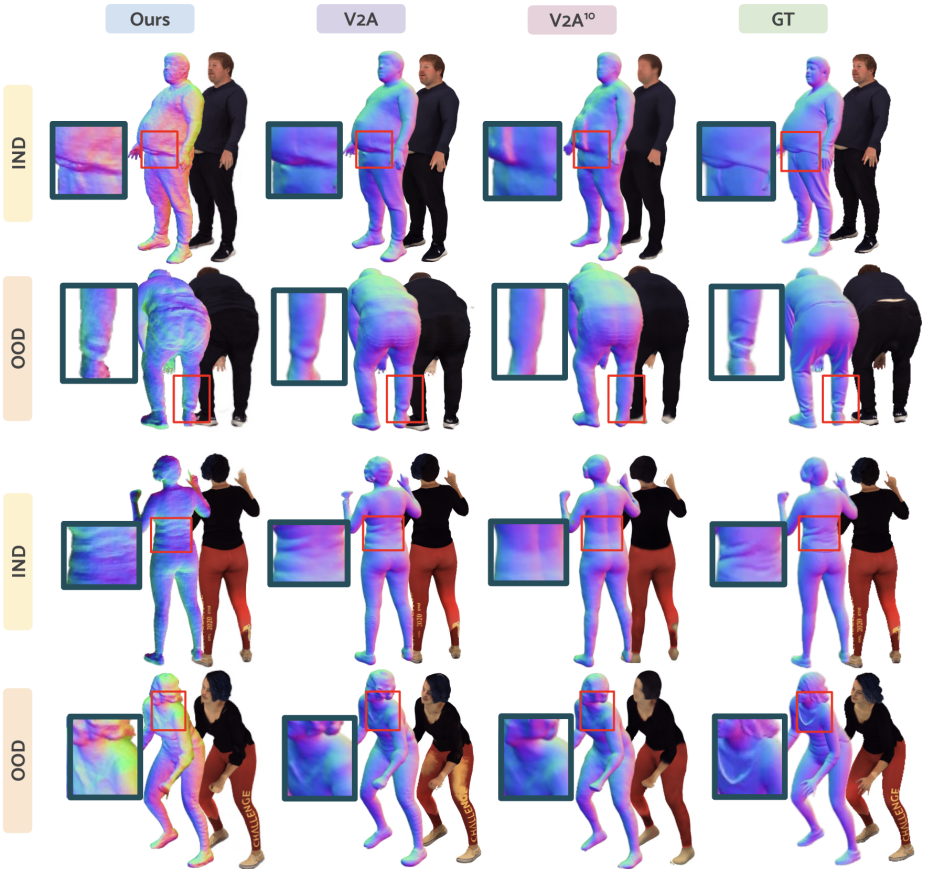
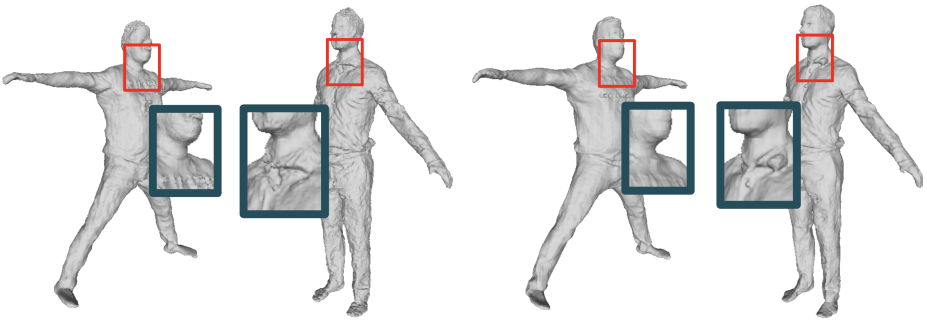


Fig. 6: Qualitative Results on Rendering and Reconstruction. V2A produces very accurate and smooth reconstructions after several hours of training, while V2A^{10} fails to represent higher frequency details like the letters on the second subject's leg. InstantGeoAvatar can produce satisfying results in less than ten minutes.



(a) Without smoothing term. **(b)** With smoothing term.

Fig. 7: Importance of Smooth Surface Regularization.

References

1. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: Proc. SIGGRAPH (2008)
2. Alldieck, T., Magnor, M., 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) (jun 2019)
3. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR Spotlight Paper
4. 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)
5. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
6. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3D human reconstruction. In: European Conference on Computer Vision (ECCV). Springer (August 2020)
7. 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. Advances in Neural Information Processing Systems **33**, 12909–12922 (2020)
8. Bozic, A., Palafox, P., Zollhofer, M., Thies, J., Dai, A., Niessner, M.: Neural deformation graphs for globally-consistent non-rigid reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1450–1459 (June 2021)
9. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless Garment Capture. ACM Transactions on Graphics (Proc. SIGGRAPH 2008) **27**(3), 99 (2008)
10. Burov, A., Nießner, M., Thies, J.: Dynamic surface function networks for clothed human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10754–10764 (2021)
11. Cai, H., Feng, W., Feng, X., Wang, Y., Zhang, J.: Neural surface reconstruction of dynamic scenes with monocular RGB-D camera. Advances in Neural Information Processing Systems **35**, 967–981 (2022)
12. Casado-Elvira, A., Comino Trinidad, M., Casas, D.: PERGAMO: Personalized 3d garments from monocular video. Computer Graphics Forum (Proc. of SCA), 2022 (2022)
13. Chen, D., Lu, H., Feldmann, I., Schreer, O., Eisert, P.: Dynamic multi-view scene reconstruction using neural implicit surface. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
14. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular RGB videos (2021)
15. Chen, M., Zhang, J., Xu, X., Liu, L., Cai, Y., Feng, J., Yan, S.: Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In: European Conference on Computer Vision. pp. 222–239. Springer (2022)
16. 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 **45**(10), 11796–11809 (2023)

17. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: International Conference on Computer Vision (ICCV) (2021)
18. Chen, Y., Wang, X., Chen, X., Zhang, Q., Li, X., Guo, Y., Wang, J., Wang, F.: UV volumes for real-time rendering of editable free-view human performance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16621–16631 (2023)
19. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3D shape reconstruction and completion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6970–6981 (2020)
20. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. In: ACM Trans. Graph. vol. 34. Association for Computing Machinery, New York, NY, USA (jul 2015)
21. Dong, Z., Guo, C., Song, J., Chen, X., Geiger, A., Hilliges, O.: PINA: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20470–20480 (2022)
22. Dou, M., Davidson, P., Fanello, S.R., Khamis, S., Kowdle, A., Rhemann, C., Tankovich, V., Izadi, S.: Motion2fusion: real-time volumetric performance capture. ACM Trans. Graph. **36**(6) (nov 2017)
23. Engelhardt, A., Raj, A., Boss, M., Zhang, Y., Kar, A., Li, Y., Sun, D., Brualla, R.M., Barron, J.T., Lensch, H., et al.: SHINOBI: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. arXiv preprint arXiv:2401.10171 (2024)
24. Fan, J., Zhang, J., Hou, Z., Tao, D.: AniPixel: Towards animatable pixel-aligned human avatar. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 8626–8634. MM '23, Association for Computing Machinery, New York, NY, USA (2023)
25. Gao, Q., Wang, Y., 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 (2023)
26. Geng, C., Peng, S., Xu, Z., Bao, H., Zhou, X.: Learning neural volumetric representations of dynamic humans in minutes. In: CVPR (2023)
27. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3789–3799. PMLR (13–18 Jul 2020), `https://proceedings.mlr.press/v119/gropp20a.html`
28. Guo, C., Chen, X., Song, J., Hilliges, O.: Human performance capture from monocular video in the wild. In: 2021 International Conference on 3D Vision (3DV). pp. 889–898 (2021)
29. 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 (CVPR) (June 2023)
30. Guo, K., Lincoln, P., Davidson, P.L., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S.R., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P.E., Izadi, S.: The relightables: volumetric performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6), 217:1–217:19 (2019)

31. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-PiFU: Geometry and pixel aligned implicit functions for single-view human reconstruction. Advances in Neural Information Processing Systems **33**, 9276–9287 (2020)
32. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: ARCH++: Animation-ready clothed human reconstruction revisited. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11026–11036 (2021). https://doi.org/10.1109/ICCV48922.2021.01086
33. Heo, H., Kim, T., Lee, J., Lee, J., Kim, S., Kim, H.J., Kim, J.H.: Robust camera pose refinement for multi-resolution hash encoding. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
34. Hilton, A., Starck, J.: Multiple view reconstruction of people. In: Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. pp. 357–364 (2004)
35. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D gaussians. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
36. 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 (CVPR) (June 2020)
37. Iqbal, U., Caliskan, A., Nagano, K., Khamis, S., Molchanov, P., Kautz, J.: RANA: Relightable articulated neural avatars. In: ICCV (2023)
38. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16922–16932 (2023)
39. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: European Conference on Computer Vision. pp. 402–418. Springer (2022)
40. Jiang, Y., Shen, Z., Wang, P., Su, Z., Hong, Y., Zhang, Y., Yu, J., Xu, L.: HiFi4G: High-fidelity human performance rendering via compact gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19734–19745 (June 2024)
41. Kanade, T., Rander, P., Narayanan, P.: Virtualized reality: constructing virtual worlds from real scenes. vol. 4, pp. 34–47 (1997). https://doi.org/10.1109/93.580394
42. Kant, Y., Siarohin, A., Guler, R.A., Chai, M., Ren, J., Tulyakov, S., Gilitschenski, I.: Invertible neural skinning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8715–8725 (2023)
43. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
44. Kim, H., Nam, H., Kim, J., Park, J., Lee, S.: LaplacianFusion: Detailed 3d clothed-human body reconstruction. ACM Trans. Graph. **41**(6) (nov 2022)
45. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: HUGS: Human gaussian splatting. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), https://arxiv.org/abs/2311.17910
46. 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)
47. 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** (2021)

48. Kwon, Y., Liu, L., Fuchs, H., Habermann, M., Theobalt, C.: DELIFFAS: Deformable light fields for fast avatar synthesis. Advances in Neural Information Processing Systems (2023)
49. Leroy, V., Franco, J.S., Boyer, E.: Volume Sweeping: Learning Photoconsistency for Multi-View Shape Reconstruction. In: International Journal of Computer Vision. vol. 129, pp. 1–16 (02 2021)
50. Li, C., Zhao, Z., Guo, X.: ArticulatedFusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
51. Li, D., Shao, T., Wu, H., Zhou, K.: Shape completion from a single RGB-D image. IEEE Transactions on Visualization and Computer Graphics **23**(7), 1809–1822 (2017)
52. Li, M., Tao, J., Yang, Z., Yang, Y.: Human101: Training 100+fps human gaussians in 100s from 1 view (2023)
53. Li, R., Dumery, C., Guillard, B., Fua, P.: Garment recovery with shape and deformation priors (2023)
54. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: TAVA: Template-free animatable volumetric actors. In: European Conference on Computer Vision. pp. 419–436. Springer (2022)
55. Li, X., Fan, Y., Xu, D., He, W., Lv, G., Liu, S.: SFNet: Clothed human 3D reconstruction via single side-to-front view RGB-D image. In: 2022 8th International Conference on Virtual Reality (ICVR). pp. 15–20 (2022)
56. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: NeuralAngelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
57. Li, Z., Zheng, Z., Liu, Y., Zhou, B., Liu, Y.: PoseVocab: Learning joint-structured pose embeddings for human avatar modeling. In: ACM SIGGRAPH Conference Proceedings (2023)
58. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable Gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
59. Li, Z., Zheng, Z., Zhang, H., Ji, C., Liu, Y.: Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In: European Conference on Computer Vision. pp. 322–341. Springer (2022)
60. Lin, L., Zhu, J.: Semantic-preserved point-based human avatar. arXiv preprint arXiv:2311.11614 (2023)
61. Lin, S., Zhang, H., Zheng, Z., Shao, R., Liu, Y.: Learning implicit templates for point-based clothed human modeling. In: European Conference on Computer Vision. pp. 210–228. Springer (2022)
62. Lin, W., Zheng, C., Yong, J.H., Xu, F.: Relightable and animatable neural avatars from videos. AAAI (2024)
63. Liu, S., Lin, S., Lu, J., Saha, S., Supikov, A., Yip, M.: BAA-NGP: Bundle-adjusting accelerated neural graphics primitives. arXiv preprint arXiv:2306.04166 (2023)
64. Liu, Y., Dai, Q., Xu, W.: A point-cloud-based multiview stereo algorithm for free-viewpoint video. In: IEEE Transactions on Visualization and Computer Graphics. vol. 16, pp. 407–418 (2010)
65. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. Association for Computing Machinery, New York, NY, USA, 1 edn. (2023)

66. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: SCALE: Modeling clothed humans with a surface codec of articulated local elements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16082–16093 (2021)
67. Ma, Q., Yang, J., Black, M.J., Tang, S.: Neural point-based shape modeling of humans in challenging clothing. In: 2022 International Conference on 3D Vision (3DV). pp. 679–689. IEEE (2022)
68. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10974–10984 (2021)
69. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: LEAP: Learning articulated occupancy of people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10461–10471 (2021)
70. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
71. Moon, G., Nam, H., Shiratori, T., Lee, K.M.: 3D clothed human reconstruction in the wild. In: European conference on computer vision. pp. 184–200. Springer (2022)
72. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022)
73. Newcombe, R.A., Fox, D., Seitz, S.M.: DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
74. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality. pp. 127–136 (2011). https://doi.org/10.1109/ISMAR.2011.6092378
75. Oechsle, M., Peng, S., Geiger, A.: UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: International Conference on Computer Vision (ICCV) (2021)
76. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: ASH: Animatable gaussian splats for efficient and photoreal human rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1165–1175 (June 2024)
77. Peng, S., Xu, Z., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Bao, H., Zhou, X.: Animatable implicit neural representations for creating realistic avatars from videos. TPAMI (2024)
78. Pesavento, M., Volino, M., Hilton, A.: Super-resolution 3D human shape from a single low-resolution image. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II. pp. 447–464. Springer (2022)
79. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4d clothing capture and retargeting. ACM Transactions on Graphics, (Proc. SIGGRAPH) **36**(4) (2017)
80. Prokudin, S., Ma, Q., Raafat, M., Valentin, J., Tang, S.: Dynamic point fields. arXiv preprint arXiv:2304.02626 (2023)

81. Qian, S., Xu, J., Liu, Z., Ma, L., Gao, S.: UNIF: United neural implicit functions for clothed human reconstruction and animation. In: European Conference on Computer Vision. pp. 121–137. Springer (2022)
82. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3DGS-Avatar: Animatable avatars via deformable 3d gaussian splatting (2024)
83. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International Conference on Machine Learning. pp. 5301–5310. PMLR (2019)
84. Ramasinghe, S., MacDonald, L.E., Lucey, S.: On the frequency-bias of coordinate-MLPs. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 796–809 (2022)
85. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PiFU: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019)
86. Saito, S., Simon, T., Saragih, J., Joo, H.: PiFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
87. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In: Computer Vision and Pattern Recognition (CVPR) (2024)
88. Shen, K., Guo, C., Kaufmann, M., Zarate, J.J., Valentin, J., Song, J., Hilliges, O.: X-avatar: Expressive human avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16911–16921 (2023)
89. Song, C., Wandt, B., Helge, R.: Pose modulated avatars from video. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
90. Song, D.Y., , Lee, H., Seo, J., Cho, D.: DIFu: Depth-guided implicit function for clothed human reconstruction (2023)
91. Starck, J., Hilton, A.: Surface capture for performance-based animation. vol. 27, pp. 21–31 (2007). https://doi.org/10.1109/MCG.2007.68
92. Su, S.Y., Bagautdinov, T., Rhodin, H.: DANBO: Disentangled articulated neural body representations via graph neural networks. In: European Conference on Computer Vision (2022)
93. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In: Advances in Neural Information Processing Systems (2021)
94. Tsiminaki, V., Franco, J.S., Boyer, E.: High resolution 3d shape texture from multiple videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
95. Tu, L.W.: Differential geometry: Connections, curvature, and characteristic classes. Springer (2017)
96. Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., Matusik, W.: Dynamic shape capture using multi-view photometric stereo. In: ACM SIGGRAPH Asia. SIGGRAPH Asia '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1661412.1618520, https://doi.org/10.1145/1661412.1618520
97. Wand, M., Adams, B., Ovsjanikov, M., Berner, A., Bokeloh, M., Jenke, P., Guibas, L., Seidel, H.P., Schilling, A.: Efficient reconstruction of nonrigid shape and mo-

tion from real-time 3d scanner data. vol. 28. Association for Computing Machinery, New York, NY, USA (may 2009)
98. Wang, C., Kang, D., Cao, Y.P., Bao, L., Shan, Y., Zhang, S.H.: Neural point-based volumetric avatar: Surface-guided neural points for efficient and photorealistic volumetric head avatar. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–12 (2023)
99. Wang, L., Zhao, X., Sun, J., Zhang, Y., Zhang, H., Yu, T., Liu, Y.: StyleAvatar: Real-time photo-realistic portrait avatar from a single video. In: ACM SIGGRAPH 2023 Conference Proceedings (2023)
100. Wang, L., Zhao, X., Yu, T., Wang, S., Liu, Y.: NormalGAN: Learning detailed 3D human from a single RGB-D image. In: ECCV (2020)
101. 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)
102. Wang, S., Antić, B., Geiger, A., Tang, S.: IntrinsicAvatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. arXiv.org **2312.05210** (2023)
103. Wang, S., Schwarz, K., Geiger, A., Tang, S.: ARAH: Animatable volume rendering of articulated human sdfs. In: European Conference on Computer Vision (2022)
104. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16210–16220 (June 2022)
105. Wu, C., Varanasi, K., Liu, Y., Seidel, H.P., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In: 2011 International Conference on Computer Vision. pp. 1108–1115 (2011). https://doi.org/10.1109/ICCV.2011.6126358
106. Xiang, T., Sun, A., Wu, J., Adeli, E., Fei-Fei, L.: Rendering humans from object-occluded monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3239–3250 (2023)
107. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: ECON: Explicit Clothed humans Optimized via Normal integration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
108. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13296–13306 (June 2022)
109. Xu, H., Alldieck, T., Sminchisescu, C.: H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. Advances in Neural Information Processing Systems **34**, 14955–14966 (2021)
110. Xu, T., Fujita, Y., Matsumoto, E.: Surface-aligned neural radiance fields for controllable 3d human synthesis. In: CVPR (2022)
111. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
112. Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y.: BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 910–919 (2017)
113. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors. In: Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5746–5756 (June 2021)
114. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7287–7296 (2018)
115. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: MonoHuman: Animatable human neural field from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16943–16953 (2023)
116. Zakharkin, I., Mazur, K., Grigorev, A., Lempitsky, V.: Point-based modeling of human clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14718–14727 (2021)
117. Zhang, H., Lin, S., Shao, R., Zhang, Y., Zheng, Z., Huang, H., Guo, Y., Liu, Y.: CloSET: Modeling clothed humans on continuous surface with explicit template decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
118. Zhang, Y., Funkhouser, T.: Deep depth completion of a single RGB-D image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 175–185 (2018)
119. Zhao, X., Hu, Y.T., Ren, Z., Schwing, A.G.: Occupancy planes for single-view RGB-D human reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3633–3641 (2023)
120. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21057–21067 (2023)
121. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: PointAvatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21057–21067 (June 2023)
122. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction (2021)
123. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: AvatarRex: Real-time expressive full-body avatars. ACM Transactions on Graphics (TOG) **42**(4) (2023)
124. Zhi, Y., Qian, S., Yan, X., Gao, S.: Dual-Space NeRF: Learning animatable avatars and scene lighting in separate spaces. In: International Conference on 3D Vision (3DV) (Sep 2022)
125. Zhu, H., Zheng, Z., Zheng, W., Nevatia, R.: CAT-NeRF: Constancy-aware tx2former for dynamic body modeling. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 6619–6628 (2023)
126. Zhu, H., Liu, F., Zhang, Q., Cao, X., Ma, Z.: RHINO: Regularizing the hash-based implicit neural representation. arXiv preprint arXiv:2309.12642 (2023)
127. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4574–4584 (2022), https://api.semanticscholar.org/CorpusID:253761096