FiGA: Fidelity-Aware Gaussian Avatars for Human Reconstruction

Zekai Li Yufen Sun
School of Computer Science and Artificial Intelligence, Wuhan University of Technology
Wuhan 430070, China
li.comp2023@whut.edu.cn, yufen@whut.edu.cn

Abstract

The reconstruction of animatable human avatars using 3D Gaussian Splatting (3DGS) demonstrates significant potential. However, most existing approaches rely on generic human templates and restrict non-rigid modeling to the canonical space. Although this strategy provides stability, it underutilizes the expressive power of 3DGS and often leads to loss of fine details in dynamic scenarios. To this end, we introduce FiGA, a two-stage framework that combines the high-fidelity representation of 3D Gaussian Splatting with the dynamic modeling power of CNNs. Our approach starts by constructing a personalized Gaussian representation through 3DGS-based differentiable rendering, with the optimized geometry and attributes embedded into dense UV maps. We then introduce a dual-space dynamic modeling strategy: pose-conditioned structural pre-adaptation in the canonical space provides global stability, followed by a refinement network in the posed space that recovers high-frequency details. This structured decomposition enhances both optimization stability and detail fidelity. Experimental results demonstrate that the proposed method significantly enhances rendering quality, yielding superior results over prior work in both visual comparisons and quantitative metrics.

Keywords: Human Reconstruction, 3D Gaussian, Neural Rendering, Avatar

1. Introduction

High-fidelity and animatable digital human avatars are pivotal for enabling immersive experiences in applications including virtual reality, film production, and the metaverse. Traditional reconstruction pipelines, which depend on expensive scanning systems and labor-intensive manual modeling, are inherently costly and difficult to scale [4]. In response, significant research efforts have shifted toward reconstructing photorealistic and drivable avatars directly from multi-view video data [1, 37, 49, 53], offering a more scalable and accessible alternative. However, the challenge of designing avatar representations that balance structural stability with pose-dependent details persists as a key issue, driving the development of diverse approaches.

The pursuit of animatable avatars has evolved via several representational paradigms. Early works relied on explicit mesh-based representations, often extending parametric models like SMPL [29] with learned vertex displacements to model clothing [2, 6, 7]. Although efficient, these methods were constrained by a fixed mesh topology, which hindered the accurate representation of complex geometries including loose garments and detailed hair [3,20]. A significant shift occurred with the emergence of implicit representations, especially Neural Radiance Fields (NeRF) [30]. By constructing a continuous volumetric field within a canonical space [8,32,46], NeRF-based methods enabled the modeling of arbitrary topologies and achieved superior rendering fidelity. However, the sampling process required by volumetric rendering prevented real-time application. Furthermore, the intrinsic spectral bias of multilayer perceptrons [40] attenuated high-frequency surface details, compromising geometric precision.

Recently, 3D Gaussian Splatting (3DGS) [21] has emerged as a prominent alternative for dynamic scene representation. The explicit, point-based nature of 3DGS preserves high visual quality while enabling real-time rendering [17, 23, 24, 31, 36, 53]. These methods demonstrate exceptional proficiency in synthesizing complex, pose-dependent details and achieving high-fidelity rendering results.

Despite the success of these 3DGS-based methods, they largely follow a common architectural strategy. Specifically, non-rigid corrections are predicted within a static canonical space, rather than directly in the dynamic posed space. Although this stability-driven approach mitigates optimization instabilities, it introduces a critical limitation: an "open-loop" learning process. In this setup, the network is required to pre-compensate for dynamic effects without receiving feedback from the final geometry. Such decoupling often forces a single network to address both global structural errors and local details, leading to optimization entanglement.

To address these challenges, our method decomposes the

problem into two primary stages. First, static personalization is separated from dynamic modeling by constructing a personalized Gaussian template. Instead of relying on a generic model, a set of subject-specific 3D Gaussians is optimized from multi-view videos. Their attributes are then embedded into continuous UV maps to establish a stable, personalized foundation. This template fixed shape constraints of generic SMPL to better capture subject-specific attire. Building upon this foundation, a dual-space framework for dynamic modeling is introduced. The first step involves a coarse, pose-conditioned pre-adaptation in the canonical space to manage large-scale structural variations. This is followed by a fine-grained refinement network operating directly in the posed space. Freed from global structural considerations, this network concentrates solely on the local post-deformation correction of high-frequency surface details. By refining geometry in the posed space, FiGA provides direct feedback for high-frequency synthesis. The structured decomposition of the problem ensures a clear separation of concerns, enhancing both modeling stability and the expressive detail of the reconstructed avatar. Our main contributions are as follows:

- A personalized canonical representation is proposed via parameterizing optimized Gaussian attributes into UV texture space. This approach disentangles the modeling of subject-specific static attributes from dynamic deformation, establishing a personalized foundation for modeling pose-dependent deformations.
- A dual-space architecture is proposed to decouple coarse structural adaptation from fine-grained detail refinement. By assigning specialized tasks to their respective coordinate spaces, this framework helps mitigate optimization entanglement and improves rendering fidelity.
- Experimental results on public datasets confirm the effectiveness of our method, with improvements in both reconstruction accuracy and rendering fidelity.

2. Related Work

Explicit representation serves as the foundational paradigm for constructing animatable digital humans, a technological preference primarily attributable to its robust compatibility with conventional graphics rendering pipelines. The canonical pipeline commences with the reconstruction of a static mesh using Multi-View Stereo (MVS) or 3D scanning techniques [9,41]. This foundational mesh subsequently undergoes a rigging process, wherein it is bound to a parametric skeleton to enable animation. To model dynamic appearance changes, some methods leverage neural networks for generating pose-conditioned textures in UV space [4, 14, 47, 48]. Despite methodological maturity, this approach faces fundamental limitations. High-fidelity geometry acquisition requires dense multi-view systems and computationally expensive offline processing [13, 39]. Furthermore, the topological invariance of polygonal meshes inherently constrains their ability to represent complex deformable structures and surface discontinuities, as exemplified in garment motion simulation.

In response to these limitations, Neural Radiance Fields (NeRF) [30] have emerged as a compelling alternative for high-fidelity human avatar reconstruction from multi-view video data [18, 25, 26, 33]. These approaches typically operate by warping points from an observed pose to a canonical space [5, 10, 16, 19, 43], frequently incorporating learned residual deformations to model dynamic appearance details [11, 28, 50]. However, this paradigm encounters fundamental challenges. From a computational perspective, volumetric rendering remains costly because dense sampling and an ill-posed inverse deformation mapping necessitate slow numerical solvers. Architectural optimizations such as the hash encodings in InstantNVR [12] and AvatarRex [52] accelerate rendering but do not resolve the intrinsic limitation of representational capacity. The underlying MLPs exhibit spectral bias [40], which restricts the reconstruction of high-frequency details, producing overly smooth or blurry outputs and motivating the pursuit of alternative representations.

Recent advances have converged on 3DGS [21], whose explicit representation, high-fidelity, and real-time rendering provide a powerful basis for dynamic avatar reconstruction. Building on this foundation, representative methods such as GauHuman [17], GoMAvatar [45], and SplattingAvatar [38] animate canonical 3D Gaussians by anchoring them to the parametric SMPL body model [29] and applying Linear Blend Skinning (LBS). However, since LBS cannot capture non-rigid deformations, two correction strategies have emerged. The first employs MLP-based residual prediction, as in GART [23], 3DGS-Avatar [36], and HUGS [22], to learn pose-dependent offsets. Despite its simplicity, this strategy inherits the spectral bias of NeRF-based models, yielding overly smoothed details. To address this limitation, an alternative approach exploits CNNs on UV maps for high-frequency synthesis, which has been validated by state-of-the-art methods with superior rendering quality [15, 27, 31]. Those methods typically focus on neural texture synthesis within a single coordinate space. FiGA, however, decouples the modeling into canonical and posed spaces. Global stability and local refinement are optimized separately to ensure high-fidelity results.

3. Method

Given multi-view RGB videos of a subject along with the corresponding SMPL registration parameters, our objective is to reconstruct an animatable 3D Gaussian avatar with high fidelity. We design a fully 3DGS-based two-stage

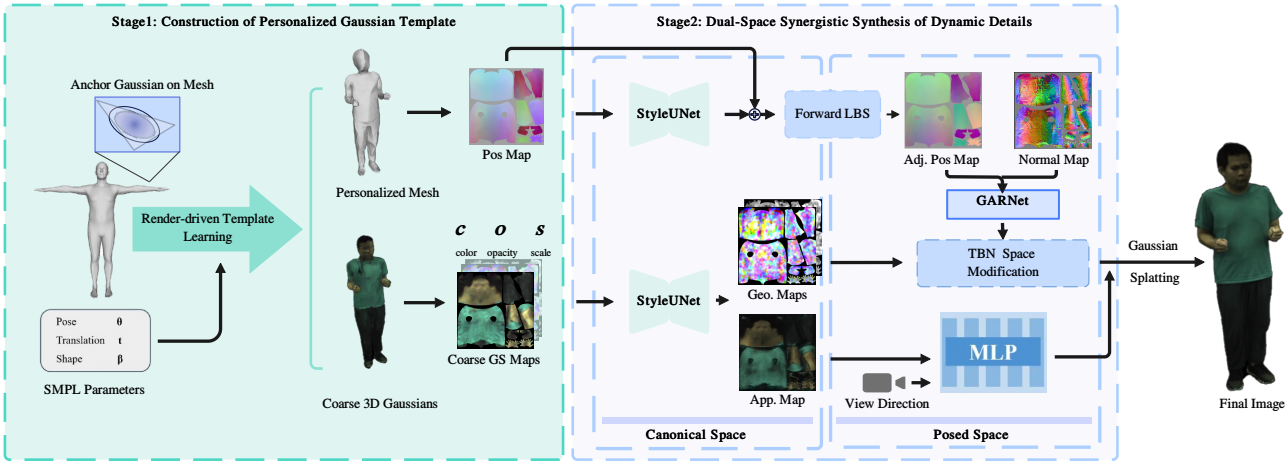


Figure 1. Overview of our two-stage reconstruction framework. In the first stage, we build a personalized template by jointly optimizing the mesh with its anchored 3D Gaussian attributes. The resulting attributes are then parameterized into continuous UV maps to provide a reliable foundation. The second stage introduces a dual-space framework that explicitly decouples the learning objective. A StyleUNet performs pose-conditioned adjustments in the canonical space; a Geometry-Aware Refinement Network (GARNet) then recovers fine-grained dynamic details in the posed space.

reconstruction method, as illustrated in Figure 1. In the first stage, the SMPL mesh is extended under the supervision of a 3DGS renderer to generate a Gaussian human representation, which is subsequently parameterized into a set of 2D Gaussian attribute maps. This texture-like representation reformulates the mapping from skeletal motion to Gaussian splatting parameters as an image-to-image translation task. In the second stage, a dual-space optimization framework models pose-dependent non-rigid deformations, effectively decoupling the optimization of global structural corrections from fine-grained detail synthesis.

3.1. Personalized Gaussian Template

The canonical SMPL model represents an average human topology and skeleton, failing to capture individual geometric variations or clothing-induced offsets. To establish a reliable foundation for high-fidelity reconstruction, we construct a personalized template that integrates subject-specific geometric and appearance priors, effectively compensating for the details omitted by the generic body model.

3.1.1 Optimization of Mesh-Anchored Gaussian Attributes

To construct a personalized Gaussian template, we jointly optimize a customized mesh $\mathcal{M}'_{\text{SMPL}}$ and mesh-anchored Gaussian attributes over the entire video sequence. Following GoMAvatar [45], all Gaussian points are rigidly anchored to the triangular faces of the standard SMPL mesh in canonical space. Diverging from standard 3DGS, the optimization indirectly controls Gaussian positions through learnable per-vertex offsets $\Delta\mathbf{v}_i$ on the SMPL mesh and represents appearance with a base RGB color $\mathbf{c}_i$ rather than spherical harmonics. These offsets $\Delta\mathbf{v}_i$ allow the template to diverge from the restricted tight-fitting body surface, enabling the capture of loose attire and identity-specific geometry.

To enable adaptability to sparse or monocular view input, all frames are leveraged for dynamic supervision. Specifically, the canonical representation is deformed into target poses using LBS. The covariance matrix $\mathbf{\Sigma}_i$ of each Gaussian primitive in the posed space is computed as:

$$\mathbf{\Sigma}_i = \mathbf{A}_i(\mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^T\mathbf{R}_i^T)\mathbf{A}_i^T, \tag{1}$$

where $\mathbf{R}_i$ and $\mathbf{S}_i$ denote the local rotation and scaling matrices for each Gaussian, and $\mathbf{A}_i$ represents the local-to-world affine transformation derived from the vertices of its anchor face in the posed mesh [45]. Eq. (1) aligns each Gaussian with its supporting surface, effectively shaping it into a thin sheet closely conforming to the mesh.

During optimization, we strictly maintain the one-to-one binding between Gaussian points and mesh faces, without altering their count by sub-division. Through the differentiable rendering and skinning pipeline, gradients from the entire dynamic sequence are propagated back to the canonical space. This gradient propagation facilitates the joint optimization of a coherent set of Gaussian attributes alongside the vertices of the underlying SMPL mesh. Refined vertex positions are computed by applying learnable per-vertex offsets $\Delta\mathbf{v}_i$ as:

$$\mathbf{v}'_i = \mathbf{v}_i + \Delta\mathbf{v}_i, \tag{2}$$

where $\mathbf{v}_i$ and $\mathbf{v}'_i$ denote the original and optimized vertex positions in canonical space, respectively.

3.1.2 2D Parameterization and Densification

Although convolutional networks excel at processing grid-structured data, optimized Gaussian attributes reside on discrete primitives. To reconcile this representational mismatch, per-face attributes are parameterized into dense 2D maps via the intrinsic UV space of SMPL. Specifically, for each attribute type a, a sparse 2D map $\mathbf{I}_a^{\text{sparse}}$ is constructed. Each per-face attribute value $\mathbf{a}_i$ is mapped to the UV coordinate $\mathbf{u}_i$ corresponding to the centroid of its anchor triangle:

$$\mathbf{I}_a^{\text{sparse}}(\mathbf{u}) = \sum_{i=1}^{N_f} \mathbf{a}_i \cdot \delta(\mathbf{u} - \mathbf{u}_i), \tag{3}$$

where $\delta(\cdot)$ denotes the 2D Dirac delta function and N_f is the total number of mesh faces. As sparse maps are unsuitable for convolution, a differentiable kernel-based densification operator $\mathcal{D}$ is applied to produce continuous attribute fields:

$$\mathbf{I}_a(\mathbf{u}) = \mathcal{D}(\mathbf{I}_a^{\text{sparse}}) = \frac{\sum_{i=1}^{N_f} w(\mathbf{u}, \mathbf{u}_i)\, \mathbf{a}_i}{\sum_{i=1}^{N_f} w(\mathbf{u}, \mathbf{u}_i)}, \tag{4}$$

where the weighting function $w(\mathbf{u}, \mathbf{u}_i)$ is a Gaussian kernel defined as $w(\mathbf{u}, \mathbf{u}_i) = \exp(-\|\mathbf{u} - \mathbf{u}_i\|^2 / 2\sigma^2)$, with σ controlling the kernel width. Differentiable densification transforms discrete primitives into a structured representation suitable for the subsequent dual-space synthesis, ensuring stable optimization of pose-dependent details.

The 3D vertex positions of the optimized mesh $\mathcal{M}'_{\text{SMPL}}$ are parameterized into UV space, producing a canonical position map. Combined with densified scale, opacity, and color maps, a set of spatially aligned coarse Gaussian attribute maps $\mathbf{I}_{\text{coarse}}$ is formed in canonical space. Each UV pixel within the mask corresponds to a potential Gaussian primitive whose canonical attributes are initialized from $\mathbf{I}_{\text{coarse}}$. Simultaneously, LBS weights from SMPL vertices are propagated to all Gaussian points through k-NN diffusion, endowing them with skeleton-driven motion.

3.2. Dual-Space Synergistic Synthesis of Dynamic Details

The personalized template provides a canonical foundation yet lacks pose-dependent details, such as non-rigid deformations and appearance variations essential for high-fidelity reconstruction. To overcome this limitation, we introduce a dual-space framework that applies global pose-conditioned structural corrections in canonical space and synthesizes fine-grained local details in the posed space. Our synergistic architecture distributes modeling tasks across dual coordinate systems. This balance maintains global stability while leveraging direct feedback from the target posture to recover high-frequency textures.

3.2.1 Pose-Conditioned Attribute Adjustment in Canonical Space

In canonical space, pose-dependent global variations relative to the personalized template are modeled, encompassing large-scale geometric deformations and surface appearance changes. Compared with direct learning in the posed space, this approach exhibits greater stability, as it operates prior to skinning and leverages the template's geometric priors. To effectively encode global pose information into the detail-rich personalized template, a StyleUNet-based network [42] is employed.

To obtain a compact pose representation, the canonical position map is first transformed into the target pose via LBS, yielding the posed position map $\mathbf{I}_{\text{pos}}^{\text{tgt}}$. Subsequently, a lightweight encoder $\mathcal{E}_{\text{pose}}$ maps this to a compact pose embedding:

$$\mathbf{z}_{\text{pose}} = \mathcal{E}_{\text{pose}}(\mathbf{I}_{\text{pos}}^{\text{tgt}}). \tag{5}$$

The pose embedding $\mathbf{z}_{\text{pose}}$ implicitly captures the temporal state of the motion sequence, driving the reconstruction without requiring an explicit time index.

The coarse attribute maps $\mathbf{I}_{\text{coarse}}$ generated in Sec. 3.1 serve as the backbone input, offering stable global geometry and surface priors. The low-dimensional pose vector $\mathbf{z}_{\text{pose}}$ is injected as a style code to guide the prediction of pose-conditioned residuals. Compared with directly using the posed position map as input, our design offers two advantages: (i) $\mathbf{I}_{\text{coarse}}$ integrates multi-view information, supplying more complete geometric and appearance priors for residual prediction; (ii) encoding pose into a compact style embedding helps disentangle pose from identity, allowing the network to focus on learning pose-dependent variations. The network produces pose-conditioned residual attribute maps:

$$\Delta\mathbf{I}_{\text{coarse}} = \text{StyleUNet}(\mathbf{I}_{\text{coarse}} \oplus \mathbf{I}_{\text{pos}}^{\text{tgt}}, \mathbf{z}_{\text{pose}}), \tag{6}$$

where $\oplus$ denotes channel-wise concatenation. The residuals are subsequently added to the canonical attributes, producing pre-adapted canonical maps that incorporate structural compensation for the target pose.

3.2.2 Dynamic Geometric Refinement in the Posed Space

Following structural compensation, the canonical attributes are deformed to the target pose using LBS. We introduce a Geometry-Aware Refinement Network (GARNet) dedicated to geometric refinement, capturing fine-grained local shape variations within the posed space.

Direct regression of final world coordinates is inherently ill-posed due to the absence of geometric constraints, which frequently leads to fragmented or noisy surfaces. To achieve stable and physically plausible refinements, displacements

are predicted in the local tangent–bitangent–normal (TBN) space of each point, constraining modifications to the surface manifold and preventing spurious deformations. TBN-based parameterization preserves local geometric consistency, a crucial constraint absent in unconstrained global coordinate regression. GARNet produces a multi-channel detail map defined as:

$$\{\Delta\mathbf{p}_{\text{tbn}}, \Delta s, \Delta\alpha\} = \text{GARNet}(\mathbf{N}_{\text{posed}} \oplus \mathbf{P}_{\text{posed}}), \quad (7)$$

where $\mathbf{P}_{\text{posed}}$ is the posed position map obtained by applying LBS to the pre-adapted canonical geometry, $\mathbf{N}_{\text{posed}}$ is the corresponding normal map, $\Delta\mathbf{p}_{\text{tbn}}$ denotes local displacement in the TBN space, and Δs, $\Delta\alpha$ are residual corrections for anisotropic scale and opacity, respectively. Because the refinement operates directly on the deformed geometry, GARNet synthesizes pose-dependent details that remain coherent with the continuous skeletal motion.

Clothing-related high-frequency details are strongly correlated with second-order surface geometry. To capture these effects, $\mathbf{P}_{\text{posed}}$ and $\mathbf{N}_{\text{posed}}$ are augmented with two additional features prior to inputting them into GARNet: (i) Fourier positional encodings derived from UV coordinates for absolute spatial awareness, and (ii) curvature maps computed from the normal field to characterize local surface bending. These features focus the network on intricate cloth folds, thereby concentrating visual gains on the most challenging high-frequency regions.

To efficiently process high-resolution feature maps, we introduce a hybrid block architecture. Lightweight residual blocks are used at high-resolution stages, and gated geometry-aware blocks at lower-resolution stages. Guided by curvature maps, these blocks combine convolution kernels with varying dilation rates, adapting the receptive field to local geometric complexity. Gated geometry-aware blocks' outputs are modulated by gating signals from global features, effectively ensuring that fine-grained local refinements remain consistent with the global articulated motion.

In the final attribute integration stage, the tangent-space basis $\mathbf{M}_{\text{tbn}}$ at each point is computed from the gradients of $\mathbf{P}_{\text{posed}}$. The final refined position $\mathbf{P}_{\text{final}}$ is obtained by adding a world-space displacement map $\Delta\mathbf{P}_{\text{map}}$ to the posed geometry:

$$\begin{aligned} \mathbf{P}_{\text{final}} &= \mathbf{P}_{\text{posed}} + \Delta\mathbf{P}_{\text{map}}, \quad (8) \\ \text{where} \quad \Delta\mathbf{P}_{\text{map}} &= \mathbf{M}_{\text{tbn}} \cdot \Delta\mathbf{p}_{\text{tbn}}. \end{aligned}$$

Direct regression of 3D rotations is challenging and prone to invalid solutions. Instead of predicting rotations directly, we deterministically determine the final rotation from the TBN basis of the refined geometry $\mathbf{P}_{\text{final}}$. This basis is computed from the gradients of the refined positions and then converted from its matrix representation to the final quaternion. Other geometric attributes are updated residually. A lightweight MLP is employed to refine canonical-space color predictions conditioned on the viewing direction, accounting for view-dependent effects such as specular highlights.

3.3. Training

Our total training loss $\mathcal{L}$ consists of two main parts: a photometric loss $\mathcal{L}_{\text{photo}}$ and a regularization loss $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{reg}}. \quad (9)$$

1) Photometric Loss: The photometric loss enforces consistency between the rendered image and the ground-truth observation. It comprises an L1 loss, a perceptual LPIPS loss, and an SSIM loss:

$$\mathcal{L}_{\text{photo}} = \lambda_{\text{l1}}\mathcal{L}_{\text{l1}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}}. \quad (10)$$

The L1 loss enforces pixel-level color fidelity. The LPIPS and SSIM losses are computed on randomly sampled image patches, emphasizing perceptual quality and high-frequency structural details, respectively.

2) Regularization Loss: To ensure training stability and maintain geometric plausibility, a regularization loss $\mathcal{L}_{\text{reg}}$ is employed:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{offset}}\mathcal{L}_{\text{offset}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}}. \quad (11)$$

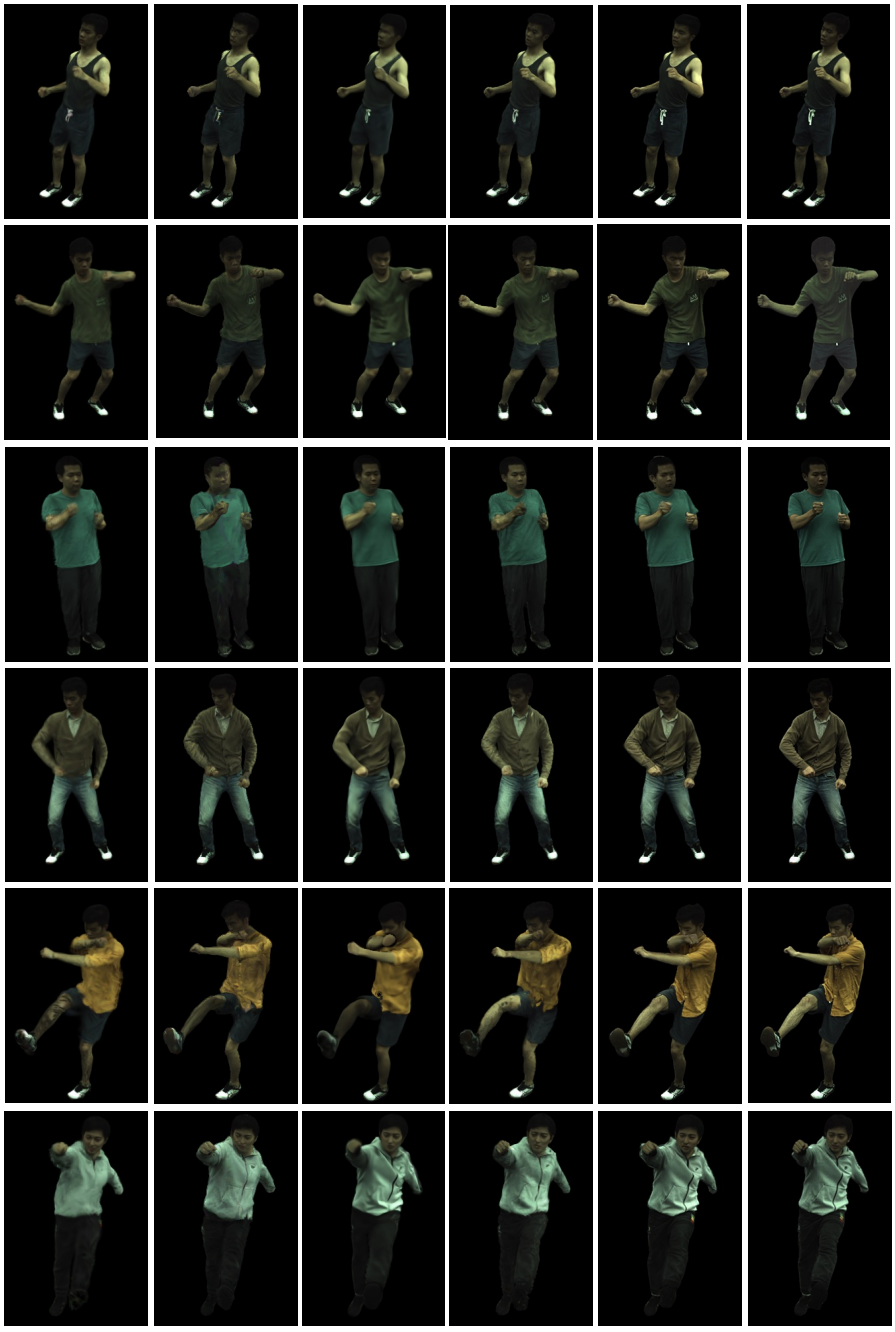
Here, $\mathcal{L}_{\text{offset}}$ is an L2 loss used to constrain the predicted offsets of the Gaussian points in the canonical space. $\mathcal{L}_{\text{lap}}$ denotes a curvature-aware regularizer that constrains the magnitude of corrections in the posed space displacement field, formulated as a weighted norm of the discrete Laplacian vectors:

$$\mathcal{L}_{\text{lap}} = \mathbb{E}_{\mathbf{u}}\left[\|\mathbf{W}(\mathbf{u}) \cdot (\nabla^2\Delta\mathbf{P}_{\text{map}})(\mathbf{u})\|_2\right], \quad (12)$$

where the weight map $\mathbf{W}(\mathbf{u})$ is designed to be adaptive to the local surface curvature. We define it as:

$$\mathbf{W}(\mathbf{u}) = \left(1 + \kappa \cdot \|\nabla^2\mathbf{N}_{\text{posed}}(\mathbf{u})\|_2\right)^{-1}. \quad (13)$$

Here, the curvature is approximated by the Laplacian of the posed normal map $\mathbf{N}_{\text{posed}}$. The hyperparameter κ balances the sensitivity to this curvature. Consequently, the penalty is stronger in flatter regions (low curvature) and reduced in high-curvature areas, such as clothing folds, preserving sharp details. Finally, $\mathcal{L}_{\text{scale}}$ penalizes large Gaussian scales to encourage compactness.



(a) GH (b) GoM (c) ARAH (d) 3DGS-Avatar (e) Ours (f) GT

Figure 2. Comparison on ZJU-MoCap. "GoM" means GoMAvatar and "GH" means GauHuman. Our method generates fewer artifacts and achieves a significant improvement in the reconstruction of clothing regions.

4. Experiments

4.1. Datasets and Evaluation Metrics

4.1.1 Datasets

To rigorously evaluate the proposed methodology, we performed comprehensive experiments on two established human modeling benchmarks: ZJU-MoCap [34] and PeopleSnapshot [3]. The ZJU-MoCap dataset comprises multi-view sequences captured by a system of over 20 synchronized cameras recording subjects performing complex motions. To evaluate the reconstruction performance of our method under sparse-view conditions, we selected six subjects for evaluation. For each subject, five camera views

were used for training, while the remaining unseen views were utilized to assess novel view synthesis quality. The PeopleSnapshot dataset consists of monocular videos featuring subjects rotating before a static camera. To maintain standardized and comparable experimental conditions, we adopted the same evaluation protocol as GauHuman and Instant-Avatar, reporting average quantitative metrics across the dataset.Although prior work like GauHuman [17] utilizes a single-view setting, we adopt a 5-view configuration to introduce stronger multi-view constraints, which are essential for recovering the high-frequency garment details and complex non-rigid deformations that FiGA is designed to address.

4.1.2 Evaluation Metrics

For quantitative assessment of rendering quality, we employed three established metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [44], and Learned Perceptual Image Patch Similarity (LPIPS) [51]. PSNR quantifies pixel-level fidelity between rendered and ground-truth images, SSIM evaluates structural consistency, and LPIPS measures perceptual similarity using deep feature representations, thereby providing better alignment with human visual assessment.To further demonstrate the temporal stability of our results, which cannot be fully captured by per-frame metrics, we provide reconstructed sequences in the supplementary video.

Method	PSNR↑	SSIM↑	LPIPS↓
GauHuman	31.74	0.964	28.97
3DGS-Avatar	32.51	0.973	27.52
GoMAvatar	32.04	0.961	27.95
ARAH	31.53	0.972	35.82
Ours	**34.65**	**0.982**	**22.43**

Table 1. Quantitative results on ZJU-MoCap.

4.2. Implementation Details

4.2.1 Network Architecture

The Geometry-Aware Refinement Network is designed as a hybrid U-Net architecture. Input position and normal maps with a resolution of 512×512 are transformed into a seven-channel Gaussian geometric attribute correction map while maintaining the original resolution. Fourier positional encodings of UV coordinates serve as auxiliary inputs to enhance global spatial awareness. Within the encoder, lightweight residual blocks are deployed at the highest-resolution layer to reduce computational cost, whereas deeper layers and the bottleneck employ gated geometry-aware blocks. Each gated geometry-aware block employs three depthwise 3×3 convolutions with dilation rates of 1, 2, and 3, respectively. These multi-scale features are dynamically fused using curvature maps derived from the input normal map. The fused feature stream is subsequently modulated through a spatial gating mechanism and a feature-wise linear modulation (FiLM) layer [35]. Spatial resolution is symmetrically restored in the decoder using transposed convolutions. Instance Normalization and LeakyReLU activation are applied to all convolutional layers except the output layer. The final output is produced by three parallel 1×1 convolution heads, yielding position with three channels, scale with three channels, and opacity with one channel.

4.2.2 Training Setup

The network is optimized using Adam with a learning rate of 5×10^{-4}. Training on the ZJU-MoCap dataset is performed for 300,000 iterations with a batch size of one. Loss weights are assigned as $\lambda_{\text{l1}} = 1.0$, $\lambda_{\text{lpips}} = 0.1$, and $\lambda_{\text{ssim}} = 0.1$ for the photometric loss, and $\lambda_{\text{offset}} = 0.01$, $\lambda_{\text{lap}} = 0.1$, and $\lambda_{\text{scale}} = 0.1$ for the regularization terms.

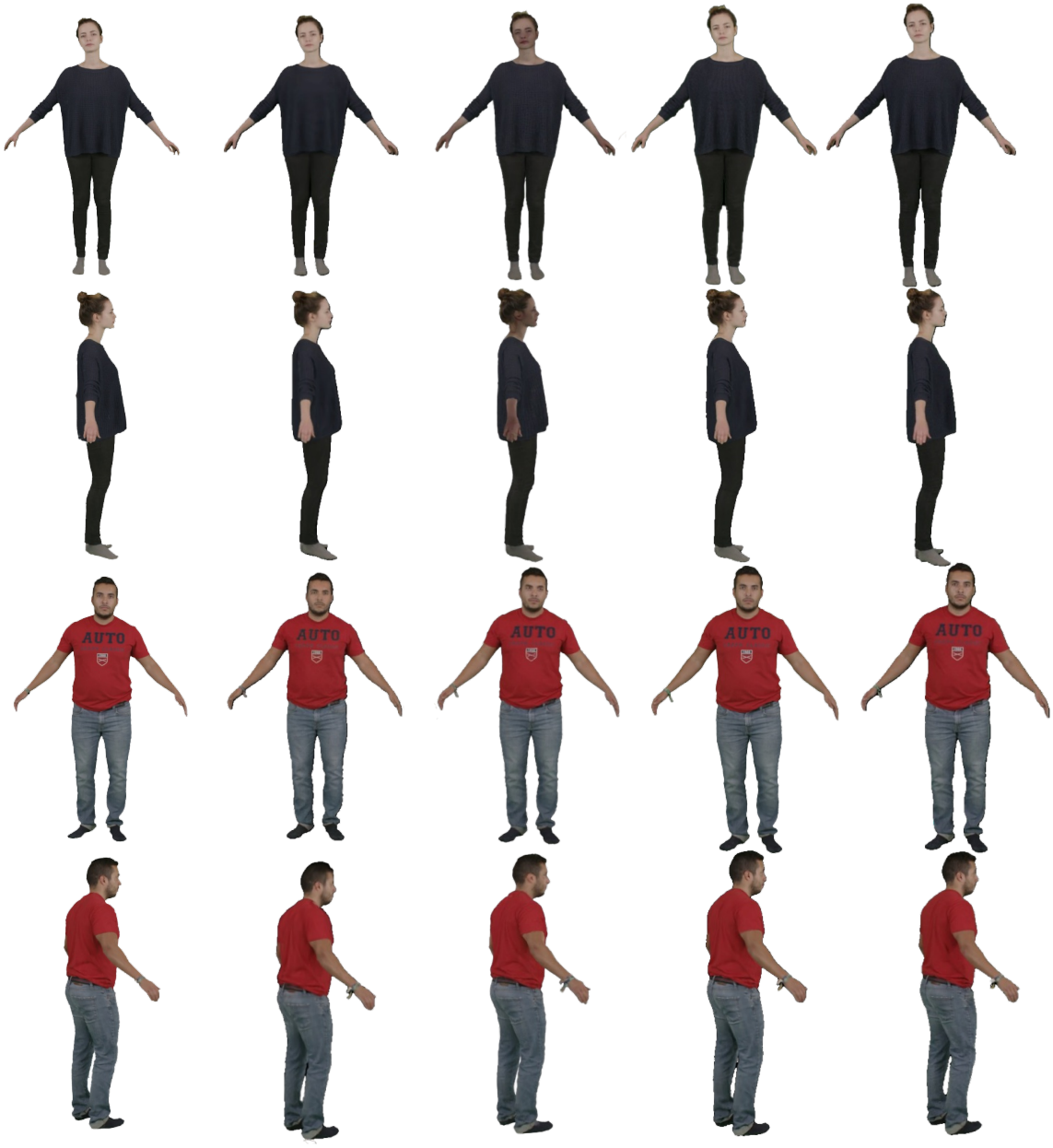
4.3. Comparisons

To comprehensively assess the proposed method, comparisons are conducted against four state-of-the-art dynamic human modeling approaches: GauHuman [17], GoMAvatar [45], 3DGS-Avatar [36], and ARAH [43]. On the ZJU-MoCap dataset, all baselines are extended to multi-view inputs under identical experimental settings to ensure fairness. Quantitative results in Tables 1 and 2 show that the proposed approach consistently outperforms the baselines in PSNR, SSIM, and LPIPS. Notably, the improvements in LPIPS indicate a stronger alignment with human perceptual judgments of visual quality. Qualitative comparisons in Figures 2 and 3 further corroborate these results. Other CNN-driven methods typically rely on dense-view inputs or body templates with more extensive prior information. Our evaluation thus focuses on representative 3DGS frameworks within a standard sparse-view SMPL configuration.

Method	PSNR↑	SSIM↑	LPIPS↓
GauHuman	30.55	0.963	22.43
3DGS-Avatar	32.53	0.974	19.75
GoMAvatar	31.28	0.957	23.31
Ours	**33.13**	**0.982**	**19.22**

Table 2. Quantitative results on PeopleSnapshot.

The neural implicit field-based method ARAH reconstructs clothing silhouettes with reasonable accuracy; however, its reliance on MLPs introduces a low-frequency bias



(a) 3DGS-Avatar (b) GH (c) GoM (d) Ours (e) GT

Figure 3. Comparison on PeopleSnapshot. “GoM” means GoMAvatar and “GH” means GauHuman.

that limits the fidelity of fine-grained textures. GauHuman omits expressive nonlinear deformation modules and instead optimizes offsets of LBS weights, enabling faster training. This simplified deformation modeling, however, shows limitations in representing complex non-rigid motions. Although GauHuman attains competitive PSNR scores, as illustrated in the second column of Figure 2, its reconstructions suffer from pronounced blurring in clothing regions. In contrast, 3DGS-Avatar estimates pose-dependent non-rigid displacements in canonical space using MLPs. However, the lack of direct optimization in the posed space substantially degrades accuracy in regions where rigid and non-rigid motions are tightly coupled, particularly in clothing folds around joints. GoMAvatar proposes a hybrid ‘Gaussians-on-Mesh’ representation that anchors Gaussians to explicit mesh patches. This design pro-

vides a stable geometric foundation at the cost of limiting the expressive capacity of the Gaussian point cloud. Consequently, the method is prone to tearing artifacts in clothing during large human motions and frequently fails to disentangle albedo from shading, as shown in Figure 3. In comparison, the proposed approach achieves superior performance in both appearance details and geometric fidelity. Notably, the reconstructions exhibit sharper and more natural rendering of clothing wrinkles and fine-grained textures.

Figure 4. Qualitative results of our method on unseen poses. Subjects are from ZJU-MoCap (top row) and PeopleSnapshot (bottom row).

Our qualitative comparisons in Figures 2 and 3 already demonstrate the high fidelity achieved by our method. To further illustrate its robustness under challenging dynamic conditions, we visualize reconstruction results on a set of novel poses unseen during training. As shown in Figure 4, our method preserves fine-grained geometric details and remains stable across highly articulated poses, highlighting its generalization to unseen motions.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o Canonical Adj	29.45	0.962	34.67
w/o Posed Ref	33.72	0.979	24.76
w/XYZ Ref	30.14	0.971	30.56
Full model	34.65	0.982	22.43

Table 3. Quantitative ablation study of the dual-space framework.

4.4. Effectiveness Analysis of the Dual-Space Framework

The core of our approach is a dual-space framework that effectively decouples global structural adjustment from local detail synthesis. To validate its efficacy, we systematically ablate key components and technical designs in this section. Qualitative results are illustrated in Figure 5, and quantitative evaluations are provided in Table 3.

1) Pose-conditioned adjustment in canonical space (`w/o Canonical Adj`). The StyleUNet module in canonical space is designed to model large-scale, pose-dependent geometric and appearance variations. To assess its contribution, we disable its optimization of geometric attributes, requiring all non-rigid deformations to be learned exclusively in the posed space. As shown in Figure 5 (c), in the absence of canonical-space correction, the Gaussian template transformed through LBS alone yields an unreliable foundation for the subsequent refinement network. This leads to significant volumetric distortions and blurred appearance in the final reconstructions. These results demonstrate that without coarse, pose-aware geometric corrections in the canonical space, the refinement network alone struggles to compensate for the large-scale structural errors introduced by LBS.

2) Dynamic geometric refinement in the posed space (`w/o Posed Ref`): The refinement network operating in the posed space is employed to synthesize high-frequency non-rigid details and correct geometric inaccuracies introduced by LBS. An ablation of this module is conducted to assess its contribution. As presented in Figure 5 (b), although the overall shape remains plausible, a loss of fine details is observed around joint regions such as shoulders and elbows. These results confirm that refinement in the posed space is essential for achieving realistic deformations in highly articulated areas.

3) Geometric refinement in local TBN space (`w/XYZ Ref`): Our method predicts geometric displacements in the local TBN space of each point, constraining deformations to the surface manifold. In comparison, a variant that regresses displacements directly in the global world coordinate system (XYZ) fails to produce continuous surfaces, as shown in Figure 5 (d), resulting in noisy artifacts and fragmented structures. This result validates the importance of local TBN space for maintaining surface continuity and geometric integrity during refinement, and highlights the inherent instability of unconstrained displacement regression.

4.5. Evaluation of Personalized Priors and Regularization Strategies

This section conducts ablation studies to evaluate two core components of our method: the personalized template and curvature-aware regularization. The personalized template serves as a critical source of identity-specific geometric and appearance priors, while the curvature-aware regularization enforces geometric smoothness and plausibility during dynamic detail synthesis.

(a) Full model (b) w/o Posed Ref (c) w/o Canonical Adj (d) w/XYZ Ref (e) GT

Figure 5. Ablation study of the dual-space framework.

(a) SMPL + Pos (b) Template + Pos (c) Full Model (d) GT

Figure 6. Ablation study of our personalized template.

Method	PSNR↑	SSIM↑	LPIPS↓
SMPL + Pos	31.45	0.968	28.76
Template + Pos	33.14	0.978	24.67
Full model	34.65	0.982	22.43

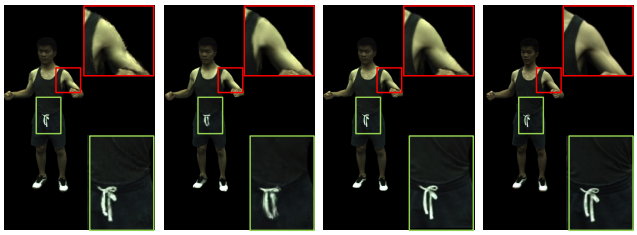
Table 4. Quantitative results for the personalized template ablation.

4.5.1 Effectiveness of the personalized template

To evaluate the contribution of the personalized template, the complete model is compared with two degraded variants. *SMPL + Pos* denotes the direct use of position maps generated from the generic SMPL model as network input. *Template + Pos* employs position maps derived from the personalized geometric template but excludes the coarse attribute map as an appearance prior. Qualitative and quantitative comparisons are presented in Figure 6 and Table 4, respectively. The baseline relying solely on generic priors produces blurry reconstructions, where hand structures are indistinguishable and T-shirt logos collapse into indistinct patches. Although the *Template + Pos* benefits from personalized geometry, the absence of a stable multi-view attribute basis forces the network to hallucinate appearance details, leading to incoherent textures and floating artifacts, particularly around the arms. These comparisons demonstrate that the complete personalized template, through $\mathbf{I}_{\text{coarse}}$, provides a robust geometric and appearance basis. This design reformulates the ill-posed direct synthesis problem as a constrained residual learning task, which is essential for suppressing artifacts and recovering high-frequency details.

Method	PSNR↑	SSIM↑	LPIPS↓
Uniform L_{lap}	32.45	0.972	25.26
w/o L_{lap}	33.66	0.981	23.43
Full model	34.65	0.982	22.43

Table 5. Quantitative evaluation of different regularization strategies.



(a) w/o L_{lap} (b) Uniform L_{lap} (c) Full Model (d) GT

Figure 7. Ablation study on our regularization strategy. Our curvature-aware weighting suppresses noise on flat surfaces while preserving sharp geometric details.

4.5.2 Role of Curvature-Aware Regularization

The proposed curvature-aware Laplacian loss $\mathcal{L}_{\text{lap}}$ regularizes the predicted displacement field while preserving high-frequency geometric details. Its effectiveness is assessed through comparisons with two alternatives, with results reported in Figure 7 and Table 5. Standard uniform-weight regularization (*Uniform* $\mathcal{L}_{lap}$) imposes indiscriminate smoothing priors on the surface, which ensures overall smoothness but undesirably removes fine details such as belt edges. Removing the loss entirely (*w/o* $\mathcal{L}_{lap}$) eliminates essential geometric constraints, resulting in severe artifacts in flat regions due to overfitting to noise. By contrast, the curvature-aware weighting adaptively modulates the regularization strength according to local geometric complexity, effectively suppressing noise in smooth regions while maintaining authentic features in high-curvature areas.

5. Conclusion

In this paper, a high-quality digital human reconstruction method based on 3D Gaussian Splatting is presented. A personalized Gaussian template is first constructed by optimizing the mesh and its anchored Gaussian attributes, which are then parameterized into UV space to provide a stable geometric and appearance foundation. Building on this template, a dual-space optimization framework powered by 2D CNNs is introduced to model pose-dependent non-rigid deformations, thereby improving both training stability and dynamic performance. Experimental results show that the proposed method outperforms existing approaches in rendering quality and multiple quantitative metrics, while also demonstrating strong potential for downstream applications. However, our design inherently couples the topology of the rendered Gaussian cloud to the fixed SMPL domain. While the proposed learnable offsets effectively capture loose attire such as hoodies, representing garments with fundamentally different topologies, such as skirts or long coats, remains a challenge. Future work will investigate hybrid representations that integrate the stable, template-based foundation with additional topology-adaptive components, thereby extending applicability to a wider range of clothing styles.

References

[1] R. Abdal, W. Yifan, Z. Shi, Y. Xu, R. Po, Z. Kuang, Q. Chen, D.-Y. Yeung, and G. Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024. 1

[2] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 1

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 1, 6

[4] T. Bagautdinov, C. Wu, T. Simon, F. Prada, T. Shiratori, S.-E. Wei, W. Xu, Y. Sheikh, and J. Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 1, 2

[5] A. Bergman, P. Kellnhofer, W. Yifan, E. Chan, D. Lindell, and G. Wetzstein. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems*, 35:19900–19916, 2022. 2

[6] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European conference on computer vision*, pages 311–329. Springer, 2020. 1

[7] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 1

[8] Y. Chen, X. Wang, X. Chen, Q. Zhang, X. Li, Y. Guo, J. Wang, and F. Wang. 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*, pages 16621–16631, 2023. 1

[9] E. De Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[10] Y. Feng, J. Yang, M. Pollefeys, M. J. Black, and T. Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[11] Q. Gao, Y. Wang, L. Liu, L. Liu, C. Theobalt, and B. Chen. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):5719–5732, 2023. 2

[12] C. Geng, S. Peng, Z. Xu, H. Bao, and X. Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 2

[13] M. Habermann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4):1–16, 2021. 2

[14] O. Halimi, T. Stuyck, D. Xiang, T. Bagautdinov, H. Wen, R. Kimmel, T. Shiratori, C. Wu, Y. Sheikh, and F. Prada. Pattern-based cloth registration and sparse-view animation. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 2

[15] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 634–644, 2024. 2

[16] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu. Sherf: Generalizable human nerf from a single image. In *2023*

IEEE/CVF International Conference on Computer Vision (ICCV), pages 9318–9330. IEEE Computer Society, 2023. 2

[17] S. Hu, T. Hu, and Z. Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20418–20431, 2024. 1, 2, 7

[18] M. Işık, M. Rünz, M. Georgopoulos, T. Khakhulin, J. Starck, L. Agapito, and M. Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[19] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 2

[20] Y. Jiang, Q. Liao, Z. Wang, X. Lin, Z. Lu, Y. Zhao, H. Wei, J. Ye, Y. Zhang, and Z. Shao. Smplx-lite: A realistic and drivable avatar benchmark with rich geometry and texture annotations. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 1

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

[22] M. Kocabas, J.-H. R. Chang, J. Gabriel, O. Tuzel, and A. Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 2

[23] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19876–19887, 2024. 1, 2

[24] M. Li, J. Tao, Z. Yang, and Y. Yang. Human101: Training 100+ fps human gaussians in 100s from 1 view. *arXiv preprint arXiv:2312.15258*, 2023. 1

[25] R. Li, J. Tanke, M. Vo, M. Zollhöfer, J. Gall, A. Kanazawa, and C. Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 2

[26] Z. Li, Z. Zheng, Y. Liu, B. Zhou, and Y. Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2

[27] Z. Li, Z. Zheng, L. Wang, and Y. Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19711–19722. IEEE Computer Society, 2024. 2

[28] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6):1–16, 2021. 2

[29] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 1, 2

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

[31] H. Pang, H. Zhu, A. Kortylewski, C. Theobalt, and M. Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1165–1175, 2024. 1, 2

[32] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7365–7375, 2020. 1

[33] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[34] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9054–9063, 2021. 6

[35] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. 7

[36] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024. 1, 2, 7

[37] S. Saito, T. Simon, J. Saragih, and H. Joo. 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*, pages 84–93, 2020. 1

[38] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 2

[39] C. Stoll, J. Gall, E. De Aguiar, S. Thrun, and C. Theobalt. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG)*, 29(6):1–10, 2010. 2

[40] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 1, 2

[41] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics*, 28(5):1–11, 2009. 2

[42] L. Wang, X. Zhao, J. Sun, Y. Zhang, H. Zhang, T. Yu, and Y. Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 4

[43] S. Wang, K. Schwarz, A. Geiger, and S. Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Eu-*

ropean conference on computer vision, pages 1–19. Springer, 2022. 2, 7

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[45] J. Wen, X. Zhao, Z. Ren, A. G. Schwing, and S. Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 2, 3, 7

[46] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 1

[47] D. Xiang, T. Bagautdinov, T. Stuyck, F. Prada, J. Romero, W. Xu, S. Saito, J. Guo, B. Smith, T. Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 2

[48] D. Xiang, F. Prada, T. Bagautdinov, W. Xu, Y. Dong, H. Wen, J. Hodgins, and C. Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 2

[49] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 1

[50] H. Xu, T. Alldieck, and C. Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[52] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4):1–19, 2023. 2

[53] W. Zielonka, T. Bagautdinov, S. Saito, M. Zollhöfer, J. Thies, and J. Romero. Drivable 3d gaussian avatars. In *2025 International Conference on 3D Vision (3DV)*, pages 979–990. IEEE, 2025. 1