JacobianAvatar: Temporally Consistent Semi-rigid Avatar Reconstruction from a Monocular Video

Changyeon Won[1], Min-Gyu Park[2,3], Seonghwan Park[2], Ju Hong Yoon[2,3], and Hae-Gon Jeon[4⋆]

[1]Department of AI Convergence, Gwangju Institute of Science and Technology
[2]Korea Electronics Technology Institute (KETI) [3]polygom
[4]Department of Artificial Intelligence, Yonsei University
cywon1997@gm.gist.ac.kr,
{mpark,shpark97,jhyoon}@keti.re.kr,
earboll@yonsei.ac.kr

Fig. 1: JacobianAvatar. A neural representation for digital human avatars that captures rigid articulated motions and non-rigid local deformations using hierarchical neural Jacobian fields, while encouraging temporal consistency with high-fidelity geometry. The top and bottom rows show the rendered color images and normal maps of an animated avatar.

Abstract. Generating realistic human avatars in complex motions—such as clothing dynamics—requires modeling of global and local deformations which remains challenging in monocular settings. We address this problem by leveraging neural Jacobian fields (NJFs) for representing semi-rigid deformations. We train self-supervised neural networks for predicting Jacobian matrices that give the pose-dependent deformations, by solving a Poisson equation. However, monocular input presents several difficulties such as self-occluded regions and invisible surfaces. To address these issues, we introduce three key components: a constrained Poisson solver, signed distance-based Jacobian regularization, and a deformation-guided residual flow loss, which together suppress boundary artifacts, recover frequently occluded regions such as armpits and thighs, and enforce

⋆ Corresponding author

temporal consistency during motion. Experiments on benchmark and in-the-wild videos demonstrate that our method generates temporally stable and geometrically coherent avatars, outperforming state-of-the-art approaches.

Keywords: Avatar Reconstruction · Temporal consistency · Semi-rigid deformations

1 Introduction

Reconstructing photorealistic and animatable human avatars from a monocular video has received a huge attention, and recent advances have significantly improved the fidelity of clothed human modeling by adopting a semi-rigid deformation paradigm [3,27,44]. This paradigm combines articulated skeletal motion governed by linear blend skinning (LBS) [25, 30] with non-rigid local deformations to capture clothing wrinkles and pose-dependent surface details.

The semi-rigid reconstruction problem naturally decomposes into two tightly coupled subproblems. The first is the recovery of a canonical model that captures the subject's intrinsic shape and appearance in a resting state. The second is the prediction of non-rigid local deformations, such as clothing wrinkles and pose-dependent correctives [44], which is crucial for realistic rendering of an avatar in motion. The majority of the studies assume the body pose is the dominant source of causing local deformations [7,10,13,15,27,39–41,48]. This pose-driven approach, therefore, usually regresses local surface offsets conditioned on estimated human pose parameters [30]. Although these methods are effective in recovering pose-induced effects, they inherently overlook history-dependent deformations arising from factors beyond instantaneous pose, such as cloth inertia, contact, or motion velocity. Time-driven methods address this limitation by incorporating temporal context, modeling deformation as a function of joint velocities or sequential latent codes to better capture dynamic surface behaviors. Since many approaches rely on Gaussian Splats [17] and NeRF [48] to represent a human avatar they heavily rely on the photometric reconstruction loss which does not encourage temporal consistency and geometric details, though rendered images look plausible. To alleviate this problem, several studies [3,38] adopt 2D vision foundation models [18,35] to improve the fidelity of geometry.

From the perspective of dynamic object reconstruction [12,28], which prioritizes underlying geometry recovery rather than synthesizing novel view images, moving objects can be reconstructed if dense correspondences can be found between consecutive frames. These correspondences can be parameterized as 6-DoF rigid transformations [12,21], optical flow between images [42,46], or can be implicitly embedded into learnable motion bases [19,43]. While effective for general dynamic scenes, these strategies remain suboptimal for avatar reconstruction because of frequent self-occlusions and non-rigid deformations.

In this paper, we propose neural avatar representation named as JacobianAvatar providing semi-rigid deformations of a human body, plausible rendering,

and temporal consistency, from a single monocular video. The core of JacobianAvatar is the use of hierarchical neural Jacobian fields (NJF) [2] that encode per-triangle deformation gradients conditioned on body poses. We first train coarse Jacobian fields to learn large and global deformations, and then, find Jacobian fields capture high-frequency details subtle local deformations such as wrinkles, enabling scalable and topology-preserving animation of human body. Furthermore, to effectively address the inherent ambiguities of sparse monocular observations, we introduce three geometric constraints. First, we modify the standard Poisson solver of NJF by incorporating a screening term to update the neural Jacobian fields from partial observations. Second, we propose a signed distance function (SDF)-based Jacobian regularizer, which preserves the global shape of the avatar by regularizing the surface normals of its canonical mesh. Finally, to enforce temporal consistency, we formulate a deformation-guided residual flow loss which minimizes the discrepancy between the projected 2D motions of our 3D vertex flows—derived from the NJF deformations—and dense 2D optical flow estimations [46]. By integrating expressive deformation modeling, effective shape constraints, and geometrically grounded temporal regularization, our approach generates fully animatable and temporally coherent avatars with fine-grained dynamic detail from limited input. We experimentally verify the performance of our method against state-of-the-art methods [7, 27, 40, 41] in geometric fidelity, deformation realism, and perceptual quality.

2 Related Work

There are many approaches to generate animatable human avatars from images or videos. One of the most popular approach is to utilize the template human models [25,31] as they give general information of articulated motion, shape, and expressions. With the aid of template human models, many studies adopt recent technologies such as neural radiance fields (NeRF), 3D Gaussian Splatting and diffusion networks, to better recover accurate and realistic avatars in motion.

For making an animatable avatar, it is essential to define an avatar in the canonical space, this can be done by inverse warping reconstructed 3D models with pre-defined LBS weights of a template model. SCANimate [34] introduces a weakly supervised framework for learning pose-conditioned clothed avatar networks from raw scans, optimizing an implicit occupancy field of a canonical model. SCALE [26] represents clothed humans as a hierarchical surface codec comprising articulated local elements, enabling decoupled modeling of body and garment dynamics. SNARF [4] proposes a differentiable forward skinning formulation that bridges neural implicit representations and LBS, allowing gradient-based optimization of non-rigid surface deformation.

The majority of studies utilize monocular videos to capture both global and local deformations using various neural scene representations. NeuMan [15] introduces neural human radiance fields from monocular video with human-centric regularizers. SelfRecon [13] proposes fully self-supervised training through canonical pose alignment and forward-backward deformation consistency. Vid2Avatar [7]

proposes a framework for learning avatars from in-the-wild monocular videos, leveraging implicit neural representation with a ray opacity sparsity regularizers for robust geometric representation. CanonicalFusion [37] generates drivable 3D human avatars from multiple images via canonical space fusion, but it limits its expressive local deformations and results in overly smooth clothing details. More recently, some works [40, 41] specifically target semi-rigid deformations. LSAvatar [40] represents the avatar through pose conditioning using graph neural networks to encode locality-sensitive deformation. FacAvatar [41] decouples deformation into coarse and fine deformations by controlling the frequencies of positional encodings. However, being built on NeRF architectures, they inevitably suffer from slow rendering and high-frequency artifacts stemming from tightly coupled geometry and texture representations.

Moreover, a growing number of studies employ 3D Gaussian Splatting (3DGS) to achieve real-time rendering speeds while maintaining high visual quality for human avatars. ExAvatar [27] extends this paradigm to expressive whole-body reconstruction by anchoring 3D Gaussians to a canonical human template mesh. PERSONA [38] generalizes the framework to generated monocular videos, optimizing identity-preserving canonical geometry from a single reference image. Vid2Avatar-Pro [8] reconstructs avatars by leveraging universal priors, achieved through the distillation of human motion and shape information from a large-scale capture dataset. GaussianAvatar [10] models dynamic clothed humans using animatable 3D Gaussians, integrating forward skinning with learned pose-dependent deformation fields to capture non-rigid garment motion. Animatable Gaussians [23] learns a compact pose-conditioned Gaussian attribute map of the avatar via a lightweight convolutional neural networks, enabling the efficient modeling of dynamic 3DGS parameters. GomAvatar [47] and SplattingAvatar [36] represent avatars by anchoring 3D Gaussians to a mesh, achieving real-time rendering while preserving high-frequency details of the avatar. Constrained by the fixed topology and intrinsic rigidity of LBS, these methods may struggle to capture complex non-rigid dynamics like wrinkles and highly deformable garments.

On the other hand, diffusion models provide powerful generative priors to resolve geometric and textural ambiguity in occluded regions. HumanGaussian [24] generates text-driven 3D clothed humans via 3D Gaussian splatting optimized using score distillation sampling (SDS) from pretrained 2D diffusion models. DreamWaltz [11] synthesizes complex animatable avatar sequences by combining diffusion-guided motion trajectory sampling with static geometry reconstruction, though dynamic deformation remains limited. PSHuman [22] integrates cross-scale multiview diffusion with SMPL-X body conditioning and explicit remeshing to enforce anatomical consistency, resolve self-occlusions, and produce watertight, animation-ready meshes. SiTH [9] employs image-conditioned diffusion with learned skinned body priors to hallucinate photorealistic textures and normal maps in unseen viewpoints. HumanRef [52] introduces reference-guided diffusion with spatially adaptive region-aware attention to preserve identity-specific details during texture synthesis. ReconFusion [49] regularizes sparse-view NeRF

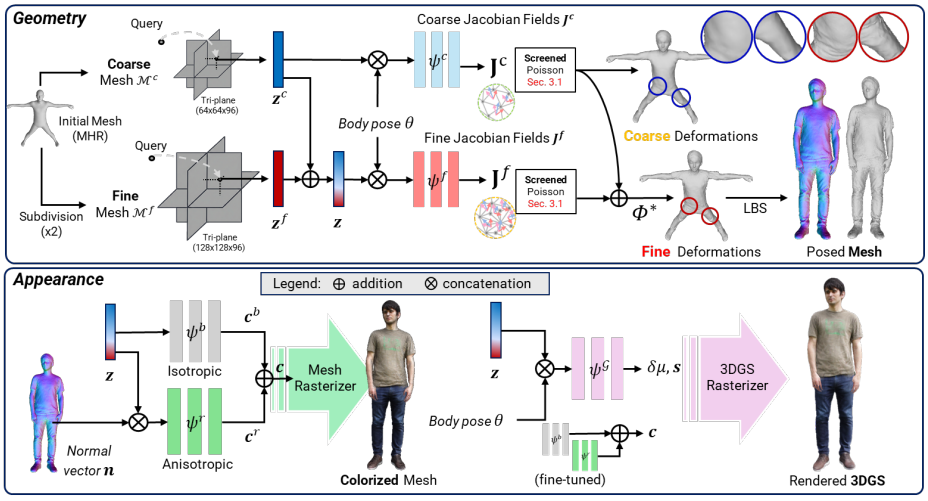

Fig. 2: Overview of our pipeline. We first initialize our canonical avatar using a human template mesh and refine it through mesh optimization. Next, we model semi-rigid deformations using two Jacobian fields integrated with a screened Poisson solver, which separately capture deformations in a coarse-to-fine manner. For appearance, we model the mesh textures as normal-conditioned colors. Finally, we incorporate 3DGS anchored to the mesh faces for a more photorealistic rendering of our avatar.

reconstruction using 3D-aware diffusion priors, enabling robust novel-view synthesis and geometry completion.

Despite substantial progress, existing methods suffer from inherent difficulties. Parametric modeling is limited by template topology and LBS rigidity, failing to represent complex semi-rigid deformations. We address these challenges by introducing a monocular framework that learns fine-grained local non-rigid deformations from sparse observations. Our approach integrates residual flow fields to enforce sub-pixel temporal consistency beyond photometric losses and regularizes deformation using learned neural Jacobian fields with screened Poisson solver that preserve local area and orientation, enabling robust, topologically flexible reconstruction of both observed and unobserved geometry from monocular video.

3 Proposed Method

We present JacobianAvatar that predicts neural Jacobian fields capturing semi-rigid deformations governed by body pose. Initially, we define a canonical mesh, a human model in a T-pose or Da-pose [15], that can be animated using body pose parameters. We initialize the canonical mesh using the momentum human rigs (MHR) [6] template mesh and existing 3D generation pipelines [50, 53] can be also employed. Prior to training neural networks, we update the canonical

mesh and body poses via differentiable rendering and mesh optimization [29] to reduce the gap between the avatar and the template model. Afterward, we train two neural networks that predict Jacobian fields. To this end, we suggest a constrained Poisson solver with signed distance-based regularizer and residual flow-based loss function. Finally, we train an additional network for predicting Gaussians attached to each face. Therefore, the avatar can be animated and rendered in real-time. The overall procedure is illustrated in Fig. 2.

Pre-processing of Canonical Mesh and Body Poses. Due to imperfections of the initial canonical mesh and body poses, we refine them by using a differentiable mesh renderer [20]. In this step, we jointly optimize the vertex positions and body poses across all video frames, guided by normal maps and segmentation masks predicted by Sapiens [18]. Formally, each vertex in the canonical mesh is transformed by LBS weights,

$$V_i^\theta = LBS(V_i^c, \theta) = \sum_{j=1}^{K} w_{i,j} G_j(\theta) V_i^c, \tag{1}$$

where V_i^c is the position of the i-th vertex in the canonical avatar, θ is the body pose of K joints, V_i^θ is the deformed V_i^c with θ, $G_j(\theta) \in \mathbb{R}^{4\times4}$ is the bone transformation matrix of the j-th joint derived from θ, $w_{i,j}$ is the skinning weight of i-th vertex with respect to the j-th joint. Then, we minimize errors between rendered and predicted normal maps and masks. After the convergence, we define the updated canonical mesh as $\mathcal{M}^c$, and subdivide the same mesh twice, denoted as $\mathcal{M}^f$. Although we used MHR as a template model, LBS weights are adopted from SMPL-X [30] which are propagated using robust skin weight inpainting [1].

3.1 Semi-rigid Deformation with Neural Jacobian Fields

Let the canonical triangle mesh obtained after the pre-processing step as $\mathcal{M} = \{\mathcal{V}, \mathcal{F}\}$, where $\mathcal{V} \in \mathbb{R}^{N\times3}$ and $\mathcal{F} \in \mathbb{Z}^{M\times3}$ are vertex positions and faces. Following the NJF [2], the mapping function Φ can transform $\mathcal{V}$ to the new positions, which is defined as

$$\Phi^* = \min_{\Phi} \sum_{a_i \in \mathcal{F}} |a_i| \|\nabla_i(\Phi) - J_i\|_2^2, \tag{2}$$

where ∇_i denotes the gradient operator for each face. $|a_i|$ and J_i are the area and a Jacobian matrix of i^{th} face. Jacobian field $\mathbf{J}$ is a set of Jacobian matrices that deforms the canonical mesh. Given $\mathbf{J}$, the optimal deformation map Φ^* can be obtained by solving the Poisson equation [2]. Here, we add a screening term [16] to the objective function,

$$\Phi^* = \min_{\Phi} \left(\sum_{a_i \in \mathcal{F}} |a_i| \|\nabla_i(\Phi) - J_i\|_2^2 + \lambda_\Phi \|\Phi - \mathcal{V}\|_2^2 \right), \tag{3}$$

where λ_Φ is a balancing parameter. The screening term tries to keep the positions of vertices, which is required for invisible and occluded surfaces. In monocular

settings, as a result, the screened Poisson behave more reliably compared to the standard Poisson. The solution to Eq. (3) can be obtained by solving the linear system,

$$(L + \lambda_{\Phi} I)\Phi^* = \mathcal{A}\nabla^T \mathbf{J} + \lambda_{\Phi}\mathcal{V}, \tag{4}$$

where L is cotangent Laplacian matrix, I is identity matrix, and $\mathcal{A}$ is the mass matrix of $\mathcal{M}$. The optimal mapping function can be solved through differentiable Cholesky decomposition. In the following section, we explain how we train neural networks to predict optimal Jacobian fields.

3.2 Self-supervised Training of Neural Jacobian Fields

We train five MLP networks, ψ^c, ψ^f, ψ^b, ψ^r, ψ^g that predict coarse and fine Jacobian fields, colors, and remaining Gaussian properties. First of all, ψ^c and ψ^f are trained to capture both coarse and fine deformations from a video observation. ψ^c predicts a Jacobian matrix for each face at coarse level,

$$\mathbf{z}_i^c = \mathcal{E}^c(x_i^c),\ J_i^c = \psi^c(\mathbf{z}_i^c, \theta), \tag{5}$$

where $\mathcal{E}^c(x_i^c)$ extracts a spatial latent code from a tri-plane, and θ is the body pose parameters. In this step, ψ^c is trained w.r.t. the coarse mesh $\mathcal{M}^c$. Afterward, we train ψ^f to capture fine details,

$$\mathbf{z}_i^f = \mathcal{E}^f(x_j^f),\ J_i^f = \psi^f(\mathbf{z}_i^f + \mathbf{z}_i^c, \theta). \tag{6}$$

Here, the network structure is same as the coarse MLP but the coarse feature $\mathbf{z}_i^c$ is added to the output of $\mathcal{E}^f(x_j^f)$. Since we train ψ^f for a subdivided mesh $\mathcal{M}^f$ with the aid of coarse latent features, this network capture both coarse and fine deformations. For training the appearance of an avatar, we train two MLP networks,

$$\mathbf{c}_i^b = \psi^b(\mathbf{z}_i),\ \mathbf{c}_i^r = \psi^r(\mathbf{z}_i, \mathbf{n}_i),\ \mathbf{c}_i = \mathbf{c}_i^b + \mathbf{c}_i^r, \tag{7}$$

where ψ^b and ψ^r are shallow MLP networks to predict isotropic and anisotropic components of colors. We adopt this idea from ExAvatar [27], instead of using spherical harmonics. Finally, the last network ψ^g predicts scale and offset parameters of Gaussians attached to each face.

Since our representation is a hybrid of mesh and 3DGS, we use both mesh renderer $\mathcal{R}_{\text{mesh}}$ and a GS rasterizer $\mathcal{R}_{\text{gs}}$. We first update the geometry and color with $\mathcal{R}_{\text{mesh}}$,

$$\hat{\mathbf{I}}^t, \hat{\mathbf{N}}^t, \hat{\mathbf{A}}^t = \mathcal{R}_{\text{mesh}}(LBS(\Phi_{\theta_t}(\mathcal{M}), \theta_t), P_t), \tag{8}$$

where the mesh rasterizer renders an image $\hat{\mathbf{I}}^t$, a normal map $\hat{\mathbf{N}}^t$, and a foreground mask $\hat{\mathbf{A}}^t$ of a posed mesh. The mapping functions Φ_θ can be calculated by solving Eq. (4) that gives a deformed mesh. Here, t indicates the time stamp for the body pose θ_t and camera projection matrix P_t because we handle a video sequence. The rendered outputs are compared to self-supervision signals using Sapiens [18]. We train (ψ^c, $\mathcal{E}^c$), (ψ^f, $\mathcal{E}^f$), and (ψ^b, ψ^r) sequentially. Afterward,

we use $\mathcal{R}_{\text{gs}}$ to update Gaussians to recover high-fidelity appearance. We define a set of 3D Gaussians $\mathcal{G}$ as

$$\mathcal{G}_i = \{\boldsymbol{\mu}_i, \delta\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{q}_i, o_i, \mathbf{c}_i\}, \tag{9}$$

where $\boldsymbol{\mu}_i$ is the face center after deformation, *i.e.*, $\Phi_\theta(x_i)$, $\delta\boldsymbol{\mu}_i$ is an offset vector, $\mathbf{s}_i$ is scale, $\mathbf{q}_i$ is quaternion, o_i is opacity, and $\mathbf{c}_i^r$ is the face color calculated in Eq (7). We predict an offset and a scale as,

$$\delta\boldsymbol{\mu}_i, \mathbf{s}_i = \psi^{\mathcal{G}}(\mathbf{z}_i, \theta) \tag{10}$$

where $\psi^{\mathcal{G}}$ is an MLP network and $\mathbf{z}_i$ is from Eq. (6). Finally, the network is updated by minimizing the input image with the rendered image,

$$\hat{\mathbf{I}}^t = \mathcal{R}_{\text{gs}}(LBS(\mathcal{G}, \theta_t), P_t), \tag{11}$$

where $LBS(\cdot)$ warps 3D Gaussians to the body pose of t^{th} frame.

3.3 Training Objectives

We propose two total loss functions for the mesh renderer and GS renderer. The first loss function is defined as,

$$\begin{aligned}\mathcal{L}_{\text{mesh}} =& \lambda_{\text{L1}}\mathcal{L}_{\text{L1}}(\hat{\mathbf{I}}^t, \mathbf{I}^t) + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}}(\hat{\mathbf{I}}^t, \mathbf{I}^t) + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}}(\hat{\mathbf{I}}^t, \mathbf{I}^t) \\ &+ \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}(\hat{\mathbf{A}}^t, \mathbf{A}^t) + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}}(\hat{\mathbf{N}}^t, \mathbf{N}^t) \\ &+ \lambda_{\text{residual}}\mathcal{L}_{\text{residual}}(\mathbf{I}^t, \mathbf{I}^{t-1}, \mathbf{W}^t) + \lambda_{\text{SDF}}\mathcal{L}_{\text{SDF}}(\Phi_{\theta_t}(\mathcal{M})) \\ &+ \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}(\mathbf{J}_{\theta_t}) + \lambda_{\text{lap}_\mathcal{V}}\mathcal{L}_{\text{lap}}(\Phi_{\theta_t}(\mathcal{V}), \mathcal{F}) \\ &+ \lambda_{\mathbf{C}^b}\mathcal{L}_{\text{dual-lap}}(\mathbf{C}^b, \mathcal{F}) + \lambda_{\mathbf{C}}\mathcal{L}_{\text{dual-lap}}(\mathbf{C}, \mathcal{F})\end{aligned} \tag{12}$$

where L_{L1}, L_{lpips}, and L_{ssim} compute L1, perceptual, and structural similarity losses between input and rendered images. L_{normal} is a normal cosine similarity loss with a pseudo ground truth normal map from a pretrained model [18]. L_{mask} is an L1 loss between a rendered foreground mask and a pseudo ground truth mask from a pretrained model [18]. L_{reg} is an L1 regularizer on the pose dependent deformation maps derived from the $\mathbf{J}^c_{\theta_t}$ and $\mathbf{J}^f_{\theta_t}$ by solving the screened Poisson equation (4). $\mathcal{L}_{\text{lap}}$ is a Laplacian regularizer and $\mathcal{L}_{\text{dual-lap}}$ is a dual Laplacian regularizer. For color, we regularize for both anisotropic and combined colors, $\mathbf{C}^b$ and $\mathbf{C}$. Note that this loss function is applied to coarse and fine meshes in a sequential manner. The second loss function is defined as,

$$\begin{aligned}\mathcal{L}_{\text{gs}} =& \lambda_{\text{L1}}L_{\text{L1}}(\hat{\mathbf{I}}^t, \mathbf{I}^t) + \lambda_{\text{lpips}}L_{\text{lpips}}(\hat{\mathbf{I}}^t, \mathbf{I}^t) + \lambda_{\text{ssim}}L_{\text{ssim}}(\hat{\mathbf{I}}^t, \mathbf{I}^t) \\ &+ \lambda_{\text{scale}}L_{\text{scale}}(\mathbf{s}) + \lambda_{\text{offset}}L_{\text{offset}}(\delta\boldsymbol{\mu}) + \lambda_{\text{laps}}L_{\text{lap}}(\mathbf{s}) \\ &+ \lambda_{\text{lap}\delta\boldsymbol{\mu}}L_{\text{lap}}(\delta\boldsymbol{\mu}) + \lambda_{\text{residual}}L_{\text{residual},}(\mathbf{I}^t, \mathbf{I}^{t-1}, \mathbf{W}^t)\end{aligned} \tag{13}$$

where L_{scale} and L_{offset} are L2 regularizers on scale $\mathbf{s}$ and offsets $\delta\boldsymbol{\mu}$, respectively. For temporal consistency, we adapt the residual flow loss on 3D Gaussians. We

obtain a 2D flow map $\mathbf{W}^t$ by assigning a displacement vector as an attribute to each 3D Gaussian and rendering the flow map using alpha-blending.

Here, we propose $\mathcal{L}_{\text{residual}}$ and $\mathcal{L}_{\text{SDF}}$ loss functions to the fidelity and temporal consistency of an avatar, which we explain in detail in the following sections.

Deformation-guided Residual Flow. We claim that relying on per-frame observations is insufficient to capture fine-grained local deformations. Therefore, we propose a temporal consistency regularizer to capture temporal changes. First, we select two frames, *e.g.*, at time t and $t-1$, and determine the difference of projected V_i at different time steps. Then, the projected point can be denoted,

$$\mathbf{u}_i^t = \pi(LBS(\Phi_{\theta_t}(V_i), \theta_t), P_t), \tag{14}$$

where vertex V_i is deformed by the predicted mapping function Φ_{θ_t} followed by the LBS function, and π is a projection function with the camera projection matrix P_t. Then, we further define a 2D displacement vector,

$$\delta\mathbf{u}_i^t = \mathbf{u}_i^t - \mathbf{u}_i^{t-1}, \tag{15}$$

which is the displacement of V_i in pixels. Here, we assign $\delta\mathbf{u}_i^t$ as the vertex attribute; therefore, it gives a 2D flow map between the two frames. We define this as $\mathbf{W}^t$ as the 2D flow map between images. Instead of comparing $\mathbf{W}^t$ to the optical flow between the two frames, we find a residual flow by using a pre-trained optical flow network $\mathcal{F}$ [45],

$$\Delta\mathbf{W}^t = \mathcal{F}(\mathbf{I}^t, \mathbf{I}^{t-1}, \mathbf{W}^t). \tag{16}$$

where $\mathbf{I}$ is a ground truth image. Here, $\mathcal{F}$ search correspondence given initial displacement map $\mathbf{W}^t$, therefore, predicted residual $\Delta\mathbf{W}^t$ becomes zero if $\mathbf{W}^t$ is perfect. To this end, we minimize the residual flow,

$$L_{\text{residual}} = ||\Delta\mathbf{W}^t||_1, \tag{17}$$

to further constrain the temporal consistency between two frames.

Signed distance-based Jacobian regularization. Sparse observations and self-occluded regions of the avatar frequently generate unwanted deformations, such as severe mesh stretching and flipped faces. To address this, we propose a SDF-based regularizer to maintain the global shape of the canonical avatar. To achieve this, we set a voxel grid containing the discrete SDF values of the initial coarse mesh $\mathcal{M}^c$. Then, we compute the spatial gradients of signed distance values at the deformed vertices and faces, and set approximate normals as flipped spatial gradients [33].

Surface normals of the initial mesh by calculating the spatial gradients within the SDF voxel grid. Using these derived normals as a geometric prior, we regularize the face and vertex normals of the deformed canonical avatar via a cosine similarity loss,

$$\mathcal{L}_{\text{SDF}} = \lambda_{\text{face}} \sum_{a_i \in \mathcal{F}} (1 - \hat{\mathbf{n}}_i^a \cdot \mathbf{n}_{\text{SDF}}(x_i)) + \lambda_{\text{vertex}} \sum_{v_j \in \mathcal{V}} \left(1 - \hat{\mathbf{n}}_j^v \cdot \mathbf{n}_{\text{SDF}}(v_j)\right), \tag{18}$$

where x_i is the center of the i-th face of the mesh, and $\mathbf{n}_i^a$ and $\mathbf{n}_j^v$ denote the face and vertex normal vectors of the deformed canonical avatar, respectively. Furthermore, $\hat{\mathbf{n}}_{\text{SDF}}(x)$ is the normal extracted from the SDF voxel grid by normalizing the spatial gradient of the SDF volume at position x. Through these geometric priors, the Jacobian matrices evolve to neighboring iso-surfaces rather than abruptly changing positions.

4 Experimental Results

4.1 Implementation details

We initialize our mesh using the MHR [6] template at Level of Detail (LoD) 3. For the screened Poisson solver, we empirically set λ_Φ to 5 and 1 for the coarse and fine meshes, respectively. The tri-plane resolutions are set to 64 for $\mathcal{E}^c$ and 128 for $\mathcal{E}^f$. Each plane has a feature dimension of 32, and we concatenate the features extracted from all three planes. For stable training, following NJF [2], we also initialize the last layer of both MLPs, ψ^c and ψ^f, with zero weights and biases. We train our avatar on a single RTX 6000 Ada GPU. Including preprocessing, the training process takes about 10 hours, depending on the video length.

4.2 Datasets

We evaluate our method on four public datasets, which include diverse identities, poses, and environments.

MonoPerfCap dataset [51]. The MonoPerfCap dataset is captured in in-the-wild environments, including indoors and outdoors, and contains natural human poses. It is recorded under stable lighting conditions, and subjects typically wear non-loose clothing, making it ideal for evaluating the reconstruction of semi-rigid deformations. This dataset also provides high frame rate videos, yielding temporally dense and continuous motion. Following Vid2Avatar-Pro [8], we split the dataset into training and testing sets, comprising the first 80% of frames and the remaining frames, respectively. This split allows us to evaluate generalization performance to novel views.

SynWild dataset [7]. The SynWild dataset is a synthetic dataset that provides ground-truth meshes, enabling the quantitative evaluation of geometric accuracy. It contains clothing deformations, making it particularly suitable for assessing semi-rigid deformations. Since this dataset lacks camera parameters, we apply Iterative Closest Point (ICP) registration to align reconstructed avatars with the ground-truth meshes, following the evaluation protocol of the previous work [7].

NeuMan dataset [15]. The NeuMan dataset consists of outdoor videos including subjects performing repetitive motions. Compared to MonoPerfCap dataset, the motions are simple and limited, and cloth deformation is less. Besides, the humans move relatively longer distance than other datasets, leading to lighting variations depending on the position. We follow the official data split, which utilize interval views for testing, making it suitable for evaluating novel view interpolation performance.

DNA-Rendering dataset [5] The DNA-Rendering dataset is captured in a multi-view studio with various types of clothing. It contains sequences featuring loose garments with challenging non-rigid deformations. By randomly sampling a single camera per frame, we adapt this dataset to a monocular setting. Using this setup, we evaluate novel view synthesis performance and qualitatively compare mesh reconstruction results with state-of-the-art (SOTA) SDF-based methods.

4.3 Comparisons to State-of-the-art Methods

Comparison methods. To demonstrate the effectiveness of our representation, we compare our method against five methods that achieve SOTA results in geometric accuracy and rendering quality. Vid2Avatar [7], LSAvatar [40] and FacAvatar [41] represent the canonical avatar using a SDF, an approach known for its robustness in capturing geometric details. ExAvatar [27] and GoMAvatar [47] utilizes 3DGS to produce photorealistic rendering results and achieves robustness in novel pose and novel view synthesis by applying a Laplacian regularizer that is computed using the connectivity of the mesh of the canonical avatar.
Evaluation Protocol. We quantitatively compare our method with SOTA methods. To assess rendering quality, we measure PSNR, SSIM, and LPIPS ($\times 100$). In this evaluation, except for DNA-Rendering dataset, we follow a protocol of previous works, which additionally optimizes the body poses for the test set using rendering losses. This evaluation protocol reduces the dependency of the metrics on the accuracy of the body pose. To evaluate geometric accuracy, we measure the Chamfer Distance (CD) ($\times 1000$), Normal Error (NE) and the F1-score at thresholds of 1 cm and 2 cm against the ground-truth meshes. For Normal Error, we render the normals of both the ground-truth and aligned meshes from four orthogonal views (front, back, left, and right), and then measure the cosine similarity in overlapping regions.
Quantitative results. As shown in Tab. 1 (a), our method outperforms the comparison methods in view extrapolation on the MonoPerfCap dataset. This is attributable to our hybrid mesh-3DGS representation, which provides coherent geometry and high-fidelity appearance in unseen views. Furthermore, the high frame rate of this dataset is beneficial for our deformation guided residual flow loss $L_{residual}$, which in turn contributes to the improved overall rendering performance.

For the DNA-Rendering dataset, we evaluate novel view synthesis results for avatars with loose clothing in Tab. 1 (b). Our pipeline effectively reconstructs these avatars by first optimizing the global shape via an initial mesh optimization step, and subsequently utilizing the network to recover fine-level geometry. Providing more accurate geometry compared to baselines, our method achieves the best rendering quality.

In Tab. 1 (c), we achieve results comparable to SOTA methods on the NeuMan dataset, despite the severe lighting changes in the scenes, which do not explicitly consider. The results for all baseline methods with the exception of ExAvatar, are directly borrowed from previous work [8]. We evaluated ExAvatar

Table 1: Quantitative comparison of rendering quality on the MonoPerfCap, DNA-Rendering, and NeuMan datasets.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓
(a) MonoPerfCap	GoMAvatar [47]	28.35	0.975	1.95
	Vid2Avatar [7]	28.74	0.977	2.06
	LSAvatar [40]	27.96	0.975	2.15
	FacAvatar [41]	28.06	0.975	1.99
	ExAvatar [27]	30.47	**0.980**	2.00
	Ours	**31.83**	0.978	**1.67**
(b) DNA-Rendering	Vid2Avatar [7]	29.46	**0.972**	2.88
	LSAvatar [40]	27.94	0.968	2.37
	ExAvatar [27]	28.54	0.970	2.85
	Ours	**29.90**	0.971	**2.19**
(c) NeuMan	HumanNeRF [48]	27.06	0.967	1.92
	InstantAvatar [14]	28.47	0.972	2.77
	NeuMan [15]	25.48	0.966	2.87
	Vid2Avatar [7]	26.87	0.969	2.41
	3DGS-Avatar [32]	29.75	0.975	1.75
	GaussianAvatar [10]	**29.94**	**0.980**	**1.24**
	ExAvatar [27]	29.37	**0.980**	1.87
	Ours	29.75	0.974	1.94

Table 2: Quantitative comparison of geometry accuracy on the SynWild dataset.

Method	Type	CD ↓	NE ↓	F1@1cm ↑	F1@2cm ↑
GoMAvatar [47]	Explicit	9.30	0.140	0.276	0.544
ExAvatar [27]	Explicit	9.39	–	0.307	0.515
Vid2Avatar [7]	Implicit	2.59	**0.091**	0.346	0.647
LSAvatar [40]	Implicit	2.55	0.095	0.343	0.646
FacAvatar [41]	Implicit	2.91	0.106	0.329	0.623
Ours	Explicit	**2.46**	**0.091**	**0.397**	**0.681**

using the pretrained weights provided with the author's reimplementation, without rendering the background to ensure a fair comparison with these baselines.

As shown in Tab. 2, our explicit approach compares favorably against recent implicit methods, achieving the state-of-the-art performance in modeling semi-rigid deformations. Unlike implicit models that struggle to guarantee temporal consistency, we directly deform the mesh via NJF, constraining this explicit deformation through a deformation-guided residual loss. Furthermore, implicit represent may suffer from artifacts during mesh extraction, which limits their ability to preserve high-frequency geometric details.

Qualitative results. We provide qualitative comparisons from both perspectives: geometry and rendering quality of the avatar. In Fig. 3, we compare ren-

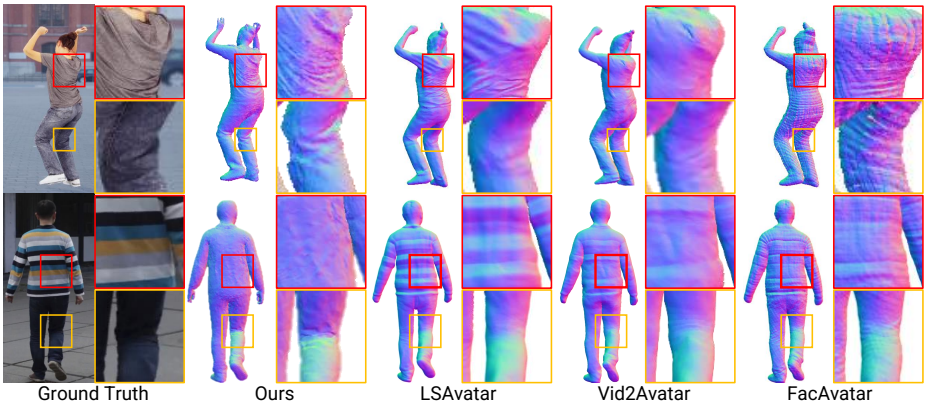

Fig. 3: Qualitative comparison of rendered normal maps. Compared methods [7, 40, 41] suffer from texture-copying artifacts on the geometry, whereas our method shows smooth surfaces in the textured regions.

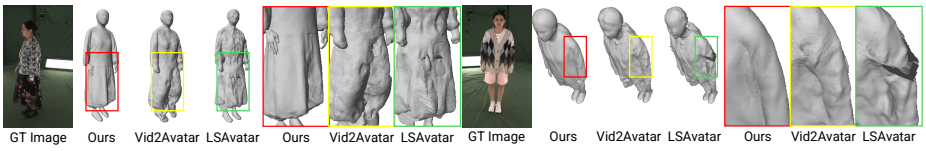

Fig. 4: Qualitative comparison of mesh reconstruction with SDF-based methods [7,40] on the DNA-Rendering dataset.

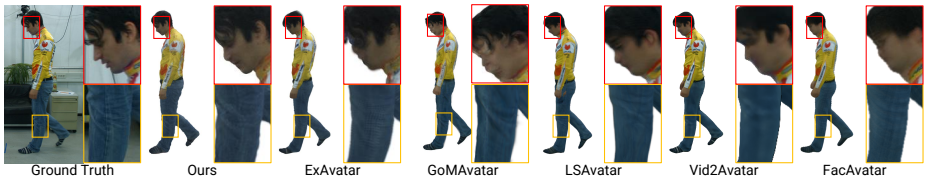

Fig. 5: Qualitative comparison of rendering quality with state-of-the-art methods [7, 27, 40, 41, 47] on the MonoperfCap Dataset.

dered normal maps using the Synwild dataset. As shown in the first row, our method recovers the sharp wrinkles in the pants and the T-shirt, with the aid of neural Jacobian fields. More importantly, we observed texture-copying artifact, where 2D textures are incorrectly baked into the 3D geometry when a human wears highly textured cloth (as shown in the bottom row). This indicates geometry and texture are entangled in the existing approach. In our case, we independently capture semi-rigid deformations without optimizing texture-related variables, which alleviated this phenomenon.

Furthermore, as shown in Fig. 4, our method produces cleaner meshes on the DNA-Rendering dataset, Specifically, the proposed pipeline is free from the

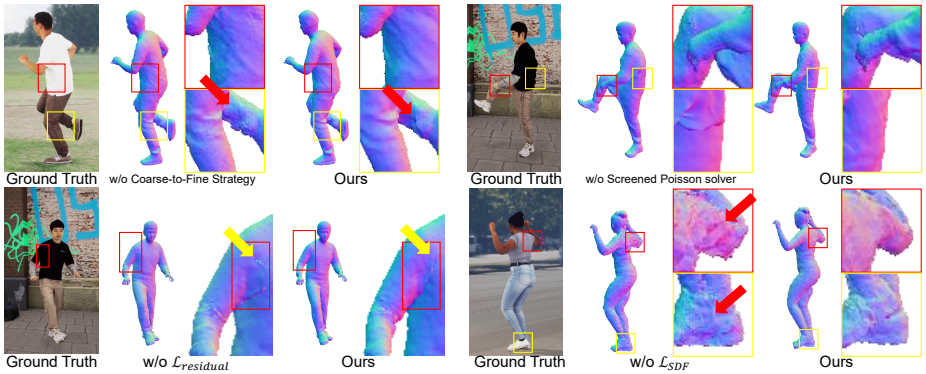

Fig. 6: Ablation study on the Synwild dataset. We visualize the effects of each component of our proposed method on the geometry quality by removing them independently.

Table 3: Ablation studies on the `00000_random` scene of the SynWild dataset. We assess geometry quality by removing each component of our method.

Method	**CD**($\times$1000) $\downarrow$	**NE** $\downarrow$
w/o $\mathcal{L}_{\text{SDF}}$	3.51	0.113
w/o $\mathcal{L}_{\text{residual}}$	<u>3.01</u>	<u>0.106</u>
w/o Screened Poisson solver	3.28	<u>0.106</u>
w/o Coarse-to-fine Strategy	**2.96**	0.107
Ours	<u>3.01</u>	**0.102**

extraction artifacts inherently caused by the Marching Cubes algorithm, which is essential for mesh reconstruction in previous SDF-based approaches.

In Fig. 5, we qualitatively compare rendered images. When the face is seen from the side view, most existing methods show blurry textures, whereas our method shows relatively clean results. This implicitly confirms that the geometry is more accurate compared to SOTA methods. Secondly, in the highlighted pants region, we can observe the proposed method shows sharp and accurate textures. We believe the Jacobian fields are good at capturing local deformations, therefore, attached 3DGS also shows sharp and clean results.

4.4 Ablation study

We analyze the impact of each component of our method qualitatively in Fig. 6 and quantitatively in Tab. 3.

Coarse-to-fine Strategy. Optimizing the fine Jacobian field without coarse Jacobian field introduces high-frequency, noisy deformations on the mesh. As shown in the figure, it also fails to capture prominent wrinkles compared to our full model. Therefore, in the table, it performs well in CD, which is relatively

insensitive to small surface noise, but performs poorly in NE where these noisy deformations are penalized.

Screened Poisson solver. Relying solely on the Poisson solver without our formulation causes large stretching artifacts due to the limited observations from monocular videos. However, regularizing the solver using the initial mesh effectively prevents these artifacts. As these large artifacts result in severe geometric distortions, they cause a degradation in the CD.

Deformation-guided Residual Flow Loss ($\mathcal{L}_{residual}$). This loss is not only helpful for capturing temporal consistency but also plays a crucial role in suppressing artifacts in heavily self-occluded areas like the armpits. Specifically, it mitigates these artifacts by regularizing the noisy flow predictions generated by pretrained models. However, these self-occluded regions occupy only a small portion of the overall body, removing this loss results in remaining a similar CD. However, the resulting local artifacts degrade the surface quality, leading to poor performance in NE.

Signed Distance-based Jacobian Regularization ($\mathcal{L}_{SDF}$). This term regularizes the global shape of the avatar, effectively preventing artifacts in regions with sparse observations. Consequently, without this regularization, the woman's back, which is a rarely visible area in the video, suffers from severe noisy deformations as shown in the figure. Moreover, omitting this loss makes the model susceptible to overfitting to specific viewpoints, leading to an overall degradation in mesh quality. As a result, the absence of this regularizer causes a performance drop in both CD and NE.

5 Conclusion

We have proposed a robust framework for generating drivable, semi-rigid avatars from monocular video. By integrating Neural Jacobian Fields (NJF) with a screened Poisson solver, and introducing a deformation-guided residual flow with SDF-based regularization, our method captures semi-rigid deformations while preventing geometric artifacts. We demonstrate effectiveness of our work through comparisons with the-state-of-the-art methods on various scenes.

Limitations and Future Work. Our method struggles to reconstruct avatars with loose-fitting clothing. However, since it relies on the LBS weights from the template model, *i.e.*, a naked body, loose clothes that behave differently from the inner body cannot be modeled. Next, our current model does not encompass facial expressions. We leave these extensions as future work, where existing facial reconstruction method can be integrated into our framework.

Acknowledgement This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.RS-2024-00398830, RS-2020-II201361 and No.RS-2025-25441838) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00338439).

References

1. Abdrashitov, R., Raichstat, K., Monsen, J., Hill, D.: Robust skin weights transfer via weight inpainting. In: SIGGRAPH Asia 2023 Technical Communications. SA '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3610543.3626180, https://doi.org/10.1145/3610543.3626180
2. Aigerman, N., Gupta, K., Kim, V.G., Chaudhuri, S., Saito, J., Groueix, T.: Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. ACM TOG **41**(4) (2022)
3. Chen, H., Peng, B., Tao, Y., Zhang, J.: Dˆ 3-human: Dynamic disentangled digital human from monocular video. In: CVPR (2025)
4. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: ICCV (2021)
5. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19982–19993 (2023)
6. Ferguson, A., Osman, A.A.A., Bescos, B., Stoll, C., Twigg, C., Lassner, C., Otte, D., Vignola, E., Prada, F., Bogo, F., Santesteban, I., Romero, J., Zarate, J., Lee, J., Park, J., Yang, J., Doublestein, J., Venkateshan, K., Kitani, K., Kavan, L., Farra, M.D., Hu, M., Cioffi, M., Fabris, M., Ranieri, M., Modarres, M., Kadlecek, P., Khirodkar, R., Abdrashitov, R., Prévost, R., Rajbhandari, R., Mallet, R., Pearsall, R., Kao, S., Kumar, S., Parrish, S., Yu, S.I., Saito, S., Shiratori, T., Wang, T.L., Tung, T., Xu, Y., Dong, Y., Chen, Y., Xu, Y., Ye, Y., Jiang, Z.: Mhr: Momentum human rig (2025), https://arxiv.org/abs/2511.15586
7. 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: CVPR (2023)
8. Guo, C., Li, J., Kant, Y., Sheikh, Y., Saito, S., Cao, C.: Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In: CVPR (June 2025)
9. Ho, I., Song, J., Hilliges, O.: Sith: Single-view textured human reconstruction with image-conditioned diffusion. In: CVPR (2024)
10. 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: CVPR (2024)
11. Huang, Y., Wang, J., Zeng, A., Cao, H., Qi, X., Shi, Y., Zha, Z.J., Zhang, L.: DreamWaltz: Make a scene with complex 3D animatable avatars. In: NeurIPS (2023)
12. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: CVPR (2021)
13. Jiang, B., Hong, Y., Bao, H., Zhang, J.: SelfRecon: Self reconstruction your digital avatar from monocular video. In: CVPR (2022)
14. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In: CVPR (2023)
15. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: NeuMan: Neural human radiance field from a single video. In: ECCV (2022)
16. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Transactions on Graphics (ToG) **32**(3), 1–13 (2013)

17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4) (2023)
18. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: ECCV (2024)
19. Kratimenos, A., Lei, J., Daniilidis, K.: Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In: ECCV (2024)
20. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM TOG **39**(6) (2020)
21. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: CVPR (2025)
22. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., Luo, W., Guo, Y.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: CVPR (2025)
23. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In: CVPR (2024)
24. Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., Liu, Z.: HumanGaussian: Text-driven 3D human generation with gaussian splatting. In: CVPR (2024)
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG **34**(6) (Oct 2015)
26. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: SCALE: Modeling clothed humans with a surface codec of articulated local elements. In: CVPR (2021)
27. Moon, G., Shiratori, T., Saito, S.: Expressive whole-body 3d gaussian avatar. In: ECCV (2024)
28. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: CVPR (June 2015)
29. Nicolet, B., Jacobson, A., Jakob, W.: Large steps in inverse rendering of geometry. ACM Transactions on Graphics (TOG) **40**(6), 1–13 (2021)
30. 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: CVPR (2019)
31. 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: CVPR (2019)
32. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In: CVPR (2024)
33. Remelli, E., Lukoianov, A., Richter, S., Guillard, B., Bagautdinov, T., Baque, P., Fua, P.: Meshsdf: Differentiable iso-surface extraction. Advances in Neural Information Processing Systems **33**, 22468–22478 (2020)
34. Saito, S., Yang, J., Ma, Q., Black, M.J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks. In: CVPR (2021)
35. Saleh, F., Aliakbarian, S., Hewitt, C., Petikam, L., Xiao, X., Criminisi, A., Cashman, T.J., Baltrusaitis, T.: David: Data-efficient and accurate vision models from synthetic data. In: ICCV (2025)
36. 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: CVPR (2024)

37. Shin, J., Lee, J., Lee, S., Park, M.G., Kang, J.M., Yoon, J.H., Jeon, H.G.: CanonicalFusion: Generating drivable 3D human avatars from multiple images. In: ECCV (2024)
38. Sim, G., Moon, G.: PERSONA: Personalized whole-body 3D avatar with pose-driven deformations from a single image. In: ICCV (2025)
39. Song, C., Wandt, B., Rhodin, H.: Pose modulated avatars from video. arXiv preprint arXiv:2308.11951 (2023)
40. Song, C., Wu, Z., Su, S.Y., Wandt, B., Sigal, L., Rhodin, H.: Locality sensitive avatars from video. In: The Thirteenth International Conference on Learning Representations (2025)
41. Song, C., Wu, Z., Wandt, B., Sigal, L., Rhodin, H.: Representing animatable avatar via factorized neural fields. In: Computer Graphics Forum. vol. 44, p. e70192. Wiley Online Library (2025)
42. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. Springer (2020)
43. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: ICCV (2025)
44. Wang, S., Simon, T., Santesteban, I., Bagautdinov, T., Li, J., Agrawal, V., Prada, F., Yu, S.I., Nalbone, P., Gramlich, M., Lubachersky, R., Wu, C., Romero, J., Saragih, J., Zollhoefer, M., Geiger, A., Tang, S., Saito, S.: Relightable full-body gaussian codec avatars. ACM TOG **2501.14726** (2025)
45. Wang, Y., Deng, J.: Waft: Warping-alone field transforms for optical flow. arXiv preprint arXiv:2506.21526 (2025)
46. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: ECCV. Springer (2024)
47. Wen, J., Zhao, X., Ren, Z., Schwing, A.G., Wang, S.: Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In: CVPR (2024)
48. 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: CVPR (2022)
49. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., Hoyński, A.: Reconfusion: 3d reconstruction with diffusion priors. In: CVPR (2024)
50. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. arXiv preprint arXiv:2412.01506 (2024)
51. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM TOG **37**(2) (2018)
52. Zhang, J., Li, X., Zhang, Q., Cao, Y., Shan, Y., Liao, J.: Humanref: Single image to 3d human generation via reference-guided diffusion. In: CVPR (2024)
53. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025)